Итан Буэно де Мескита
Энтони Фаулер

СТАТИСТИКА без подвоха

Методы критического анализа данных и причинного вывода

ДМК ПРЕСС
ИЗДАТЕЛЬСТВО

Итан Буэно де Мескита
Энтони Фаулер

Статистика без подвоха

Методы критического анализа данных и причинного вывода

Thinking Clearly with Data

A Guide to Quantitative Reasoning and Analysis

Ethan Bueno De Mesquita
Anthony Fowler

PRINCETON UNIVERSITY PRESS
Princeton and Oxford

Статистика без подвоха

Методы критического анализа данных и причинного вывода

Итан Буэно де Мескита
Энтони Фаулер

ДМК Пресс
издательство
Москва, 2023

УДК 311.1
ББК 60.6
М53

Итан Буэно де Мескита, Энтони Фаулер
М53 Статистика без подвоха: Методы критического анализа данных и причинного вывода / пер. с англ. В. С. Яценкова. – М.: ДМК Пресс, 2023. – 454 с.: ил.

ISBN 978-5-93700-240-2

Увлекательное введение в науку о данных, в котором упор делается на критическое мышление, а не на статистические методы. Введение в науку о данных или статистику не должно начинаться с доказательства сложных теорем или запоминания терминов и формул, но именно так устроены многие учебники по статистике. В книге показано, как инструменты критического анализа применяются к проблемам в самых разных областях, включая выборы, гражданские конфликты, преступность, терроризм, финансовые кризисы, здравоохранение, спорт, музыка и космические путешествия.

Издание предназначено широкому кругу читателей, которые хотят быть вдумчивыми потребителями и аналитиками тех видов информации и аргументов, с которыми они будут сталкиваться на протяжении всей своей жизни.

УДК 311.1
ББК 60.6

ISBN (анг.) 978-0-69121-436-8
ISBN (рус.) 978-5-93700-240-2

Оглавление

Итан посвящает книгу Эйбу и Ханне.
Энтони посвящает книгу Глории

Предисловие от издательства

Отзывы и пожелания

Мы всегда рады отзывам наших читателей. Расскажите нам, что вы думаете об этой книге – что понравилось или, может быть, не понравилось. Отзывы важны для нас, чтобы выпускать книги, которые будут для вас максимально полезны.

Вы можете написать отзыв на нашем сайте *www.dmkpress.com*, зайдя на страницу книги и оставив комментарий в разделе «Отзывы и рецензии». Также можно послать письмо главному редактору по адресу *dmkpress@gmail.com*; при этом укажите название книги в теме письма.

Если вы являетесь экспертом в какой-либо области и заинтересованы в написании новой книги, заполните форму на нашем сайте по адресу *http://dmkpress.com/authors/publish_book/* или напишите в издательство по адресу *dmkpress@gmail.com*.

Список опечаток

Хотя мы приняли все возможные меры для того, чтобы обеспечить высокое качество наших текстов, ошибки все равно случаются. Если вы найдете ошибку в одной из наших книг – возможно, ошибку в основном тексте или программном коде, – мы будем очень благодарны, если вы сообщите нам о ней. Сделав это, вы избавите других читателей от недопонимания и поможете нам улучшить последующие издания этой книги.

Если вы найдете какие-либо ошибки в коде, пожалуйста, сообщите о них главному редактору по адресу *dmkpress@gmail.com*, и мы исправим это в следующих тиражах.

Нарушение авторских прав

Пиратство в интернете по-прежнему остается насущной проблемой. Издательство «ДМК Пресс» очень серьезно относится к вопросам защиты авторских прав и лицензирования. Если вы столкнетесь в интернете с незаконной публикацией какой-либо из наших книг, пожалуйста, пришлите нам ссылку на интернет-ресурс, чтобы мы могли применить санкции.

Ссылку на подозрительные материалы можно прислать по адресу *dmkpress@gmail.com*.

Мы высоко ценим любую помощь по защите наших авторов, благодаря которой мы можем предоставлять вам качественные материалы.

Предисловие

Наш мир удивительно изменился. Мы утопаем в океане данных, которые сами же и генерируем. Количественная информация пронизывает наши разговоры обо всем: от политики и здравоохранения до поиска работы, спорта, образования, свиданий и национальной безопасности.

Поэтому навыки количественных рассуждений больше не должны быть прерогативой только тех, кто обладает математическими способностями или собирается сделать техническую карьеру. Способности к базовым количественным рассуждениям жизненно необходимы каждому образованному человеку и гражданину. Совершенно очевидно, что обучение людей новым навыкам требует новых методов преподавания.

Именно с этой целью мы решили написать книгу, которую вы сейчас читаете. Но мы начали не с книги. Большая часть материала и идей, которые в конечном итоге нашли отражение в последующих главах, сначала была разработана для курсов, которые помогали учащимся, почти не имеющим технического образования, стать серьезными, вдумчивыми и скептически настроенными потребителями количественной информации. В состав этих курсов входят традиционные темы, такие как введение в количественное мышление, которое преподают как студентам, так и аспирантам Чикагского университета. Но, кроме этого, мы перенесли в них многое из специальных методик обучения руководителей, политиков, военных офицеров, экспертов по национальной безопасности, аналитиков разведки и журналистов.

Занимаясь преподаванием, мы и сами извлекли много уроков, которые повлияли на выбор, сделанный нами при планировании и написании этой книги. Возможно, самым важным было создание *общедоступного* языка. Мы твердо знали, что не хотим преподавать традиционный курс статистики. Для многих студентов (особенно гуманитариев) такие курсы являются слишком техническими и не дают ответов на наиболее интересные вопросы, которые действительно важны для использования количественной информации в повседневной жизни и работе. Поэтому было заманчиво как можно быстрее перейти к интересным темам, например почему корреляция не означает причинно-следственную связь. Но это было бы ошибкой. Человек не может понять, почему корреляция не предполагает причинно-следственной связи, пока не поймет, что такое корреляция и причинность.

Поэтому первая часть книги посвящена поиску общего языка с читателем. Мы начинаем не с математических формул (хотя они тоже будут), а с ответов на вопросы о том, что мы имеем в виду, когда говорим о корреляции и причинно-следственной связи. Что сложного в корреляции и причинно-следственной связи? Почему их нужно разделять? Где и как применяются понятия корреляции и причинно-следственной связи?

А как же проблема мотивации? Считается, что для вовлечения читателей нужно начинать свой рассказ с демонстрации *полезности* обсуждаемой темы. Так вот, понимание сути корреляции – чрезвычайно полезная вещь. Но, что более важно, наш подход таков: если вы хотите, чтобы читатели были вовлечены, сделайте материал *интересным*. Для нас это означает несколько вещей.

Первый способ привлечь внимание – рассказывать истории. Вскоре вы обнаружите, что каждое концептуальное понятие дополнено по крайней мере одним развернутым, подлинным примером из реальной жизни. Некоторые примеры касаются научных исследований. Многие из них будут посвящены нашему личному опыту, когда взвешенное размышление о количественных показателях повлияло на принимаемые нами решения. Другие относятся к использованию данных и доказательств в новостях, спорте, политике, здравоохранении и культуре. От правильного толкования данных напрямую зависит, как люди проживают свою жизнь и принимают решения во всех областях человеческой деятельности. Мы хотим, чтобы вы постоянно имели это в виду. Вот почему, хотя авторы книги в свое время много занимались политологией, многие примеры взяты не из политики.

Второй способ привлечь читателей – сначала сделать акцент на идеях, а затем на технических вопросах. Мы любим технические детали. Но техничность часто бывает врагом понимания. Когда разговор переходит на технические детали, многие люди перестают думать и начинают запоминать. Мы горячо хотим этого избежать. Поэтому всегда сначала говорим об идеях и о том, почему они важны. Мы используем как можно больше визуальных представлений. И как можно меньше математики. Но как можно меньше – это не ноль, по крайней мере, по двум причинам.

Понимание технических деталей является частью критического мышления. Вы не сможете понять возврат к среднему, если не знаете, что такое среднее значение или шум. Вы не сможете понять предвзятость публикаций и кризис репликации, если не знаете, что такое *статистическая значимость*, или *p*-значение. И трудно понять проблему смещения оценки эффекта или решения, предлагаемые различными способами исследования, не имея возможности интерпретировать регрессию.

Более того, иногда для ясности и точности требуется немного математики. Мы проводим много времени, рассуждая о контрфактическом подходе и причинно-следственных связях. Но без математики разговор о контрфактичности приобретает мистический оттенок. Строгая запись определений и ожидаемых эффектов вносит дополнительную ясность. Поэтому нам не обойтись без математики. Но мы всегда делаем акцент на понимании сути.

Третий способ вовлечения читателей заключается в том, что после каждой главы или урока мы оставляем небольшую недосказанность для самостоятельных размышлений. Это очень важный момент: принятие правильных решений в наш век, основанный на данных, требует навыка критического мышления от каждого из нас. Мы не можем просто оставить это на усмотрение экспертов, поскольку – и это может шокировать – многих экспертов никогда не учили правильному толкованию количественной информации. Мы вынуждены делать это сами, иначе нас часто будут вводить в заблуждение, и мы можем совершить непоправимые ошибки.

Как построена эта книга

Как мы уже отмечали, мы начинаем часть I с поиска общего языка, фокусируясь на идеях корреляции и причинно-следственной связи как краеугольных камнях количественного анализа.

Развивая эти идеи, часть II фокусируется на том, как мы используем данные и свидетельства, чтобы выяснить, существует ли причинно-следственная связь между явлениями окружающего мира. Назначение этой части книги – показать читателям, что из анализа доступной информации можно извлечь много полезного даже до причинно-следственных выводов. Глава 4 объясняет невероятно распространенную ошибку выбора зависимой переменной, показывая невозможность установить корреляцию без вариации и приводя в пример ошеломляющее количество случаев, когда эта ошибка действительно имеет значение. Глава 5 посвящена измерению корреляций с упором на графическое объяснение регрессии. В главе 6 рассказывается о статистической значимости и проверке гипотез с точки зрения нашего любимого уравнения, которое повторяется на протяжении всей книги:

$$\text{Оценка} = \text{Оцениваемая величина} + \text{Смещение} + \text{Шум}.$$

Если в главе 4 еще не была достигнута эта цель, то в главе 7 становится ясно, как много зависит от выявления взаимосвязей в данных. Мы обсудим проблемы *p*-манипуляций, публикационного смещения и связанных с ними проблем. Наконец, в главе 8 рассматривается редко обсуждаемая тема возврата к среднему значению, а затем мы объединяем ее с ранее упомянутой темой публикационного смещения, чтобы отразить кризис репликации и распространенное явление, заключающееся в сокращении научных оценок с течением времени.

Часть III посвящена причинному выводу, напоминая читателям о том, насколько важно понимание причинности для принятия решений о том, как взаимодействовать с миром. В главе 9 объясняется, почему корреляция не обязательно подразумевает причинно-следственную связь, обсуждаются как искажающие факторы, так и обратная причинно-следственная связь. Глава 10 посвящена вопросу статистического контроля и содержит графические пояснения относительно регрессии. В главах 11–13 представлен обзор того, как ученые используют специальные исследовательские приемы, чтобы попытаться узнать о причинно-следственных связях. В главе 11 рассматриваются как рандомизированные, так и естественные эксперименты, в которых представлены инструментальные переменные как метод решения проблем несоответствия. Главы 12 и 13 посвящены экспериментальным методам «разрывной регрессии» и «разности разностей» соответственно. Глава 14 завершает эту часть книги обсуждением трудностей изучения причинно-следственных механизмов.

Часть IV указывает на то, что мы еще не закончили работу, когда занялись причинно-следственными связями. Даже надежное знание причинно-следственных связей само по себе недостаточно, чтобы гарантировать ясное понимание того, как использовать количественную информацию для принятия правильных решений. В главе 15 показано, как легко обмануть себя, думая, что часть количественной информации отвечает на один вопрос, когда на самом

деле она отвечает на совершенно другой вопрос. Мы призываем читателей избегать этой ошибки, извлекая из технической информации суть. Далее мы вводим правило Байеса. Глава 16 посвящена вопросам измерения, внешней валидности и экстраполяции, что также приводит нас к обсуждению смещения выборки. И наконец, в главе 17 рассматриваются фундаментальные ограничения, с которыми сталкивается количественный анализ при принятии обоснованных решений независимо от того, насколько четко он продуман.

В конце каждой главы есть упражнения, которые читатели могут выполнить самостоятельно, чтобы убедиться, что они усвоили материал. Некоторые из этих упражнений включают анализ данных, с ним могут справиться читатели и студенты, которые научились (или учатся) использовать статистическое программное обеспечение, такое как Stata или R. В конце каждой главы также есть раздел «Дополнительное чтение и ссылки», где любознательные читатели найдут источники, упомянутые в основном тексте, и смогут более глубоко погрузиться в ту или иную тему.

Кому адресована эта книга?

Мы надеемся, что эта книга будет полезна всем, кто хочет научиться видеть суть данных, фактов и количественных рассуждений. Как уже было сказано, мы использовали эти материалы для обучения очень широкой аудитории: от студентов до высококвалифицированных специалистов.

По нашему мнению, чтобы подготовиться к жизни в наш век данных, каждый студент должен познакомиться с подобным материалом, в идеале в первые пару лет обучения в вузе. Поэтому мы написали книгу в надежде, что она будет полезна преподавателям различных дисциплин, обучающих количественному анализу, будь то общеобразовательные курсы или вводные занятия в рамках специального курса. Мы считаем, что это особенно актуально для преподавателей, которые хотят использовать более концептуальный подход, чем традиционно основанный на статистических или математических методах, но при этом опирающийся на технический фундамент.

Мы считаем, что книга так же хорошо подойдет профессионалам, заинтересованным в повышении квалификации. Например, мы обучаем этому аспирантов, получающих степень магистра в области государственной политики. Некоторые продолжают посещать дополнительные технические курсы по эконометрике или анализу данных. Но для многих важно научиться критически осмысливать количественную информацию. Наш подход соответствует потребностям этих студентов и в то же время обеспечивает понятийную основу, которая понадобится более технически подкованным студентам на будущих курсах.

Коллеги из других университетов также использовали эти материалы в более продвинутых курсах для специалистов по общественным наукам, которые, например, должны изучать количественные методы при подготовке к написанию диссертации. В этом контексте наша книга только выигрывает от совместного использования с другими изданиями, которые являются более техническими или уделяют больше внимания вопросам статистических вычислений. Мы надеемся, что во всех этих случаях будут полезными упражнения в конце

каждой главы. Особенно это относится к прикладному анализу, примеры наборов данных для которого можно скачать в интернете.

Наконец, мы также считаем, что эта книга будет полезна многим аспирантам. Часто в аспирантуре статистику преподают быстро и на высоком техническом уровне. Это может быть продуктивно; овладение передовыми методами одновременно сложно и важно. Но, по нашему опыту, даже лучшие аспиранты могут упустить из виду то, что действительно важно, – то, как мы узнаем о мире из данных, – поскольку они сосредотачиваются на доказательстве теорем и программировании алгоритмов. Мы очень надеемся, что эта книга послужит руководством для таких студентов, помогая четко видеть общую картину, даже когда они усердно работают над техническими деталями.

Благодарности

Как мы уже упоминали, некоторые материалы в этой книге были разработаны совместными усилиями для студенческого курса Энтони и учебного курса для руководителей, который совместно разработали Джейк Шапиро, Лиам Коллинз, Кэти Фетелл и Итан. Мы выражаем огромную благодарность и глубокую признательность Джейку, Лиаму и Кэти.

Хотелось бы поблагодарить Скотта Эшворта, Криса Берри, Криса Блаттмана, Мэтта Бремса, Брюса Буэно де Мескиту, Кервина Чарльза, Девина Чесни, Линдси Кормак, Энди Эггерса, Натана Фаверо, Алекса Фуирне, Мэтта Гейбела, Джеффа Гроггера, Энди Холла, Косуке Имаи, Ренана Левина, Эндрю Литтла, Йенса Людвига, Мордехая Магенси, Эндрю Минса, Пабло Монтаня, Эмили Риттер, Стива Шваба, Майка Спагата, Дастина Тингли, Стефана Уолтона и Остина Райта за их невероятно полезные отзывы.

Том Будеску, Гаутам Наир, Том Насет, Джефф Рафф, Ванита Вируудачалам, Бекки Ван и Синъюй Инь оказали потрясающую помощь в исследованиях на ранних этапах создания этого проекта. Было очень интересно работать с ними, и мы благодарны за их вклад.

Мы хотели бы поблагодарить наших студентов за то, что они заметили многочисленные ошибки и опечатки в черновиках книги. А. К. Алилон, Дениз Азаде, Элли Ратки и Аль Шах обнаружили их поистине смущающее количество. Спасибо!

Команда издательства Princeton University Press была потрясающей. Мы особенно признательны Бриджит Флэннери-Маккой за веру в проект и ее руководство, а также Алене Чекановой за контроль над процессом. И мы в долгу перед Данной Локвуд за ее невероятную работу, которая помогла улучшить процесс написания и структуру книги. Мы также очень ценим потрясающий контроль производства Мелоди Негрон и всегда отличную индексацию терминов Дэвида Люляка.

Итан благодарит своих коллег из школы Харриса и многочисленных сотрудников, соавторов и учеников, которые являются интеллектуальным источником вдохновения и за годы героических усилий улучшили ясность его собственного мышления. Он глубоко благодарен своей жене Ребекке, которая мирилась с нервозностью автора, неизбежно сопровождающей завершение книги, и воспринимала ее с неизменной поддержкой, любовью и терпением,

с которыми она встречает все радости жизни с Итаном. Итан посвятил эту книгу своим детям, Ханне и Эйбу, доставляющим ему только радость и удовольствие. Его самая искренняя надежда состоит в том, что книга широко разойдется по учебным заведениям и однажды его дети будут учиться по учебнику, посвященному им самим. Это был бы статистически значимый успех.

Энтони благодарит своих консультантов, соавторов и коллег, которые помогают ему мыслить яснее. Он благодарен своим родителям, поощряющим и поддерживающим его на протяжении всей жизни, даже когда он не планировал поступать на юридический или медицинский факультет. И самое главное, он благодарит Глорию, свою жену и лучшую подругу, которая прочитала бесчисленные черновики научных статей, выдержала слишком много разговоров о регрессиях, развлекалась вводом данных, проверяла каждую идею мужа, вносила непропорционально большую долю своих собственных идей и украшала его жизнь способами, которые не поддаются количественной оценке.

Глава 1

Критическое мышление в эпоху данных

О чем эта глава

- Навык ясного и концептуального восприятия количественной информации важен по многим причинам, даже если вы не заинтересованы в карьере аналитика данных.
- Даже хорошо обученные профессионалы часто допускают серьезные ошибки при обработке данных.
- Мышление и данные дополняют, а не заменяют друг друга.
- Навыки, которые вы приобретете в этой книге, помогут вам использовать фактические данные для принятия более эффективных решений в личной и профессиональной жизни и стать более вдумчивыми и хорошо информированными людьми.

Введение

Мы живем в век данных. По словам бывшего генерального директора Google Эрика Шмидта, современный мир каждые два дня создает столько же новых данных, сколько было создано с начала времен до 2003 г. Предполагается, что вся эта информация призвана улучшить нашу жизнь, но, чтобы использовать ее скрытую мощь, мы должны научиться искусству здравомыслия в мире, основанном на данных. Это сложная задача, особенно если не уметь абстрагироваться от технических деталей, которыми обычно окружены данные и их анализ.

Ясное мышление в эпоху данных – это, прежде всего, сосредоточенность на идеях и вопросах. Формулы и алгоритмы должны служить этим идеям и вопросам. К сожалению, курсы по статистике и количественному анализу, на которых большинство людей изучают науку о данных, делают прямо противоположное – т. е. фокусируются на технических деталях. Студенты изучают математические формулы, запоминают названия статистических процедур и начинают манипулировать числами, хотя никто не позаботился о ясном и принципиальном понимании того, что они делают и почему они это делают. Такой подход может сработать среди людей, для которых математическое мышление является естественным. Но мы считаем, что это контрпродуктивно для подавляющего большинства из нас. Когда формальные курсы подталкива-

ют учащихся к тому, чтобы перестать думать и начать запоминать, они теряют лес за деревьями. И это совсем не весело.

Эта книга, напротив, сосредоточена на концептуальном понимании. Какие признаки мира вы сравниваете, когда анализируете данные? На какие вопросы отвечают различные виды сравнений? У вас есть правильный вопрос и сравнение для задачи, которую вы пытаетесь решить? Почему ответ, звучащий убедительно, на самом деле может вводить в заблуждение? Какие творческие подходы можно использовать, чтобы дать более информативный ответ?

Мы не стремимся преуменьшить значение технических знаний, но полагаем, что технические навыки без концептуального понимания или критического мышления – это путь к катастрофе. По нашему мнению, как только у вас появится интуитивное понимание количественного анализа и как только вы поймете, почему так важно задавать аккуратные и точные вопросы, технические навыки придут сами собой. К тому же этот способ веселее.

Вдохновленные этими соображениями, мы написали книгу, которая не требует предварительного знакомства с анализом данных, статистикой или количественными методами. Считая, что концептуальное мышление важнее, мы минимизировали (хотя, конечно, не исключили) технический материал в пользу объяснений на простом разговорном языке, где это возможно. Надеемся, что эта книга будет использоваться в качестве введения и руководства к тому, как правильно мыслить и проводить количественный анализ. Мы считаем, что любой из нас может стать искушенным потребителем (и даже производителем) количественной информации. Требуется лишь немного терпения, настойчивости, тяжелой работы и твердой решимости никогда не позволять технике заменять критическое мышление.

Большинство людей не становятся профессионалами количественного анализа. Но, независимо от того, встанете вы на этот путь или нет, мы уверены, что вы будете использовать навыки, полученные в этой книге, самыми разными способами. У многих из вас будут количественные аналитики, работающие на вас или вместе с вами. И все вы будете читать исследования, новостные репортажи и брифинги, на которых кто-то пытается убедить вас в правильности выводов, используя количественный анализ. Эта книга даст вам навыки критического мышления, необходимые для того, чтобы задавать правильные вопросы, проявлять скептицизм, когда это необходимо, и отличать полезные свидетельства от вводящих в заблуждение.

Поучительные истории

Чтобы разжечь ваш аппетит к предстоящей тяжелой работе, давайте начнем с нескольких поучительных историй, которые подчеркивают важность критического мышления в эпоху данных.

Поспешный диагноз Эйба

Первый ребенок Итана, Эйб, родился в июле 2006 г. В младенчестве он кричал и плакал почти без перерыва по ночам в течение первых пяти месяцев. В остальном Эйб был счастлив и здоров, хотя и весил немного меньше нормы. Когда ему исполнился год, семья переехала в Чикаго, а иначе вы не читали бы

эту книгу. (Последнее предложение представляет собой пример особого вида утверждений, называемых *контрфактическими* (counterfactual). Контрфактические утверждения действительно важны, и вы узнаете о них больше в главе 3.) Заметив, что Эйб был маловат для своего возраста и рос медленнее, чем полагалось, его педиатр решил провести несколько анализов.

После ряда лабораторных исследований врачи были почти уверены, что у Эйба целиакия – заболевание пищеварительной системы, характеризующееся непереносимостью глютена. Хорошая новость: целиакия не опасна для жизни и даже не очень серьезна, если ее правильно лечить. Плохая новость: в 2007 г. варианты безглютеновой диеты для маленьких детей были довольно убогими.

Оказывается, Эйбу на самом деле провели два анализа крови на целиакию. Один результат оказался положительным (что указывает на наличие заболевания), другой – отрицательным (указывает на отсутствие заболевания). По словам врачей, точность положительного теста превышает 80 %. «Это достоверный диагноз», – сказали они. Предложенный план действий заключался в том, чтобы посадить Эйба на безглютеновую диету на пару месяцев и посмотреть, не увеличится ли его вес. При таком исходе врачи могли бы сделать более точные дополнительные анализы или просто оставить Эйба без глютена на всю оставшуюся жизнь.

Итан захотел взглянуть на распечатку анализов крови Эйба. Врачи сказали, что вряд ли это будет полезно, поскольку Итан не врач. В таком ответе нет ничего удивительного, и его легко объяснить. Люди, особенно эксперты и авторитетные лица, часто не любят признавать ограниченность своих знаний. Но Итан хотел принять правильное решение для своего сына, поэтому настаивал на информации. Одна из важных целей этой книги – дать вам уверенность в себе, чтобы вы могли защищать себя, используя информацию для принятия жизненно важных решений.

Эффективность любого диагностического теста характеризуют два числа. Во-первых, это доля *ложноотрицательных* результатов, т. е. насколько часто тест показывает, что больной человек здоров. Во-вторых, это доля *ложноположительных* результатов, т. е. как часто тест показывает, что здоровый человек болен. Чтобы правильно интерпретировать результаты диагностического теста, вам необходимо знать *оба* числа. Поэтому заявление врачей Эйба о том, что положительный анализ крови был точным на 80 %, было не очень информативным. Означало ли это, что у теста 20 % ложноотрицательных результатов? 20% ложноположительных результатов? Или тест говорит о том, что у 80 % людей с положительным результатом теста есть целиакия?

К счастью, беглый поиск в Google позволил найти значения ложноположительных и ложноотрицательных результатов для обоих тестов Эйба. Вот что узнал Итан. Тест, который у Эйба дал положительный результат на целиакию, имеет долю ложноотрицательных результатов около 20 %. То есть, если 100 человек с целиакией пройдут тест, около 80 из них получат правильный положительный результат, а остальные 20 – ошибочно отрицательный. Мы предполагаем, что именно на этом факте и основано заявление врача о 80-процентной точности. Тест, однако, имеет 50 % ложноположительных результатов! Люди, у которых нет целиакии, с одинаковой вероятностью могут получить как положительный, так и отрицательный результат. (Следует отметить, что этот тест больше не рекомендуется для диагностики целиакии.) Напротив, тест, согласс-

но которому у Эйба нет целиакии, имел гораздо меньшую долю ложноотрицательных и ложноположительных результатов.

До изучения результатов анализов разумная оценка вероятности заболевания Эйба целиакией с учетом его отставания в наборе веса составляла примерно 1 к 100. То есть примерно у одного из каждых 100 детей с отставанием набора веса есть целиакия. Вооружившись лабораторными отчетами и показателями ложноположительных и ложноотрицательных результатов, Итан смог подсчитать вероятность того, что при имеющихся результатах анализов и наличии отставания веса у Эйба действительно есть целиакия. Удивительно, но сочетание положительного результата менее точного теста и отрицательного результата более точного теста на самом деле означало, что вероятность заболевания целиакией у Эйба гораздо ниже, чем 1 из 100. Фактически, как мы покажем вам в главе 15, наибольшая оценка вероятности заболевания Эйба целиакией, учитывая результаты анализов, составляла примерно 1 из 1000. Анализы крови, которые, как были уверены врачи Эйба, подтверждали диагноз целиакии, на самом деле убедительно указывали на противоположный вывод. Эйб почти наверняка не был болен целиакией.

Итан позвонил врачам, чтобы объяснить, что он выяснил, и предположить, что перевод его сына, одержимого макаронами, на безглютеновую диету, возможно, на всю жизнь не является разумным следующим шагом. На это врачи ответили: «Просто вам трудно принять диагноз». Тогда Итан нашел нового педиатра.

Каков же результат? У Эйба не было целиакии. Это был просто маленький ребенок, который в силу индивидуальных особенностей рос немного медленнее, чем другие. Сегодня это обычный ребенок с ненасытным аппетитом. Но если бы его отец не умел работать с количественными показателями или ему не хватало уверенности, чтобы бросить вызов самоуверенному эксперту, бедняга Эйб провел бы свое детство, питаясь рисовыми лепешками. Рисовые лепешки отвратительны, так что Эйб рисковал никогда не набрать вес.

Гражданское сопротивление

Многим людям на своем опыте довелось хоть раз испытать глубокое несогласие со своим правительством. Когда дела идут особенно плохо, они иногда решают провести акции протеста. Если вы когда-нибудь займетесь организацией подобной акции протеста, то столкнетесь с необходимостью принимать важные решения. Например, вам придется сделать выбор, какое движение предпочесть – с мягкой ненасильственной стратегией или допускающее более жесткие формы конфронтации? Размышляя над этим затруднительным положением, вы наверняка прислушаетесь к своим внутренним этическим убеждениям. Но вам будет полезно узнать, что говорят фактические данные о недостатках и преимуществах каждого подхода. Какая организация, скорее всего, добьется изменений в политике правительства? Какой из этих подходов с большей вероятностью приведет вас в тюрьму, больницу или морг?

Существуют количественные данные, которые вы можете использовать для обоснования своих решений. Во-первых, если сравнить антиправительственные движения по всему миру на протяжении достаточно длительного времени, то станет видно, что правительства чаще идут на уступки полностью не-

насильственным группам, чем группам, применяющим насилие. И даже если рассматривать только группы, применяющие насилие, можно сделать вывод, что правительства чаще идут на уступки группам, применяющим насилие против военных и правительственных объектов, а не против гражданского населения. Во-вторых, личные риски, связанные с насильственным протестом, выше, чем риски, связанные с ненасильственным протестом. Правительства подавляют насильственные протесты чаще, чем ненасильственные, что делает опасения по поводу тюрьмы, больницы и морга еще более острыми.

Эти аргументы звучат весьма убедительно. Ненасильственная стратегия кажется очевидным выбором. Очевидно, что это и более эффективно, и менее рискованно. И действительно, на основе такого рода данных политологи Эрика Ченовет и Эван Перкоски приходят к выводу, что «планирование и подготовка ненасильственных акций протеста имеют ключевое значение, особенно (и это парадоксально) при противостоянии жестоким режимам».

Но давайте присмотримся к доказательствам. Начнем с вопроса: «В какой ситуации группа активистов выберет ненасильственный протест вместо насильственных действий?» Нам приходит в голову несколько мыслей. Возможно, люди охотнее примут участие в ненасильственном протесте, когда имеют дело с правительством, которое, по их мнению, с большей вероятностью прислушается к требованиям своих граждан. Или, возможно, люди скорее выберут ненасильственный протест, если они имеют широкую поддержку среди своих сограждан, представляют влиятельную группу в обществе, которая может привлечь внимание средств массовой информации, или сталкиваются с менее жестким правительством.

Если что-то из сказанного верно, нам следует беспокоиться по поводу утверждения о том, что проведение ненасильственных акций является ключом к построению успешного антиправительственного движения. (Это вовсе не значит, что мы призываем к насилию!) Давайте разберемся, в чем дело.

Эмпирические исследования показывают, что в среднем правительства чаще идут на уступки в тех местах, где прошли именно ненасильственные протесты. Данный вывод основан на буквальной интерпретации разницы, а именно на том, что более высокая частота уступок со стороны государства вызвана использованием ненасильственной тактики. Иными словами, при прочих равных условиях, если бы некое движение, использующее насильственные методы, перешло на использование ненасильственных методов, правительство с большей вероятностью пошло бы на уступки. Но действительно ли такая причинная интерпретация оправдана фактами?

Предположим, что протестные движения с большей вероятностью прибегнут к насилию, если они не имеют широкой поддержки среди своих сограждан. Далее, когда мы сравниваем места, где были насильственные протесты, с местами проведения мирных акций, все остальные условия (кроме тактики протеста) нельзя считать равными. Эти места отличаются как минимум по двум причинам. Во-первых, они различаются по факту наличия насильственных и ненасильственных протестов. Во-вторых, они различаются по степени поддержки протестного движения общественностью.

Это второе отличие представляет собой проблему для причинной интерпретации. Вполне логично предположить, что общественное мнение оказывает

независимое влияние на готовность правительства пойти навстречу протестующим. То есть при прочих равных условиях (включая тактику протеста) правительство с большей готовностью пойдет на уступки протестным движениям, пользующимся широкой общественной поддержкой. Если это так, то мы не можем однозначно утверждать, что правительства идут на уступки из-за разницы в тактике протеста. Возможно, все дело в широкой общественной поддержке более миролюбивых движений. Это классическая проблема ошибочного принятия корреляции за причинно-следственную связь.

Стоит отметить несколько моментов. Во-первых, если уступки правительства на самом деле обусловлены общественным мнением, то с учетом этого влияния может оказаться, что мирные протесты не эффективнее насильственных (они могут быть даже менее эффективными). Располагая ограниченными доказательствами, мы просто не можем знать этого наверняка.

Во-вторых, если не заставлять себя мыслить критически, то в этом примере мы приходим к выводу, который нам больше нравится. Кто из нас не хотел бы жить в мире, где миролюбие всегда предпочтительнее насилия? Но весь смысл использования доказательств, помогающих нам принимать решения, состоит в том, чтобы заставить нас признать, что мир не всегда устроен так, как нам хочется или как мы верим. На самом деле именно в ситуациях, когда кажется, что факты говорят то, что вы хотели бы услышать, особенно важно заставить себя сохранить ясность мышления.

В-третьих, мы указали лишь на одну проблему в оценке последствий мирного и насильственного протеста, но есть и другие. Например, подумайте о другом эмпирическом утверждении, которое мы обсуждали: насильственные протесты с большей вероятностью спровоцируют правительство на репрессивные меры, чем мирные акции. Напомним, мы предположили, что люди с большей вероятностью будут участвовать в мирных протестах, когда они меньше злятся на свое правительство, возможно, потому что правительство ведет себя менее жестко. Спросите себя, почему, если это правда, у нас возникает аналогичная проблема интерпретации? Почему тот факт, что после насильственных протестов правительство применяет больше репрессий, чем после мирных акций, не обязательно означает, что переход от насилия к ненасилию снизит риск репрессий? Этот аргумент следует той же логике, что и ранее рассмотренная в отношении уступок. Если вы пока не понимаете, как работает этот аргумент, ничего страшного. Все станет ясно к концу главы 9.

Теория разбитых окон

В 1982 г. криминолог Джордж Келлинг и социолог Джеймс Уилсон опубликовали в The Atlantic статью, где предложили новую теорию преступности и полицейской деятельности, которая оказала огромное и долгосрочное влияние на криминальную политику в Соединенных Штатах и за их пределами.

Это знаменитая *теория разбитых окон*. Она была вдохновлена успехом программы в Ньюарке, штат Нью-Джерси, в соответствии с которой полицейские вышли из машин и патрулировали улицы пешком. По словам Келлинга и Уилсона, программа снизила уровень преступности за счет повышения «уровня общественного порядка». Общественный порядок важен, утверждают они, потому что его отсутствие запускает порочный круг:

> «Участок заброшен и зарос сорняками, окна в доме разбиты. Родители перестают ругать непослушных детей… Семьи выезжают, заселяются одинокие взрослые. Перед магазином на углу собираются подростки. Владелец просит их уйти; они отказываются. Происходят драки. Растут горы мусора. Люди начинают пить алкоголь прямо у входа в продуктовый магазин…
>
> Жители замечают рост количества преступлений, особенно насильственных… Они стараются реже выходить на улицы… Такой район подвержен дальнейшей криминализации».

Идея о том, что работа полиции, сосредоточенная на минимизации любых проявлений беспорядка, может снизить уровень насильственных преступлений, оказала большое влияние на тактику полиции. Теория разбитых окон легла в основу стратегии муниципалитета Нью-Йорка в 1990-е гг. В своей речи 1998 г. тогдашний мэр Нью-Йорка Рудольфо Джулиани сказал:

> «Мы сделали теорию разбитых окон неотъемлемой частью нашей правоохранительной стратегии…
>
> Мы концентрируемся на мелочах и посылаем всем четкий сигнал о том, что этот город заботится о поддержании закона и порядка… в итоге город в целом станет безопаснее».

И действительно, преступность в Нью-Йорке ощутимо снизилась, когда полиция начала концентрироваться «на мелочах». Согласно исследованию Хоуп Корман и Наси Мокана, в 1990-е гг. количество арестов за правонарушения увеличилось на 70 %, а количество насильственных преступлений снизилось более чем на 56 %, что вдвое лучше, чем средний показатель по стране.

Чтобы оценить, в какой степени политика борьбы с разбитыми окнами повлияла на падение уровня преступности, Келлинг и Уильям Соуза изучили взаимосвязь между насильственными преступлениями и применением стратегии борьбы с «разбитыми окнами» на территории различных полицейских участков Нью-Йорка. Если наведение порядка ведет к сокращению насильственных преступлений, утверждали они, то нам следует ожидать, что наибольшее снижение преступности произойдет в округах, где полиция больше всего сосредоточилась на стратегии «разбитых окон». Именно это они и обнаружили. На участках, где арестов за мелкие правонарушения было больше, насильственная преступность снизилась сильнее. Они подсчитали, что «для среднего полицейского округа Нью-Йорка… снижение на одно насильственное преступление приходится примерно на 28 дополнительных арестов за мелкие правонарушения».

Звучит довольно убедительно. Но давайте не будем спешить с выводом, что арест людей за мелкие правонарушения – это рецепт прекращения насильственных преступлений. Два других ученых, Бернард Харкорт и Йенс Людвиг, призывают нас более взвешенно отнестись к тому, что говорят данные.

Проблема, на которую указывают Харкорт и Людвиг, – это так называемый *возврат к среднему значению* (о котором мы подробнее поговорим в главе 8). В любой конкретный год количество преступлений на участке определяется множеством факторов, включая работу полиции, наркотики, экономику, погоду и т. д. Многие из этих факторов нам неизвестны. Некоторые из них мимолетны;

они возникают и исчезают из года в год. Поэтому мы можем предположить, что на любом конкретном участке существует некоторый «базовый» уровень преступности, причем в некоторые годы преступность случайным образом выше, а в некоторые годы – ниже (относительно базового уровня для данного участка).

Если в определенный год на участке был высокий уровень преступности (относительно его базового уровня), то ему не везло с неизвестными и мимолетными факторами, которые способствуют совершению преступлений. Вероятно, следующий год будет не таким плохим (вот что значит мимолетность), так что на этом участке, скорее всего, снизится преступность. А если в этом году на участке уровень преступности был ниже базового уровня, то ему повезло с неизвестными и мимолетными факторами, а в следующем году, вероятно, будет хуже (преступность снова пойдет вверх). Таким образом, год за годом уровень преступности на участке имеет тенденцию возвращаться к среднему (т. е. базовому уровню).

Теперь представьте себе участок, в котором в конце 1980-х гг. был действительно высокий уровень насильственных преступлений. В отношении этого участка, вероятно, справедливы две вещи. Во-первых, это, вероятно, участок с высоким базовым уровнем насильственных преступлений. Во-вторых, это также, вероятно, участок, в котором год или два были неудачными, т. е. по уникальным и мимолетным причинам уровень преступности в конце 1980-х гг. был высоким по сравнению с базовым уровнем этого участка. То же самое, конечно, верно и наоборот для участков, где в конце 1980-х гг. был низкий уровень преступности. У них, вероятно, низкий базовый уровень преступности, и вдобавок выдалась пара хороших лет.

Почему это проблема для выводов Келлинга и Соузы? Благодаря возвращению к среднему значению мы ожидаем, что в наиболее криминальных районах конца 1980-х гг. в среднем будет наблюдаться снижение насильственных преступлений, даже без каких-либо изменений в работе полиции. И вполне логично с точки зрения полиции, но, к несчастью для исследования, именно на участках, где в конце 1980-х гг. был высокий уровень преступности, чаще всего в начале 1990-х гг. полицейские работали по программе борьбы с «разбитыми окнами». Итак, когда мы видим снижение количества насильственных преступлений на участках, где активно велась полицейская деятельность по борьбе с мелкими преступлениями, мы не знаем, что сработало на самом деле – полицейская стратегия или возврат к среднему значению.

Харкорт и Людвиг пошли еще дальше, пытаясь найти более убедительные доказательства. Они изучили то, как изменения в количестве арестов за мелкие правонарушения связаны с изменениями в уровне насильственных преступлений на участках, где в конце 1980-х гг. наблюдался аналогичный уровень насильственных преступлений. Сравнивая участки с аналогичным начальным уровнем насильственных преступлений, они в некоторой степени устраняют проблему возврата к среднему значению. Удивительно, но это простое изменение на самом деле меняет отношения! Вместо того чтобы подтвердить вывод Келлинга и Соузы о том, что аресты за мелкие правонарушения ведут к снижению количества насильственных преступлений, Харкорт и Людвиг обнаружили, что на участках, которые больше внимания уделяли арестам за мелкие правонарушения, на самом деле, похоже, наблюдался рост

насильственных преступлений. Это полная противоположность нашим ожиданиям от теории разбитых окон.

Впрочем, это изменение не опровергает эффективность борьбы с разбитыми окнами. Связь между арестами за мелкие проступки и насильственными преступлениями, которую обнаружили Харкорт и Людвиг, может существовать по множеству причин. Например, возможно, районы с увеличением количества правонарушений в целом становятся менее безопасными и в них будет больше насильственных преступлений независимо от стратегий полиции. На самом деле все эти рассуждения свидетельствуют лишь об одном – что данные, если их правильно рассмотреть, определенно не дают однозначного подтверждения теории разбитых окон, как можно было подумать, исходя из открытия Келлинга и Соузы. И увидеть это можно только в том случае, если у вас есть способность критически анализировать некоторые тонкие нюансы.

Ошибочное мышление сыграло свою роль. Основанные на фактических данных выводы Келлинга и Соузы убедили политиков и городскую администрацию в том, что работа полиции с мелкими правонарушениями является правильным путем вперед, хотя на самом деле она могла отвлечь ресурсы от предотвращения и расследования насильственных преступлений и, возможно, создать более враждебные отношения между полицией и непропорционально бедным населением и национальными меньшинствами, которых часто обвиняли в мелких нарушениях.

Дополнение или замена?

Наш количественный мир наполнен новыми интересными данными и аналитическими инструментами для анализа этих данных с причудливыми названиями, такими как алгоритмы машинного обучения, искусственный интеллект, случайные леса и нейронные сети. Мы все чаще слышим, что новые технологии позволят машинам думать за нас. Но это не так. Как подчеркивают наши поучительные истории, никакой анализ данных, каким бы футуристическим ни было его название, не будет работать, если мы не задаем правильные вопросы, если мы не проводим правильные сравнения, если основные предположения неверны или используемые данные не подходят. Тот факт, что аргумент опирается на сложный количественный анализ данных, не означает, что этот аргумент строгий или правильный. Чтобы использовать возможности данных для принятия более эффективных решений, мы должны сочетать количественный анализ с критическим мышлением.

Наши истории также показывают, как интуиция может сбить нас с пути. Требуется много внимания и практики, чтобы научиться критически воспринимать «очевидные» доказательства. Интуитивные выводы врачей о том, что у Эйба была целиакия, из-за теста с 80-процентной точностью и выводы исследователей о том, что полицейская стратегия борьбы с «разбитыми окнами» работает, потому что преступность снизилась в тех местах, где она была применена, кажутся разумными. Но оба интуитивных вывода оказались ошибочными, и мы предположили, что необходимо скептически относиться к первоначальным догадкам. Хорошая новость заключается в том, что критическое мышление можно довести до автоматизма, если его регулярно применять.

Данные и инструменты количественного анализа не заменяют критического мышления. На самом деле навыки работы с количественными данными без критического мышления весьма опасны. Читая следующие главы, вы наверняка будете потрясены тем, в какой степени косное мышление влияет даже на самые важные решения, принимаемые людьми. В ходе прочтения этой книги вы увидите, как неверно истолкованная информация ведет к ошибочным решениям, от которых зависит жизнь больного, как она искажает национальную и международную политику борьбы с терроризмом, вносит путаницу в деловые и филантропические решения, принимаемые самыми богатыми людьми мира, как мы ошибочно устанавливаем приоритеты в образовании наших детей и заблуждаемся во множестве других вопросов, от банальных до глубоких. По сути, ни один аспект жизни не застрахован от критических ошибок в понимании и интерпретации количественной информации.

По нашему опыту, это происходит потому, что искаженное восприятие доказательств глубоко укоренилось в человеческой психологии. Разумеется, наша собственная интуиция, оставленная без контроля, часто подвержена фундаментальным ошибкам. Мы предполагаем, что и ваша тоже. Самое тревожное – что эксперты, от чьих советов вы зависите, – будь то врачи, бизнес-консультанты, журналисты, учителя, финансовые консультанты, ученые или кто-то еще, – склонны к совершению таких же ошибок. Поскольку они считаются экспертами, мы безоговорочно доверяем их суждениям, а они не сомневаются в собственных умозаключениях. Вот почему так важно научиться критически воспринимать количественные доказательства в первую очередь для себя. Это единственный способ научиться задавать правильные вопросы, которые приведут вас и тех, от чьих советов вы зависите, к наиболее надежным и продуктивным выводам.

Как получается, что эксперты в столь многих областях так часто допускают существенные ошибки? Экспертное знание в любой области приходит в результате обучения, практики и опыта. Никто не рассчитывает стать экспертом в области инженерии, финансов, медицины или сантехники без обучения и многих лет работы. Но, несмотря на фундаментальную значимость вопроса, почти никто не прикладывает аналогичные усилия, чтобы научиться корректно и непредвзято работать с данными. И даже когда люди стремятся к этому, их преподаватели склонны преувеличивать технические аспекты и недооценивать концептуальные, хотя фундаментальные проблемы почти всегда связаны с концептуальными ошибками в мышлении, а не с техническими ошибками в расчетах.

Отсутствие опыта критического мышления ставит перед нами две проблемы. Во-первых, если столько экспертных советов и выводов ненадежны, откуда нам знать, чему верить? Во-вторых, как выделить мнения экспертов, которые действительно отражают критическое мышление?

В это книге мы закладываем основу решения упомянутых проблем. Каждая из последующих глав объясняет и иллюстрирует на множестве примеров фундаментальные принципы критического мышления в мире, управляемом данными. Первая часть помогает найти общий язык с читателем, поясняя, что мы подразумеваем под корреляцией и причинно-следственной связью и в каких случаях применяются эти понятия. Во второй части рассказано, как мож-

но определить, является ли статистическая взаимосвязь подлинной. В третьей части вы узнаете, как установить, отражают ли эти отношения причинность. А в четвертой части обсуждается, как нам следует и не следует включать количественную информацию в процесс принятия решений.

Мы надеемся, что чтение этой книги поможет вам усвоить принципы критического мышления настолько глубоко, что они станут вашей второй натурой. Вы поймете, что находитесь на правильном пути, когда обнаружите, что замечаете основные ошибки в том, как люди думают и говорят о значении доказательств, куда бы вы ни обратились – когда вы смотрите новости, листаете журналы, разговариваете с деловыми партнерами, посещаете врача, слушаете комментарии во время спортивных соревнований, читаете научные статьи или участвуете в общественных мероприятиях. Мы подозреваем, что поначалу будет трудно поверить, сколько чепухи вам регулярно говорят самые разные эксперты. Когда у вас наступит прозрение, постарайтесь оставаться скромными и конструктивными в своей критике. Но не стесняйтесь поделиться этой книгой с теми, чьи аргументы, по вашему мнению, особенно в ней нуждаются. Или, еще лучше, предложите им купить собственный экземпляр!

Дополнительное чтение и ссылки

Эссе о ненасильственном протесте Эрики Ченоуэт и Эвана Перкоски, которое мы цитируем, можно найти по адресу *https://politicalviolenceataglance.org/2018/05/08/states-are-far-less-likely-to-engage-in-mass-violence-against-nonviolent-uprisings-than-violent-uprisings/*.

Следующая книга содержит дополнительные исследования взаимосвязи между миролюбием и эффективностью протестов:

Erica Chenoweth and Maria J. Stephan. 2011. *Why Civil Resistance Works: The Strategic Logic of Nonviolent Conflict.* Columbia University Press.

Следующие статьи упоминались в указанном порядке на тему проверки теории разбитых окон:

George L. Kelling and James Q. Wilson. 1982. *Broken Windows: The Police and Neighborhood Safety*. The Atlantic. March. https://www.theatlantic.com/magazine/archive/1982/03/broken-windows/304465/;

Archives of Rudolph W. Giuliani. 1998. *The Next Phase of Quality of Life: Creating a More Civil City.* February 24. *http://www.nyc.gov/html/rwg/html/98a/quality.html*;

Hope Corman and H. Naci Mocan. 2005. *Carrots, Sticks, and Broken Windows.* Journal of Law and Economics 48 (1): 235–66;

George L. Kelling and William H. Sousa, Jr. 2001. *Do Police Matter? An Analysis of the Impact of New York City's Police Reforms.* Civic Report for the Center for Civic Innovation at the Manhattan Institute;

Bernard E. Harcourt and Jens Ludwig. 2006. *BrokenWindows: New Evidence from New York City and a Five-City Social Experiment.* University of Chicago Law Review 73: 271–320. Опубликованная версия имеет опечатку в таблице ключей. Исправление см. в *Errata*, 74 U. Chi. L. Rev. 407 (2007).

ЧАСТЬ I

В поиске общего языка

Глава 2

Корреляция: что это такое и для чего она нужна?

О чем эта глава

- Корреляция говорит нам о том, в какой степени два явления мира имеют тенденцию возникать вместе.
- Чтобы измерить корреляцию двух величин, мы должны иметь данные с изменяющимися значениями двух переменных.
- Корреляции могут быть потенциально полезны для описания, прогнозирования и выявления причинно-следственных связей. Но мы должны четко понимать, когда они подходят для каждой из этих задач.
- Корреляции представляют собой линейные отношения, но это не такое строгое ограничение, как вы думаете.

Введение

Корреляция не подразумевает причинно-следственную связь. Предельно понятно. Однако, по нашему опыту, от этого краткого высказывания мало пользы. Хотя многие люди запомнили, что корреляция не то же самое, что причинно-следственная связь, на самом деле у них нет четкого понимания определений того и другого.

В первой части книги мы собираемся потратить некоторое время на создание общего словарного запаса. Абсолютно важно убедиться, что мы с вами одинаково понимаем и используем эти и некоторые другие ключевые термины для обозначения одного и того же понятия, если хотим подвергнуть их критическому обсуждению в последующих главах.

Эта глава посвящена изучению корреляции. Мы постараемся пояснить, что это такое и для чего она нужна. Корреляция – это основной инструмент, с помощью которого аналитики количественных данных описывают мир, прогнозируют будущие события и отвечают на научные вопросы. Аккуратные аналитики не избегают и не игнорируют корреляции. Но они должны хорошенько подумать о том, на какие вопросы корреляции могут дать ответы, а от каких лучше воздержаться.

Что такое корреляция?

Корреляция между двумя явлениями мира – это степень, в которой они склонны происходить вместе. Из этого определения следует, что *корреляция* – это связь между двумя явлениями (которые иногда называют *переменными*). Если два явления мира имеют тенденцию проявляться вместе, они *положительно коррелированы*. Если в рассматриваемом мире возникновение одного события никак не связано с возникновением другого, они *некоррелированы*. А если при возникновении одного события мира другое имеет тенденцию не проявляться, такие события называются *отрицательно коррелированными*.

Давайте уточним, что означает высказывание «два явления мира проявляются вместе». Начнем с самого простого примера. Предположим, мы хотим оценить корреляцию между двумя признаками мира, и для каждого из них есть только два возможных значения (мы называем их *бинарными* переменными). Например, деление суток на «после полудня» и «до полудня» характеризуется бинарной переменной (напротив, время, измеряемое в часах, минутах и секундах, не является бинарным; оно может принимать намного больше двух значений).

Политологи и экономисты иногда говорят о *ресурсном проклятии*, или *парадоксе изобилия*. Идея состоит в том, что страны с обилием природных ресурсов часто менее экономически развиты и менее демократичны, чем страны с меньшим количеством природных ресурсов. Природные ресурсы могут сделать страну менее склонной инвестировать в другие формы развития или более склонной к насилию и автократии[1].

Чтобы убедиться в наличии или отсутствии ресурсного проклятия, нам придется оценить корреляцию между природными ресурсами и некоторыми особенностями экономической или политической системы. Этот процесс начинается со сбора данных, что мы и сделали. Чтобы измерить природные ресурсы, мы посмотрели, какие страны являются крупнейшими производителями нефти. Мы относим страну к крупному производителю нефти, если она экспортирует более сорока тысяч баррелей в день на миллион человек. Что касается политической системы, мы рассмотрели, какие страны считаются автократиями, а какие демократиями, по мнению авторов проекта Polity IV. В табл. 2.1 показано, сколько стран попадает в каждую из четырех возможных категорий: демократия и крупный производитель нефти, демократия и некрупный производитель нефти, автократия и крупный производитель нефти, автократия и некрупный производитель нефти.

Таблица 2.1. Объем добываемой нефти и тип политической системы

	Некрупный производитель	Крупный производитель	Итого
Демократия	118	9	127
Автократия	29	11	40
Итого	147	20	167

[1] Здесь и далее подобные оценочные суждения выражают личное мнение авторов цитируемых исследований и приведены исключительно в иллюстративных целях. – *Прим. издат.*

Мы можем выяснить, коррелируют ли эти две бинарные переменные (крупный/некрупный производитель и автократия/демократия), путем сравнения. Например, мы могли бы задаться вопросом, являются ли крупные производители нефти более склонными к автократии, чем страны, которые относятся к небольшим производителям. Или аналогичным образом мы могли бы задаться вопросом, являются ли автократии более крупными производителями нефти, чем демократии. Если одно из этих утверждений верно, то и другое должно быть верным. И эти сравнения говорят нам, имеют ли эти два свойства – крупный производитель нефти и автократия – тенденцию встречаться вместе.

В табл. 2.1 добыча нефти и автократия действительно положительно коррелируют. 55 % крупнейших производителей нефти являются автократиями (11/20 = 0.55), тогда как среди стран, не являющихся крупными производителями нефти, только около 20 % относятся к автократиям (29/147 ≈ 0.20). Аналогично крупными производителями нефти являются 27.5 % автократий (11/40 = 0.275) и лишь около 7 % демократий (9/127 ≈ 0.07). Другими словами, крупные производители нефти с большей вероятностью будут автократиями, чем страны, которые не являются крупными производителями нефти, и тогда, естественно, автократии с большей вероятностью будут крупными производителями нефти, чем демократии.

Эта положительная корреляция интересна с описательной точки зрения. Кроме того, она может пригодиться для прогнозирования. Предположим, что за пределами наших данных существуют другие страны, в политической системе правления которых мы не уверены. Знание того, являются ли они крупными производителями нефти, поможет предсказать наиболее вероятную форму правления.

В определенном смысле такие знания могут быть полезны для причинно-следственных выводов. Возможно, в какой-то стране обнаружены большие запасы нефти, и политологам интересно, какое влияние это может оказать на политическую систему страны. Однако, как будет подробно рассмотрено в главе 9, мы должны быть очень осторожны, давая корреляциям такого рода причинную интерпретацию.

Мы можем оценить корреляции, даже если имеющиеся данные не позволяют составить таблицу всех возможных комбинаций, как мы это сделали выше. Предположим, например, что мы хотим оценить взаимосвязь между преступностью и температурой в Чикаго. Мы могли бы составить электронную таблицу, в которой каждая строка соответствует дню, а каждый столбец – признаку каждого дня. Мы часто называем строки *наблюдениями*, а признаки, перечисленные в столбцах, – *переменными*. В этом случае наблюдения проводятся в разные дни. Первой переменной может быть средняя температура в тот день, измеренная в аэропорту Мидуэй. Второй переменной может быть количество преступлений, зарегистрированных в тот день во всем Чикаго. Третья переменная может указывать на то, была ли в тот день на первой полосе газеты Chicago Tribune статья о преступлениях. Как видите, переменные могут принимать бинарные значения (была статья или нет), дискретные, но не бинарные (количество преступлений) или непрерывные (средняя температура). Мы собрали подобные данные для Чикаго в 2018 г. и хотели бы оценить корреляцию между преступностью и температурой. Но как мы можем оценить корреляцию между двумя небинарными переменными?

Одной из отправных точек является построение простого графика, называемого *диаграммой рассеяния*. На рис. 2.1 показаны наши данные по Чикаго за 2018 г. Каждая точка соответствует одному наблюдению в наших данных; здесь это означает, что каждая точка – это день в Чикаго в 2018 г. Горизонтальная ось нашего рисунка – средняя температура в аэропорту Мидуэй в этот день. По вертикальной оси указано количество преступлений, зарегистрированных в городе в этот день. Таким образом, расположение каждой точки отражает среднюю температуру и уровень преступности в данный день.

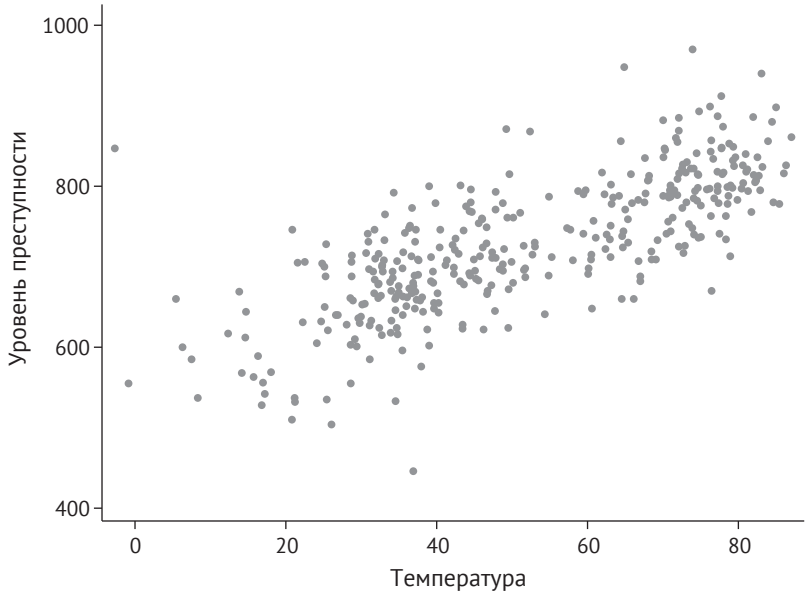


Рис. 2.1. Преступность и температура (в градусах по Фаренгейту) в Чикаго по дням в 2018 г.

Достаточно взглянуть на рисунок, чтобы сделать вывод о наличии положительной корреляции между температурой и преступностью. Точки в левой части диаграммы рассеяния (более холодные дни) также имеют тенденцию располагаться довольно низко по вертикальной оси (низкий уровень преступности), а точки в правой части диаграммы (более теплые дни) расположены довольно высоко по вертикальной оси (более высокий уровень преступности).

Но как нам количественно оценить это визуальное первое впечатление? На самом деле в количественной оценке корреляции можно задействовать различные статистические показатели. Один из таких показателей называется наклоном. Предположим, мы нашли линию, наилучшим образом отражающую расположение точек данных. Под *линией наилучшего соответствия* мы упрощенно подразумеваем линию, для которой среднее расстояние между точками данных и линией является минимальным. (Подробнее об этом мы поговорим в главе 5.) Наклон линии наилучшего соответствия – это один из способов описания корреляции между двумя непрерывными переменными.

На рис. 2.2 показана диаграмма рассеяния с добавленной линией наилучшего соответствия. Наклон линии говорит нам о характере взаимосвязи между этими двумя переменными. Если наклон отрицательный, корреляция отрицательная. Если наклон равен нулю, температура и преступность не коррелируют. Если наклон положительный, корреляция положительная. А крутизна наклона говорит нам о силе корреляции между этими двумя переменными. Здесь мы видим, что переменные положительно коррелируют: в теплые дни преступность обычно выше. В частности, наклон равен 3.1, т. е. в среднем на каждый дополнительный градус температуры (по Фаренгейту) приходится на 3.1 больше преступлений.

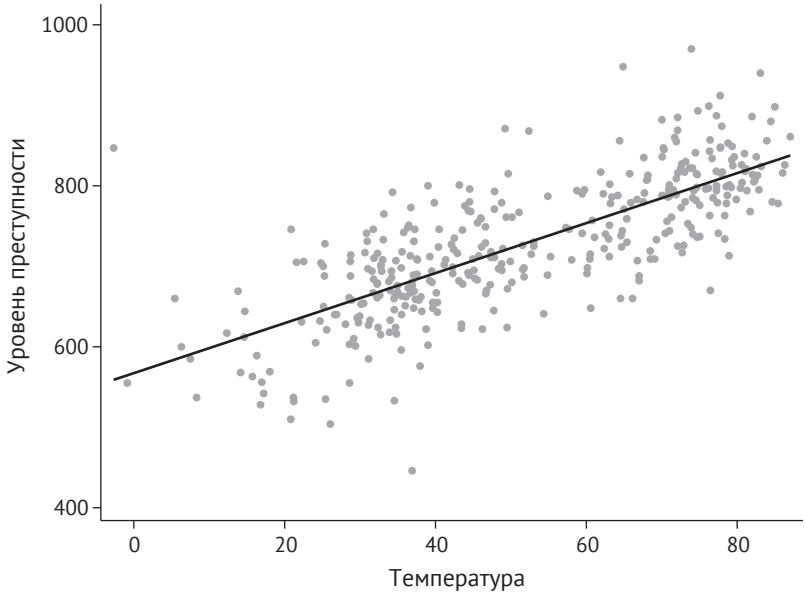


Рис. 2.2. Линия наилучшего соответствия, обобщающая взаимосвязь между преступностью и температурой (в градусах по Фаренгейту) в Чикаго по дням в 2018 г.

Обратите внимание: интерпретация наклона зависит от того, какая переменная находится на вертикальной оси, а какая – на горизонтальной. Если бы мы нарисовали график наоборот (как на рис. 2.3), мы бы описывали взаимосвязь между теми же двумя переменными. Но на этот раз мы бы узнали, что на каждое дополнительное зарегистрированное преступление температура в среднем на 0.18 °F выше. Знак наклона (положительный или отрицательный) не зависит от того, какая переменная находится на горизонтальной или вертикальной оси, поскольку перестановка переменных между осями не меняет их положительную или отрицательную корреляцию. Но числовое значение наклона и его содержательная интерпретация изменились.

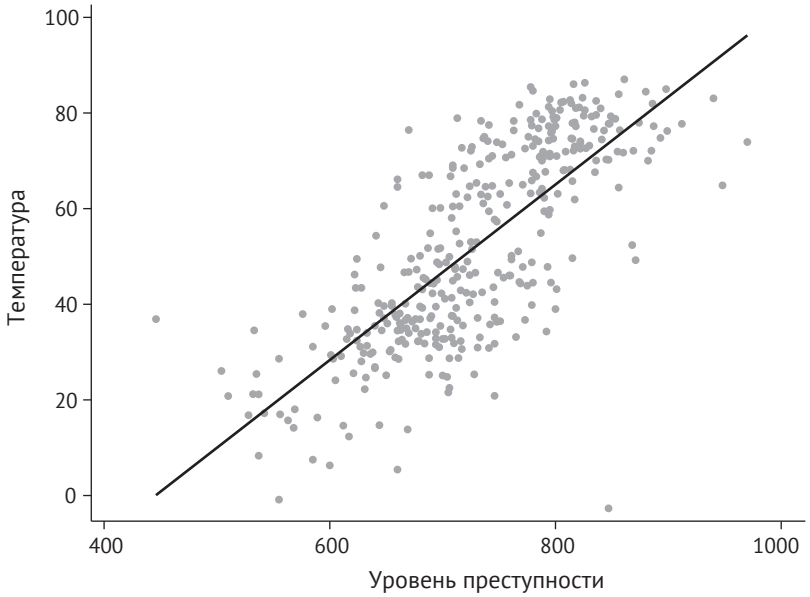


Рис. 2.3. Линия наилучшего соответствия, обобщающая взаимосвязь между температурой и преступностью в Чикаго по дням в 2018 г.

Факт или корреляция?

Чтобы установить, существует ли корреляция, мы всегда должны проводить какое-то сравнение. Например, чтобы узнать о корреляции между температурой и преступностью, нам нужно сравнить теплые и холодные дни и посмотреть, различаются ли уровни преступности, или, как вариант, мы можем сравнить дни с высоким и низким уровнем преступности, чтобы увидеть, различаются ли их температуры. Это означает, что для оценки корреляции между двумя переменными нам необходимо иметь вариации обеих переменных. Например, если бы мы собирали данные только в те дни, когда средняя температура составляла 0 °F, у нас не было бы возможности оценить корреляцию между температурой и преступностью. И то же самое верно, если бы мы рассматривали только дни, в которые зарегистрировано 500 преступлений.

Теперь давайте остановимся и проверим, насколько хорошо вы понимаете, что такое корреляция и как о ней можно узнать. Не волнуйтесь, если почувствуете неуверенность. Обнаружение истинной корреляции – непростая задача. Этой теме мы посвятим всю главу 4. Тем не менее полезно провести предварительную проверку знаний прямо сейчас. Итак, давайте попробуем.

Подумайте над следующими утверждениями. Какие из них описывают корреляцию, а какие – нет?

1. Люди, дожившие до 100 лет, обычно принимают витамины.
2. В городах с высоким уровнем преступности, как правило, нанимают больше полицейских на душу населения.

3. Успешные люди потратили не менее десяти тысяч часов на оттачивание своего мастерства.
4. Большинство политиков, столкнувшихся со скандалом, переизбираются на следующий срок.
5. Пожилые люди голосуют чаще, чем молодые.

Хотя каждое из этих утверждений сообщает о факте, не все они описывают корреляцию, т. е. свидетельство того, что два явления наблюдаемого мира имеют тенденцию проявляться вместе. В частности, утверждения 1, 3 и 4 не описывают корреляцию, а утверждения 2 и 5 – описывают. Давайте разберемся почему.

Утверждения 1, 3 и 4 являются *фактами*. Они исходят из данных. Они звучат научно. И если бы мы добавили к этим утверждениям конкретные цифры, мы могли бы назвать их *статистикой*. Но не все факты и статистика описывают корреляции. Ключевой момент заключается в том, что эти утверждения не описывают, склонны ли два явления мира возникать вместе, т. е. они не сравнивают разные значения двух переменных.

Чтобы лучше понять это, обратите внимание на утверждение 4:

«Большинство политиков, столкнувшихся со скандалом, переизбираются на второй срок».

Здесь фигурируют два явления. Первое – участие политика в скандале. Второе – победа политика на следующих выборах. Утверждение намекает на положительную корреляцию между участием в скандале и победой на выборах. Но на самом деле из этой констатации факта мы не можем узнать, имеют ли эти два явления тенденцию возникать вместе, т. е. мы не сравнивали частоту переизбрания тех, кто столкнулся со скандалом, с частотой переизбрания тех, кто не замешан в скандале. Да, мы могли бы оценить эту корреляцию, но только не с помощью данных, упомянутых в утверждении 4. Чтобы оценить корреляцию, нам нужны вариации обеих переменных – количества скандалов и побед на последующих выборах. Просто ради интереса давайте рассмотрим эту корреляцию на реальных данных о действующих членах Палаты представителей США, претендовавших на переизбрание в период с 2006 по 2012 г. Скотт Бейсингер из Университета Хьюстона систематически собирал данные о скандалах в конгрессе. Используя его данные, давайте посмотрим, сколько случаев попадает в четыре соответствующие категории: политики, которые столкнулись со скандалом и были переизбраны, политики, которые столкнулись со скандалом и не были переизбраны, политики без скандалов и переизбраны, политики без скандалов и не переизбраны.

Из табл. 2.2 мы видим, что утверждение 4 действительно является фактом: 62 из 70 (около 89 %) членов конгресса, столкнувшихся со скандалом и добивавшихся переизбрания, победили на выборах. Но мы также видим, что большинство членов конгресса не замешаны в скандалах и также добились переизбрания. Фактически были переизбраны 1192 из 1293 (около 92 %) политиков, которых не коснулись скандалы. Теперь мы видим, что на самом деле существует небольшая отрицательная корреляция между участием в скандале и победой на следующих выборах.

Таблица 2.2. Большинство членов конгресса, столкнувшихся со скандалом, побеждают на выборах, но между скандалом и переизбранием существует отрицательная корреляция

	Без скандала	Со скандалом	Итого
Не переизбраны	101	8	109
Переизбраны	1192	62	1254
Итого	1293	70	1363

Мы надеемся, что теперь понятно, почему утверждение 4 не содержит достаточно информации, чтобы понять, существует ли корреляция между скандалом и переизбранием. Проблема в том, что это утверждение касается только политиков, столкнувшихся со скандалом. Оно говорит нам лишь о том, что среди этой части политиков победителей больше, чем проигравших. Но чтобы выяснить, существует ли корреляция между скандалом и победой на выборах, нам нужно сравнить долю политиков, столкнувшихся со скандалом и выигравших переизбрание, с долей политиков, не сталкивавшихся со скандалом и тоже выигравших переизбрание. Если бы только 85 % членов конгресса, не замешанных в скандалах, выиграли переизбрание, между скандалом и переизбранием была бы положительная корреляция. Если бы выиграли 89 %, корреляции не было бы. Но, поскольку теперь мы знаем, что доля политиков, добившихся переизбрания и не столкнувшихся со скандалом, составляет 92 %, мы видим, что существует отрицательная корреляция. Аналогичный анализ покажет, что утверждения 1 и 3 сами по себе также не несут достаточно информации для оценки корреляции.

Утверждения 2 и 5 действительно описывают корреляции. Обратите внимание, что оба они содержат сравнение. Утверждение 2 говорит нам, что в городах с более высоким уровнем преступности в среднем более крупные полицейские силы, чем в городах с меньшим уровнем преступности. А утверждение 5 говорит, что пожилые люди (большего возраста), как правило, голосуют чаще, чем молодые (меньшего возраста). В обоих случаях мы сравниваем различия в одной переменной (численность полиции или уровень голосования) с различиями в другой переменной (уровень преступности или возраст). Это та информация, которая вам нужна для установления корреляции.

Как мы говорили вначале, не волнуйтесь, если почувствуете замешательство. Иногда бывает сложно понять, какая информация необходима для установления корреляции, а не просто факта. Мы посвятим четвертую главу тому, чтобы убедиться, что вы действительно все поняли.

Для чего нужна корреляция?

Теперь, когда у вас есть общее понимание того, что такое корреляция, давайте поговорим о том, для чего она нужна. Мы отметили, что корреляции – пожалуй, самый важный инструмент количественного анализа. Но почему? В общих чертах – потому что корреляция позволяет нам предсказать изменение какого-либо показателя или свойства на основании известных изменений других показателей или свойств.

Есть как минимум три варианта использования такого рода знаний: (1) описание, (2) прогнозирование и (3) причинно-следственная связь. Каждый раз, когда вы собираетесь использовать корреляцию, следует четко понимать, какую из этих трех задач вы пытаетесь решить и какими должны быть достоверные знания о мире, чтобы корреляция была применима в конкретных условиях.

Описание

Описание отношений между свойствами или признаками объектов мира – самый простой способ использования корреляций.

Почему у нас может возникнуть необходимость в описании взаимосвязи между свойствами объектов? Предположим, вы опасаетесь, что молодые люди недостаточно представлены в опросах на конкретных выборах относительно их доли в населении. В таком случае может пригодиться описание взаимосвязи между возрастом и голосованием. На рис. 2.4 показана диаграмма рассеяния данных о возрасте и средней явке избирателей на выборах в Конгресс США в 2014 г. На этом рисунке наблюдение представляет собой возрастную когорту. Для каждого возраста на вертикальной оси показана доля избирателей, имеющих право голоса и принявших участие в голосовании.

На рисунке также изображена линия наилучшего соответствия данным. Эта линия имеет наклон 0.006. Другими словами, в среднем на каждый дополнительный год возраста вероятность того, что человек проголосовал в 2014 г., увеличивается на 0.6 процентных пункта. Молодые люди действительно выглядят недостаточно представленными на выборах, поскольку их показатели ниже, чем у пожилых людей.

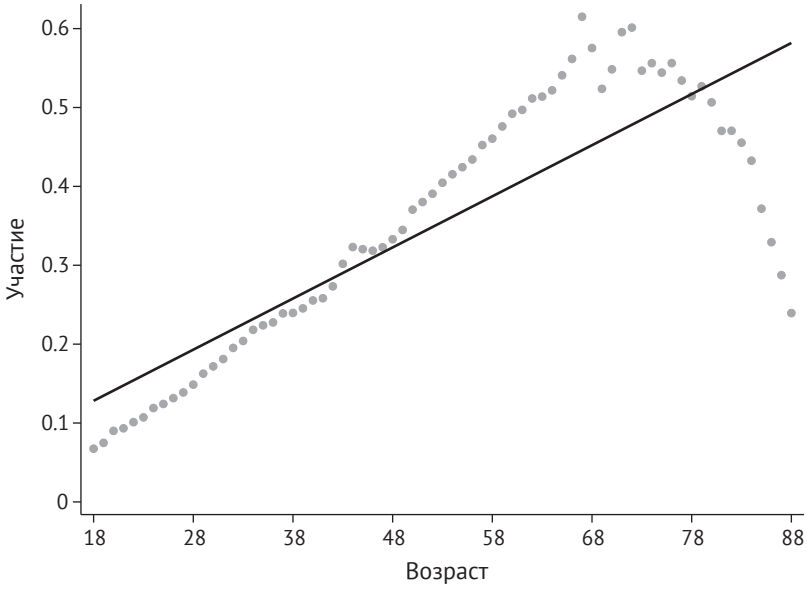


Рис. 2.4. Явка и возраст избирателей на выборах 2014 г.

Этот вид описательного анализа может быть интересен сам по себе. Важно знать, что молодые люди с меньшей вероятностью голосовали в 2014 г., чем по-

жилые, и поэтому были недостаточно представлены в избирательном процессе. Эти отношения могут повлиять на то, как вы рассматриваете исход данных выборов. Более того, знание этой корреляции может побудить вас к дальнейшему изучению причин и последствий феномена низкой рождаемости среди поколения более молодых людей.

Конечно, эта описательная связь не обязательно означает, что молодые люди продолжат меньше голосовать на предстоящих выборах. Поэтому вы вряд ли сможете использовать эти знания для прогнозирования будущей явки избирателей. И это также не означает, что с возрастом бывшие молодые люди обязательно начнут чаще голосовать. Поэтому здесь вряд ли получится обнаружить причинно-следственную связь. Этот описательный анализ всего лишь говорит нам, что на выборах 2014 г. пожилые избиратели в среднем голосовали с большей вероятностью, чем молодые. Чтобы развить интерпретацию дальше, вам нужно сделать более сильные предположения о мире, который мы сейчас исследуем.

Прогнозирование

Другой причиной, побуждающей нас изучать корреляцию, является *прогнозирование* или *предсказание* – два термина, которые мы будем использовать как синонимы. Прогнозирование предполагает использование информации из некоторой выборочной совокупности для прогнозирования другой совокупности.

Например, вы можете использовать данные об избирателях на прошлых выборах, чтобы делать прогнозы об избирателях на будущих выборах. Или вы можете использовать данные избирателей в одном штате, чтобы делать прогнозы об избирателях в другом штате. Предположим, вы проводите избирательную кампанию, у вас ограниченные ресурсы и вы пытаетесь выяснить, на кого из ваших сторонников следует нацелиться, стуча в дверь и напоминая им, что пора идти проголосовать. Если вы заранее уверены в том, что человек пойдет голосовать без вашего вмешательства, незачем тратить время своих волонтеров на посещение его дома. Таким образом, точное прогнозирование явки избирателей может повысить эффективность вашей кампании.

Для такого рода прогнозирования могут быть полезны корреляции, подобные приведенной выше связи возраста и явки избирателей. Поскольку возраст сильно коррелирует с явкой, он может быть полезной переменной для прогнозирования того, кто намерен проголосовать, а кто – нет. Например, если вы можете предсказать на основе возраста, что некоторая группа избирателей практически наверняка явится на выборы даже без усилий ваших волонтеров, можете сосредоточить свои мобилизационные ресурсы на других избирателях.

Чтобы использовать таким способом корреляцию между возрастом и явкой избирателей для прогнозирования, вам незачем знать, *почему* они коррелируют. Но, в отличие от случая, когда вы просто хотите описать взаимосвязь между возрастом и явкой избирателей на выборах 2014 г., если вы хотите делать прогнозы, нужно быть готовым сделать некоторые дополнительные предположения о мире.

Это поднимает две важные проблемы, о которых вы должны хорошенько подумать, чтобы ответственно использовать корреляцию для прогнозирования. Во-первых, является ли взаимосвязь, которую вы обнаружили в своей выборке, отражением более широкого явления, или она является результатом случайных изменений в ваших данных? Ответ на этот вопрос требует статистического

вывода, который будет рассмотрен в главе 6. Во-вторых, даже если вы убеждены, что обнаружили реальную взаимосвязь в своей выборке, следует подумать о том, является ли ваша выборка репрезентативной для генеральной совокупности, относительно которой вы пытаетесь сделать прогнозы. Мы рассмотрим репрезентативность более подробно при обсуждении выборок и внешней валидности в главах 6 и 16.

Давайте вернемся к использованию информации о возрасте и явке избирателей на одних выборах, чтобы сделать прогнозы насчет следующих выборов. Это имеет смысл только в том случае, если можно предположить, что взаимосвязь между этими двумя переменными не меняется слишком быстро. То есть корреляция между возрастом и явкой избирателей, например, на выборах 2014 г. будет полезна для выяснения того, на каких избирателей ориентироваться на выборах 2016 г., если кажется вероятным, что связь между возрастом и явкой в 2016 г. будет приблизительно такая же, как в 2014 г. Аналогично, если бы у вас были данные только о возрасте и явке избирателей на выборах 2014 г. в 25 штатах, вы могли бы использовать корреляцию между возрастом и явкой в этих штатах, чтобы получить информацию стратегии в остальных 25 штатах. Но это было бы разумно только при наличии оснований полагать, что взаимосвязь между возрастом и явкой в остальных штатах будет приблизительно такой же.

Соответственно, при использовании для прогнозирования некоторых статистических данных, таких как наклон линии наилучшего соответствия, нам нужно подумать о том, является ли зависимость на самом деле линейной. В противном случае линейное представление отношений может ввести в заблуждение. Мы обсудим это более подробно ниже.

Стоит отметить, что на практике было бы странно пытаться делать прогнозы на основании корреляции всего лишь между двумя переменными. Более разумно попытаться предсказать явку избирателей, используя ее связь с множеством переменных, таких как пол, раса, доход, образование и явка на предыдущие выборы. Мы обсудим такие многомерные и условные корреляции в главе 5.

Использование данных для прогнозирования и предсказания – быстро растущая область деятельности аналитиков в политике, бизнесе, полиции, спорте, правительстве, разведке и многих других областях. Например, предположим, что вы руководите отделом общественного здравоохранения вашего города. Каждый раз, когда вы отправляете санитарного инспектора в ресторан, это стоит времени и денег. Но нарушения санитарных норм в ресторанах наносят вред жителям вашего города. Поэтому вам бы очень хотелось направить инспекторов именно в те рестораны, которые с наибольшей вероятностью нарушают санитарные нормы, чтобы не тратить время и деньги на проверки, которые в конечном итоге не улучшат общественную безопасность. Чем точнее вы сможете спрогнозировать, какие рестораны нарушают правила, тем эффективнее вы сможете задействовать своих инспекторов. Допустим, можно использовать данные о ресторанах, которые нарушали и не нарушали нормы здравоохранения в прошлом, чтобы попытаться предсказать такие нарушения на основе их корреляции с другими наблюдаемыми характеристиками[1] ресторана. Вероятно,

[1] В машинном обучении и анализе данных такие характеристики называют *признаками* (feature). – *Прим. перев.*

полезные признаки ресторана могут включать обзоры Yelp, информацию о посещениях больниц при пищевых отравлениях, местоположении, ценах и т. д. Затем, имея на руках эти корреляции, вы можете использовать будущие обзоры Yelp и другую информацию, чтобы предсказать, какие рестораны, скорее всего, нарушают нормы здравоохранения, и направить в эти рестораны проверку.

Этот пример указывает еще на одну сложную проблему. Сам факт использования корреляций для прогнозирования иногда может привести к тому, что корреляции, существовавшие в прошлом, перестанут действовать в будущем. Например, предположим, что департамент здравоохранения наблюдает сильную корреляцию между ресторанами, открытыми 24 часа в сутки, и нарушениями санитарных норм. На основе этой корреляции они могут начать непропорционально часто отправлять инспекторов здравоохранения в круглосуточные рестораны. Наблюдательный владелец ресторана, заметивший эту закономерность, может приспособиться, чтобы обмануть департамент здравоохранения, – скажем, закрывать свой ресторан с 2:00 до 3:00 каждую ночь. Это небольшое изменение в часах работы вряд ли поможет навести порядок в ресторане. Однако оно позволяет вывести ресторан из категории круглосуточных и тем самым обмануть систему, основанную на корреляции. Мы обсудим эту общую проблему адаптации более подробно в главе 16.

Прогнозирование также пригодится политику, который хотел бы знать ожидаемую продолжительность экономического спада для планирования госбюджета, банкиру, который хочет знать кредитоспособность потенциальных заемщиков, или страховой компании, желающей знать, какова вероятность попадания конкретного клиента в аварию. Менеджеры наших любимых Chicago Bears хотели бы предсказать, какие футболисты колледжа смогут увеличить шансы команды на победу в Суперкубке. Но, учитывая их прошлый послужной список, мы не питаем особых надежд. Данные не могут творить чудеса.

Также стоит подумать о потенциальных этических последствиях использования прогнозов для управления поведением. Например, исследования показывают, что жалобы потребителей на чистоту в онлайн-обзорах ресторанов положительно коррелируют с нарушениями санитарного кодекса. Это потенциально полезная прогностическая информация: правительственные органы могли бы использовать данные, собранные с обзорных сайтов, чтобы выяснить, куда направить проверяющих. Основываясь на этом предположении, статья в The Atlantic заявила: «Yelp может навести порядок в ресторанной индустрии». Но исследование Кристен Альтенбургер и Дэниела Хо показывает, что онлайн-рецензенты предвзято относятся к азиатским ресторанам: сравнивая рестораны, получившие одинаковую оценку от инспекторов по безопасности общественного питания, они обнаружили, что рецензенты с большей вероятностью жалуются на чистоту в азиатских ресторанах. Это означает, что, если правительства будут использовать прогнозирующую корреляцию между онлайн-обзорами и нарушениями санитарного кодекса, это приведет к непреднамеренной дискриминации азиатских ресторанов, непропорционально часто подвергая их проверкам. Вы действительно хотите, чтобы ваше правительство использовало такую информацию? Или этические издержки проверки ресторанов перевешивают преимущества более точных прогнозов? Мы вернемся к некоторым из этих этических вопросов в конце книги.

Причинный вывод

Еще одна причина, по которой нас могут заинтересовать корреляции, – это изучение причинно-следственных связей. Многие из наиболее интересных вопросов, с которыми сталкиваются количественные аналитики, по своей сути являются причинно-следственными. Это вопросы о том, как изменение какого-то признака мира повлечет за собой изменение другого признака мира. Приведет ли снижение стоимости обучения в колледже к уменьшению неравенства доходов? Поможет ли введение всеобщего базового дохода уменьшить количество бездомных? Повысит ли новая маркетинговая стратегия прибыль? Это вопросы, относящиеся к причинно-следственным связям. Как мы увидим на протяжении всей книги, использование корреляций для вывода о причинно-следственных связях является обычным явлением. Но на этом пути кроется множество ловушек для рационального мышления. (Понимание причинности станет темой следующей главы.)

Использование корреляции для причинных выводов сопряжено со всеми потенциальными проблемами, которые мы только что обсуждали, когда говорили об использовании корреляции для прогнозирования, но есть и специфические проблемы. Главная из них заключается в том, что *корреляция не обязательно подразумевает причинно-следственную связь*. То есть корреляция между двумя признаками мира не означает, что одна из них является причиной другой.

Предположим, вас интересует, как школьное обучение математике повлияет на последующую успеваемость в колледже. Это важный вопрос, если вы учащийся средней школы, родитель или репетитор старшеклассника или политик, устанавливающий образовательные стандарты. Насколько больше будет вероятность, что старшеклассники поступят в колледж и благополучно закончат его, если будут изучать углубленную математику в старшей школе?

Как оказалось, корреляция между углубленным изучением математики и окончанием колледжа положительна и довольно сильна – например, люди, которые изучают математический анализ в средней школе, имеют гораздо больше шансов окончить колледж, чем те, кто этого не делает. И корреляция еще сильнее для общей алгебры, тригонометрии и углубленного математического анализа. Но это не значит, что занятия математическим анализом заставят студентов закончить колледж.

Конечно, одним из возможных источников этой корреляции является то, что математический анализ действительно готовит студентов к поступлению в колледж и повышает вероятность его окончания. Но это не единственный возможный источник корреляции. Например, возможно, в среднем дети, которые изучают математический анализ, более академически мотивированы, чем дети, которые этого не делают. И возможно, мотивированные дети с большей вероятностью закончат колледж независимо от того, изучают ли они математический анализ в старшей школе или нет. Если это так, мы увидим положительную корреляцию между изучением математического анализа и окончанием колледжа, даже если математический анализ сам по себе не влияет на окончание колледжа. Скорее, тот факт, что студент изучал в школе математический анализ, был бы просто косвенным показателем мотивации, которая коррелирует с окончанием колледжа.

Что здесь поставлено на карту? Получается, если причинный вывод верен, то изучение основ математического анализа в школе поможет окончить колледж тем студентам, которые в противном случае не доучились бы до конца. Но если все дело в мотивации, то изучение математического анализа в школе не поможет окончить колледж. В этой истории матанализ – всего лишь *индикатор* мотивации. Если заставить школьника заниматься математическим анализом, это не делает его волшебным образом более мотивированным. Может даже оказаться, что требование заниматься математическим анализом может повлечь за собой реальные затраты – с точки зрения самооценки, мотивации или времени, затрачиваемого на другие виды деятельности – без каких-либо компенсирующих выгод.

Ошибка, которую мы только что описали, была допущена в рецензируемой научной статье. Исследователи сравнили успеваемость в колледже людей, которые посещали и не посещали различные интенсивные курсы математики в средней школе. На основании положительной корреляции они предложили школьным методистам «использовать результаты этого исследования, чтобы информировать учащихся, их родителей и опекунов о важной роли, которую школьные курсы математики играют в последующем получении степени бакалавра». То есть они приняли корреляцию за причинно-следственную связь. На основании этих корреляций они рекомендовали студентам, которые в противном случае не планировали этого делать, записаться на интенсивные курсы математики, чтобы увеличить свои шансы на окончание колледжа.

Мы вернемся к проблеме ошибочного принятия корреляции за причинно-следственную связь в части III. А пока вам следует запомнить, что в целом неправильно делать вывод о причинно-следственной связи на основе корреляций, хотя многие эксперты занимаются этим постоянно.

Измерение корреляций

Существует несколько общих разновидностей статистических данных, которые можно использовать для описания и измерения корреляции между переменными. Здесь мы обсудим три из них: *ковариацию*, *коэффициент корреляции* и *наклон линии регрессии*. Но, прежде чем перейти к этим трем различным способам измерения корреляций, нам нужно поговорить о средних значениях, дисперсиях и стандартных отклонениях – статистических показателях, которые помогают понимать переменные.

Среднее значение, дисперсия и стандартное отклонение

Давайте вернемся к нашим данным о преступности и температуре в Чикаго. Напомним, что в этом наборе данных каждое наблюдение соответствует дню в 2018 г. И для каждого дня мы наблюдаем две переменные: количество зарегистрированных преступлений и среднюю температуру, измеренную в градусах по Фаренгейту в аэропорту Мидуэй. Мы не будем воспроизводить здесь весь набор данных, поскольку он состоит из 365 строк (по одной на каждый день 2018 г.). В табл. 2.3 показано, как выглядят данные за январь. В оставшейся части обсуждения мы будем рассматривать дни января 2018 г. как интересующую нас совокупность данных.

Таблица 2.3. Средняя температура в аэропорту Чикаго Мидуэй и количество преступлений, зарегистрированных в Чикаго, за каждый день января 2018 г.

День	Температура (°F)	Преступления
1	–2.7	847
2	–0.9	555
3	14.2	568
4	6.3	600
5	5.4	660
6	7.5	585
7	25.4	535
8	33.9	618
9	30.1	653
10	44.9	709
11	51.7	698
12	21.6	705
13	12.3	617
14	15.7	563
15	16.8	528
16	14.6	612
17	14.7	644
18	25.6	621
19	34.8	707
20	40.4	724
21	42.9	716
22	48.9	722
23	32.3	716
24	29.2	610
25	35.5	640
26	46.0	759
27	45.6	754
28	35.0	668

Окончание табл. 2.3

День	Температура (°F)	Преступления
29	25.2	650
30	24.7	632
31	37.6	708
Среднее	26.3	655.6
Дисперсия	220.3	5183.0
Стандартное отклонение	14.8	72.0

Для наблюдений i примем обозначение переменной преступности crime_i и переменной температуры temperature_i. В нашей таблице данных i может принимать любое значение от 1 до 31, соответствующее 31 дню января 2018 г. Так, например, 13 января воздух прогрелся до $\text{temperature}_{13} = 12.3$, а количество преступлений, зарегистрированных 24 января, составляло $\text{crime}_{24} = 610$.

Переменная также характеризуется *распределением* – описанием частоты, с которой она принимает разные значения. Нам часто бывает нужно иметь возможность обобщить распределение переменной с помощью нескольких ключевых статистических показателей. Здесь мы поговорим о трех из них.

Нам потребуются некоторые обозначения. Символ Σ (заглавная греческая буква сигма) обозначает суммирование. Например, $\sum_{i=1}^{31} \text{crime}_i$ – это сумма всех значений переменной crime с 1-го по 31-й день. Чтобы найти ее, мы берем значения преступности за 1-й, 2-й, 3-й день и т. д. до 31-го дня и складываем их вместе. То есть мы суммируем $\text{crime}_1 = 847$, $\text{crime}_2 = 555$ и $\text{crime}_3 = 568$ и т. д. вплоть до $\text{crime}_{31} = 708$. Конкретные значения для переменной преступности в определенный день извлекаются из данных в табл. 2.3.

Теперь можно вычислить *среднее значение* распределения каждой переменной. (Иногда его называют просто средним значением переменной, без упоминания распределения.) Среднее значение обозначается греческой буквой μ (мю). Среднее значение распределения – это всего лишь известное вам со школы арифметическое среднее значение. Мы находим его, суммируя значения наблюдений (для которых у нас теперь есть удобные обозначения) и деля сумму на количество наблюдений. На январь 2018 г. средние значения наших двух переменных равны

$$\mu_{\text{crime}} = \frac{\sum_{i=1}^{31} \text{crime}_i}{31} = \frac{847 + 555 + \cdots + 708}{31} = 655.6$$

и

$$\mu_{\text{temperature}} = \frac{\sum_{i=1}^{31} \text{temperature}_i}{31} = \frac{-2.7 + -0.9 + \cdots + 37.6}{31} = 26.3.$$

Второй заслуживающий внимания показатель – это дисперсия, которую мы обозначаем σ^2 (строчная греческая буква сигма). Чуть позже вы узнаете, поче-

му оно возведено в квадрат. *Дисперсия* – это способ измерения того, насколько далеки от среднего значения отдельные значения переменной. Можно даже сказать, что дисперсия измеряет, насколько изменчива переменная. (Вы также можете приближенно рассматривать это понятие как меру ширины распределения переменной.)

Дисперсию рассчитывают следующим образом. Предположим, у нас есть некоторая переменная X (например, crime или temperature). Для каждого наблюдения вычислите отклонение значения X этого наблюдения от среднего значения X. Например, для наблюдения i отклонение равно значению X для наблюдения i (X_i) минус среднее значение X по всем наблюдениям (μX), т. е. $X_i - \mu X$. 13 января 2018 г. температура составила 12.3 °F. Средняя температура в январе 2018 г. составила 26.3 °F. Следовательно, 13 января отклонение от январского среднего значения составило 12.3 – 26.3 = –14. То есть 13 января 2018 г. было на 14 °F холоднее, чем в среднем в январе 2018 г. Напротив, отклонение 23 января 2018 г. составило 32.3 – 263 = 6. Следовательно, 23 января было на 6 °F теплее, чем в среднем в этом месяце.

Обратите внимание, что эти отклонения могут быть положительными или отрицательными, поскольку наблюдения могут быть больше или меньше среднего значения. Но для измерения изменчивости наблюдений не имеет значения, является ли данное отклонение положительным или отрицательным. Мы просто хотим знать, насколько далеко каждое наблюдение находится от среднего значения в любом направлении. Значит, нам нужно преобразовать отклонения в положительные числа, которые измеряют только расстояние от среднего значения без учета знака. Для этого можно было бы взять абсолютное значение отклонений. Но по причинам, которые мы обсудим позже, обычно отклонения делают положительными, возводя их в квадрат. Дисперсия представляет собой среднее значение этих квадратов отклонений. Итак, если имеется N наблюдений (в нашем случае $N = 31$), дисперсия равна

$$\sigma_X^2 = \frac{\sum_i^N (X_i - \mu_X)^2}{N}.$$

Для двух переменных в наших данных дисперсии равны

$$\sigma^2_{\text{crime}} = \frac{\sum_{i=1}^{31} (\text{crime}_i - \mu_{\text{crime}})^2}{31}$$

$$= \frac{(847 - 655.6)^2 + (555 - 655.6)^2 + \cdots + (708 - 655.6)^2}{31} \approx 5183$$

и

$$\sigma^2_{\text{temperature}} = \frac{\sum_{i=1}^{31} (\text{temperature}_i - \mu_{\text{temperature}})^2}{31}$$

$$= \frac{(-2.7 - 26.3)^2 + (-0.9 - 26.3)^2 + \cdots + (37.6 - 26.3)^2}{31} \approx 220.3.$$

Благодаря тому, что мы рассматриваем квадраты отклонений, а не среднее абсолютного значения отклонений, дисперсия придает больший вес наблюдениям, которые находятся дальше от среднего значения. Если самый богатый человек в обществе становится богаче, это увеличивает дисперсию богатства больше, чем если бы умеренно богатый человек стал богаче на ту же сумму. Например, предположим, что среднее богатство равно 1. Если кто-то с богатством 10 получает еще одну единицу богатства, дисперсия увеличивается на $(10^2 - 9^2)/N = 19/N$. Но если кто-то с богатством 100 получит еще одну единицу богатства, дисперсия увеличится на $(100^2 - 99^2)/N = 199/N$.

Дисперсия является точной мерой того, насколько изменчивой является переменная. Но, поскольку мы все возвели в квадрат, в некотором смысле она не выражается в тех же единицах измерения, что и переменная. Иногда нам нужна мера изменчивости на той же шкале, что и сама переменная. В этом случае мы используем *стандартное отклонение*, которое представляет собой квадратный корень дисперсии. Обозначим стандартное отклонение через σ (строчная греческая буква сигма):

$$\sigma_X = \sqrt{\sigma_X^2} = \sqrt{\frac{\sum_i^N (X_i - \mu_X)^2}{N}}.$$

Стандартное отклонение, которое также является мерой того, насколько широк разброс распределения переменной, примерно соответствует тому, насколько далеко, по нашим ожиданиям, наблюдения будут находиться от среднего значения. Однако, как мы уже отмечали, по сравнению со средним абсолютным значением отклонений этот критерий придает дополнительный вес наблюдениям, которые находятся дальше от среднего значения.

Для двух переменных в наших данных стандартные отклонения равны

$$\begin{aligned}\sigma_{\text{crime}} &= \sqrt{\frac{\sum_{i=1}^{31} (\text{crime}_i - \mu_{\text{crime}})^2}{31}} \\ &= \sqrt{\frac{(847 - 655.6)^2 + (555 - 655.6)^2 + \cdots + (708 - 655.6)^2}{31}} \approx 72\end{aligned}$$

и

$$\begin{aligned}\sigma_{\text{temperature}} &= \sqrt{\frac{\sum_{i=1}^{31} (\text{temperature}_i - \mu_{\text{temperature}})^2}{31}} \\ &= \sqrt{\frac{(-2.7 - 26.3)^2 + (-0.9 - 26.3)^2 + \cdots + (37.6 - 26.3)^2}{31}} \approx 15.1.\end{aligned}$$

Теперь, когда вы знаете, что такое среднее значение, дисперсия и стандартное отклонение, мы можем обсудить три важных способа измерения корреляций: *ковариацию*, *коэффициент корреляции* и *наклон линии регрессии*.

Ковариация

Предположим, у нас есть две переменные, такие как crime и temperature, и мы хотим измерить корреляцию между ними. Один из способов сделать это – вычислить их *ковариацию* (обозначаемую cov). Чтобы упростить обозначения, назовем эти две переменные X и Y. Предположим, что у нас есть совокупность размером N.

Ковариацию находят так. Для каждого наблюдения рассчитайте отклонения – т. е. насколько далеко значение X находится от среднего значения X и насколько далеко значение Y находится от среднего значения Y. Теперь для каждого наблюдения перемножьте отклонения между собой, получив, таким образом, произведения отклонений $(X_i - \mu_X)(Y_i - \mu_Y)$ для каждого наблюдения i. Наконец, чтобы найти ковариацию X и Y, вычислите среднее значение этих произведений:

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^{N}(X_i - \mu_X)(Y_i - \mu_Y)}{N}.$$

Давайте убедимся, что ковариация является мерой корреляции. Рассмотрим особенно сильную версию положительной корреляции: предположим, что всякий раз, когда X больше среднего $(X_i - \mu_X > 0)$, Y также больше среднего $(Y_i - \mu_Y > 0)$ и всякий раз, когда X меньше среднего $(X_i - \mu_X < 0)$, Y также меньше среднего $(Y_i - \mu_Y < 0)$. В этом случае произведение отклонений будет положительным для каждого наблюдения – либо оба отклонения будут положительными, либо отрицательными. Таким образом, ковариация будет положительной, отражая положительную корреляцию. Теперь рассмотрим особенно сильную версию отрицательной корреляции: предположим, что всякий раз, когда X больше среднего, Y меньше среднего, а всякий раз, когда X меньше среднего, Y больше среднего. В этом случае произведение отклонений будет отрицательным для каждого наблюдения, поскольку одно отклонение всегда отрицательное, а другое – положительное. Таким образом, ковариация будет отрицательной, отражая отрицательную корреляцию. Конечно, ни один из этих крайних случаев не является обязательным. Но если переменная X, превышающая среднее значение, обычно сочетается с переменной Y, тоже превышающей среднее значение, то ковариация будет положительной, отражая положительную корреляцию. Если X больше среднего обычно сочетается с Y меньше среднего, то ковариация будет отрицательной, отражая отрицательную корреляцию. А если значения X и Y не связаны друг с другом, ковариация будет равна нулю, что отражает тот факт, что переменные не коррелируют.

Коэффициент корреляции

Со знаком ковариации все ясно, но интерпретация ее величины может вызвать затруднения, поскольку произведение отклонений зависит от того, насколько изменчивы переменные. Мы можем получить более легко интерпретируемый статистический показатель, который по-прежнему измеряет корреляцию, учитывая дисперсию переменных.

Коэффициент корреляции (обозначаемый как corr) – это просто ковариация, деленная на произведение стандартных отклонений:

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}.$$

Когда мы делим ковариацию на произведение стандартных отклонений, мы выполняем *нормирование*. То есть ковариация в принципе может принимать любое значение. Но коэффициент корреляции всегда принимает значение от –1 до 1. Значение 0 по-прежнему указывает на отсутствие корреляции. Значение 1 указывает на положительную корреляцию и идеальную линейную зависимость – если вы построите диаграмму рассеяния двух переменных, то сможете провести через все точки направленную вверх прямую линию. Значение –1 указывает на отрицательную корреляцию и идеальную линейную зависимость. Значение от 0 до 1 указывает на положительную корреляцию, но не идеальную линейную зависимость. А значение от –1 до 0 указывает на отрицательную корреляцию, но не идеальную линейную зависимость.

Коэффициент корреляции часто обозначается буквой r. Иногда коэффициент корреляции возводят в квадрат, чтобы вычислить показатель r^2. Его значение всегда находится между 0 и 1.

Одной из потенциально привлекательных особенностей показателя r^2 является то, что его можно интерпретировать как пропорцию. Его часто интерпретируют как долю изменчивости Y, объясняемую X, или, что то же самое, долю X, объясняемую Y. Как мы обсудим в последующих главах, слово «объясняемый» здесь может ввести в заблуждение. Оно *не* означает, что изменение X вызывает изменение Y или наоборот. Оно также не учитывает возможность того, что наблюдаемая корреляция могла возникнуть случайно.

Наклон линии регрессии

Одна из потенциальных проблем, связанных с коэффициентом корреляции и показателем r^2, заключается в том, что они ничего не говорят вам о существенной важности или размере связи между X и Y. Предположим, что нас интересуют две переменные – преступность и температура в Чикаго. Коэффициент корреляции 0.8 говорит нам о том, что между двумя переменными существует сильная положительная связь, но не говорит нам, в чем она заключается. Возможно, каждый градус температуры добавляет 0.1 дополнительного преступления, а может быть, каждый градус температуры соответствует 100 дополнительным преступлениям. И то и другое возможно при коэффициенте корреляции 0.8, но разница между этими ситуациями очень велика.

По этой причине мы не тратим много времени на размышления о вышеупомянутых способах измерения корреляции. Как мы уже говорили, основное внимание сосредоточено на наклоне линии наилучшего соответствия. Более того, имеет смысл сосредоточиться на одном конкретном способе определения линии наилучшего соответствия. Помните, что линия наилучшего соответствия минимизирует среднее расстояние между точками данных и линией. Обычно измеряют квадрат расстояния от точки данных до линии (поэтому каждое значение является положительным, как и в случае возведения в квадрат отклонений) и ищут линию наилучшего соответствия, которая минимизирует сумму этих квадратов расстояний (или *сумму квадратов ошибок*). Эта конкретная

линия наилучшего соответствия называется линией регрессии наименьших квадратов (ordinary least squares, OLS), и обычно, когда кто-то просто говорит «линия регрессии», он имеет в виду линию регрессии OLS. Все линии наилучшего соответствия, которые мы упоминали ранее в этой главе, были линиями регрессии OLS.

Оказывается, наклон линии регрессии можно рассчитать с помощью ковариации и дисперсии. Наклон линии регрессии (также иногда называемый *коэффициентом регрессии*), когда значение Y находится на вертикальной оси, а X – на горизонтальной оси, равен

$$\frac{\text{cov}(X, Y)}{\sigma_X^2}.$$

Это число наглядно показывает нам, насколько в среднем изменяется Y при увеличении X на одну единицу. Если поместить в знаменатель σ_Y^2 вместо σ_X^2, то полученный коэффициент покажет нам, на сколько в среднем изменяется X увеличении Y на одну единицу. Как вы уже видели, это могут быть разные числа.

Линиям регрессии будут посвящены главы 5 и 10.

Совокупности и выборки

Прежде чем двигаться дальше, рассмотрим еще один вопрос. Каждый статистический показатель, о котором мы говорили, – среднее значение, дисперсию, ковариацию, коэффициент корреляции, наклон линии регрессии, – можно рассматривать двояко. Для каждого из этих статистических показателей существует значение, которое соответствует *всей* интересующей нас совокупности. И есть значение этого показателя, которое соответствует *выборке* имеющихся у нас данных. Эти значения могут существенно различаться. До сих пор мы избегали этой проблемы, сосредоточившись на случае, когда выборка и совокупность совпадают, – у нас есть преступность и температура для каждого дня в январе 2018 г. Но так будет не всегда. Например, нас могла бы заинтересовать взаимосвязь между преступностью и температурой в январе на протяжении многих лет, но у нас имеется только выборка данных за 2018 г. Это породило бы всевозможные вопросы о том, что мы можем узнать о январе 2019 г. по данным за 2018 г. Вернемся к этим вопросам в главе 6.

Откровенно о линейности

Все способы измерения корреляции, которые мы обсуждали, сосредоточены на оценке линейных связей между переменными. Мы углубимся в эту тему позже, особенно в главе 5, когда вернемся к теме возраста и явки избирателей в контексте нашего обсуждения регрессии. А пока отметим, что линейные связи зачастую интересны и важны, но не все интересные и важные связи линейны. Рассмотрим, например, две возможные связи между переменными X и Y, показанные на рис. 2.5.

Как видно из линий регрессии, на обоих графиках корреляция между X и Y равна 0. Но эти отношения явно различны не в том смысле, который отражает линия регрессии.

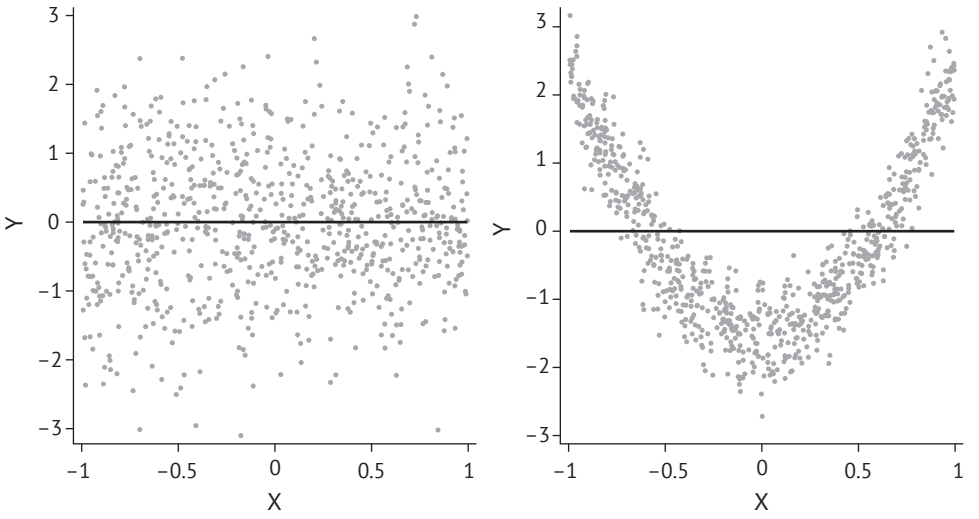


Рис. 2.5. Нулевая корреляция может означать многое

На левом графике нет никакой корреляции между *X* и *Y*, а также, похоже, нет каких-либо интересных отношений. Вы действительно не можете предсказать значение *Y* по *X* или наоборот. На правом графике также нет корреляции между *X* и *Y* – в среднем более высокие значения *X* не обязательно связаны с высокими значениями *Y*, а низкие значения *X* не имеют тенденции возникать вместе с низкими значениями *Y*. Но между этими двумя переменными, безусловно, существует связь. Фактически переменная *X* на правом графике весьма полезна для прогнозирования *Y*. Это преподает нам важный урок. Рациональный анализ данных требует большего, чем просто вычисление корреляций. Помимо прочего, важно смотреть на свои данные с разных точек зрения (например, с помощью диаграмм рассеяния, подобных этой), чтобы не пропустить интересные нелинейные зависимости.

Существует множество статистических подходов к решению проблемы нелинейности, и некоторые из них мы обсудим в этой книге. Но, как оказалось, линейные инструменты описания данных могут быть полезны, даже если переменные связаны нелинейным образом. Например, в правой части рис. 2.5 наблюдается сильная отрицательная корреляция между *X* и *Y*, когда *X* меньше 0, и сильная положительная корреляция между *X* и *Y*, когда *X* больше 0. С помощью инструментов линейной регрессии мы могли бы построить две линии наилучшего соответствия: одну для случая $X < 0$ и одну для случая $X > 0$. На рис. 2.6 показано, как это будет выглядеть.

А еще можно преобразовать одну из переменных, чтобы связь выглядела более линейной. Например, хотя корреляция между *Y* и *X* отсутствует, между *Y* и X^2 существует сильная линейная связь. На рис. 2.7 мы откладываем X^2 по горизонтальной оси и *Y* по вертикальной оси. Когда мы преобразуем *X* в X^2, отрицательные значения *X* становятся положительными значениями X^2 (например, −1 становится 1), а положительные значения остаются положительными (например, 1 остается 1). Это похоже на то, как если бы мы согнули фигуру в точке $X = 0$, а затем немного скрутили и растянули ее, так что *X* становится X^2 (ноль остается нулем, единица остается единицей, 0.5 становится $0.5^2 = 0.25$ и т. д.).

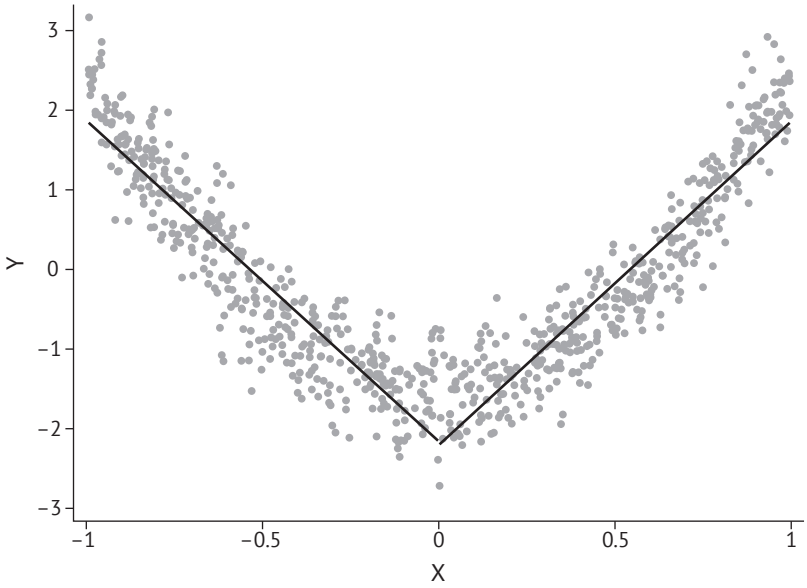


Рис. 2.6. Подгонка двух отдельных линий регрессии к нелинейной зависимости

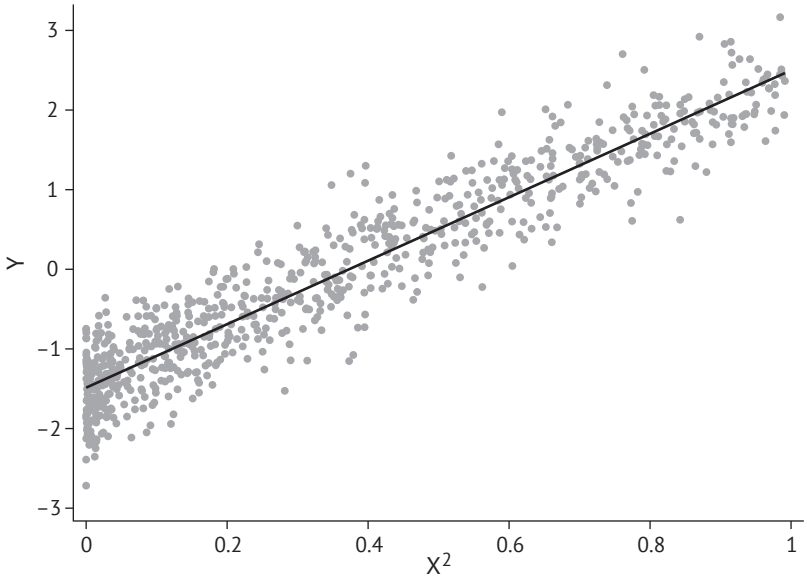


Рис. 2.7. Создание линейной зависимости путем преобразования переменной

Благодаря этому преобразованию наша линия регрессии показывает сильную положительную связь между Y и X^2, и мы можем хорошо описать взаимосвязь между этими переменными с помощью наших линейных инструментов.

Также стоит отметить, что описание связи между двумя переменными с помощью линейной функции всегда уместно, когда мы имеем дело с бинарными переменными.

Например, давайте вернемся к взаимосвязи между добычей нефти и автократией. На рис. 2.8 представлены данные. Диаграмма рассеяния не очень интересна и информативна, поскольку существует только четыре возможных комбинации наших двух переменных. Соответственно, все точки данных лежат на одной из этих четырех точек (хотя мы попытались сделать диаграмму рассеяния более информативной, сделав размер точек пропорциональным количеству стран в каждом наборе значений). Однако все равно можно построить наклон линии регрессии. Наклон этой линии представляет собой просто долю основных нефтедобывающих стран, являющихся автократиями, минус доля неосновных нефтедобывающих стран, являющихся автократиями. Другими словами, из этого рисунка мы узнаем то же самое, что узнали из таблицы в начале главы.

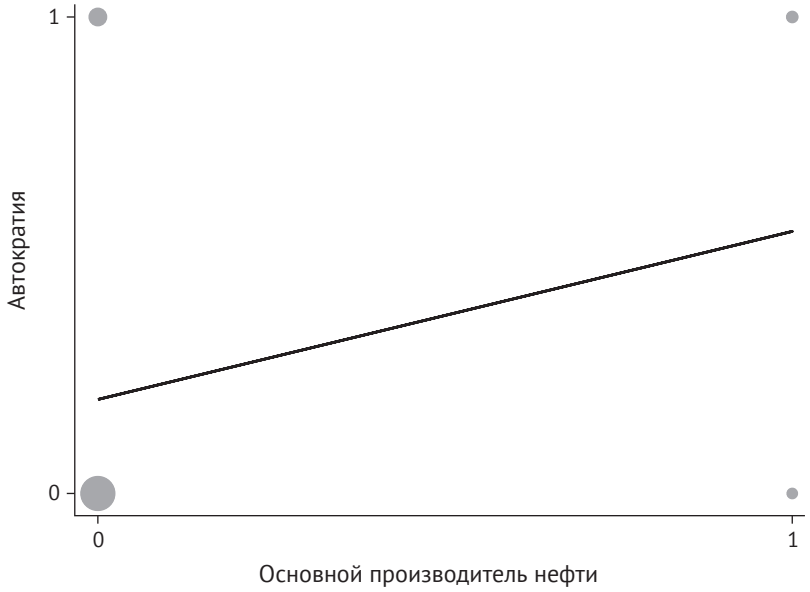


Рис. 2.8. Линия регрессии по данным с бинарной переменной дает разницу в средних значениях

Одна из причин, по которой мы так много внимания уделяем линейным зависимостям, заключается в том, что даже нелинейные зависимости начинают выглядеть приблизительно линейными, если вы достаточно увеличите масштаб, т. е. если вас интересует достаточно узкий диапазон значений переменной *X*. Но к такой экстраполяции нужно относиться очень осторожно. По мере удаления от диапазона данных, в котором взаимосвязь приблизительно линейна, наши описания взаимосвязи (и, как следствие, любые прогнозы, которые мы делаем) будут все менее и менее точными.

Чтобы лучше убедиться в опасностях экстраполяции, рассмотрим пример. Политические аналитики обнаружили, что действующая партия на президентских выборах в США имеет тенденцию получать около 46 % голосов при нулевом росте доходов и дополнительные 3.5 процентных пункта голосов на каждый процентный пункт роста доходов. Разумеется, они измерили эту взаимосвязь, используя данные о фактически случившемся росте доходов. Озна-

чает ли это, что доля голосов за действующую партию составит 81 %, если рост доходов составит 10 %? Скорее всего, нет. И доля голосов за действующего президента определенно не составила бы 116 % при росте доходов на 20 % – это невозможно! Но это не означает, что линейное описание данных бесполезно для диапазона роста доходов, который мы действительно наблюдаем.

Подведение итогов

Корреляции составляют основу анализа данных. Именно на их языке мы говорим о взаимоотношениях между признаками мира. Различные статистические показатели, с помощью которых мы измеряем корреляцию, такие как ковариация, коэффициент корреляции или наклон линии регрессии, – это способ количественной оценки подобных взаимоотношений.

Как мы уже говорили, корреляцию можно использовать для различных целей, включая описание, прогнозирование и причинно-следственные выводы. В главе 3 мы сосредоточимся на причинности и обстоятельно разберемся, что она означает. Вы начнете глубже понимать афоризм, с которого мы начали книгу: корреляция не обязательно означает причинность. Однако более полное понимание взаимосвязи между корреляцией и причинностью придется отложить до главы 9.

Ключевые термины

- **Корреляция** между двумя явлениями мира – это мера того, насколько они имеют тенденцию возникать вместе.
- **Положительная корреляция:** когда более высокие (более низкие) значения одной переменной преимущественно сочетаются с более высокими (более низкими) значениями другой переменной, мы говорим, что эти две переменные положительно коррелированы.
- **Отрицательная корреляция:** когда более высокие (более низкие) значения одной переменной преимущественно сочетаются с более низкими (более высокими) значениями другой переменной, мы говорим, что эти две переменные отрицательно коррелированы.
- **Некоррелированные переменные:** когда между двумя переменными нет корреляции, т. е. когда более высокие (более низкие) значения одной переменной систематически не сочетаются с более высокими или более низкими значениями другой переменной, мы говорим, что они не коррелированы.
- **Линия наилучшего соответствия:** линия, которая сводит к минимуму среднее расстояние между точками данных и линией в соответствии с некоторой мерой расстояния.
- **Среднее (μ):** среднее значение переменной.
- **Отклонение от среднего значения:** расстояние между значением текущего наблюдения для некоторой переменной и средним значением этой переменной.
- **Дисперсия (σ^2)**: мера того, насколько изменчивой является переменная. Это усредненное значение квадрата отклонений от среднего значения.

- **Стандартное отклонение (σ):** еще один показатель того, насколько изменчивой является переменная. Стандартное отклонение – это квадратный корень дисперсии. Его преимущество состоит в том, что оно измеряется в том же масштабе, что и сама переменная, и примерно соответствует тому, насколько далеко типичное наблюдение отстоит от среднего значения (хотя, как и дисперсия, оно придает больший вес наблюдениям, далеким от среднего значения).
- **Ковариация (cov):** мера корреляции между двумя переменными. Она рассчитывается как усредненное произведение отклонений от среднего значения.
- **Коэффициент корреляции (r):** еще один показатель корреляции между двумя переменными. Он рассчитывается как ковариация, деленная на произведение дисперсий. Коэффициент корреляции принимает значение от –1 до 1, где –1 отражает идеальную линейную отрицательную зависимость, 0 отражает отсутствие корреляции и 1 отражает идеальную линейную положительную зависимость.
- r^2: квадрат коэффициента корреляции. Он принимает значения от 0 до 1 и часто интерпретируется как доля дисперсии одной переменной, объясняемой другой переменной. Следует четко понимать, что подразумевается под «объяснением». Этот термин *не* означает, что изменение одной переменной вызывает изменение другой.
- **Сумма квадратов ошибок:** сумма квадратов расстояния от каждой точки данных до заданной линии наилучшего соответствия. Это дает нам один из способов измерения того, насколько хорошо линия подогнана / описывает / объясняет данные.
- **Линия регрессии OLS:** линия, которая лучше всего соответствует данным, где наилучшее соответствие означает, что она минимизирует сумму квадратов ошибок.
- **Наклон линии:** показывает, насколько значение на линии изменяется по вертикальной оси при перемещении на одну единицу по горизонтальной оси. Полностью горизонтальная линия имеет наклон 0. Линия, наклоненная вверх под углом 45°, имеет наклон 1, нисходящая линия под углом 45° имеет наклон –1 и т. д.
- **Наклон линии регрессии или коэффициент регрессии:** наклон описывает, как в среднем изменяется значение одной переменной при изменении другой переменной. Наклон линии регрессии представляет собой ковариацию двух переменных, деленную на дисперсию одной из них, и иногда также называется коэффициентом регрессии.

Упражнения

2.1. Обдумайте следующие три утверждения. Какие из них описывают корреляцию, а какие нет? Почему?

a) Большинство профессиональных аналитиков данных прошли курс статистики в колледже.

b) Среди игроков Высшей бейсбольной лиги питчеры, как правило, имеют усредненные показатели ударов ниже среднего среди всех игроков. (Мы узнаем, почему это так, в главе 16.)

c) Какой бы кандидат в президенты ни победил, Огайо, как правило, побеждает в коллегии выборщиков.

2.2. Рассмотрим последнее утверждение об Огайо и президентских выборах. Как вы думаете, оно полезно для описания? Для прогнозирования? Для причинно-следственного вывода? Поясните свой ответ.

2.3. В таблице ниже показаны некоторые данные о том, какие страны являются основными производителями нефти и какие страны пережили гражданскую войну в период с 1946 по 2004 г. Какова корреляция между статусом крупного производителя нефти и гражданской войной: положительная, отрицательная или отсутствует? Поясните свой ответ.

	Была гражданская война	**Не было гражданской войны**
Производитель нефти	7	12
Не производитель нефти	55	94

2.4. В таблице ниже представлены данные о росте и доходах американских мужчин, взятые из национального статистического опроса. Для выполнения этого задания можно использовать калькулятор, но не электронные таблицы или специализированное программное обеспечение для расчета ответов.

Рост (дюймы)	**Средний доход, долл.**
60	39 428
61	35 087
62	40 575
63	39 825
64	55 508
65	56 377
66	59 746
67	66 699
68	59 787
69	66 176
70	79 202
71	70 432
72	77 975
73	72 606
74	71 063
75	80 330

a) Вычислите среднее значение каждой из этих переменных.
b) Рассчитайте дисперсию каждой из этих переменных.
c) Рассчитайте стандартное отклонение каждой из этих переменных.
d) Рассчитайте ковариацию между этими двумя переменными.
e) Рассчитайте коэффициент корреляции для этих переменных.
f) Являются ли эти две переменные положительно коррелированными, отрицательно коррелированными или некоррелированными? Поясните свой ответ.

Дополнительное чтение и ссылки

Дополнительную информацию о данных о коррупции, которые мы обсуждали, см. здесь:

Scott J. Basinger. 2013. *Scandals and Congressional Elections in the Post-Watergate Era.* Political Research Quarterly 66 (2): 385–398.

Для получения дополнительной информации о проекте Polity IV, который классифицирует страны как демократические или автократические, см. *https://www.systemicpeace.org/polity/polity4.htm*.

Мы обсудили две статьи об использовании онлайн-обзоров для прогнозирования нарушений норм здравоохранения:

Emily Badger. 2013. *HowYelpMightCleanUp theRestaurant Industry.* The Atlantic. July/August;

Kristen M. Altenburger and Daniel E. Ho. 2018. *When Algorithms Import Private Bias into Public Enforcement: The Promise and Limitations of Statistical Debiasing Solutions.* Journal of Institutional and Theoretical Economics 174 (1): 98–122.

Связь между изучением высшей математики и окончанием колледжа:

Jerry Trusty and Spencer G. Niles. 2003. *High-SchoolMath Courses and Completion of the Bachelor's Degree.* Professional School Counseling 7 (2): 99–107.

Если вас интересуют примеры растущего использования прогнозирования и предсказаний при решении важных политических проблем:

Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. *Prediction Policy Problems.* American Economic Review 105 (5): 491–95.

Глава 3

Причинно-следственная связь: что это такое и для чего она нужна?

О чем эта глава

- Причинный эффект – это изменение какой-либо характеристики мира, которое может возникнуть в результате изменения какой-либо другой характеристики мира.
- Оценка причинно-следственных связей имеет решающее значение для стратегии поведения и принятия решений.
- «Какое влияние это оказало на результат?» – это более концептуально правильный вопрос, чем «Что стало причиной явления?»
- Причинно-следственные связи заключаются в сравнении *контрфактических* миров. В результате они принципиально не наблюдаемы. Но в определенных ситуациях мы можем узнать о их существовании из данных.

Введение

Как было сказано в главе 2, знание корреляции полезно для многих целей. Одной из наиболее важных, но и самых неприятных целей является изучение причинно-следственных связей.

Мы постоянно заявляем о причинном знании. Я плохо сдал экзамен, потому что не выспался. Поступление в колледж улучшит мои перспективы трудоустройства. Политический кандидат проиграл выборы из-за агрессивной рекламы. Уровень насильственных преступлений снизился благодаря новой полицейской стратегии.

Наличие навыка критического мышления о существовании причинно-следственных связей, является, пожалуй, самым важным условием использования информации для принятия более эффективных решений. Знание причин и следствий является ключом к пониманию того, как ваши решения и действия влияют на мир вокруг вас. Если вы предлагаете новую налоговую политику, стратегию подготовки к экзаменам, план тренировок или рекламную кампанию, вы делаете это не потому, что надеетесь на корреляцию с лучшими

результатами. Вы верите, что принятие вашего предложения послужит непосредственной причиной достижения лучших результатов.

Наша цель в этой главе – прояснить, что именно мы имеем в виду, когда говорим о причинно-следственных связях. Причинность – глубокая и запутанная тема, которой уделяют много внимания ученые из самых разных областей. Мы не будем пытаться разрешить все острые философские вопросы. Вместо этого мы поставили перед собой более скромные цели. Во-первых, убедимся, что понимаем друг друга, определив, как мы будем использовать причинный язык на протяжении всей этой книги. Затем объясним, почему принятое нами определение причинности особенно ценно. Наконец, обсудим некоторые другие подходы к причинности и объясним, почему, с нашей точки зрения, они менее полезны, чем тот, которого мы придерживаемся.

Что такое причинно-следственная связь?

Говоря о причинно-следственной связи, мы говорим о влиянии одной вещи на другую. Если выражаться простым языком, *причинный эффект* – это изменение какой-либо характеристики мира вследствие изменения другой его характеристики. Так, например, можно сказать, что ставка налога оказывает причинное влияние на государственные доходы, если изменение ставки налога приведет к изменению государственных доходов.

Мы дали весьма свободное разговорное определение причинного эффекта, поэтому вы, возможно, не заметили, что мы немного коснулись философии. Что подразумевается под следствием или результатом? Ведь мир такой, какой он есть. Откуда взялось это изменение какой-то другой характеристики мира?

Это хороший вопрос. Фактически наше определение причинно-следственной связи основано на мысленном эксперименте, который необходимо четко объяснить. Начнем с примера.

Кинозвезда Гвинет Пэлтроу руководит компанией Goop, которая продвигает наклейки Body Vibes, призванные способствовать укреплению здоровья, хорошему самочувствию и чистой коже. Вот что сайт Goop говорит о Body Vibes:

> «Человеческие тела работают на идеальной энергетической частоте, но повседневные стрессы и беспокойство могут нарушить наш внутренний баланс, истощая наши энергетические запасы и ослабляя иммунную систему. Наклейки Body Vibes заранее запрограммированы на идеальную частоту, что позволяет им устранять дисбаланс. Пока вы их носите – у сердца, на левом плече или руке, – они восполнят недостатки ваших резервов, создавая успокаивающий эффект, сглаживая как физическое напряжение, так и тревогу. Создатели наклеек – опытные косметологи – утверждают, что они помогают очистить кожу, уменьшая воспаление и ускоряя обновление клеток».

Предположим, вы заплатили 6 долл. за наклейку, потому что очень хотите иметь чистую кожу. Но потом ваши друзья начали смеяться над вами за то, что вы такой легковерный простофиля. Защищая себя, вы утверждаете, что Body Vibes действительно влияет на чистоту вашей кожи. Но что именно вы имели в виду под «влиянием»?

Вот как можно рассуждать об этом. Представьте себе альтернативный мир, где в тот самый момент, когда вы решили наклеить стикеры Body Vibes, один из друзей без вашего ведома заменил их такими же на вид наклейками, которые стоили десять центов вместо шести долларов и не были «заранее запрограммированы на идеальную частоту». Если бы в этом альтернативном мире чистота вашей кожи была бы хуже, мы бы сказали, что Body Vibes положительно влияет на чистоту вашей кожи. Если бы чистота вашей кожи в этом альтернативном мире была такой же, мы бы пришли к выводу, что Body Vibes не оказывают заявленного влияния на чистоту кожи. А если бы чистота вашей кожи оказалась лучше в этом альтернативном мире, мы бы пришли к выводу, что Body Vibes имеют отрицательное влияние.

Мы можем расширить этот мысленный эксперимент. В реальном мире нет ничего особенного. Раз уж мы подумали об одном альтернативном мире, можно с тем же успехом подумать и о двух. Например, подумаем об эффекте десятицентовых наклеек по сравнению с магическими кристаллами, даже если вы никогда не пробовали ни один из этих подходов к уходу за кожей. Нам просто нужно сравнить два воображаемых мира: один – где ваши друзья тайно наклеили наклейки на ваше левое плечо ближе к сердцу, и другой – где они спрятали кристаллы в ваши карманы. Подобные сравнения называются *контрфактическими* (counterfactual) мысленными экспериментами, потому что по крайней мере один из миров, которые мы сравниваем, не является реальным, фактическим миром – он находится в нашем воображении. Сравнение результатов такого мысленного эксперимента является *контрфактическим сравнением*.

Теперь становится ясно, что, говоря о «следствии» или «результате», мы на самом деле имеем в виду определение причинно-следственной связи. Мы говорим о сравнении между результатом в реальном мире и результатом в контрфактическом мире, который полностью идентичен реальному миру до тех пор, пока не изменится характеристика мира, которая, как утверждается, имеет причинный эффект.

Идея контрфактуальности философски тонка. Чтобы удостовериться в наличии критического мышления, мы введем математический механизм описания контрфактических явлений, называемых *потенциальными исходами* (potential outcome). Использование понятия потенциальных исходов требует некоторых обозначений, но это не слишком сложно. И как только вы овладеете обозначениями, то гораздо глубже поймете, что такое причинность на самом деле. Итак, давайте попробуем.

Потенциальные исходы и контрфактические сравнения

Нас интересует влияние какого-либо *воздействия* (скажем, приклеивания Body Vibes) на какой-то *исход* (скажем, здоровье кожи). Обозначим воздействие буквой T. Это бинарная переменная, принимающая значение 0 или 1. Если для какого-то человека $T = 1$, это означает, что он подвергся воздействию Body Vibes. Если для какого-то человека $T = 0$, это означает, что он не подвергался воздействию Body Vibes. Иногда говорят, что объект (здесь: человек) с $T = 1$ обработан, а объект с $T = 0$ не обработан, хотя это условная классификация (например, мы могли бы с таким же успехом говорить об эффекте *отсутствия* Body Vibes).

Аналогичным образом обозначим интересующий нас исход буквой *Y*. В нашем примере *Y* описывает здоровье кожи человека. Сделаем метафизическое предположение, что существуют два уровня здоровья кожи – один был бы у человека, если бы он использовал Body Vibes, и другой, если бы он не использовал Body Vibes. Это потенциальные исходы данного человека. Однако в любой момент времени мы можем наблюдать только один из них: каждый человек либо использует, либо не использует Body Vibes. Тем не менее размышления о потенциальных исходах помогают нам составить контрфактические представления:

Y_{1i} = исход для объекта *i*, если $T = 1$,

Y_{0i} = исход для объекта *i*, если $T = 0$.

Влияние ношения Body Vibes на здоровье кожи человека – это всего лишь разница в состоянии здоровья кожи с Body Vibes и без них. В наших обозначениях потенциальных исходов это выглядит так:

Влияние Body Vibes на здоровье кожи объекта *i* равно $Y_{1i} - Y_{0i}$.

В табл. 3.1 показан более конкретный пример. Мы наблюдаем десять объектов эксперимента. Относительно каждого человека мы смотрим, применял ли он Body Vibes и стала ли его кожа чистой. Если человек *i* применял Body Vibes, его статус воздействия $T_i = 1$; если не применял – $T_i = 0$. И если человек *i* подвергся воздействию *T*, мы записываем его исход как $Y_{Ti} = 1$, если его кожа стала чистой, и $Y_{Ti} = 0$, если его кожа не изменилась.

Таблица 3.1. Потенциальные исходы состояния кожи с применением Body Vibes и без него. Для каждого человека фактический исход, который мы можем наблюдать, выделен жирным шрифтом. Контрфактический исход, которого мы не наблюдаем, записан обычным шрифтом

		Здоровье кожи при наличии Body Vibes Y_{1i}	**Здоровье кожи при отсутствии Body Vibes** Y_{0i}	**Эффект воздействия на человека *i*** $Y_{1i} - Y_{0i}$
Получили Body Vibes	Человек 1	**1**	1	0
	Человек 2	**0**	0	0
	Человек 3	**0**	0	0
	Человек 4	**1**	1	0
	Человек 5	**1**	1	0
Не получали Body Vibes	Человек 6	0	**0**	0
	Человек 7	0	**0**	0
	Человек 8	1	**1**	0
	Человек 9	1	**1**	0
	Человек 10	0	**0**	0

Фактический результат для каждого человека в таблице выделен жирным шрифтом. Люди 1–5 получили Body Vibes, поэтому их фактический исход – Y_{1i}. Таблица также сообщает нам, какими были бы исходы этих людей, если бы они не получали Body Vibes, Y_{0i}. Однако в реальном мире никто не может наблюдать эти контрфактические исходы, поскольку на самом деле они не случаются. Люди 6–10 не получают Body Vibes. Таким образом, их фактический исход – Y_{0i}. Опять же, хотя таблица говорит нам, какими были бы их исходы, если бы они получили Body Vibes, Y_{1i}, эти контрфактические исходы не наблюдаются в реальном мире.

Поскольку таблица показывает нам потенциальные исходы в реальном и контрфактическом мирах, мы можем найти эффект применения Body Vibes для каждого человека, рассчитав $Y_{1i} - Y_{0i}$. Это показывает, что Body Vibes на самом деле не оказывает никакого влияния на здоровье кожи любого человека. У участников 1, 4, 5, 8 и 9 кожа чистая. Но для всех этих людей это было бы верно независимо от того, получили ли они Body Vibes или нет. У участников 2, 3, 6, 7 и 10 кожа болезненная. Опять же, это будет верно как с наклейками Body Vibes, так и без них. Важно отметить (как мы обсудим позже), это *отсутствие* причинного эффекта на самом деле невозможно наблюдать в реальном мире, потому что мы наблюдаем только фактический результат для каждого человека, а не потенциальный результат в контрфактическом мире, где у него другой статус воздействия.

Мы говорим, что причинность связана с контрфактическими сравнениями, потому что можем наблюдать только одну из двух величин, Y_{1i} или Y_{0i}, для любого человека в любой конкретный момент времени. Это означает, что невозможно напрямую измерить влияние ношения Body Vibes на здоровье кожи человека. Мы подозреваем, что этот факт является ключевым в их бизнес-модели.

Зачем нужно знать причинно-следственную связь?

Знание причинно-следственной связи необходимо для понимания последствий действия, изменяющего какую-либо характеристику мира. В частности, чтобы взвесить затраты и выгоды от решения, вам необходимо знать, как ваше действие повлияет на результаты, которые вас интересуют.

Например, вы не можете знать, стоит ли тратить деньги на препарат для лечения сердечно-сосудистых заболеваний, не зная причинно-следственной связи – снижает ли препарат риск сердечных заболеваний. То же самое касается многих решений. Когда вы решаете, стоит ли каким-либо образом вмешиваться в текущую жизнь (с помощью политических решений, плана упражнений, стратегии воспитания, нового вида онлайн-обучения или чего-то еще), вам следует знать, как вмешательство влияет на интересующий вас исход.

Хотя примеры, которые мы обсуждали, легко понять с точки зрения контрфактических сравнений, иногда контрфактическое мышление может показаться раздражающим или сбивающим с толку. В следующих разделах мы рассмотрим некоторые проблемы.

Фундаментальная проблема причинного вывода

Обсуждая табл. 3.1, мы указали на важную проблему: причинные явления в том виде, в каком мы их определили, никогда не могут наблюдаться напрямую.

Каждый человек либо получает наклейки Body Vibes, либо нет. Следовательно, для каждого человека мы можем наблюдать только один потенциальный исход. Но причинно-следственное явление – это разница в потенциальных исходах для одного человека. Эта присущая причинным явлениям принципиальная ненаблюдаемость называется *фундаментальной проблемой причинного вывода*. Давайте посмотрим, почему мы не можем наблюдать причинные явления и что это означает для нашей способности изучать причинность.

Влияние поступления в колледж на ваш доход – это разница в ваших доходах в мире, в котором вы учитесь в колледже, и в мире, в котором вы остаетесь таким же до принятия решения о поступлении в колледж, но не поступаете в него. Как минимум один из этих миров контрфактичен. Вы не можете одновременно поступить и не поступить. То есть у вас есть два потенциальных результата – Y_{college} и $Y_{\text{no college}}$. Но реальный исход у вас только один: либо вы поступили в колледж, либо нет. Учитывая это, мы никогда не сможем наблюдать влияние обучения в колледже на ваш доход, поскольку наблюдаем ваш доход только в реальном, а не в контрфактическом мире.

Таким образом, фундаментальная проблема причинного вывода заключается в том, что в любой момент времени мы наблюдаем любую данную единицу анализа (например, человека, баскетбольную команду или страну) только в одном состоянии. Мы не можем наблюдать влияние на эту единицу бытия в текущем состоянии по сравнению с каким-то другим состоянием, потому что все остальные состояния контрфактичны. Мы не можем знать $Y_{\text{college}} - Y_{\text{no college}}$, потому что придерживаемся только одного из двух значений переменной. Мы обнаружили этот факт ранее в табл. 3.1, где заметили, что можем наблюдать для конкретного человека только фактический исход; другой потенциальный исход был контрфактическим.

Но как добиться ответов на причинно-следственные вопросы, если последствия воздействия принципиально ненаблюдаемы? К счастью, существует множество ситуаций, когда нам не обязательно знать эффект для каждого отдельного объекта анализа. Вместо этого достаточно знать средний эффект для многих людей.

Предположим, например, что Управление по санитарному надзору за качеством пищевых продуктов и медикаментов решает, одобрить ли новый препарат. Чтобы узнать о влиянии препарата на здоровье, ученые проводят рандомизированное исследование, в ходе которого одним людям назначают препарат (группа, подвергнутая воздействию), а другим – плацебо (группа, не подвергнутая воздействию). Из-за фундаментальной проблемы причинно-следственных связей ученые не могут наблюдать эффект от приема препарата на отдельном человеке. Каждый человек либо принимает препарат, либо нет. Но, сравнивая средние показатели здоровья людей в группе, не получавшей препарат, со средними показателями здоровья людей в группе, получавшей препарат, они могут оценить средний эффект от применения препарата. (Мы подробнее поговорим о том, как это работает, во второй и третьей частях.) Это позволит ученым ответить на ключевой вопрос, лежащий в основе решения комиссии: как повлияет одобрение нового препарата на состояние здоровья популяции в среднем?

Одобрение лекарств – это ситуация, в которой знания усредненного эффекта достаточно для принятия ключевых решений. Но есть ситуации, когда это

не так, и фундаментальная проблема причинного вывода представляет собой настоящую проблему. Например, оценка юридической ответственности включает в себя ответы эксперта на вопросы типа «что, если?». От эксперта ждут ответов на такие вопросы, как «Был бы причинен вред Энтони, если бы не действия Итана?». Фундаментальная проблема причинного вывода гласит, что мы никогда не сможем знать этого наверняка, поскольку мир, в котором Итан не предпринял своих действий, контрфактичен, и мы не можем увидеть, что произошло с Энтони в другом мире. Как мы только что сказали и рассмотрим более подробно в оставшейся части книги, существуют методы ответа на несколько иной вопрос, например: «В среднем, когда люди совершают действия, подобные тем, которые совершил Итан, они способствуют причинению вреда другим людям?» Убедительный ответ на последний вопрос может быть, а может и не быть убедительным для суда, который ждет ответа на первый вопрос.

Критическое мышление о причинно-следственных связях предполагает признание того, что иногда мы не можем с полной уверенностью ответить на определенные вопросы, даже если эти вопросы очень важны.

Принципиальные вопросы

Причинность – глубокая и трудная тема. Контрфактическое определение причинности не дает ответов на все вопросы. Однако оно помогает нам более ясно воспринимать сложные принципиальные вопросы. Давайте поговорим о некоторых из них.

В чем причина?

Одно из разочарований, которые люди иногда испытывают в отношении контрафактического подхода, заключается в том, что некоторые из причинно-следственных вопросов, которые мы привыкли задавать, кажутся бессмысленными в рамках контрфактического подхода. Возьмем, к примеру, такие вопросы: «Почему цены на жилье упали во время последнего финансового кризиса? Почему Chicago Blackhawks выиграли Кубок Стэнли? Что стало причиной Первой мировой войны?» Подобные вопросы о причинной атрибуции встречаются часто. Но когда причинно-следственная связь определяется в терминах контрфактических сравнений, они не имеют особого смысла.

Поговорим о Первой мировой войне. Расхожее утверждение гласит, что Первая мировая война была вызвана убийством в 1914 г. эрцгерцога Фердинанда, наследника престола Австро-Венгрии. Убийцы были частью движения, которое хотело, чтобы Сербия взяла под свой контроль Южные Балканы, включая Боснию и Герцеговину, которые Австро-Венгрия аннексировала в 1908 г. Правительство Австро-Венгрии ответило на убийство июльским ультиматумом. Условия ультиматума были настолько обременительны, что сербское правительство их отвергло. Когда ультиматум был отклонен, Австро-Венгрия объявила войну Сербии, что побудило Россию мобилизовать свою армию для защиты Сербии. В ответ Германия (союзница Австро-Венгрии) объявила войну России, Франция (союзница России) объявила войну Германии, и вся эта неразбериха переросла в Первую мировую войну. Поэтому принято говорить, что убийство эрцгерцога Фердинанда вызвало Первую мировую войну.

Нам не составит труда применить к этому утверждению контрфактический подход. Можно задаться вопросом: случилась бы Первая мировая война в контрфактическом мире, в котором Фердинанд не был убит? Если бы в этом контрфактическом мире Первая мировая война не началась, то было бы правильно сказать, что убийство *повлияло* на начало войны. Но это далеко не утверждение, что убийство эрцгерцога *стало* причиной войны. Ведь существует множество факторов, которые, будь они иными, предотвратили бы начало Первой мировой войны. Конечно, если бы эрцгерцог Фердинанд не был убит, возможно, война (в том виде, как мы ее знаем из истории) не началась бы. Но кроме того, если бы Австро-Венгрия не аннексировала Боснию и Герцеговину, возможно, Фердинанд никогда бы не был убит и война никогда бы не началась, поэтому аннексия была такой же причиной, как и убийство. Точно так же, если бы сербское правительство приняло июльский ультиматум, возможно, войны удалось бы избежать, поэтому несоблюдение ультиматума также было причиной. И чтобы еще раз проиллюстрировать, сколько таких причин существует, если бы какое-нибудь рыбоподобное существо в палеозойскую эру плыло влево, а не вправо, возможно, человеческая раса в том виде, в каком мы ее знаем, не существовала бы, и – опять же – Первая мировая война никогда бы не началась. Или, если обратиться к авторитетным личностям, французский математик XVII в. Блез Паскаль, размышляя о пристрастии Марка Антония к длинным носам, пошутил: «Если бы нос Клеопатры был короче, изменилось бы лицо всего мира»[1]. Это побудило Джеймса Фирона в эссе о контрфактических рассуждениях задаться вопросом: «Означает ли это, что ген, определивший длину носа Клеопатры, был причиной Первой мировой войны?» Как видите, проблема не в том, что утверждение об убийстве эрцгерцога Фердинанда, ставшего причиной Первой мировой войны, является ложным.

Как только мы начинаем мыслить контрфактически, становится совершенно ясно, что все явления имеют множество причин. Это мешает нам получить прямой ответ на вопрос «В чем причина?». Вместо этого приходится задавать вопросы «Входит ли данное действие в число причин?» или «Могло ли это событие иметь последствия?». Наверное, это разочаровывает.

Вы можете возразить, что некоторые причины, несомненно, более важны или более близки по времени, чем другие. Если это так, возможно, мы можем рассуждать о *важных* или *непосредственных* причинах Первой мировой войны. Как это сделать?

Подход, которого придерживаются некоторые философы, выглядит примерно так. Представьте себе все контрфактические миры, в которых не произошла Первая мировая война. Некоторые из этих контрфактических миров сильно отличаются от реального мира – например, Первая мировая война, вероятно, не происходит во многих контрфактических мирах, в которых нет гравитации. Другие очень похожи на реальный мир – возможно, Первая мировая война

[1] Любовные отношения Антония и Клеопатры имели огромные последствия для мировой истории. Например, историки обычно полагают, что конец Римской республики и создание Римской империи были обеспечены, когда Антоний и Клеопатра потерпели поражение от Октавиана (позже императора Августа) в битве при Акциуме. Если бы этого не произошло, кто знает, насколько иначе могла бы сложиться остальная часть западной истории?

происходит не в мире, идентичном нашему до 27 июня 1914 г., а в котором эрцгерцог Фердинанд проспал поездку 28 июня. О непосредственных причинах Первой мировой войны мы узнаем из сравнения реального мира с контрфактическим миром, в котором не произошла Первая мировая война и который наиболее похож на реальный мир. Этот вид анализа способен дать разумные ответы на вопросы «В чем причина?», не отказываясь при этом от определения причинности, основанного на контрфактических сравнениях. Например, кажется разумным думать, что убийство эрцгерцога Фердинанда является более непосредственной причиной Первой мировой войны, чем нос Клеопатры, законы гравитации или прихоти палеозойских рыб.

Определенно, в этом подходе что-то есть. Но тем не менее часто бывает трудно принципиально оценить важность или близость одной причины к другой. Если вы немного знакомы с историей, то наверняка сможете найти и другие причины Первой мировой войны, которые кажутся столь же близкими. Например, многие ученые утверждают, что свою роль в возникновении Первой мировой войны сыграли военные доктрины начала XX в., отдающие предпочтение наступательным стратегиям, а не оборонительным. Является ли мир, в котором была принята несколько иная военная доктрина, более близким к нашему миру, чем тот, в какой эрцгерцог Фердинанд не был убит? Действительно ли мир, в котором одна палеозойская рыба повернула в другую сторону, значительно отличается от нашего? Сложно сказать.

Чтобы изучить проблему в несколько менее возвышенной и, возможно, более знакомой обстановке, рассмотрим женский баскетбольный матч третьего дивизиона NCAA между Chicago Maroons (где некоторые из наших звездных учениц также являются звездными спортсменами) и Emory Eagles. Предположим, что Maroons отстают от Eagles на одно очко, и у Maroons осталось ровно столько времени, чтобы сделать последний бросок. Бросок был удачным, и команда выиграла с преимуществом в одно очко (в баскетболе попадание в корзину с игры приносит минимум два очка). На следующий день газета Chicago Maroon[1] сосредоточится на этом последнем броске, и репортер может даже написать, что именно последний бросок стал причиной победы Chicago Maroons. Десятки бросков за всю игру сыграли решающую роль. Фактически *каждый* удачный бросок, сделанный Chicago Maroons, был решающим – в контрфактическом мире, в котором они бы промахнулись раньше, а все остальные броски состоялись, как и в реальном мире, они бы проиграли, а не выиграли. Точно так же каждый промах Eagles имел решающее значение: в контрфактическом мире, в котором они хотя бы раз вместо промаха попали в корзину, а все остальное прошло как в реальном мире, они бы выиграли. Так что же такого особенного в этом последнем броске? Одно из объяснений заключается в том, что всем заранее было точно известно, что последний бросок будет иметь решающее значение. Но очень немногие другие причины соответствуют этому критерию, и уж точно не убийство эрцгерцога Фердинанда. Так что, на наш взгляд, нет очевидных оснований считать, что последний бросок стал

[1] Да, тут можно запутаться. Баскетболисты – это Maroons, газета – это Maroon, и, вероятно, ни спортивные команды, ни газеты не должны называться в честь цвета кожи (*maroon* – беглый негр-раб). Наш университет никогда не славился успехами в легкой атлетике или удачным брендингом.

более важной причиной победы Maroons, чем остальные броски. Мы считаем, что этот пример иллюстрирует фундаментальный, хотя и разочаровывающий факт жизни: отдельные события могут иметь множество одинаково важных и последовательных причин.

Еще одно удивительное следствие контрфактического подхода заключается в том, что, по крайней мере, в принципе, некоторые события могут вообще не иметь причин. Предположим, авторы этой книги придумали идеальное преступление. Мы оба стреляем и убиваем нашего заклятого врага одновременно, зная, что любая пуля сама по себе будет смертельной. На допросе Энтони говорит: «Очевидно, что меня нельзя обвинить в преступлении. Мои действия не имели прямого эффекта. Если бы я не выстрелил, жертва все равно умерла бы». И точно так же Итан возражает: «Я тоже не мог быть причиной смерти жертвы. Если бы я не выстрелил из ружья, он бы все равно умер». Хотя наша логика, возможно, и не впечатлит систему правосудия, контрфактически она верна. Некоторые события могут быть результатом совпадения нескольких факторов, при этом наличие или отсутствие отдельного фактора не способно изменить исход. Эта теоретическая возможность является еще одной причиной того, что, возможно, не имеет особого смысла задавать вопросы типа «Что стало причиной Первой мировой войны?». Вполне возможно, что, несмотря на все факторы, о которых мы любим говорить, устранения любого из них на самом деле было бы недостаточно для предотвращения войны.

Причинность и контрпримеры

Одной из распространенных скептических реакций на доказательства существования среднего эффекта является указание на контрпримеры. Возможно, у вас был подобный опыт на семейном собрании. Вы прочитали исследование, показывающее, что прививки от гриппа в среднем снижают риск заражения гриппом. Вы упоминаете об этом за ужином в честь Дня благодарения, призывая своих близких сделать прививку. Но ваш родственник, скептически относящийся к вакцинам, говорит: «Ерунда, я сделал прививку в прошлом году и все равно заболел гриппом». Многие люди кивают и соглашаются, возможно, указывая на то, что кто-то в их окружении тоже сделал прививку от гриппа и все равно заболел.

В основе подобного возражения в виде контрпримера лежит примерно такая логика: «Если прививки от гриппа действительно предотвращают грипп, то никто, получивший прививку от гриппа, не заболеет гриппом. Таким образом, мой единственный контрпример означает, что вакцина не работает».

Этот аргумент противоречит критическому мышлению. Факты говорят, что прививка от гриппа привела к снижению риска гриппа в среднем среди множества людей, каждый из которых имеет свою уникальную биологию, уровень воздействия вируса, окружающую среду и т. д. Никто не говорит, что прививка устраняет риск заболеть гриппом для каждого человека. Но для того, чтобы риск гриппа снизился в среднем, прививка от гриппа должна была предотвратить грипп (т. е. иметь причинный эффект), по крайней мере, для некоторых людей. Мы просто не знаем точно, на кого подействовала прививка.

Давайте вернемся к нашим обозначениям потенциальных исходов. Потенциальными исходами будем считать наличие и отсутствие заболевания грип-

пом. Мы будем писать $Y = 1$, если вы остались здоровы, и $Y = 0$, если вы заболели гриппом. Воздействием считается прививка от гриппа: $T = 1$ означает, что вы сделали прививку, а $T = 0$ означает отказ от прививки.

Допустим, есть три разных типа людей – назовем их *всегда болеющими*, *никогда не болеющими* и *реагирующими на вакцину*. Первые два типа демонстрируют исходы, которые никогда не зависят от воздействия. Всегда болеющие заболевают гриппом независимо от того, получили ли они прививку от гриппа, а никогда не болеющие никогда не заболевают гриппом. В наших обозначениях

$$Y_{1,\text{всегда болеет}} = 0 \qquad Y_{0,\text{всегда болеет}} = 0$$

и

$$Y_{1,\text{никогда не болеет}} = 1 \qquad Y_{0,\text{никогда не болеет}} = 1.$$

Но люди, реагирующие на вакцину, отличаются; они заболеют гриппом, если не сделают прививку, и не заболеют, если сделают:

$$Y_{1,\text{реагирует}} = 1 \qquad Y_{0,\text{реагирует}} = 0.$$

В популяции, состоящей из этих трех групп людей, прививка от гриппа снижает вероятность того, что вы заболеете гриппом. То есть в среднем эффект лечения положительный. Вы не знаете, в какой группе вы находитесь. Есть вероятность, что вы реагируете на вакцину. Таким образом, прививка от гриппа снижает вероятность заболеть и для вас тоже.

Давайте рассмотрим это на примере. Предположим, есть группа из 10 человек. Участники 1–5 получают прививку от гриппа, а участники 6–10 – нет. Участники 1, 3, 4, 5 и 8 постоянно болеют, поэтому они заболевают гриппом. Участники 5, 6, 7 и 10 никогда не болеют, поэтому остаются здоровыми. Участники 2 и 9 реагируют на вакцину. Участнику 2 делают прививку от гриппа, поэтому он остается здоровым. Но участник 9 не получил прививку, поэтому он заболел.

В табл. 3.2 показаны потенциальные исходы и результаты воздействия. Как мы видим, не у всех представителей этой группы населения наблюдается положительный эффект воздействия. Но средний эффект воздействия на этих 10 человек равен 2/10, потому что двое из десяти реагируют на вакцину. Таким образом, для любого человека, не знающего, к какому типу он принадлежит, существует вероятность 20 % того, что прививка от гриппа предотвратит заболевание гриппом.

Важно отметить, что упоминание единичного контрпримера в данном случае ничего не доказывает. Возможно, вашим невезучим родственником был человек, подобный участнику 1, 3 или 4, у которого обстоятельства сложились так, что прививка от гриппа не дала эффекта (т. е. он болеет в любом случае). Это не значит, что прививка не влияет на других людей.

И это даже не означает, что прививка от гриппа не предотвратит грипп у того же родственника в следующем году или не поможет вам. В отсутствие какой-либо дополнительной информации о принадлежности к группе лучшее предположение любого человека состоит в том, что прививка снизит его шансы заразиться гриппом, поскольку в среднем это так. И мы даже не обсудили более сложную проблему: результаты на самом деле не являются бинарными, поэтому прививка может оказывать причинное влияние на тяжесть протекания гриппа.

Таблица 3.2. Потенциальные исходы заболевания гриппом с прививкой и без нее. Для каждого человека фактический результат, который мы можем наблюдать, выделен полужирным шрифтом. Контрфактический результат, которого мы не наблюдаем, показан обычным шрифтом

		Y_{1i}	Y_{0i}	$Y_{1i} - Y_{0i}$
С прививкой	**Участник 1** (всегда болеет)	**0**	0	0
	Участник 2 (реагирует на вакцину)	**1**	0	1
	Участник 3 (всегда болеет)	**0**	0	0
	Участник 4 (всегда болеет)	**0**	0	0
	Участник 5 (никогда не болеет)	**1**	1	0
Без прививки	**Участник 6** (никогда не болеет)	1	**1**	0
	Участник 7 (никогда не болеет)	1	**1**	0
	Участник 8 (всегда болеет)	0	**0**	0
	Участник 9 (реагирует на вакцину)	1	**0**	1
	Участник 10 (никогда не болеет)	1	**1**	0

Конечно, возможность существования разного эффекта для разных людей порождает еще один набор важных концептуальных проблем. Существуют так называемые *гетерогенные эффекты* воздействия, особенно если они связаны с наблюдаемыми категориями (например, мужчины и женщины, пожилые и молодые, здоровые и больные). Чтобы выявить такие гетерогенные эффекты, мы могли бы провести отдельный эксперимент для каждой группы, который бы показал нам средний эффект для каждой группы, а не для всей популяции. Но что, если эффекты у разных людей различаются по сложным или неясным причинам, которые могут никогда не прийти нам в голову? Далее, когда мы приступаем к рассмотрению эффекта какого-либо воздействия, очень важно помнить, что мы изучаем средний эффект. У некоторых людей эффект может быть намного выше среднего. Другие могут демонстрировать эффект намного меньший, чем в среднем. У некоторых людей эффект может отсутствовать вообще или быть противоположным среднему. Если мы не знаем источник этой неоднородности, мы можем лишь что-то говорить о среднем значении. Такое знание, как мы уже говорили, тоже имеет свою ценность.

Причинность и закон

Как мы кратко упомянули ранее, одним из мест, где философские вопросы о причинности приобретают серьезное практическое значение, является право. Отправление правосудия требует возложения вины и оценки ответственности. Если мы хотим знать, должен ли, скажем, Итан нести ответственность за какой-то вред, причиненный Энтони, нам, конечно же, нужно знать, причинили ли действия Итана этот вред. Но, как уже было сказано, подобные рассуждения чреваты опасными заблуждениями. Многие вещи, от поведения палеозойской рыбы до предполагаемой халатности Итана, могли оказать причинное влияние на вред, нанесенный Энтони. Рыба тоже несет ответственность?

Закон осознает философскую загадку. Но в конечном итоге он должен прийти к какой-то прагматичной резолюции, которая позволит судьям и адвокатам продолжить работу по отправлению правосудия.

В «Общем праве» причинность рассматривается с точки зрения двух условий, тесно связанных с вещами, о которых мы говорили. Их называют *фактической* (cause-in-fact) и *непосредственной* причинностью (proximate causality).

Фактическая причинность – это, по сути, зеркальное отражение контрфактической причинности. Являются ли действия Итана причиной страданий Энтони, зависит от того, пострадал бы Энтони, если бы не действия Итана.

Конечно, как вы уже знаете, контрфактический стандарт не очень строгий. Допустим, Первой мировой войны не было бы, если бы палеозойские рыбы повернули в другую сторону. Означает ли это, что мы должны винить бедную рыбу в Первой мировой войне?

Ответ закона – нет. Вот тут-то и возникает понятие *непосредственности*. Закон требует, чтобы для возникновения ответственности какая-то причина в действительности находилась достаточно близко в причинно-следственной цепочке. Эта мысль также знакома – например, из нашего аргумента о том, что убийство эрцгерцога Фердинанда является более непосредственной причиной Первой мировой войны, чем нос Клеопатры.

Следовательно, юридическая оценка причинно-следственной связи может выглядеть примерно так. Предположим, вы заказываете доставку еды, а курьер неосторожно врезается в машину вашего соседа. Несете ли вы ответственность за ущерб, причиненный соседу? Вполне вероятно, что, если бы не ваше решение заказать доставку, курьер не оказался бы в этом районе и машина вашего соседа не была бы повреждена. Так что ваши действия, вероятно, являются причиной страданий вашего соседа. Но в причинно-следственной цепочке между вашими действиями и столкновением есть много ступеней, и все они находятся вне вашей власти. Таким образом, закон не возлагает на вас ответственность за повреждение машины вашего соседа.

Конечно, как мы уже говорили, точно знать, как применять условия фактической и непосредственной причинности, непросто. Чтобы применить контрфактический тест, мы должны знать, каков правильный контрфактический мир. И определение того, насколько контрфактический мир близок к реальному, является сложной проблемой, полной субъективных суждений. Все это означает, что эти вопросы о причинности неприятны и имеют большое практическое значение.

Может ли причинно-следственная связь распространяться вспять во времени?

Одно из устойчивых убеждений заключается в том, что причинность должна распространяться только вперед во времени. То есть событие, которое происходит сейчас, может повлиять на события, которые произойдут в будущем. Но, как утверждает здравый смысл, события будущего не могут повлиять на события в прошлом. Действительно, одна из распространенных стратегий установления причинно-следственной связи состоит в том, чтобы показать, что предполагаемая причина обычно возникает раньше предполагаемого следствия.

Давайте проверим это рассуждение, подумав об поздравительных открытках. Вот корреляция, которая, как мы надеемся, верна во всем мире: количество поздравительных открыток, которые приходят вам по почте в течение конкретной недели, сильно коррелирует с тем, что они приходятся на неделю после вашего дня рождения. То есть за неделю до вашего дня рождения вам будет отправлено гораздо больше поздравительных открыток, чем за любую другую неделю в году.

Итак, хотя корреляция не обязательно подразумевает причинно-следственную связь, мы подозреваем, что в данном случае существует причинно-следственная связь, но явно не в традиционном понимании. Получение поздравительных открыток не означает, что ваш день рождения наступил благодаря отправке поздравлений. В контрфактическом мире, в котором эти открытки были отправлены в другое время, или даже в контрфактическом мире, где поздравительные открытки не существуют, ваш день рождения все равно никуда не денется и будет приходиться на дату вашего рождения. Наоборот, это ваш день рождения влияет на отправку поздравительных открыток. В контрфактическом мире, в котором ваш день рождения приходится на другой месяц, вам будет отправлено меньше поздравительных открыток за неделю, предшествующую вашему дню рождения в этом мире. Таким образом, согласно нашему контрфактическому определению, ваш день рождения оказывает причинное влияние на поздравительные открытки. Выглядит так, будто эта причинно-следственная связь распространяется назад во времени.

Против этой линии рассуждений существуют очевидные возражения. Например, можно возразить, что не ваш будущий день рождения, а его ожидание оказывает причинное влияние на отправку поздравительных открыток. Если бы мы изменили мнение людей о том, приближается ли ваш день рождения, мы бы изменили их поведение при отправке открыток. Но если бы мы изменили ваш фактический день рождения, не изменив их убеждений, открытки все равно были бы отправлены. Согласно этому аргументу причинность распространяется во времени вперед, как и должно быть.

Но даже этот аргумент не обязательно служит последней точкой в споре. Ведь откуда взялось ожидание вашего дня рождения? Вероятно, оно возникло из-за вашего настоящего дня рождения. Если бы мы изменили дату вашего фактического дня рождения в будущем, мы бы изменили ожидание людьми вашего дня рождения сейчас (что, в свою очередь, изменило бы их поведение при отправке открыток). Возможно, мы вернулись к причинно-следственной связи, идущей назад во времени. Или нет? Действительно ли изменение дня рождения в будущем влияет на ожидания людей сегодня? Или это говорит им

об изменении вашего будущего дня рождения, и в этом случае мы возвращаемся к причинно-следственной связи, направленной вперед во времени?

Как вы можете разумно заметить, наша книга не то место, где следует погружаться в подобные споры. Но мы хотим, чтобы вы ясно увидели две вещи. Во-первых, свидетельство того, что одно событие произошло раньше другого, само по себе не является убедительным доказательством того, что одно стало причиной другого. Во-вторых, независимо от того, считаете ли вы, что причинность может или не может идти вспять во времени, мы всегда можем определить причинные связи с контрфактической точки зрения.

Требует ли причинно-следственная связь физической связи?

Еще одна идея, которую разделяют многие люди, заключается в том, что причинно-следственная связь обязательно сопровождается физической связью, – точка зрения, которую мы будем называть *физикализмом*. Один бильярдный шар влияет на другой, сталкиваясь с ним. Возможно, такие физические связи всегда лежат в основе причинно-следственных связей.

Хотя, конечно, существует множество примеров причинных эффектов, возникающих посредством физической связи, есть веские аргументы в пользу того, что физическая связь требуется не всегда. Представьте человека, которого от ограбления банка удерживает страх перед тюремным заключением. На поведение такого человека влияет наличие полиции, судов, уголовного кодекса и тюремной системы. Система уголовного правосудия влияет на то, совершит ли этот человек преступление, даже если между ними нет физической связи.

Действительно, вспомните наше предыдущее обсуждение влияния дней рождения на отправку поздравительных открыток. Дни рождения сами по себе не являются чем-то физическим. Трудно представить, что вообще может означать причинно-следственная связь между днями рождения и отправкой поздравительных открыток в чисто физическом смысле.

Защитник физикализма мог бы сказать, что, проявив достаточно творческого подхода, мы можем описать влияние системы уголовного правосудия на преступность в чисто физических терминах. Возможно, прошлые аресты и осуждения людей, совершивших преступления, побудили репортеров написать об этой деятельности в газетах. Прочтение об этих арестах в газете через сложную последовательность колебаний света, попадающего в глаза человека, привело к возникновению множества химических и электрических связей в мозгу этого человека, удерживающих его от совершения преступления. Вы можете проделать аналогичное упражнение для дней рождения и поздравительных открыток.

Опять же, мы не дадим однозначного ответа. У обеих сторон дебатов по поводу физикализма могут быть разумные аргументы. Здесь важно лишь то, что мы можем спокойно рассматривать причинные отношения, которые не зависят от очевидных, буквальных физических связей, как в примере с бильярдным шаром.

Причинно-следственная связь не обязательно подразумевает корреляцию

Мы пришли к убеждению, что корреляция не обязательно подразумевает причинно-следственную связь. Но, что, возможно, еще более удивительно, при-

чинно-следственная связь не обязательно подразумевает корреляцию, и уж тем более корреляцию в ожидаемом направлении. Существует множество ситуаций, в которых какая-то характеристика мира оказывает (скажем) *негативное* влияние на какую-то другую характеристику мира, но эти две характеристики мира *положительно* коррелируют (или наоборот).

Вероятно, вы обнаружите сильную положительную корреляцию между количеством пожарных, недавно посетивших дом, и размером ущерба, нанесенного этому дому пожаром. Но нам хорошо известно, что пожарные в среднем уменьшают ущерб от пожара. Другими словами, если бы к дому приехало меньше пожарных, ущерб наверняка был бы еще больше. Так почему же корреляция *положительная*? Дело в том, что пожарные обычно посещают горящие дома. Поэтому, хотя пожарные в некоторой степени уменьшают ущерб от пожара, дома, которые посетило много пожарных, как правило, имеют больший ущерб от пожара. Не следует ожидать, что существование отчетливой причинно-следственной связи автоматически означает существование простой корреляции в том же направлении.

Подведение итогов

Понимание особенностей причинно-следственных связей является одной из фундаментальных целей количественного анализа. Но если мы собираемся это сделать, нужно хорошо разбираться в том, что означает причинность.

Мы считаем, что лучший способ выявления причинности – это мысленный эксперимент, включающий контрфакты. Воздействие оказывает причинное влияние на исход, если при отсутствии воздействия исход был бы другим. Конечно, в реальном мире воздействие было таким, каким оно было. Мы не можем наблюдать воображаемый контрфактический мир, в котором воздействие было другим, чтобы выяснить, отличается ли исход. Это фундаментальная проблема причинного вывода.

Тот факт, что причинные эффекты не наблюдаемы, не означает, что анализ данных не может помочь нам узнать о них. В частности, мы можем измерить средний эффект в некоторой популяции, хотя не можем напрямую наблюдать какие-либо отдельные эффекты.

Для этого необходимо осторожно использовать количественные знания о таких вещах, как корреляции. Во второй части мы переходим к более подробному обсуждению того, как можно обнаружить и количественно оценить корреляцию. Это позволит нам применить критическое мышление в части III об оценке причинно-следственных связей.

Ключевые термины

- **Причинно-следственная связь:** говоря упрощенно, это изменение какой-либо характеристики мира, которое может произойти в результате изменения другой характеристики мира. Более формально это разница в потенциальных исходах для некоторого наблюдаемого объекта при двух разных статусах воздействия.
- **Body Vibes:** наклейки, которые, как утверждает компания Goop, вызывают очищение кожи. Авторы этой книги не одобряют Body Vibes, главным

образом потому что собираются выпустить конкурирующий товар: Brain Vibes. Одной наклейки на лбу достаточно для появления критического мышления.

- **Контрфактическое сравнение:** сравнение вещей в двух разных мирах или положениях дел, по крайней мере одно из которых на самом деле не существует.
- **Воздействие:** термин, которую мы используем для описания любого вмешательства в явления мира. Обычно мы используем этот термин, когда рассматриваем причинно-следственную связь, потому что стремимся сравнить исходы с воздействием и без него. Воздействие может принимать самые разнообразные формы и применяться ко всему, что происходит в мире.
- **Система потенциальных исходов:** математическая основа для представления контрфактических явлений.
- **Потенциальный исход:** для некоторого объекта при определенном статусе воздействия это – исход, который объект получит при данном (возможно, контрфактическом) статусе воздействия.
- **Фундаментальная проблема причинного вывода:** отражает тот факт, что, поскольку в любой момент времени мы наблюдаем данный объект только в одном статусе воздействия, мы никогда не можем напрямую наблюдать причинную связь с воздействием.
- **Гетерогенный отклик на воздействие:** когда отклик на воздействие не одинаков для всех наблюдаемых объектов (как в случае прививки от гриппа и практически любого другого интересного примера причинно-следственной связи), мы говорим, что наблюдается гетерогенность (неоднородность) отклика. Иногда нас по-прежнему интересует средний отклик, хотя мы знаем, что отклик на воздействие не однороден, а иногда мы хотим подробно изучить природу неоднородности. (Напротив, при обсуждении маловероятной возможности того, что отклики на воздействие будут одинаковыми для всех наблюдаемых объектов, мы будем иметь в виду однородность отклика.)

Упражнения

3.1. Сара говорит, что голодна. Джон протягивает ей кусок пиццы. Сара съедает пиццу, а затем заявляет, что больше не голодна.

a) Фундаментальная проблема причинного вывода, по-видимому, заключается в том, что вы не можете знать, что именно из-за поедания пиццы Сара больше не голодна. Верно ли это утверждение? Объясните свой ответ.

b) Считаете ли вы, что у вас так или иначе есть веские основания полагать, что именно употребление пиццы повлияло на то, что Сара больше не голодна? Объясните свой ответ.

c) Есть ли у вас веские основания полагать, что передача пиццы от Джона к Саре оказала причинное влияние на исчезновение чувства голода? Как вы думаете, какое событие имеет большее причинное влияние на исчезновение чувства голода – передача пиццы или ее поедание?

3.2. Допустим, правительство рассматривает возможность запрета употребления алкоголя в рамках кампании общественного здравоохранения. Мы будем рассматривать появление закона о запрете алкоголя как воздействие T. Мы принимаем $T = 1$, если правительство делает алкоголь незаконным, и $T = 0$, если правительство оставляет алкоголь легальным.

Для каждого человека мы рассматриваем бинарный исход: либо он пьет алкоголь, либо нет. Если человек i пьет при статусе воздействия T, мы записываем его потенциальный исход как $Y_{Ti} = 1$, а если он не пьет, записываем его как $Y_{Ti} = 0$.

Предположим, общество состоит из трех групп: всегда пьющих, умеренно пьющих и никогда не пьющих. Всегда пьющие будут пить независимо от того, разрешен алкоголь или нет. Умеренно пьющие будут пить тогда и только тогда, когда алкоголь является легальным. Никогда не пьющие не будут пить независимо от того, разрешен алкоголь или нет.

a) Запишите в форме потенциальных исходов и в форме числа (0 или 1) каждый из двух потенциальных исходов для каждой из трех групп.
b) Запишите в форме потенциальных исходов и в форме числа (0 или 1) причинное влияние закона о запрете алкоголя на употребление алкоголя для каждой из трех групп.
c) Есть ли в среднем эффект от запрета алкоголя в этом обществе?
d) Предположим, вы обедаете с друзьями, и один из них говорит: «Мой дядя живет в стране, где запрещен алкоголь, и все его друзья продолжают пить, поэтому я не думаю, что запрет на что-то влияет». Используя условия этого задания, поясните, почему это неубедительный аргумент.

3.3. Национальный комитет Республиканской партии (RNC) нанял трех консультантов и попросил их выяснить причину поражения республиканцев на президентских выборах 2020 г. Первый консультант говорит, что они недостаточно занимались телевизионной рекламой. Второй консультант заключил, что республиканцам следовало больше призывать своих сторонников голосовать, а не углубляться в критику голосования по почте. Третий консультант утверждает, что Дональду Трампу следовало бы лучше реагировать на пандемию COVID-19 и проявить больше сострадания в ходе предвыборной кампании. Смущенный явно противоречивой информацией, RNC нанимает вас как количественного аналитика, чтобы сделать выбор между этими тремя вариантами. Что бы вы им сказали? Как бы вы поступили?

3.4. На турнире по гольфу US Open 2016 г. Дастин Джонсон лидировал в финальном раунде, и его мяч находился на пятом грине. Готовясь к предстоящему удару, он постучал клюшкой по земле рядом с мячом, и мяч сдвинулся с места. В правилах того времени говорилось: если очевидно, что игрок заставил свой мяч сдвинуться с места, даже если это было непреднамеренно, он должен понести штраф. Поскольку вы являетесь экспертом в области причинно-следственных связей, представители судейской коллегии вызывают вас, чтобы оценить ситуацию. Судьи приводят следующие аргументы. Пожалуйста, дайте свой экспертный ответ на каждый из них.

a) Джонсон не мог заставить мяч сдвинуться с места, потому что он (и его клюшка) никогда не касался его.
b) Джонсон не должен получить штраф, поскольку истинной причиной перемещения мяча был садовник. Если бы в то утро садовник не подстригал траву, мяч бы не сдвинулся с места.
c) Судья, мыслящий эмпирически, вышел на тот же грин, положил мяч, постучал клюшкой по земле рядом с мячом, и он не сдвинулся с места. Следовательно, действия Джонсона не могли привести к перемещению мяча.
d) Один судья внимательно наблюдал за инцидентом и практически уверен, что, если бы Джонсон не постучал клюшкой рядом с мячом, тот не сдвинулся бы с места. Следовательно, игрок заставил мяч сдвинуться с места и должен получить штраф.

Дополнительное чтение и ссылки

Цитата Блеза Паскаля о носе Клеопатры взята из его сборника XVII в. под названием «Pense´es».

Эссе о контрфактических рассуждениях, обсуждающих ген, контролирующий длину носа Клеопатры: James D. Fearon. 2011. *Counterfactuals and Hypothesis Testing in Political Science.* World Politics 43 (2): 169–195.

Если вы хотите узнать больше о контрфактическом определении причинности, потенциальных исходах и сопутствующих дискуссиях и дебатах, взгляните на эти статьи:

David Lewis. 1973. *Causation.* Journal of Philosophy 70: 556–67;

Paul W. Holland. 1986. *Statistics and Causal Inference.* Journal of the American Statistical Association 81 (396): 945–60;

Stephen Mumford and Rani Lill Anjum. 2014. *Causality: A Very Short Introduction.* Oxford University Press.

Есть также хорошая статья Питера Мензиса и Хелен Биби в Стэнфордской энциклопедии философии: *https://plato.stanford.edu/entries/causation-counterfactual/*.

ЧАСТЬ II

Существует ли взаимосвязь?

Глава 4

Не бывает корреляции без вариаций

О чем эта глава

- Вы не можете узнать о корреляции без изменения обеих интересующих переменных.
- Во многих сферах жизни – от образования до медицины и ракетостроения – люди попадают в ловушку, пытаясь делать заявления о корреляциях, не меняя переменные.
- Особенно часто люди совершают эту ошибку, выбирая зависимую переменную – исследуя только случаи, когда какое-то явление имело место, а не сравнивая случаи, когда оно имело место, со случаями, когда оно не происходило.
- Многие институциональные процедуры заставляют нас выбирать зависимую переменную без нашего ведома.

Введение

В главе 2 мы обсуждали идею о том, что корреляция между двумя характеристиками мира – это степень, в которой они имеют тенденцию проявляться вместе. Мы начали обсуждение корреляции с размышлений о том, коррелируют ли добыча нефти и автократия. Чтобы выяснить это, мы рассмотрели данные на уровне страны, представленные в табл. 4.1.

Таблица 4.1. Объем добываемой нефти и тип политической системы

	Некрупный производитель	Крупный производитель	Итого
Демократия	118	9	127
Автократия	29	11	40
Итого	147	20	167

Чтобы определить, существует ли корреляция между добычей нефти и автократией, мы сравнили долю крупных производителей нефти, являющихся автократиями, с долей стран, не являющихся крупными производителями нефти,

но являющихся автократиями. Чтобы провести это сравнение, нам потребовалась информация, состоящая из четырех частей: количества автократий, которые являются крупными производителями нефти, количества демократий, являющихся крупными производителями нефти, количества автократий, которые не являются крупными производителями нефти, и количества демократий, являющихся крупными производителями нефти. Если бы у нас не было какой-либо из этих частей информации, мы бы не смогли выяснить, коррелированы ли добыча нефти и автократия.

Чтобы понять, почему так, предположим, что мы не знаем, сколько демократий являются крупными производителями нефти. (Конечно, в таком случае мы не должны знать общее количество стран, потому что иначе могли бы вычесть количество стран в трех других категориях из общего числа стран.) Мы все равно знаем, что около 20 % (29) стран, которые не являются крупными производителями нефти, – автократии. Но сейчас мы не можем выяснить, какая часть крупнейших производителей нефти является автократиями. Это может быть что угодно. Если бы число демократических стран, являющихся крупными производителями нефти, оказалось (допустим) 11, то 50 % (11/22) крупнейших производителей нефти были бы автократиями, и была бы положительная корреляция. Если бы число демократических стран, являющихся крупными производителями нефти, оказалось (допустим) 99, то только 10 % (11/110) крупнейших производителей нефти были бы автократиями, поэтому была бы отрицательная корреляция. Если бы число демократий, являющихся крупными производителями нефти, оказалось равным 44, то 20 % (11/55) крупнейших производителей нефти были бы автократиями – так же, как и для стран, которые не являются крупными производителями нефти, – и не было бы никакой корреляции вообще. Как мы дополнительно убедились при обсуждении скандалов и выборов в конгресс в главе 2, чтобы выяснить корреляцию, нам необходимо изучить все четыре части информации.

Именно это мы имеем в виду, когда говорим, что корреляция требует вариаций: если вы хотите выяснить, коррелируют ли две переменные, вам нужно наблюдать вариации обеих из них. Вы должны рассматривать количество стран, которые являются и не являются крупными производителями нефти. И вы должны рассматривать количество автократий и демократий в каждой группе. Просто наблюдать за изменением той или иной переменной недостаточно. В главе 2, когда мы спросили, какое из пяти фактических утверждений описывает корреляцию, проблема с тремя утверждениями, которые не описывали корреляцию, заключалась в отсутствии вариаций в одной из переменных.

Хотя на основе нашего простого бинарного примера может показаться очевидным, что корреляция требует вариаций, по нашему опыту это совсем не так. На самом деле неспособность найти вариацию той или иной переменной при попытке установить корреляцию является исключительно распространенной ошибкой.

В данной главе мы исследуем эту ошибку и попытаемся понять, почему она так распространена. В целом мы считаем, что существуют две тесно связанные причины, по которым люди так часто пытаются найти корреляцию без вариаций. Первая причина заключается в выборе зависимой переменной. Вторая – в том, что мир часто устроен таким образом, что подталкивает нас совершать эту ошибку.

Эта глава больше, чем остальные главы в книге, построена на примерах. Мы делаем это не просто так. Мы обнаружили, что, как только мы начинаем объяснять, что корреляция требует вариаций, люди охотно кивают головой в знак согласия, делая вид, что понимают. Действительно, поскольку эта мысль кажется очевидной, если изложить ее простым языком, многие люди скептически относятся к тому, что это может быть большой проблемой. И тем не менее они снова и снова наступают на те же грабли. Мы надеемся, что, показав вам множество примеров того, как очень умные люди совершают эту ошибку в условиях высоких ставок, мы убедим вас, что это реальная проблема и чтобы избежать этой ошибки, требуется критическое мышление, искренние усилия и концентрация внимания.

Выбор зависимой переменной

Если вы хотите спрогнозировать или объяснить какое-то явление, естественным побуждением будет начать с изучения предыдущих случаев возникновения этого явления. Это называется *выбором зависимой переменной*. Но если вы смотрите только на случаи, когда явление имело место, вы пытаетесь оценить корреляцию без вариаций, поскольку у вас нет различий в том, произошло это явление или нет. Это все равно, что искать корреляты автократии, не исследуя никаких демократий. Это не сработает.

Термин «зависимая переменная» относится к переменной, представляющей явление, которое вы пытаетесь спрогнозировать или объяснить. Эта ошибка называется выбором зависимой переменной, поскольку вы выбираете, какие случаи рассматривать, основываясь исключительно на значении зависимой переменной (например, рассматривая только автократии), а не рассматривая вариации зависимой переменной.

Рассмотрим несколько примеров. После финансового кризиса 2008 г. ученые и журналисты, которые хотели понять, как предсказывать будущие финансовые кризисы, потратили огромное время и энергию на изучение исторических данных, чтобы найти закономерности в предыдущих кризисах. Малкольм Гладуэлл в своей книге *Outliers* («Аномалии») пытается понять корреляты личного успеха, рассматривая жизни весьма успешных людей в поисках сходства. Конгресс, рассматривая возможность изменения американской стратегии борьбы с повстанцами в Афганистане, заслушал свидетельства о коррелятах террористов-смертников от академического эксперта, который провел исчерпывающее исследование всех кампаний террористов-смертников с 1980 г. в поисках общих характеристик.

Каким бы естественным ни казался поиск общих признаков в прошлых событиях, на основе которых вы строите предсказание, на самом деле это ошибка. Корреляция требует вариаций. Каждое из только что описанных исследований было бы гораздо более информативным, если бы в них присутствовали вариации зависимой переменной.

Утверждение о том, что мы не можем узнать о коррелятах финансовых кризисов или террористов-смертников, ища общие черты между историческими случаями подобных событий, может показаться нелогичным. Но поскольку мы знаем, что корреляция требует вариаций, ошибку на самом деле довольно

легко уловить. Говоря языком нашего предыдущего примера, каждый из этих примеров аналогичен поиску коррелятов добычи нефти без каких-либо данных о странах, *не* производящих нефть!

Чтобы взглянуть на ключевой концептуальный недостаток всех этих аргументов с другой стороны, давайте начнем с рассмотрения центрального утверждения в книге Гладуэлла – так называемого *правила 10 000 часов*.

Правило 10 000 часов

Идея Гладуэлла состоит в том, что для освоения любого сложного навыка требуется около 10 000 часов серьезной практики. Талант тоже важен, но прежде всего, если вы ищете выдающегося человека, присмотритесь к тем, кто посвятил этому 10 000 часов практики.

Разумеется, Гладуэлл не просто заинтересован в прогнозировании большого успеха. Он считает, что правило 10 000 часов может иметь причинное значение. Если это правда, то следование правилу обещает далеко идущие последствия. При достаточной практике, возможно, любой из нас сможет достичь практически чего угодно.

Но говорить о причинно-следственной связи преждевременно. Прежде чем мы сможем задуматься о причинно-следственной связи, нам нужно выяснить, убедительны ли аргументы Гладуэлла в пользу утверждения о корреляции между 10 000 часов практики и большим успехом.

Гладуэлл спрашивает: «Является ли правило 10 000 часов общим правилом успеха?» Ответ – да, заключает он. Доказательство? «Если мы заглянем под поверхность каждого великого человека», мы увидим общую картину. «Практически каждая история успеха… предполагает, что кто-то или некая группа работают усерднее, чем их коллеги». В каждом случае, от Билла Гейтса до Beatles, Гладуэлл показывает, что великие люди тратят на практику в своем деле не менее 10 000 часов – по его мнению, неопровержимое доказательство того, что практика приносит успех.

Давайте попробуем более критически взглянуть на доказательства Гладуэлла. Что показал нам Гладуэлл? Конечно, он на самом деле не рассматривал всех великих людей. Но он продемонстрировал нам доказательства того, что многие выдающиеся люди практикуются в своем деле не менее 10 000 часов. Главная проблема в том, что он ничего не рассказал нам обо всех людях, которые *не* достигли больших успехов. Таблица доказательств для книги *Outliers* будет выглядеть примерно так, как показано в табл. 4.2.

Таблица 4.2. Великие достижения требуют более 10 000 часов практики

	Выдающаяся личность	**Невыдающаяся личность**	**Итого**
Есть 10 000 часов практики	Многие	?	
Нет 10 000 часов практики	Очень редко	?	
Итого			

Даже если допустить, что Гладуэлл прав в том, что большинство выдающихся людей тратят на практику 10 000 часов, это не говорит нам о том, коррелируют

ли 10 000 часов практики с большим успехом. Корреляция требует вариаций. Поскольку он выбрал зависимую переменную, в данных Гладуэлла отсутствуют различия в достижениях. Если вы хотите знать, коррелируют ли 10 000 часов практики с успехом, недостаточно заметить, что большинство выдающихся людей практикуются по 10 000 часов. Нам также необходимо знать о поведении тех, кто *не* достигает выдающихся успехов.

Конечно, анализ Гладуэлла действительно дает некоторую информацию, которой у нас раньше не было. Давайте на мгновение предположим, что Гладуэлл не подбирал истории так, чтобы они соответствовали его повествованию (хотя, конечно, он это делал: он рассказчик, а не ученый). В этом случае мы узнали, что наиболее успешные люди тратят 10 000 часов практики, прежде чем добиться большого успеха.

Хотя этой информации недостаточно для измерения корреляции, Гладуэлл и его защитники могут утверждать, что у нас уже есть интуитивное ощущение, что большинство представителей широкой общественности, не являющихся выдающимися людьми, не потратили 10 000 часов практики. В этом случае, возможно, анализ Гладуэлла существенно меняет наши представления о корреляции между практикой и большим успехом, даже если он явно не измерял эту корреляцию. В тех случаях, когда у нас уже есть хорошее представление о распространенности чего-либо среди населения в целом, возможно, полезно показать, что распространенность различна для определенных групп.

Может быть. Но мы по-прежнему скептически относимся к тому, что анализ Гладуэлла многому нас учит. Дело в том, что многие люди (если не большинство) посвятили чему-то как минимум 10 000 часов практики. Энтони провел на поле для гольфа 10 000 часов, и он не Тайгер Вудс. Итан потратил 10 000 часов, играя на гитаре, и он не Джими Хендрикс. Если вы проработали где-то полный рабочий день в течение пяти лет, но не являетесь самым успешным человеком в своей области, то вы один из многих, многих людей в верхней правой ячейке табл. 4.2, которых Гладуэлл никогда не принимал во внимание.

Следует также помнить, что Гладуэлл – одаренный рассказчик. В крайне маловероятном сценарии, когда Энтони выиграет Кубок мастеров, Гладуэлл мог бы написать вдохновляющую и убедительную историю о том, как, несмотря на то что Энтони был штатным профессором колледжа, многие годы практики, неудач и новых тренировок позволили ему добиться величайших результатов. История Золушки в мире спорта (давайте на минутку помечтаем). Но гораздо более вероятно, что Энтони с радостью продолжит оставаться одним из миллионов, если не миллиардов, людей, которые что-то любят, усердно над этим работают, но никогда не достигают огромных успехов и которые никогда не учитываются в анализе Гладуэлла.

Чтобы проверить ваше понимание, давайте рассмотрим проблему с утверждениями, подобными утверждениям Гладуэлла, в другом контексте. Мы собираемся повторить его точную аргументацию, но на вымышленном примере, который, как мы надеемся, сделает проблему еще яснее.

Предположим, в городе с населением 10 000 человек наблюдается неожиданная волна заболеваний. В течение месяца с одинаковыми симптомами заболевают 500 человек. Местные работники здравоохранения хотят определить причину заболевания. Они изучают истории болезни 500 больных людей

и ищут общие черты. В ходе расследования они обнаружили, что все 500 человек употребляли один и тот же напиток из одного и того же источника за день до госпитализации.

В табл. 4.3 показаны данные, соответствующие нашей вымышленной истории.

Таблица 4.3. Что пили больные (вымышленные данные)

	Заболели	Не заболели	Итого
Пили напиток	500		
Не пили напиток	0		
Итого	500		

Факты о напитке и болезни в точности соответствуют фактам о практике и успехе в книге *Outliers*. Все, кто заболел (добился успеха), выпили один и тот же напиток (за 10 000 часов). Несомненно, употребление этого напитка (10 000 часов практики) выглядит важным предиктором болезни (большого успеха). Если мы хотим знать, кто еще может заболеть, нам следует обследовать весь город и выяснить, кто еще пил тот же напиток. Верно?

Предположим, рассматриваемый напиток – это водопроводная вода. Утверждение о том, что «характеристика» заболевания предполагает корреляцию между напитком и заболеванием, сразу становится сомнительным. Почему? Потому что почти все люди потребляют водопроводную воду каждый день. Действительно, в нашем вымышленном городе все 500 заболевших человек пили воду из-под крана, но то же самое делали и 9500 человек, которые не заболели. Как видно из табл. 4.4, на самом деле нет никакой корреляции между употреблением напитка и заболеванием: напиток пили 100 % больных и 100 % здоровых людей.

Таблица 4.4. Что пили больные и здоровые люди (вымышленные данные)

	Заболели	Не заболели	Итого
Пили напиток	500	9 500	10 000
Не пили напиток	0	0	0
Итого	500	9 500	10 000

Аналогичным образом данные, предоставленные Гладуэллом, не подтверждают правило 10 000 часов. Да, многие успешные люди очень усердно практикуются в своем деле. То же самое делают и многие менее успешные люди. Подумайте обо всех группах, которые репетировали бесчисленные часы, отыграли бесчисленное количество концертов и так и не стали Beatles.

Деградация молодежи

Американские дети, которые любили рок-музыку в 1980-х гг., возможно, помнят Центр музыкальных ресурсов для родителей (PMRC). PMRC представлял собой группу лоббистов, члены которой выступали против того, что они счи-

тали чрезвычайно неуместным содержанием рок-музыки. Самой известной среди основателей PMRC была Типпер Гор, жена тогдашнего сенатора, а затем вице-президента Эла Гора, которая основала PMRC после того, как была шокирована текстом песни Принса.

PMRC заявил, что откровенные тексты песен развращают молодежь, провоцируя самоубийства, сексуальное насилие и даже убийства. Они осудили «порно-рок» – категорию, в которую попал даже Брюс Спрингстин, поскольку песня «I'm on Fire» содержала сексуальный подтекст – и потребовали размещать на альбомах предупреждающие надписи. В 1985 г. Комитет Сената по торговле, науке и транспорту провел слушания. Музыканты всего музыкального спектра, от кантри-певца Джона Денвера до Ди Снайдера из Twisted Sister, выступили против позиции PMRC. Но PMRC победил.

Давайте рассмотрим некоторые аргументы. Вот заявление Джеффа Линга, консультанта PMRC:

> «Многие альбомы сегодня содержат песни, поощряющие самоубийства, агрессию, сексуальное насилие и насилие просто ради насилия… Вот, например, Стив Баучер. Стив умер во время прослушивания песни AC/DC "Shoot to Thrill". Стив выстрелил себе в рот из отцовского пистолета… Несколько дней назад я выступал с речью в Сан-Антонио. За день до моего приезда похоронили юного старшеклассника. Этот молодой человек взял свой кассетный магнитофон на футбольное поле. Он повесился, слушая песню AC/DC "Shoot to Thrill". Суицид стал в нашей стране эпидемией среди подростков. В этом году около 6000 человек покончат с собой. Многие из этих молодых людей находят поддержку у некоторых рок-звезд, которые представляют смерть как позитивную, почти привлекательную альтернативу… Конечно, AC/DC не новички в производстве жестокого контента… Один из их сторонников, о котором я знаю, – это пресловутый Night Stalker».

Аргумент Линга, типичный для участников крестового похода против развращения молодежи, сводится к следующему:

1) некоторые молодые люди ведут себя отвратительно;
2) вся молодежь, которая ведет себя отвратительно, слушает эту ужасную рок-музыку;
3) музыка является причиной отвратительного поведения.

Конечно, разговоры о причинно-следственной связи вновь преждевременны. Мы сосредоточимся на том, предполагают ли такие данные хотя бы корреляцию.

Тридцатью годами ранее, в 1954 г., Сенат услышал поразительно похожие свидетельства о биче молодежи тогдашнего поколения – комиксах. Вот фрагмент речи невролога и психиатра Фредрика Вертама, выступавшего перед подкомитетом Сената:

> «В городе штата Нью-Йорк есть школа, где процветают воровство и насилие. Некоторое время назад дети напали на одного мальчика и так сильно вывернули ему руку, что она сломалась в двух местах и, как в комиксах, кость пробила кожу.
>
> Примерно через 10 дней в той же школе семь мальчиков набросились на другого мальчика и ударили его головой о бетон, так что мальчик потерял со-

знание, и его пришлось доставить в больницу. У него было сотрясение мозга.

В этом же городе в средней школе за один год забеременели 26 девочек. В этом году, по-моему, восемь. Возможно, сейчас уже девять.

Господин председатель, это то, что я называю этическим и моральным разложением. Я не думаю, что кто-то из этих мальчиков или девочек по отдельности сильно отличается от других детей. Общую картину невозможно объяснить индивидуальными причинами.

Здесь налицо всеобщая моральная дезориентация, и я думаю, что этих девушек соблазнили морально задолго до того, как соблазнили физически, и, конечно, все эти люди очень и очень склонны – не все, но большинство из них – к постоянному чтению комиксов».

Подобные аргументы встречаются и в наше время. Мы все слышали и, возможно, даже сами делали подобные заявления о коварном влиянии телевидения, видеоигр или социальных сетей. Например, после ужасающей стрельбы в средней школе Columbine Министерство образования США и Секретная служба создали совместную рабочую группу, чтобы определить, какие факторы позволят школьным властям предвидеть и предотвратить школьное насилие. Рабочая группа изучила все 37 случаев школьного насилия с 1974 по 2000 г. Придя к выводу, что не существует единого профиля школьного стрелка, они также сообщили следующее (среди прочего):

1) «До нападения многие нападавшие чувствовали себя запуганными, преследуемыми или уязвленными со стороны других школьников»;
2) «Известно, что большинству нападавших было трудно справиться со значительными потерями или личными неудачами»;
3) «Большинство злоумышленников до инцидента совершали какие-то поступки, которые вызывало беспокойство у других или указывали на необходимость в помощи»;
4) «Более половины нападавших продемонстрировали определенный интерес к насилию в фильмах, видеоиграх, книгах и средствах массовой информации».

Похожая комиссия была созвана в 2018 г. Хоть она и меньше фокусировалась на конкретных развратителях молодежи, но и эта комиссия порой ограничивалась отбором по зависимой переменной. Например, в главе, в которой рекомендуется уделять больше внимания воспитанию личности, комиссия отмечает, что многие школьные стрелки испытали социальную изоляцию, не сравнивая ее с уровнем социальной изоляции среди тех, кто не участвует в насилии:

«После стрельбы в Паркленде многочисленные сообщения указывали на то, что предполагаемый стрелок в годы, предшествовавшие нападению, испытывал чувство изоляции и депрессии. Виновные в предыдущих расстрелах в школах разделяли это чувство отстраненности. Например, стрелок из школы Columbine был охарактеризован как депрессивный и затворнический тип. Члены семьи и знакомые стрелка из Технологического института Вирджинии сказали, что, по мере того как его изоляция росла в последний год обучения, его “внимание к учебе и учебному времени упало”. То же самое было и в Sandy Hook».

Иногда комиссия избегает выбора зависимой переменной. В главе о психическом здоровье они пишут:

> «Лица, совершающие массовые расстрелы, могут иметь или не иметь серьезное психическое заболевание. Существует недостаточно данных на уровне популяции, подтверждающих мнение о том, что люди с диагнозом психического заболевания чаще, чем кто-либо другой, совершают преступления с применением огнестрельного оружия».

Но вскоре после этого они возвращаются к аргументам, из которых следует, что они ищут корреляцию без вариаций:

> «Анализ Министерства образования США и Секретной службы США показал, что около четверти лиц, совершивших массовые расстрелы, лечились от психических заболеваний. Такие люди часто чувствуют себя обиженными и крайне злыми и питают фантазии о жестокой мести».

Это не единственные подобные правительственные отчеты; такой анализ кажется неизбежным после актов насилия среди молодежи. Но по причинам, которые мы уже видели, эти выводы, как и приведенные выше показания в Сенате, вводят в заблуждение. Даже если бы было правдой, что практически каждый молодой человек, который вызывает беспокойство, также слушает рок, читает комиксы или играет в видеоигры, это не установило бы корреляцию между таким поведением и предполагаемыми развратителями молодежи. Корреляция требует вариаций. Доказательства предположения о том, что дети, которые занимаются такими видами деятельности, с большей вероятностью будут проявлять насилие, чем дети, которые не участвуют в таких действиях, должны включать сравнение этих двух типов детей.

Если мы хотим знать, существует ли связь между каким-то сомнительным увлечением молодежи и насилием, мы не должны выбирать зависимую переменную, т. е. мы должны сравнивать жестоких детей с обычными и смотреть, являются ли жестокие дети более вероятными сторонниками сомнительных занятий, чем остальные. (И даже в этом случае мы не можем сказать, что эта связь – причинно-следственная!) Тот факт, что даже эксперты не могут критически осмыслить собственные выводы, означает, что, несмотря на все экспертные мнения, высказанные по этой теме, мы знаем гораздо меньше, чем могли бы, о корреляции насилия среди молодежи.

Уход из средней школы

Давайте пока прервем обсуждение проблемной молодежи. В Америке начала XXI в. возникла проблема с окончанием средней школы. Хотя экономическая отдача от образования находится на рекордно высоком уровне, почти треть учащихся государственных школ не успевают закончить среднюю школу вовремя. Более 10 % никогда не заканчивают обучение.

В 2006 г. Фонд Билла и Мелинды Гейтс решил выделить средства на решение этой проблемы. В качестве одного из шагов в поисках решения они заказали исследование причин прекращения учебы в средней школе. Основная идея доклада заключается в том, что отчисление из школы не связано в первую

очередь с тем, о чем вы могли подумать, – с проблемами дома, отсутствием академической подготовки или прослушиванием рок-музыки. Напротив, главная проблема заключается в том, что дети не вовлечены в образовательный процесс и находят школу скучной.

Как говорится в отчете, «почти половина (47 %) прервавших обучение заявили, что основной причиной ухода из школы были неинтересные занятия». И «почти 7 из 10 респондентов (69 %) заявили, что у них нет мотивации или вдохновения усердно работать».

К сожалению, поскольку корреляция требует вариаций, данные в этом исследовании Фонда Гейтса, как и доказательства, представленные до него PMRC и лобби против комиксов, довольно неинформативны.

Тот факт, что половина бросивших школу считает учебу скучной, не означает, что скучность школы коррелирует с уходом из нее. Поскольку корреляция требует вариаций, ее измерение должно включать сравнение тех, кто бросил школу, с теми, кто не бросил, чтобы увидеть, насколько велика доля бросивших школу среди тех, кто считает учебу скучной. Исследование Фонда Гейтса, поскольку оно рассматривает только тех, кто бросил школу, не может провести такое сравнение.

Этот момент не просто придирка. На секунду задумайтесь вот о чем. Оба автора этой книги учились в средней школе. Ни один из них не бросил учебу. Однако оба автора вспоминают, что некоторые занятия казались им весьма скучными. Вероятно, вы можете сказать о себе то же самое.

Наш личный опыт также не является убедительным доказательством. Итак, давайте посмотрим, сможем ли мы добиться большего, чтобы выяснить, действительно ли скучная учеба является ключевым фактором, предсказывающим уход из школы. Исследователи из Университета Индианы провели общенациональный репрезентативный опрос старшеклассников в 2009 г. Большинство из этих учеников не собираются бросать учебу, однако исследователи сообщают, что «двум из трех респондентов (66 %) в 2009 г. бывает скучно в школе практически каждый день». Это даже больше, чем 50 % бросивших школу, которые считают школу скучной, согласно исследованию Фонда Гейтса.

Но давайте будем осторожны. Есть много причин, по которым нельзя сравнивать опрос Фонда Гейтса и опрос Университета Индианы. Они выбирают разные группы студентов, задают разные вопросы и рассматривают учеников разных классов. Поэтому мы не хотим делать поспешных выводов. Но, по крайней мере, опрос Университета Индианы должен заставить вас задуматься о том, что на самом деле школьная рутина является обычным явлением для подавляющего большинства старшеклассников, а не только для тех, кто бросает учебу.

Будущее школьного образования – это серьезно. Замечательно, что Фонд Гейтса пытается улучшить образование. Но их исследование игнорирует ключевой принцип критического отношения к данным; они пытаются узнать о корреляции неоконченного образования без анализа различий между вариантами. Этот подход не может сработать.

Атаки смертников

В 2009 г. профессор Чикагского университета и известный эксперт по терроризму Роберт Пейп выступил в подкомитете Палаты представителей по противо-

действию терроризму. Темой слушаний было предложение генерала Стэнли Маккристала о наращивании сорокатысячного военного контингента для борьбы с повстанческим движением Талибана в Афганистане. Вот что сказал Пейп:

> «Картина ясна: чем больше западных войск входит в Афганистан, тем больше местные жители считают себя находящимися под иностранной оккупацией и используют самоубийства и другие акты террора, чтобы противостоять ей… Как показывает мое исследование террористов-смертников во всем мире с 1980 г., ими движет не существование террористического убежища, а присутствие иностранных сил на земле, которую они ценят. Поэтому неудивительно, что американские войска производят антиамериканских террористов-смертников».

Пейп далее рекомендует серьезно переосмыслить американскую военную стратегию в Афганистане. Его аргумент основан на утверждении, что нападения террористов-смертников в первую очередь мотивированы иностранной оккупацией. Его доказательствами являются данные, которые он собрал и проанализировал в статьях и двух книгах о каждой террористической кампании смертников в мире, начиная с 1980 г.

Аргумент звучит правдоподобно. В Афганистане силы США подверглись нападению террористов-смертников, которые хотели, чтобы Соединенные Штаты покинули страну. Террористы-смертники из группировки «Тамильские тигры» атаковали правительство в Шри-Ланке, которое, по их мнению, оккупировало их родину. Палестинские террористы-смертники нападают на израильтян, утверждая, что они являются иностранными оккупантами. Кажется, что оккупация является основным коррелятом атак террористов-смертников.

Тем не менее мы считаем спорным утверждение о том, что практически каждая атака террористов-смертников направлена против иностранных оккупантов. (Например, хотя Усама бен Ладен утверждал, что американские войска, дислоцированные на военных базах в Саудовской Аравии по приглашению саудовского правительства, были оккупационными силами, можно ли сказать, что Америка оккупировала Саудовскую Аравию?) Но в целях рассуждения давайте предположим, что основной фактический принцип утверждения верен. Означает ли это, что существует корреляция между иностранной оккупацией и нападениями смертников?

Разумеется, нет. Корреляция требует вариаций. Чтобы понять корреляты атак террористов-смертников, вы не можете просто изучать каждый отдельный случай атак террористов-смертников и искать общие черты. Это выбор зависимой переменной. Вы должны сравнить конфликты с атаками террористов-смертников с конфликтами без них.

В этом случае проще всего просто взглянуть на каждую страну и задаться вопросом: действительно ли оккупированные иностранцами страны чаще подвергаются нападениям террористов-смертников, чем страны, которые не оккупированы иностранцами? Оказывается, в недавнем исследовании социологи провели именно это сравнение и обнаружили, что ответ отрицательный. В частности, если мы сравним оккупированные и неоккупированные страны, разница в вероятности стать жертвой самоубийственного террора составит менее одного процентного пункта!

В чем же дело? Все те примеры террористов-смертников, которые мы перечислили, связаны с нападениями на иностранных оккупантов. Как могло случиться, что между иностранной оккупацией и нападениями террористов-смертников почти нет связи?

Чтобы достичь понимания, нужно подумать о том, сколько было иностранных оккупаций, которые *не* привели к атакам террористов-смертников. Британская оккупация Ирландии, несмотря на десятилетнюю кампанию активного сопротивления, никогда не приводила к террористическим атакам смертников. Баскские сепаратисты в Испании вели кампанию, продолжавшуюся десятилетиями, и никогда не прибегали к атакам смертников. В разные периоды Холодной войны (и после нее) Соединенные Штаты размещали войска в Германии, Японии, Южной Корее, Гренаде, Панаме и Гаити (вероятно, все они были такими же оккупационными, как и предполагаемая оккупация Саудовской Аравии), но не пострадали даже от единственного теракта террориста-смертника в любом из этих мест. Если оккупация предвещает атаки самоубийц, что же происходило во всех этих местах?

У этого примера есть еще одна показательная особенность. Он не только иллюстрирует ошибку поиска корреляции без вариаций. Он показывает, до какой степени можно впасть в заблуждение, рассматривая только случаи, когда происходит интересующее явление (в данном случае атаки террористов-смертников), т. е. путем выбора зависимой переменной. Чтобы убедиться в этом, полезно вернуться немного в историю.

Предположим, вы начали собирать данные о насилии, связанном с самоубийствами, в начале 1980-х гг. К 1986 г. было бы зарегистрировано 33 нападения и более тысячи смертей. По сути, каждое из этих нападений было совершено вооруженным шиитским ополчением «Хезболла» против американских, израильских и французских объектов в Ливане, включая нападение на казармы морской пехоты США в Бейруте, в результате которого погибло 320 человек.

Если бы вы искали общие черты среди всех нападений террористов-смертников, когда-либо совершенных в 1986 г., вы могли бы заметить, что все они были совершены мусульманами на Ближнем Востоке. Используя ту же логику, которая привела к выводу, что оккупация является основным предиктором нападений террористов-смертников, вы могли бы прийти к выводу, что ислам является ключевым коррелятом.

Конечно, если бы вы провели правильное сравнение, вы бы не пришли к такому выводу. В мире существует множество стран с мусульманским большинством. В 1986 г. почти никто из них не подвергался насилию, связанному с терактами.

Более того, если бы вы пытались предсказать, где может произойти следующая атака террориста-смертника, этот вывод 1986 г. сильно ввел бы вас в заблуждение. В 1987 г. мир стал свидетелем первого нападения террориста-смертника, совершенного «Тиграми освобождения Тамил Илама» («Тамильские тигры») – группой светских сепаратистов в Шри-Ланке, не связанных с исламом. Это нападение ознаменовало начало того, что впоследствии стало крупнейшей кампанией террористического насилия, которую когда-либо видел мир. Пытаясь установить корреляцию без вариаций, вы можете совершить колоссальную ошибку.

Мир заставляет нас выбирать зависимую переменную

Как вы убедились, невероятно легко попасть в ловушку выбора зависимой переменной, просто не умея критически мыслить. Но дело обстоит еще хуже. Иногда кажется, что мир устроен таким образом, что заставляет нас искать корреляцию без вариаций. В этом разделе мы рассмотрим три сценария, где обычно так и происходит: особенности некоторых профессий, практика анализа после катастроф и то, как мы ищем жизненный совет.

Врачи чаще наблюдают за больными людьми

Любой, кто страдал от сильной боли в спине, знает, насколько это тяжело. Когда у многих из вас неизбежно возникнут боли в спине, вы, скорее всего, обратитесь к врачу, который отправит вас на МРТ. Обычно МРТ показывает протрузию (выпячивание) или грыжу межпозвоночных дисков. Эти выпячивания дисков каким-то не до конца понятным образом являются причиной болей в спине (возможно, из-за защемления нерва).

Рекомендации после этого диагноза могут сильно различаться. Одни врачи предложат операцию. Другие направят в специализированную клинику, где вам вонзят в спину иглы с лекарством, заглушающим боль и уменьшающим воспаление. Третьи предложат попробовать физиотерапию и принять много обезболивающих.

Для нас это звучит как вызов. Насколько мы можем судить, существует очень мало доказательств того, что протрузия межпозвоночного диска коррелирует с болями в спине. Вот факты. У людей с болями в спине весьма вероятно наличие грыжи межпозвоночных дисков. Действительно, в британском исследовании 2011 г., опубликованном в журнале Pain, около двух третей пациентов с болями в спине, направленных на МРТ, имели сдавление нервов в результате протрузии или грыжи диска. Это похоже на свидетельство того, что деформация межпозвоночных дисков действительно представляет собой корень проблемы.

Но помните, корреляция требует вариаций. Вы должны спросить себя: а как насчет людей, у которых нет болей в спине? Как выглядят их диски? Хороший вопрос. Ответ в том, что в среднем они выглядят точно так же, как диски людей, страдающих болями в спине! Исследование 1994 г., опубликованное в New England Journal of Medicine, показало, что около двух третей людей, не страдающих болями в спине, также имеют протрузию или грыжу диска. Как только вы сравните обе интересующие переменные, очевидная связь между протрузиями дисков и болями в спине исчезнет.

Легко понять, почему врачи могли связать протрузию межпозвоночных дисков с болью в спине. Даже критически мыслящий врач в силу особенностей своей профессии почти обречен не смотреть на вариации. Больные люди идут к врачу. Здоровые люди, как правило, этого не делают. Типичный врач-вертебролог просто не имеет возможности изучить МРТ людей, которые не жалуются на боли в спине.

Анализ постфактум

Другая разновидность мироустройства, заставляющая нас искать корреляцию без вариаций, – это институциональные правила или процедуры. Особенно

распространенным примером является реакция организаций как на большие неудачи, так и на большие успехи.

После кризиса или катастрофы организации хотят знать, что пошло не так, чтобы избежать подобных ошибок в будущем. Аналогичным образом после больших успехов они хотят знать, что было сделано правильно для внедрения лучших методик. Достижению этих целей служит анализ событий постфактум. Пристальное изучение случаев большой неудачи или большого успеха само по себе не является ошибкой. Действительно, это очень разумная отправная точка. Но если вы научились мыслить критически, вы сможете увидеть, что подобной процедуры самой по себе недостаточно для установления корреляции между тем, что пошло неправильно (или правильно), и существующими практиками.

Оценивая уроки, извлеченные из кризиса, вы должны попытаться ответить на вопрос: какие решения следовало принять иначе, чтобы избежать кризиса, учитывая знания, которые мы имели *в то время*? Однако, оценивая полученные уроки, мы часто скатываемся к ответу на несколько иной вопрос: какие решения следовало принять иначе, чтобы избежать кризиса, учитывая то, что мы знаем *сейчас*?

На последний вопрос не так уж полезно отвечать по причинам, о которых мы уже говорили в этой главе. Предположим, вы обнаружили какое-то решение, которое, как выяснилось, привело непосредственно к катастрофе. После этого легко сказать: «Если бы мы не предприняли эти действия, катастрофы не произошло бы». Но значит ли это, что не стоит предпринимать подобные действия в будущем? Чтобы узнать ответ на этот вопрос, необходимо знать, насколько больше вероятность катастрофы при наличии таких действий, чем при их отсутствии. То есть необходимо знать, существует ли корреляция между такими действиями и происходящими бедствиями. Чтобы установить корреляцию, вам нужны вариации. Но анализ постфактум почти по определению не имеет вариаций. Вы всего лишь рассматриваете конкретный случай произошедшей катастрофы.

Чтобы лучше понять, что мы имеем в виду, давайте начнем с вымышленного примера. Затем обратимся к реальным случаям.

Представьте, что вы руководитель школьной музыкальной группы и готовитесь к региональному конкурсу, который начнется через неделю. Вам придется решить, нагружать ли детей изнурительным графиком репетиций или дать им отдых, чтобы они спокойно подошли к соревнованию. Вы взвешиваете все за и против, решив, что подготовка важнее психического состояния. Итак, вы запланировали неделю дополнительных репетиций. К сожалению, на конкурсе группа играет не очень хорошо, и вы выбываете в первом туре.

В своем анализе постфактум вы задаете вопрос: «Что мне следовало сделать, чтобы избежать провала?» Вы вспоминаете, что видели, как многие группы проигрывают соревнования таким же образом (т. е. репетируя до изнеможения за неделю до конкурса), поэтому решаете собрать некоторые данные. Вы изучаете историю всех конкурсов, в которых ваша группа выбывала в первых раундах. Как и в случае с конкурсом этого года, вы обнаружите, что почти на каждом из этих соревнований у вас был запланирован плотный график репетиций на неделе, предшествующей конкурсу.

Допустим, вы провели неделю интенсивных репетиций перед тем, как проиграть 80 конкурсов из 88. Вывод кажется очевидным. Более чем в 90 % случаев

ваша группа выбывала досрочно после недели изнурительных репетиций. Теперь вы чувствуете еще большую уверенность: интенсивные репетиции – неправильная стратегия. В табл. 4.5 суммировано то, что вам известно на данный момент из анализа постфактум.

Таблица 4.5. Стратегии репетиций за неделю до соревнований, на которых ваша группа выступила плохо (вымышленные данные)

	Выступили успешно	Выступили плохо	Итого
Много репетировали	?	80	?
Отдыхали неделю	?	8	?
Итого	?	88	?

Но из собранных вами данных не обязательно следует именно этот вывод. Фактически, исходя только из этих данных, невозможно узнать, связаны ли репетиции с провальным выступлением, потому что вы ответили не на тот вопрос.

Вам не нужно знать, проводили ли группы дополнительные репетиции перед большинством конкурсов, на которых они выступили плохо. Вам нужно узнать, какая корреляция – положительная или отрицательная – существует между репетициями в последнюю неделю и успехом на конкурсе (и существует ли она вообще). Ответ на этот вопрос поможет вам понять, являются ли дополнительные репетиции хорошей идеей для следующих соревнований.

Чтобы ответить на этот вопрос, следует посмотреть на корреляцию между дополнительными репетициями и *удачными* выступлениями на соревнованиях. Но вы не можете узнать такую корреляцию на основе анализа постфактум. Корреляция требует вариаций. Ваш анализ, сосредоточенный только на плохих результатах, гарантирует, что у вас не будет вариаций, необходимых для установления корреляции.

Чтобы лучше выполнять свою работу, вы можете просмотреть историю всех соревнований, в которых участвовали, чтобы увидеть, хорошо вы выступили или плохо. Теперь у вас есть вариации обеих переменных, и вы располагаете всеми данными, как показано в табл. 4.6.

Таблица 4.6. Стратегия репетиций за неделю до конкурсов, на которых ваша группа выступила хорошо или плохо (вымышленные данные)

	Выступили успешно	Выступили плохо	Итого
Много репетировали	300	80	380
Отдыхали неделю	12	8	20
Итого	312	88	400

Из этой таблицы видно, что на самом деле существует сильная положительная корреляция между дополнительными репетициями и хорошим выступлением. Вероятность того, что ваша группа выступит хорошо, если вы усердно репетируете, составляет около 79 % (300/380 ≈ .79). Напротив, вероятность того,

что ваша группа выступит хорошо, если вы расслабились за неделю до соревнований, составляет всего 60 % (12/20 ≈ .60). Единственная причина, по которой почти каждому плохому выступлению предшествовали интенсивные репетиции, заключается в том, что эти дополнительные репетиции настолько эффективны, что разумные руководители оркестров назначают их почти всегда.

Найдя вариацию, необходимую для установления корреляции, действительно имеющей отношение к рассматриваемому вопросу, вы приходите к совершенно противоположному выводу, чем при первоначальном анализе. После провала на конкурсе казалось, что интенсивные репетиции – плохая идея. Но с учетом всей имеющейся информации усердные репетиции выглядят совершенно правильным решением. Столкнувшись снова с той же ситуацией, вам, вероятно, следует принять аналогичное решение.

Эта проблема свойственна процессу расследования после катастроф. Мы склонны смотреть на факторы, которые, как нам кажется, способствовали катастрофе, спрашивать, присутствовали ли они и в прошлых катастрофах, и, если это так, приходить к выводу, что нам следует устранить эти факторы в будущем. Но при этом мы совершаем ту же ошибку, что и руководитель оркестра. Без анализа всех ситуаций, включая те, в которых катастрофы не было, мы не можем на самом деле узнать, коррелирует ли наличие этих факторов с возникновением катастрофы. Поэтому мы не знаем, есть ли чему поучиться.

Мы собираемся показать вам, что мы имеем в виду, на двух примерах послеаварийного расследования, последовавших за крупными катастрофами – взрывом космического корабля «Челленджер» в 1986 г. и финансовым кризисом 2008 г. В каждом случае, как вы скоро увидите, были допущены некоторые серьезные и очевидные ошибки, однако менее очевидно, что лица, принимающие решения, могли заранее знать, что они совершают ошибки. Как только это станет ясно, мы сможем более четко представить себе анализ, извлекающий по-настоящему информативные уроки.

Катастрофа «Челленджера»

28 января 1986 г. космический челнок «Челленджер» развалился на части у берегов мыса Канаверал менее чем через две минуты после запуска. Семь членов экипажа погибли. В ночь перед взрывом «Челленджера» небольшая группа инженеров подрядчика НАСА, ответственного за твердотопливные ускорители шаттла, предсказала, что холодная погода приведет к катастрофическому отказу, который вполне может поставить корабль под угрозу разрушения. Беспокойство вызывал тот факт, что критически важные уплотнительные кольца, ответственные за удержание газов, образующихся при сгорании ракетного топлива, не были сертифицированы для работы при низких температурах, которые предшествовали этому конкретному запуску. Инженеры предупреждали, что, если уплотнительные кольца выйдут из строя, горячий газ под давлением может прожечь корпус ракеты, что приведет к катастрофе.

Эти прогнозы, отвергнутые менеджерами НАСА и фирмой, в которой работали инженеры, оказались трагически верными. Многие послеаварийные расследования были сосредоточены на неспособности НАСА серьезно отнестись к этим опасениям. Большинство наблюдателей пришли к выводу, что катастрофа была вызвана организационными и культурными пробелами в НАСА,

которые способствовали размыванию ответственности и заставляли менеджеров систематически игнорировать важные возражения экспертов. Например, в отчете Президентской комиссии по катастрофе космического корабля «Челленджер» (Комиссия Роджерса) сделан вывод: «Недостатки во внутренних коммуникациях… привели к решению о запуске 51-L на основании неполной, а иногда и вводящей в заблуждение информации, конфликта между инженерными данными и суждениями руководства. Этому способствовала структура управления НАСА, которая позволяет держать в неведении ключевых менеджеров проекта относительно внутренних проблем безопасности полетов».

Случай с Челленджером интересен. Никто не подвергает сомнению физические явления, лежащие в основе вывода о том, что уплотнительные кольца вышли из строя из-за низких температур. В комиссию Роджерса был включен лауреат Нобелевской премии физик Ричард Фейнман – именно для того, чтобы авторитетно сказать, правы ли инженеры с научной точки зрения. Они действительно правы. Так что в этом смысле запуск шаттла был явной ошибкой.

Поскольку научная часть произошедшего совершенно очевидна, кажется вполне логичным задаться вопросом, что же такого было в процессе, который заставил лиц, принимающих решения, игнорировать инженеров, приводящих хорошие научные аргументы. Вот тут-то наши знания о подводных камнях анализа постфактум должны заставить нас остановиться и задуматься. Мы знаем, что решение о запуске было трагически ошибочным. Но это знание постфактум. Мы должны оценить, было ли это решение плохим на момент его принятия. Для этого нам необходимо знать о корреляции между наличием научно обоснованных инженерных проблем и успехом запуска шаттлов. И чтобы знать об этой корреляции, нам нужны вариации; мы должны сравнивать катастрофические запуски с успешными.

Мы не инженеры, поэтому не собираемся взвешивать, было ли решение о запуске «Челленджера» разумным на момент его принятия. Но мы видим, что для анализа этого происшествия комиссиям по расследованию инцидента необходимо задавать вопросы, которые они не привыкли задавать. Комиссии выясняют, что привело к катастрофе, высказывали ли люди соответствующие опасения, и если да, то почему к этим опасениям не прислушались. Кроме того, комиссиям необходимо задаться вопросом, высказывали ли инженеры научно обоснованные опасения перед множеством *успешных* запусков. Это вполне обоснованный вопрос. Запуски космических кораблей – невероятно сложное и опасное мероприятие. Возможно, почти всегда найдется научно обоснованная причина для серьезного беспокойства. Если это так, то на самом деле не будет большой (если вообще будет) корреляции между наличием таких опасений и успехом запуска. В таком случае, если только вы не готовы просто закрыть космическую программу, было бы несправедливо утверждать, что решение о запуске после научно обоснованного возражения инженера всегда является ошибкой. Хотелось бы получить от комиссий ответ именно на этот вопрос, прежде чем делать выводы об изменении организационной культуры или методов управления НАСА.

Финансовый кризис 2008 года

Финансовый кризис, потрясший мировую экономику в 2007 и 2008 гг., начался с краха рынка вторичного кредитования жилья в США. Этот крах прошелся

взрывной волной по банковскому сектору и в конечном итоге распространился по всему миру. Понятно, что после этого кризиса – на тот момент худшего со времен Великой депрессии – политики и общественность были заинтересованы в выявлении индикаторов раннего предупреждения, которые могли бы помочь им прогнозировать и предотвращать будущие кризисы.

Возможно, самым важным анализом, пытающимся предоставить такие индикаторы раннего предупреждения, была книга экономистов Кармен Рейнхарт и Кеннета Рогоффа «На этот раз все будет иначе». Рейнхарт и Рогофф собрали и проанализировали данные обо всех крупных финансовых кризисах за последние 800 лет. По их мнению, таким образом они смогут определить несколько ключевых показателей, которые почти всегда предшествуют подобному кризису. К ним относятся необычно большое отрицательное сальдо текущего торгового баланса (т. е. разности между совокупными стоимостями экспорта и импорта), пузыри цен на активы и чрезмерные заимствования. Например, в 2006 г. в Соединенных Штатах дефицит торгового баланса составил около 7 % ВВП, присутствовали пузырь на рынке недвижимости и растущий федеральный долг. Поэтому Рейнхарт и Рогофф заключают: «Мы знали об этом раньше». Они хотят сказать, что в 2008 г. финансовый кризис в США можно было предсказать благодаря наличию тех же факторов, которые, судя по всему, характеризуют финансовые кризисы на протяжении всего времени наблюдений и во всем мире. Подобные признаки наблюдались и раньше, перед финансовыми кризисами в Латинской Америке в начале 2000-х гг., в Восточной Азии в 1990-х гг., в странах Северной Европы в 1980-х гг. и т. д. в истории.

Проблема с этим аргументом та же, что и в наших предыдущих примерах. Индикаторы раннего предупреждения должны коррелировать с финансовыми кризисами. Поскольку корреляция требует вариаций, чтобы знать, коррелируют ли отрицательное сальдо торгового баланса, растущие цены на недвижимость и крупные заимствования с финансовыми кризисами, нам нужны вариации кризисов. То есть нам необходимо знать не только то, что эти факторы обычно присутствуют, когда происходят кризисы, но также и то, как часто они присутствуют, когда кризисы *не* происходят. Без таких вариаций мы не можем установить корреляцию.

Изучение каждого крупного финансового кризиса за 800 лет не может ответить на этот вопрос. И есть причины для беспокойства по поводу выводов. Как отмечает политолог Массачусетского технологического института Дэвид Сингер, достаточно взглянуть на современную историю, чтобы поставить под сомнение выводы Рейнхарт и Рогоффа. Например, в конце 1990-х гг. в США наблюдались все признаки предстоящего финансового кризиса. В результате массовых иностранных инвестиций в доткомы возник большой дефицит текущего торгового баланса. Более того, когда пузырь доткомов лопнул, «он уничтожил около 5 трлн долларов рыночной капитализации». Однако финансового кризиса не произошло. Это, конечно, только один эпизод в мировой истории. Но он должен заставить вас задаться вопросом, являются ли факторы, на которые указывают Рейнхарт и Рогофф, действительно хорошими предикторами финансовых кризисов или просто общими свойствами мира, которые существуют без ощутимой связи с кризисами.

Жизненные советы

Мы утверждаем, что наш мир устроен таким образом, что мы интуитивно пытаемся выяснить корреляты успеха или неудачи, не обращая внимания на вариации, даже если это не сработает. Важно понимать, что эта проблема не ограничивается крупными институциональными рамками. Мы все становимся жертвами этого заблуждения каждый день во множестве мелких ситуаций.

Одним из простых примеров является то, как мы ищем жизненный совет: почти всегда мы спрашиваем успешных людей, как им удалось добиться успеха. В нашем деле, например, аспирантам полагается спрашивать старших профессоров, что они сделали, чтобы добиться успеха в своей области. Мы полагаем, что нечто подобное происходит и в других профессиях. Естественно, нет недостатка в книгах по самосовершенствованию, описывающих привычки успешных людей.

Но такая «мудрость» страдает именно от тех проблем, на которые мы указывали. Успешные люди, размышляя о своей жизни, склонны выделять несколько принятых ими решений или несколько личных качеств, которые кажутся важными, и предлагать их в качестве совета следующему поколению. Но эти успешные люди обычно понятия не имеют, принимали ли подобные решения многие другие, менее успешные люди или обладали такими же характеристиками. То есть их самоанализ по поводу коррелятов успеха не отличается разнообразием. Успешные люди на самом деле не знают, являются ли уроки, извлеченные из их персональной истории, коррелятами успеха или нет. А теперь мы поделимся собственной житейской мудростью: остерегайтесь жизненных советов. Большинство из них, вероятно, полная ерунда.

Подведение итогов

Корреляция требует вариаций. Но отсутствие опыта критического мышления и общепринятые требования часто заставляют нас выбирать зависимую переменную – попытка установить корреляты какого-либо явления, зная только случаи, когда оно произошло. Требуется острое критическое мышление, чтобы не попасть в эту ловушку, независимо от того, проводите ли вы количественный анализ или просто пытаетесь неформально рассуждать о доказательствах. Даже умение задать себе вопрос о том, сможете ли вы заполнить все четыре ячейки одной из наших таблиц, является хорошей отправной точкой для того, чтобы избежать поиска корреляции без вариаций.

Вы можете добиться большей строгости выводов, используя количественные методы измерения корреляций. Самый важный из таких методов называется регрессией (тема главы 5).

Ключевые термины

- **Выбор зависимой переменной:** изучение только тех случаев, когда интересующее явление имело место, а не сравнение случаев, когда оно имело место, со случаями, когда его не было.

Упражнения

4.1. В главе 2 мы обсудили различия между утверждениями о корреляциях и другими фактическими утверждениями, которые не передают информацию о корреляции. Теперь, когда у вас есть более глубокое понимание того, что корреляция требует вариаций, рассмотрите следующие утверждения. Какие из них описывают корреляцию, а какие нет?

a) В большинстве наиболее успешных школ небольшая численность учащихся.
b) Женатые люди обычно счастливее, чем неженатые.
c) Среди профессионалов более высокие баскетболисты, как правило, имеют более низкий процент штрафных бросков, чем низкие игроки.
d) В Соединенных Штатах местами с самым высоким уровнем заболеваемости раком обычно являются небольшие города.
e) В старых домах чаще используется свинцовая краска, чем в новых.
f) Большинство случаев простуды в округе Кук приходится на холодные дни.

4.2. По меньшей мере 20 миллиардеров бросили колледж, прежде чем заработать свое состояние, включая Билла Гейтса и Марка Цукерберга.

a) Означает ли это, что досрочное прекращение учебы в колледже коррелирует с тем, чтобы стать миллиардером? Поясните свой ответ.
b) Нарисуйте таблицу 2×2, которая позволит вам оценить, коррелирует ли досрочное прекращение учебы в колледже с тем, чтобы стать миллиардером. Предположим, что ровно 20 человек бросили колледж и стали миллиардерами, и вы знаете, что записать в одну из четырех ячеек таблицы. Сделайте предположения для других ячеек. На момент написания статьи в мире проживает около 7.8 млрд человек и около двух тысяч миллиардеров. Как вы думаете, существует положительная или отрицательная корреляция между тем, чтобы бросить колледж и стать миллиардером?
c) Учитывая ваши предположения из части (b), какую долю бросившие колледж должны составлять от числа обычных людей (не миллиардеров), чтобы корреляция была отрицательной? Какова должна быть доля бросивших колледж от числа обычных людей, чтобы корреляция была положительной?
d) Если вы в настоящее время студент колледжа и решаете, стоит ли бросить учебу в надежде стать миллиардером, вы можете ограничить внимание людьми, которые сперва поступили в колледж. Считаете ли вы, что корреляция между тем, чтобы бросить колледж и стать миллиардером, будет более или менее положительной, если мы ограничим внимание только людьми, которые начинали учиться в колледже?
e) Около 7 % населения мира имеет высшее образование. И около трети людей, поступивших в колледж, заканчивают его. Если мы предположим, что каждый, кто становится миллиардером, поступил в колледж, теперь у вас должна быть вся информация, необходимая для оценки корреляции между тем, как стать миллиардером и бросить колледж, среди тех, кто поступает в колледж. Какую корреляцию вы наблюдаете – положительную, отрицательную или нулевую?

4.3. Назовите один недавний случай, когда аналитик допустил ошибку, обсуждаемую в этой главе. То есть найдите случай, когда кто-то (по крайней мере, неявно) утверждает о корреляции, но у него нет вариаций ни в одной из переменных. Ваш пример может быть взят из газетной статьи, научного исследования, аналитического доклада или заявления политика.

   a) Кратко изложите выдвинутое утверждение (возможно, высказанное косвенно) и объясните, почему доказательства не обязательно подтверждают это утверждение.
   b) Объясните, какие дополнительные данные нужно собрать и проанализировать, чтобы оценить корреляцию.
   c) Нарисуйте таблицу 2×2, иллюстрирующую ваш аргумент, и определите, какими должны быть неизвестные числа в таблице, чтобы корреляция была положительной, отрицательной или нулевой.

Дополнительное чтение и ссылки

Мы подробно обсудили книгу:

Malcolm Gladwell. 2008. *Outliers: The Story of Success.* Little, Brown.

Для получения дополнительной информации о Фредрике Вертаме и его ошибочных аргументах в отношении комиксов мы рекомендуем прочесть книгу:

David Hajdu. 2009. T*he Ten-Cent Plague: The Great Comic-Book Scare and How It Changed America.* Picador.

И хотя выводы Вертама не следуют из его эмпирических данных, поскольку корреляция требует вариаций, оказывается, есть подозрение, что Вертам сфабриковал свои данные. Прочтите статью:

Carol L. Tilley. 2012. *Seducing the Innocent: Fredric Wertham and the Falsifications That Helped Condemn Comics.* Information & Culture: A Journal of History 47 (4): 383–413.

Мы обсудили два доклада Министерства образования и Секретной службы США о безопасности в школах.

- Отчет за 2002 г. находится здесь: *https://www.govinfo.gov/content/pkg/ERIC-ED466024/pdf/ERIC-ED466024.pdf*.
- Отчет за 2018 г. можно найти здесь: *https://www2.ed.gov/documents/school-safety/school-safety-report.pdf*.

Доклад о проблеме ухода из средней школы, подготовленный для Фонда Гейтса:

JohnM. Bridgeland, John J.Dilulio, Jr., andKarenBurkeMorison. *The Silent Epidemic: Perspectives of High School Dropouts*. *https://docs.gatesfoundation.org/documents/thesilentepidemic3-06final.pdf*.

Опрос о скучных занятиях в школе называется «Опрос вовлеченности учащихся в старших классах». Он проводился Центром оценки и образовательной политики Университета Индианы. Выдержка из их исследования 2010 г. Для получения дополнительной информации см. *http://newsinfo.iu.edu/news-archive/14593.html*.

Дополнительную информацию о нападениях террористов-смертников, в том числе о том, почему мы не можем узнать о причинах или коррелятах

насилия со стороны террористов-смертников, изучая только случаи, когда оно происходит, можно получить в следующих работах:

Robert A. Pape. 2003. *The Strategic Logic of Suicide Terrorism.* American Political Science Review 97 (3): 343–61;

Scott Ashworth, Joshua D. Clinton, Adam Meirowitz, and Kristopher W. Ramsay. 2008. *Design, Inference, and the Strategic Logic of Suicide Terrorism.* American Political Science Review 102 (2): 269–73;

Robert A. Pape. 2008. *Methods and Findings in the Study of Suicide Terrorism.* American Political Science Review 102 (2): 275–77;

Scott Ashworth, Joshua D. Clinton, Adam Meirowitz, and Kristopher W. Ramsay. 2008. *Design, Inference, and the Strategic Logic of Suicide Terrorism: A Rejoinder.* Неопубликованные заметки: http://home.uchicago.edu/~sashwort/rejoinder3.pdf.

Чтобы узнать о высокой частоте протрузий и грыжи дисков среди людей с болями в спине и без них (и, следовательно, об отсутствии корреляции между этими характеристиками и болью в спине), см. публикации:

Michael J. DePalma, Jessica M. Ketchum, and Thomas Saullo. 2011. *What Is the Source of Chronic Low Back Pain and Does Age Play a Role?* Pain Medicine 12 (2): 224–33;

Maureen C. Jensen, Michael N. Brant-Zawadzki, Nancy Obuchowski, Michael T. Modic, Dennis Malkasian, and Jeffrey S. Ross. 1994. *Magnetic Resonance Imaging of the Lumbar Spine in People without Back Pain.* New England Journal of Medicine 331: 69–73.

Два исследования, на которые мы ссылались, посвященные финансовому кризису 2008 г.:

Carmen M. Reinhart and Kenneth S. Rogoff. 2009. *This Time Is Different: Eight Centuries of Financial Folly.* Princeton University Press;

David A. Singer. 2010. *Is This Time Different?* The Political Economist. Fall, pp. 4–5.

Глава 5

Применение регрессии в описании и прогнозировании

О чем эта глава

- Регрессия означает поиск линии наилучшего соответствия на основе некоторых данных. Возможно, это самый важный инструмент для описания взаимосвязи между двумя или более переменными.
- При определенных условиях регрессия может быть полезна для прогнозирования.
- Регрессия может не получиться, особенно если у нас небольшой объем данных. Среди наиболее важных проблем, которые могут возникнуть, – переобучение.
- Откуда взялась регрессия?

Введение

В главе 2 мы дали определение *корреляции* и обсудили три ее применения: описание, прогнозирование и причинно-следственный вывод. Мы также говорили о различных способах количественной оценки корреляций, включая наклон линии регрессии, ковариацию и коэффициент корреляции. Линии регрессии являются наиболее распространенными и полезными из них. В этой главе вы ближе познакомитесь с регрессией, чтобы достичь полного понимания этого важного метода.

Основы регрессии

Давайте вернемся к данным о преступности и температуре в Чикаго, которые мы обсуждали в главе 2. Рисунок 5.1 напоминает вам, как выглядит диаграмма рассеяния этих данных.

Как вы можете видеть, просто взглянув на данные, вообще говоря, в теплые дни происходит больше преступлений. Но иногда хочется уточнить отношения между переменными. Представьте, что вы работаете в полицейском управлении Чикаго и ваш начальник попросил вас обобщить взаимосвязь между температурой и преступностью. Вряд ли он будет доволен, если вы просто принесете ему этот график. Скорее всего, ему понадобится простое описание отношений между переменными факторами, которое будет легко понять и донести до людей, принимающих политические решения. Здесь на помощь приходит линейная регрессия.

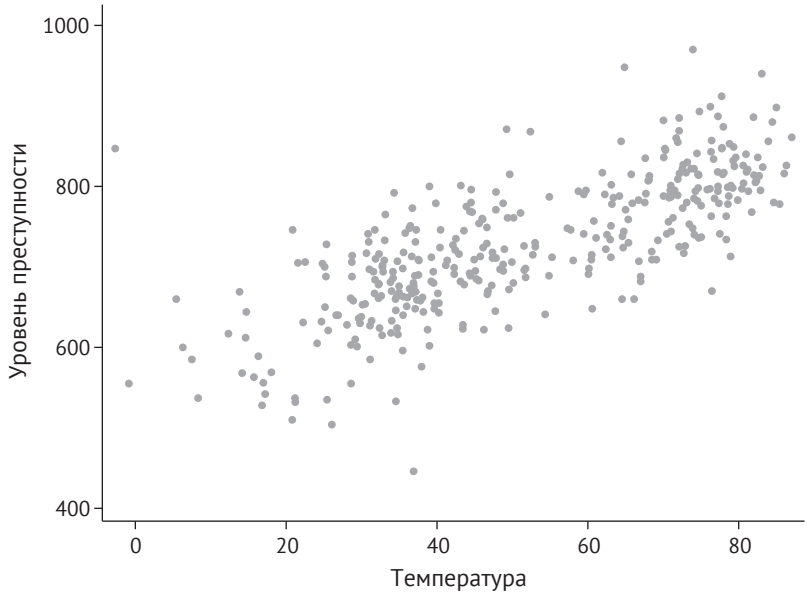


Рис. 5.1. Количество зарегистрированных преступлений и температура в градусах Фаренгейта в Чикаго по дням в 2018 г.

Линия наилучшего соответствия дает именно то доступное обобщение взаимосвязи между температурой и преступностью, которое мы ищем. Такая линия, если она построена правильно, открывает перед нами две возможности. Во-первых, для любой заданной температуры эта линия дает нам разумное приближение (или прогноз) количества преступлений. И во-вторых, как мы обсуждали в главе 2, наклон линии говорит нам кое-что о знаке и величине корреляции между двумя переменными, т. е. примерно говорит нам, насколько сильно меняется преступность при изменении температуры. Итак, давайте выясним, как определить линию наилучшего соответствия, чтобы мы могли подумать о том, как интерпретировать и пересказать то, что она нам сообщает.

За исключением полностью вертикальных линий (которые в любом случае не дают полезного описания или прогноза), все гипотетические линии, которые мы могли бы нарисовать на графике рис. 5.1, могут быть описаны так называемым *уравнением регрессии* следующего вида:

$$\text{Прогнозируемая преступность} = \alpha + \beta \cdot \text{Температура}.$$

Уравнение регрессии выражает линейную связь между *зависимой* (или *выходной*) переменной в левой части уравнения и *независимой* (или *объясняющей*) переменной в правой части уравнения. (Как мы увидим позже в этой главе, в правой части может быть более одной объясняющей переменной.) Зависимая переменная соответствует результату, который мы пытаемся описать, предсказать или объяснить. Независимая переменная соответствует тому, при помощи чего мы пытаемся описать, предсказать или объяснить зависимую переменную.

Уравнение регрессии связывает зависимые и независимые переменные по линейному закону через параметры регрессии. Параметры регрессии опреде-

ляют конкретную линию, которую мы строим. В нашем уравнении регрессии, приведенном выше, параметрами регрессии являются α и β. Параметр α называется *точкой пересечения*; это прогнозируемое количество преступлений в день, когда средняя температура составляет 0 °F. Параметр регрессии β – это наклон; это количество ожидаемых преступлений увеличивается с каждым градусом по Фаренгейту. Любая возможная линия на графике соответствует одной конкретной комбинации α и β. (Как вы увидите позже в этой главе, может быть более двух параметров регрессии, если имеется более одной независимой переменной. И, как вы увидите позже в книге, вы можете обозначать параметры регрессии буквами, отличными от α и β, если это удобно.)

Конечно, нам не хотелось бы описывать или предсказывать преступность на основе температуры, используя какую-либо произвольную линию. Неправильная линия даст очень плохие прогнозы. Мы хотим использовать линию, которая лучше всего соответствует данным.

Чтобы найти значения α и β, которые задают линию, лучше всего соответствующую данным, нам нужно начать с определения «наилучшего соответствия». Мы делаем это количественно, выбирая меру того, насколько хорошо та или иная линия резюмирует или аппроксимирует данные. Затем мы находим (или просим наш компьютер найти) значения α и β, которые приводят к наилучшему возможному значению этой меры. Найденные значения α и β описывают линию наилучшего соответствия данных в соответствии с выбранной нами мерой.

Важно выбрать правильную меру соответствия. Как мы кратко упомянули в главе 2, наиболее часто используемой мерой (и той, которую мы будем использовать) является сумма квадратов ошибок. Итак, давайте разберемся, что означает эта мера.

Для любых выбранных α и β наша линия дает прогноз уровня преступности в день с любой заданной температурой. Например, предположим, что мы выбрали $\alpha = 650$ и $\beta = 2$. Тогда в определенный день (например, 26 января 2018 г.), когда средняя температура составляла 46 °F, наш прогноз количества преступлений будет таким:

$$\text{Прогнозируемая преступность} = 650 + 2 \cdot 46 = 742.$$

Конечно, предсказание линии будет не совсем верным – мы жертвуем некоторой точностью, чтобы получить вычислительно недорогую сводку данных. Например, на самом деле число преступлений 26 января 2018 г. составило 759. Разница между истинным значением зависимой переменной и предсказанием нашей линии для любого данного наблюдения называется *ошибкой* этого наблюдения:

$$\text{Ошибка}_i = \text{Преступность}_i - \text{Прогноз преступности}_i.$$

Так, например, при нашем выборе α и β ошибка на 26 января 2018 г. составит $759 - 742 = 17$.

Иными словами, для любой выбранной нами линии (т. е. значений α и β) мы можем описать любое наблюдение i следующим образом:

$$\text{Преступность}_i = \underbrace{\alpha + \beta \cdot \text{Температура}_i}_{\text{Прогноз преступности}_i} + \underbrace{\text{Ошибка}_i}_{\text{Преступность}_i - \text{Прогноз преступности}_i}.$$

На рис. 5.2 поверх данных построена линия с $\alpha = 650$ и $\beta = 2$ и показано, как измеряются ошибки. (Как мы увидим позже, это не линия наилучшего соответствия.) Ошибки – это линии, идущие от точки данных вертикально к линии регрессии. Мы нарисовали ошибки только для нескольких точек данных, чтобы рисунок не получился слишком запутанным. Однако, чтобы оценить соответствие линии данным, мы фактически начали бы с расчета ошибки для каждой отдельной точки данных. Ошибка для любой заданной точки данных может быть положительной (если точка данных находится выше линии) или отрицательной (если точка данных находится ниже линии). Но нам нужна лишь мера того, насколько далеко точка данных находится от линии. Нам все равно, вверху она или внизу. Чтобы получить такую меру, мы затем возводим в квадрат ошибку для каждой точки данных. Квадрат ошибки положителен независимо от того, находится ли точка данных выше или ниже линии. Это всего лишь мера того, насколько далеко точка данных находится от линии.

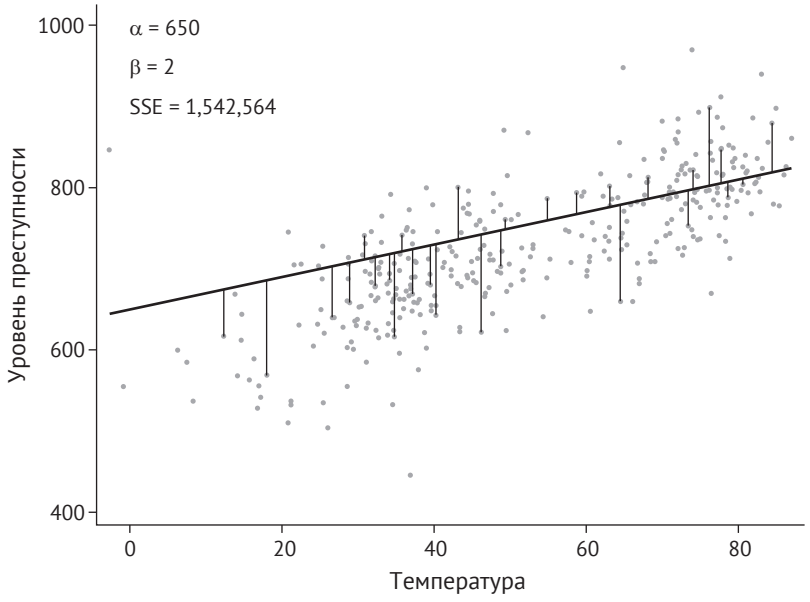


Рис. 5.2. Линия регрессии для зависимости преступности от температуры (в градусах по Фаренгейту) с указанием некоторых ошибок

Затем все эти квадраты ошибок нужно просуммировать, чтобы получить *сумму квадратов ошибок* (sum of squared errors, SSE). В левом верхнем углу рисунка показано значение SSE для этой конкретной линии.

Мы можем повторить эту процедуру, чтобы получить сумму квадратов ошибок для линии, определяемой любыми другими значениями α и β. Разные линии имеют разные SSE. Чем больше сумма квадратов ошибок, тем дальше в среднем данные от линии.

Искомая линия регрессии – это линия с наименьшей суммой квадратов ошибок. То есть находим (ваш компьютер умеет это делать) такие значения параметров α и β, которые минимизируют сумму квадратов ошибок. Этот про-

цесс называется регрессией по методу *наименьших квадратов* (ordinary least squares, OLS). Мы обозначаем значения параметров, которые минимизируют сумму квадратов ошибок, как α^{OLS} и β^{OLS}. Эти значения параметров называются *коэффициентами регрессии* метода наименьших квадратов. Линия, построенная с этими параметрами, является линией регрессии OLS. Это и есть искомая линия наилучшего соответствия.

Существует множество терминов, описывающих поиск α и β, при которых сумма квадратов минимальна. Иногда мы просто говорим, что «строим регрессию преступности от температуры». Когда у нас болтливое настроение, мы говорим, что «находим регрессию методом наименьших квадратов, где преступность является зависимой переменной, а температура – независимой».

На рис. 5.3 показаны данные о преступности и температуре, через которые проведены четыре разные линии, соответствующие различным комбинациям α и β. Для каждой строки на рисунке показаны значения α, β и сумма квадратов ошибок. Некоторые ошибки показаны вертикальными черными линиями. В нижнем правом сегменте рисунка показана линия регрессии OLS, которая минимизирует сумму квадратов ошибок. Глядя на график, мы видим, что эта линия лучше аппроксимирует данные, чем три других варианта. На практике нам не приходится искать эту линию методом проб и ошибок. Вместо этого мы попросим наш компьютер сделать всю работу за нас, и он, используя линейную алгебру, найдет значения α и β, которые минимизируют сумму квадратов ошибок, прежде чем вы успеете моргнуть.

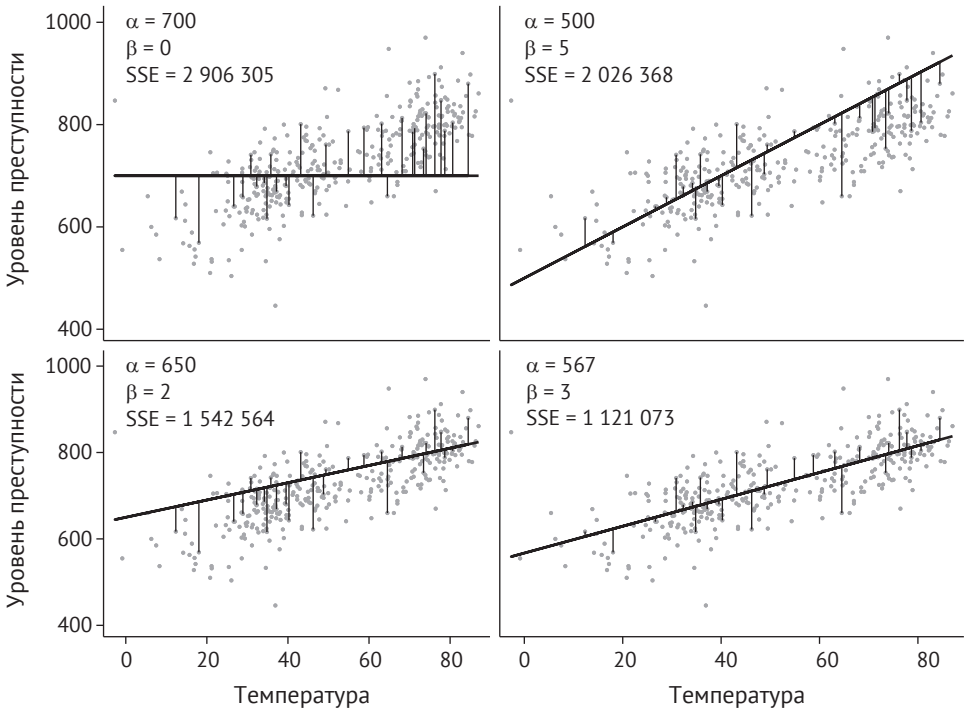


Рис. 5.3. Подгонка различных линий к данным о преступности и температуре с указанием некоторых ошибок

Как интерпретировать линию регрессии OLS? Как мы видим на рисунке, округляя до ближайшего целого числа, точка пересечения (α^{OLS}) равна 567, а наклон (β^{OLS}) равен 3. Другими словами, линия регрессии OLS говорит нам, что в 2018 г. в дни, когда среднее значение температуры в аэропорту Мидуэй составляло 0 °F, в среднем совершалось около 567 преступлений, причем на каждый дополнительный градус по Фаренгейту среднее количество преступлений увеличивалось примерно на 3. Так, например, прогнозируемое количество преступлений в день, когда температура была 46 °F (как и 26 января 2018 г.), составит:

$$\text{Прогнозируемая преступность} = 567 + 3 \cdot 46 = 705.$$

Нам не обязательно выбирать линию регрессии путем минимизации суммы квадратов ошибок. В зависимости от наших целей мы могли бы вместо этого минимизировать сумму абсолютных значений ошибок. Или мы могли бы минимизировать сумму ошибок, возведенную в четвертую степень. Возможности безграничны.

Нам нравится сумма квадратов ошибок по нескольким причинам. Во-первых, минимизация суммы квадратов ошибок обеспечивает наилучшую линейную аппроксимацию другой полезной функции – *условного среднего* (conditional mean function). Функция условного среднего сообщает вам среднее значение некоторой переменной в зависимости от значения некоторых других переменных. Здесь нас интересует конкретная функция условного среднего, которая дает среднее число преступлений, зависящее от температуры.

Предположим, что для каждого градуса по Фаренгейту вы рассчитали среднее количество преступлений в дни с такой температурой и нанесли их на график. Вы получили график функции условного среднего – для каждого градуса температуры он показывает среднее количество преступлений. На рис. 5.4 светло-серые точки – это наши необработанные данные о преступности и температуре, а большие черные точки – это среднее количество преступлений, обусловленное нахождением в интервале 5 °F (0–5 °F, 6–10 °F и т. д.) Условное среднее – еще один разумный способ предсказать преступность на основе температуры. Однако функция условного среднего не так экономна в вычислительном отношении, как линия регрессии: чтобы суммировать функцию условного среднего, вам нужен список среднего уровня преступности для каждого температурного интервала, тогда как линия определяется по двум параметрам. Но, как вы можете видеть, линия регрессии, помимо того, что она является линией наилучшего соответствия необработанным данным, также является очень хорошей аппроксимацией этих условных средних – более того, это лучшая линейная аппроксимация из них. Итак, если вас интересуют условные средние, то линия, минимизирующая сумму квадратов ошибок, – хороший способ аппроксимации.

Конечно, вас могут не интересовать средние значения. Возможно, вместо этого вы хотите описать или предсказать условную медиану. В этом случае, оказывается, вам нужно провести линию, минимизирующую сумму абсолютных значений ошибок. Как мы уже говорили, существует множество разумных вариантов.

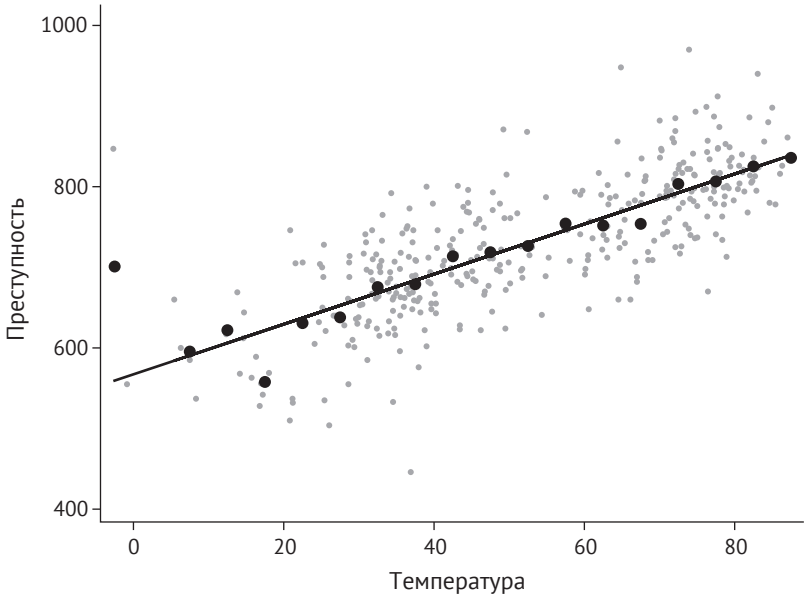


Рис. 5.4. Линия регрессии также является лучшим линейным приближением к условным средним

Вторая причина, по которой люди стремятся минимизировать сумму квадратов ошибок, носит исторический характер. Как указано выше, для вашего компьютера существует простой способ вычислить значения α и β, который минимизирует сумму квадратов ошибок с использованием линейной алгебры; в результате коэффициенты OLS можно рассчитать довольно быстро. Но в те времена, когда люди делали это вручную или даже когда компьютеры были намного медленнее, это было важным фактором. По мере увеличения скорости вычислений это соображение утратило свою актуальность.

Линейная регрессия при нелинейных данных

Что делать, если мы хотим использовать линейную регрессию, но наши данные плохо описываются линией? Чтобы найти ответ на этот вопрос, давайте вернемся к данным, которые мы рассматривали при обсуждении явки избирателей в главе 2. Помните, там мы хотели описать взаимосвязь между возрастом и явкой избирателей – возможно, чтобы узнать, достаточно ли молодые люди представлены в политике или, возможно, для того, чтобы решить, на кого нацелить кампанию за участие в голосовании.

На рис. 5.5 показана явка избирателей для каждого года в возрасте от 18 до 68 лет на промежуточных выборах 2014 г. Обратите внимание: в этих данных наблюдение не является индивидуальным; это возрастная группа. Как и в случае с температурой и преступностью, взаимосвязь между возрастом и явкой потенциально весьма сложна. Что, если мы хотим просто выявить среднюю взаимосвязь между возрастом и явкой? Что, если у нас нет данных по 31-летним (не показаны на рисунке) и мы хотели бы получить наилучшее предпо-

ложение об их уровне явки? Или если бы мы захотели спрогнозировать явку на основе возраста на выборах 2018 г? Линейная регрессия может быть полезна для всех этих целей.

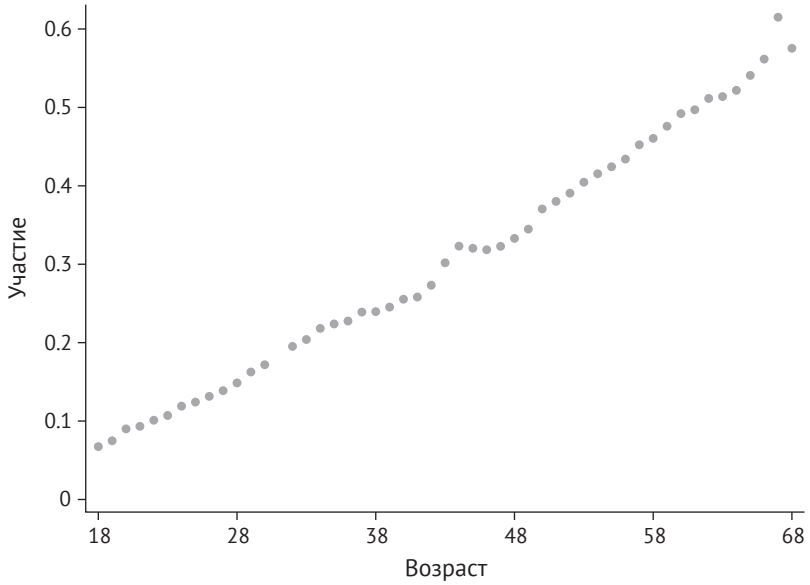


Рис. 5.5. Явка избирателей по возрасту на промежуточных выборах в США в 2014 г.

При беглом взгляде на график взаимосвязь между возрастом и явкой выглядит примерно линейной, по крайней мере, для этого диапазона данных. Другими словами, мы, вероятно, могли бы построить на этом графике линию, которая проходит достаточно близко к каждой точке данных. И если бы мы провели такую линию, это было бы весьма полезно и для описания, и для прогнозирования.

Давайте попробуем применить регрессию OLS к нашим данным о явке избирателей. Мы могли бы снова описать любую линию с помощью следующего уравнения регрессии:

$$\text{Прогнозируемая явка} = \alpha + \beta \cdot \text{Возраст}.$$

Наша статистическая программа сообщает нам, что для этих данных $\alpha^{OLS} = -0.1381$ и $\beta^{OLS} = 0.0103$. С помощью этих двух чисел мы можем построить линию наилучшего соответствия данным и сгенерировать прогнозируемую явку для любого заданного возраста. На рис. 5.6 показано, как выглядит линия регрессии OLS.

Теперь сделаем паузу и критически подумаем о содержательном значении линии регрессии.

Число α^{OLS} соответствует пересечению линии регрессии с осью y. Это говорит нам о том, что прогнозируемая явка людей нулевого возраста составляет –0.1381, или около –14 %. Это довольно странный прогноз. Явка не может быть отрицательной. А младенцы не могут голосовать. Мы знаем, что явка избирателей с нулевым возрастом равна нулю.

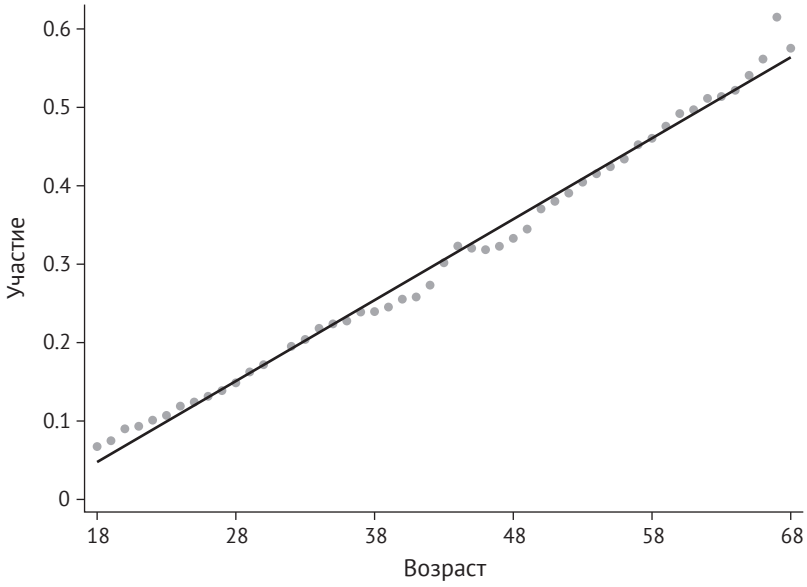


Рис. 5.6. Линия регрессии OLS для зависимости явки избирателей от возраста

Означает ли это, что регрессия бессмысленна или неверна? Нет. Просто регрессия не очень полезна для описания или прогнозирования поведения младенцев в роли избирателей. Это не удивительно. Наша линия регрессии была выбрана для хорошей аппроксимации имеющихся данных. Мы не должны ожидать, что она будет очень хорошо аппроксимировать поведение людей, возраст которых значительно выходит за рамки этих данных. У нас вообще нет данных о людях моложе 18 лет.

Число β^{OLS} представляет собой наклон. Оно говорит нам о том, что в среднем в пределах наших данных каждый дополнительный год жизни соответствует увеличению явки чуть более чем на один процентный пункт. Другими словами, в среднем в возрасте от 19 до 68 лет люди голосуют примерно на один процентный пункт чаще, чем люди, которые всего на год моложе их самих. Это интересно. Причем этот эффект накапливается с годами, а это означает, что вероятность голосования 68-летних примерно на 50 процентных пунктов больше, чем 18-летних, и это именно то, что мы видим в данных.

Линия регрессии неплохо справляется со своей задачей. Он дает нам довольно простое и краткое представление о взаимосвязи между возрастом и явкой людей в возрасте от 18 до 68 лет на выборах 2014 г. На этих конкретных выборах 18-летние проголосовали с приблизительной нормой 4.8 % ($(-0.1381 + 0.0103 \cdot 18) \cdot 100 \approx 4.8$), а затем явка увеличивается чуть более чем на один процентный пункт на каждый дополнительный год возраста. Хотя этот обзор не совсем точно отражает явку для каждой возрастной группы, он весьма близок к этому. И, на наш взгляд, потеря точности (по сравнению, скажем, с простым перечислением явки по возрасту) с лихвой компенсируется экономностью вычислений и простотой обобщения.

Мы также можем использовать α^{OLS} и β^{OLS} для прогнозирования уровня явки избирателей, возраст которых не указан в наших данных. По причинам, кото-

рые уже обсуждали, мы не можем экстраполировать слишком далеко. Мы не можем экстраполировать регрессию на младенцев или даже на 17-летних, поскольку они не имеют права голоса. Вероятно, также не стоит бездумно экстраполировать результаты на людей старше 68 лет.

Наши прогнозы наверняка будут довольно хорошими для 69- и 70-летних, для которых мы прогнозируем явку примерно на уровне 57.3 % ((−0.1381 + 0.0103 · 69) · 100) и 58.3 % ((−0.1381 + 0.0103 · 70) · 100) соответственно. Но чем дальше мы уходим от диапазона наших фактических данных, тем больше нам следует беспокоиться о надежности наших прогнозов.

Наиболее уверенными в своих прогнозах мы можем быть в отношении 31-летних. По какой-то причине на нашем графике нет данных по этой возрастной группе. (В данном случае мы намеренно опустили этот возраст в иллюстративных целях. Но если вы начнете работать с данными, вы обнаружите, что подобные вещи происходят постоянно. Возможно, окружной секретарь пролил кофе на результаты голосования 31-летних избирателей.) Но у нас есть много данных о людях, возраст которых расположен по обе стороны от 31 года. Так что мы, вероятно, можем сделать довольно хорошие прогнозы относительно явки 31-летних. Давайте проверим.

Наше уравнение регрессии предсказывает, что явка избирателей в возрасте 31 года составит чуть более 18 % (−0.1381 + 0.0103 · 31) · 100 = 18.12. Поскольку на самом деле у нас есть данные, мы можем увидеть, насколько хорошо сбывается наш прогноз, просто добавив 31-летних на график.

На рис. 5.7 изображена та же линия регрессии, соответствующая данным о людях в возрасте от 18 до 68 лет, исключая 31-летних. Но он вводит некоторые ранее исключенные точки данных, отображая их в виде полых кругов. Новые данные включают 31-летних, а также людей в возрасте 69–88 лет.

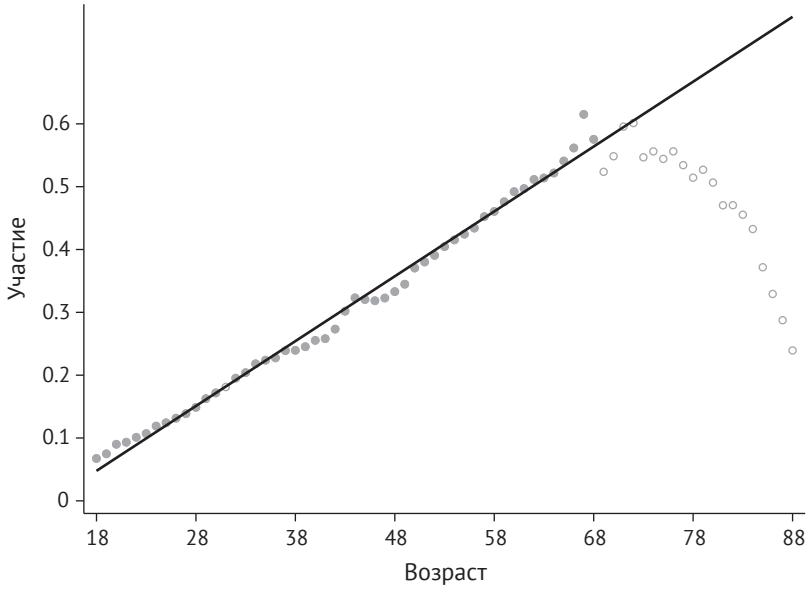


Рис. 5.7. Использование линии регрессии для прогнозирования уровня явки избирателей для возрастов, не входящих в выборку

С 31-летними мы попали в цель почти идеально: мы прогнозировали явку в 18.12 %, а истинный показатель составил 18.11 %. С возрастом от 69 до 72 лет у нас тоже все неплохо, хотя и не так хорошо. Но наши прогнозы начинают работать очень плохо для самых старших избирателей.

Это потому, что связь между возрастом и явкой для пожилых людей, похоже, совершенно иная. Среди молодых людей явка увеличивается с возрастом. Но как только люди достигают возраста 70 лет или около того, явка, похоже, начинает уменьшаться с возрастом. В результате попытка предсказать разницу в явке избирателей между 80- и 88-летними, используя данные о явке избирателей в возрасте от 18 до 68 лет, работает плохо. Как и в случае с нашим прогнозом явки младенцев на уровне –14 %, этот результат иллюстрирует, что может пойти не так, когда мы попытаемся экстраполировать наши прогнозы за пределы диапазона данных, использованных для создания линии регрессии.

Предположим, мы захотели проанализировать взаимосвязь между возрастом и явкой для всех избирателей в возрасте от 18 до 88 лет. Взглянув на диаграмму распределения, мы видим, что зависимость явно не является линейной. Как нам следует учитывать эту нелинейность?

Один из подходов – подобрать новую линейную регрессию, используя все имеющиеся данные. Даже если сами данные не располагаются на прямой линии, мы все равно можем найти линию, которая минимизирует сумму квадратов ошибок. Как вы можете видеть на рис. 5.8, теперь ошибок стало намного больше, поскольку мы подгоняем линию к данным, которые имеют явно нелинейную связь.

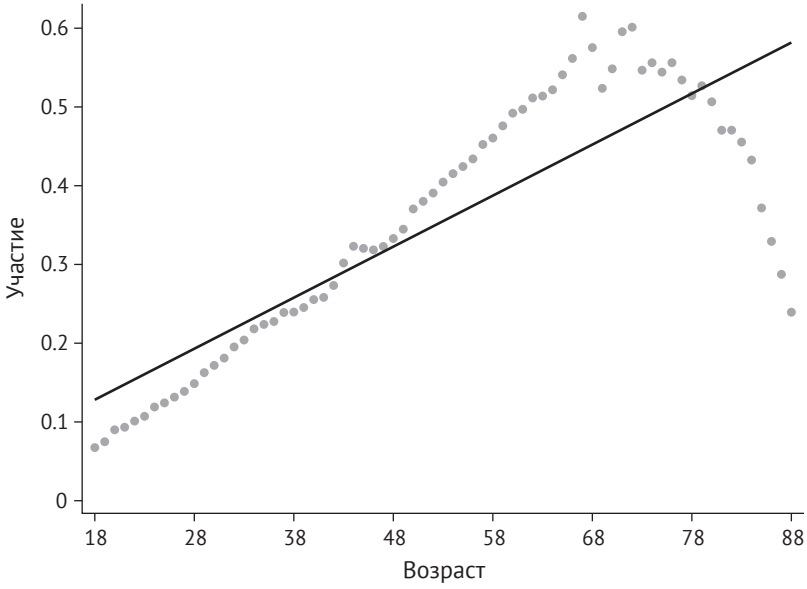


Рис. 5.8. Регрессия явки избирателей по возрасту для всех возрастов

Второй подход – сохранить наилучшие линии регрессии, но использовать разные линии для разных частей данных. Например, мы могли бы найти ли-

нию, которая минимизирует сумму квадратов ошибок для данных о людях в возрасте от 16 до 68 лет, вторую линию, которая минимизирует сумму квадратов ошибок для данных о людях в возрасте 69–78 лет, и третью линию для данных о людях в возрасте от 69 до 78 лет. Это было бы не так экономно и легко, как запуск одной регрессии, – вместо двух параметров (α и β) у нас было бы шесть параметров (отдельные α и β для каждой линии регрессии). Но, как вы можете видеть на рис. 5.9, выгодой, которую мы получаем от недостатка экономии, является более точное соответствие данным (т. е. меньшая ошибка).

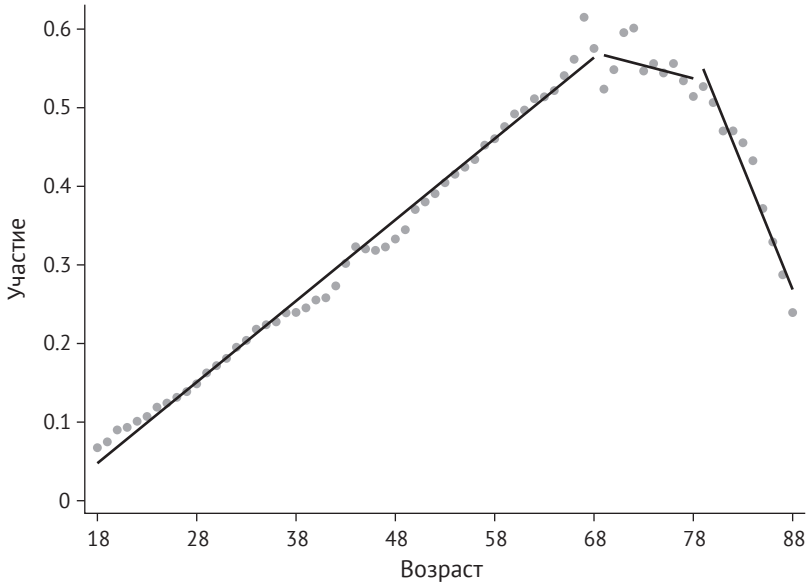


Рис. 5.9. Отдельные линии регрессии для явки избирателей и категорий возраста 16–68, 69–78 и 79–88 лет

Еще в главе 2 мы намекнули на третий способ борьбы с нелинейностью. Нет причин, по которым наше уравнение регрессии должно иметь только одну объясняющую переменную. Если мы знаем, что существует нелинейная связь между явкой и возрастом, возможно, имеет смысл рассмотреть преобразование переменной возраста в квадрат возраста, куб возраста и т. д.

Применяя этот подход в главе 2, мы сохранили простоту регрессии. Мы регрессировали результирующую переменную по квадрату объясняющей переменной. Но мы можем применить более общий подход. Вместо того чтобы искать регрессию явки избирателей только по возрасту или по квадрату возраста, мы можем найти регрессию по обоим показателям. Это позволяет нам подогнать под наши данные более гибкую функцию, чем уравнение прямой линии. Конечно, с каждой новой переменной, которую мы включаем в уравнение, добавляется новый коэффициент, который придется варьировать для минимизации суммы квадратов ошибок. Но наш компьютер с этим справится.

В принципе, мы не обязаны ограничиваться различными преобразованиями переменной возраста. Мы могли бы также включить другие факто-

ры – средний доход или средний статус регистрации избирателей, – которые наверняка еще больше улучшат наши прогнозы. Мы вернемся к этой возможности в главе 10. А пока давайте ограничимся преобразованиями переменной возраста. Имея всего одну объясняющую переменную, легко визуализировать то, что мы делаем, когда запускаем регрессию. Мы просто строим линию через данные в двухмерном пространстве – в частности, линию, которая минимизирует сумму квадратов ошибок.

С двумя объясняющими переменными все становится немного более абстрактно, но в разумных пределах. Теперь мы можем подумать о поиске линии, проходящей через наши данные в трехмерном пространстве. Просто представьте, как на наших графиках вы добавляете третью ось, выходящую из страницы к вам. Эта ось будет иметь масштаб второй объясняющей переменной (возможно, квадрата возраста). Теперь данные образуют облако в этом трехмерном пространстве. Регрессия по-прежнему представляет собой линию, которая минимизирует сумму квадратов ошибок, но теперь линия проходит через облако трехмерных точек данных. Для описания этой линии требуются три параметра вместо двух: точка пересечения (α), наклон по отношению к изменениям первой объясняющей переменной (мы можем назвать ее β_1) и наклон по отношению к изменениям второй объясняющей переменной (мы можем назовем это β_2).

Как только мы выходим за пределы двух объясняющих переменных, становится трудно визуализировать линию регрессии, поскольку большинство из нас не может мыслить в четырех или более измерениях. Но можно провести аналогию. Вы понимаете, что значит найти линию, минимизирующую сумму квадратов ошибок с одной или двумя объясняющими переменными. Нет причин, по которым мы не можем сделать то же самое с десятью. Вашему компьютеру не составит труда вычислить сумму квадратов ошибок и найти коэффициенты регрессии OLS в многомерном пространстве.

Давайте посмотрим, как это работает на практике. Мы воспроизвели регрессию явки избирателей по возрасту, но добавили в уравнение квадрат возраста в качестве объясняющей переменной. То есть мы получили следующее уравнение:

$$\text{Прогнозируемая явка} = \alpha + \beta_1 \cdot \text{Возраст} + \beta_2 \cdot \text{Возраст}^2.$$

Как только наш компьютер рассчитает соответствующие коэффициенты регрессии, мы можем подставить любое значение возраста и соответствующее значение квадрата возраста, чтобы получить прогнозируемый уровень явки. Так, например, если бы мы хотели узнать прогнозируемую явку 31-летних, мы бы подставили 31 вместо возраста и $31^2 = 961$ вместо квадрата возраста.

Мы не обязаны ограничиваться возрастом и квадратом возраста. На рис. 5.10 показана прогнозируемая явка из различных регрессий: одна с возрастом и квадратом возраста в качестве объясняющих переменных (это называется полиномом второго порядка); другая с возрастом, квадратом возраста и кубом возраста в качестве объясняющих переменных (полином третьего порядка); далее с полиномом четвертого порядка; и, наконец, с полиномом десятого порядка!

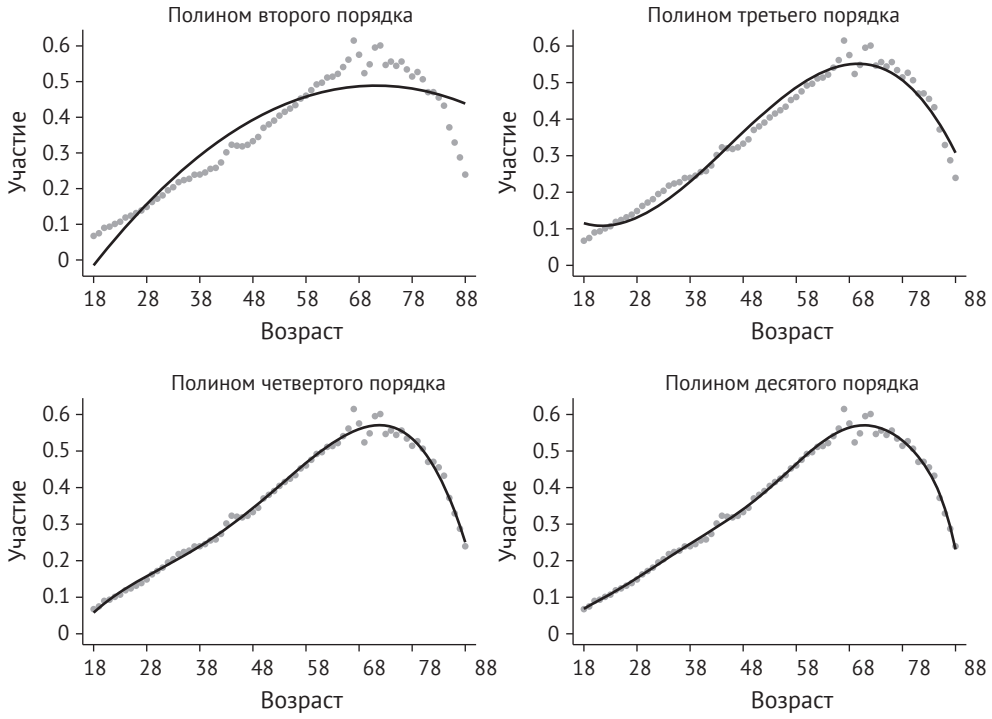


Рис. 5.10. Аппроксимация регрессии явки избирателей с различными полиномами возраста

В целом взаимосвязь между возрастом и явкой довольно сложна. Она примерно линейна от 18 до 68, но через некоторое время после этого происходит крутой поворот. В результате если мы ограничимся только переменной возраста, то не сможем точно подобрать линию регрессии. Точно так же мы видим, что регрессия с возрастом и квадратом возраста также не работает достаточно хорошо, потому что взаимосвязь в данных плохо аппроксимируется квадратичной кривой. Наши прогнозы становятся все лучше и лучше по мере того, как мы включаем все больше и больше объясняющих переменных, поскольку у нас появляется все больше и больше параметров, с которыми мы можем экспериментировать, чтобы настроить кривую на соответствие данным. Когда мы добираемся до полинома четвертого порядка, соответствие выглядит уже довольно хорошо.

Конечно, полином десятого порядка аппроксимирует данные еще лучше: чем больше объясняющих переменных включено в регрессию, тем лучше соответствие. Но это не обязательно означает, что вы должны использовать как можно больше объясняющих переменных. Нужно искать компромисс.

Прежде всего помните, что частью нашей цели является описание данных в простой и лаконичной форме, которую легко понять и объяснить. Описывать данные с помощью 11 параметров (α плюс $\beta_1 \ldots \beta_{10}$) в этом отношении немногим лучше, чем просто перечислять в таблице уровень явки для каждой возрастной группы.

Кроме того, часто возникает необходимость делать прогнозы за пределами выборки, прогнозируя явку избирателей для возрастных групп, которые фак-

тически не наблюдаются в наших данных (например, 90-летние). Добавление все большего количества членов уравнения часто приводит к ухудшению прогнозов за пределами выборки. Причина в том, что, по мере того как используемая нами функция становится все более и более гибкой, она может начать воспринимать каждый небольшой скачок и сбой в данных как значимый эффект, даже если это не так.

Чтобы проиллюстрировать этот момент, мы повторили приведенный выше анализ, но построили регрессии только с использованием данных о людях в возрасте 18–78 лет. Тогда мы сможем увидеть, насколько хорошо получаются вневыборочные прогнозы явки для людей в возрасте старше 78 лет. (Эти прогнозы выходят за пределы выборки, потому что мы намеренно исключили из наших данных избирателей старше 78 лет.) На рис. 5.11 показаны результаты. Данные, использованные для аппроксимации регрессии, показаны закрашенными кружками. Данные, которые мы пытаемся предсказать, показаны пустыми кружками. Кривая линия представляет прогнозируемые значения регрессии. Как видите, полином четвертого порядка хорошо прогнозирует явку самых старых избирателей. Но полином десятого порядка – это катастрофа!

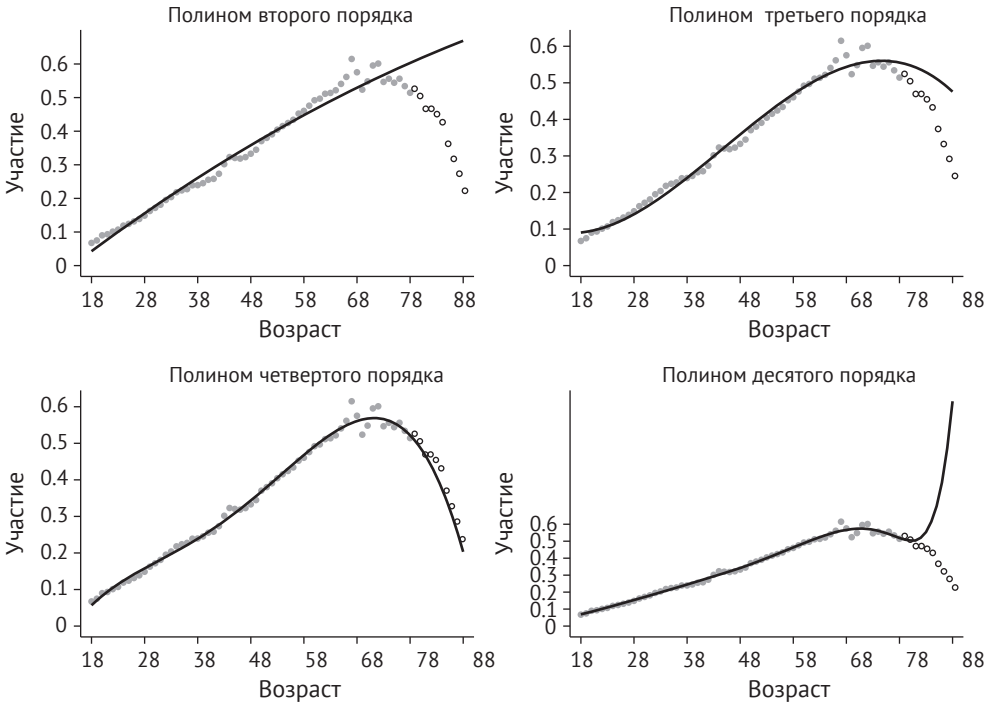


Рис. 5.11. Использование регрессий на основе явки избирателей и различных полиномов возраста для прогнозирования явки избирателей с возрастом, не входящим в выборку

Проблема переобучения

Пример, который мы только что видели, где полином десятого порядка работает хуже, чем полином четвертого порядка при прогнозировании вне выборки,

является примером более общего явления, называемого *переобучением* (overfitting). Если мы проверим достаточное количество объясняющих переменных, мы обязательно обнаружим те, которые случайно коррелируют с результатом в наших данных. Полиномиальная регрессия десятого порядка использовала бессмысленные корреляции между преобразованиями высокого порядка переменной возраста и явкой избирателей среди одной группы избирателей, чтобы попытаться предсказать явку среди другой группы избирателей. Неудивительно, что эти бессмысленные корреляции больше не сохранялись. Чтобы лучше понять суть переобучения, давайте поговорим о несколько более реалистичной задаче прогнозирования.

Прогнозирование президентских выборов

Американцы поголовно заинтересованы в прогнозировании результатов предстоящих президентских выборов. Когда мы говорим людям, что мы политологи, самый распространенный вопрос, который мы получаем, – это «Кто победит на следующих выборах?». Мы склонны разочаровывать своими ответами, поскольку большинство политологов тратят свое время не на прогнозирование выборов.

Однако по сравнению с большинством сложных политических явлений президентские выборы на самом деле довольно предсказуемы. Даже за несколько месяцев до выборов мы часто имеем довольно хорошее представление о том, кто победит, исходя из состояния экономики. А в последние недели перед днем выборов среднее значение опросов обычно находится в пределах одного или двух процентных пунктов от окончательной доли голосов. Журналист Нейт Сильвер зарекомендовал себя как гигант в области анализа политических данных, по сути, усредняя результаты опросов.

Конечно, тот факт, что мы обычно можем предсказать долю голосов с точностью до одного или двух процентных пунктов, не означает, что мы всегда знаем, кто победит. Большинство президентских выборов сопровождаются острой конкуренцией, и коллегия выборщиков позволяет некоторым кандидатам побеждать на выборах, даже теряя голоса избирателей. В условиях равной гонки, как в 2000 или 2016 гг., располагая доступной информацией утром перед выборами, честный количественный аналитик, вероятно, не мог быть уверен более чем на 90 % в победе какого-либо конкретного кандидата.

Хотя мы сказали, что большинство политологов не тратят много времени на предсказание результатов выборов, некоторые занимаются именно этим. Академический журнал PS: Political Science & Politics обычно публикует перед каждыми президентскими выборами подборку различных попыток предсказать результат с использованием количественных данных и анализа. Зачастую цель такого анализа – увидеть, насколько хорошо исследователи могут предсказать результаты предстоящих выборов без использования данных опросов. Например, мы могли бы увидеть, насколько хорошо можно предсказать долю голосов, если учитывать только фундаментальные факторы, такие как экономический рост и срок пребывания в должности действующего президента.

Чтобы сделать такой прогноз, исследователь может построить регрессию, используя исторические данные, в которых каждое наблюдение представляет собой выборы, результирующая переменная – это доля голосов действующей

партии на двухпартийных выборах, а различные объясняющие переменные – это характеристики конкретных выборов, такие как экономический рост в год выборов, добивается ли действующий президент переизбрания, количество военных потерь за последние четыре года и т. д. Получив коэффициенты регрессии на основе данных предыдущих выборов, исследователь может затем подставить значения объясняющих переменных из текущих выборов и получить прогноз предстоящей доли голосов двух партий. Поскольку многие другие аналитики делают то же самое, их цель часто состоит в том, чтобы найти какую-то новую переменную, которую можно включить в собственную регрессию, чтобы улучшить ее предсказательную силу.

Подражая этому подходу, мы построили регрессию, предсказывающую долю голосов за действующего президента на президентских выборах в период с 1948 по 2012 гг. Для большей точности мы включили десять различных независимых переменных, каждая из которых была определена политологами как факторы, которые могут помочь нам предсказать результаты выборов. В частности, мы включили индикатор того, является ли действующий президент демократом или республиканцем; индикатор того, претендует ли действующий президент на переизбрание; рост ВВП в 1-й, 2-й, 3-й и 4-й годы последнего президентского срока; показатель того, была ли страна вовлечена в крупную войну за время текущего президентского срока; подсчет количества сроков подряд, в течение которых одна и та же партия находилась у власти (многие ожидают, что избиратели с большей вероятностью заменят партию, которая находилась у власти в течение длительного времени); уровень безработицы и изменение уровня безработицы за последние четыре года.

У нас есть веские основания ожидать, что эти десять переменных помогут нам предсказать результаты президентских выборов, и на первый взгляд кажется, что так оно и есть. Статистический показатель r^2 регрессии равен 0.83, т. е. 83 % изменений в доле голосов за действующего президента, по-видимому, объясняются этими переменными. Более того, когда мы рассчитываем прогнозируемые значения на основе этой регрессии, они расходятся с фактической долей голосов в среднем лишь на 1.7 процентных пункта.

Однако кажущийся успех нашей регрессии обманчив. Оказывается, если бы мы просто сгенерировали десять случайных величин (что мы и сделали в компьютерном моделировании) и запустили ту же регрессию, используя эти бессмысленные числа в качестве объясняющих переменных, мы бы получили в среднем показатель r^2 около 0.67 и среднюю ошибку 2.4 процентных пункта. Это почти так же хорошо, как наши прогнозы, основанные на реальных данных, хотя наши десять случайно сгенерированных переменных вообще не должны содержать никакой информации о вероятном исходе выборов.

Это весьма удивительно. Что произошло? Когда вы генерируете множество абсолютно случайных переменных, некоторые из них случайно окажутся коррелирующими с вашим результатом. В регрессии эти бессмысленные переменные будут предсказывать результат. Но на самом деле это, конечно, не так. Если вы попытаетесь использовать прогнозы, основанные на взаимосвязи между этими бессмысленными переменными и прошлыми результатами, для прогнозирования будущих результатов, вы потерпите неудачу. Их предсказательная сила – всего лишь иллюзия, созданная случайно.

Один из способов попытаться оценить и смягчить переобучение – исключить некоторые данные из вашего регрессионного анализа и провести тесты вне выборки, как мы сделали с избирателями старше 78 лет в предыдущем разделе. В контексте прогнозирования результатов выборов при составлении прогноза доли голосов в 2012 г. мы могли бы исключить данные за 2012 г. из выборки, построить регрессию, используя все остальные выборы, сгенерировать прогнозируемое значение для 2012 г., используя эти коэффициенты регрессии и истинные значения объясняющих переменных на 2012 г. и посмотреть, как оправдаются наши прогнозы. В принципе, мы могли бы сделать это для каждого года в нашем наборе данных – удалить одно наблюдение, запустить нашу регрессию, сгенерировать прогнозируемое значение для этого наблюдения, проверить наш прогноз на истинность и повторить это для каждого наблюдения.

Когда мы подвергаем нашу регрессию с десятью объясняющими переменными тестированию вне выборки, она оказывается намного хуже, чем казалось на первый взгляд. Средняя ошибка прогноза подскочила с 1.7 до 5.6 процентных пункта. Мы сомневаемся, что какая-либо кампания наймет консультанта по статистике, который мог бы только пообещать предсказать результаты выборов в среднем с точностью до 5 или 6 процентных пунктов. Еще более смущает то, что наивный прогноз, основанный на простом среднем значении других показателей в выборке, находится в пределах 4.6 процентных пункта. Другими словами, переобученная регрессия, которая, как мы думали, дает нам такие точные прогнозы, на самом деле хуже, чем анализ, который вообще не использует объясняющие переменные.

Конечно, когда аналитикам удается избежать переобучения, они могут генерировать полезные прогнозы. Простая регрессия, в которой в качестве объясняющей переменной используется только рост ВВП за четвертый год, дает ошибку прогнозирования вне выборки в 3.8 процентных пункта, превосходя модель без объясняющих переменных. А если бы мы включили результаты опросов, как это делает Нейт Сильвер, мы бы сделали еще лучше. Тем не менее легко обмануть себя, заставив думать, что вы делаете хорошие прогнозы, хотя на самом деле это не так. Аккуратные аналитики включают в свою регрессию только те переменные, которые, по их мнению, действительно коррелируют с результатом, они избегают слишком большого количества переменных в своей регрессии по сравнению с количеством наблюдений и подтверждают свою прогностическую стратегию с помощью тестирования вне выборки.

Как представляют выводы регрессии

Иногда выводы регрессии представляют графически, как мы делали до сих пор. Но наиболее распространенной формой представления выводов регрессии является таблица. Например, в табл. 5.1 показано, как могут быть представлены выводы регрессии явки избирателей по возрасту.

Возможно, вы еще не совсем понимаете значение всех чисел в этой таблице (к числу в скобках, означающему стандартную ошибку, мы еще вернемся в главе 6), но почти все должно быть вам знакомо. Число в строке «Константа» – это точка пересечения α^{OLS}. Число в строке «Возраст» представляет собой наклон линии регрессии β^{OLS}. Мы также уже обсуждали идею r^2 в главе 2: это величина отклонения в явке избирателей, которую можно предсказать по воз-

расту. А Root-MSE – это квадратный корень из среднеквадратической ошибки, который дает вам некоторое представление о том, насколько в среднем наши прогнозы регрессии далеки от реальных точек данных.

Таблица 5.1. Вывод регрессии средней явки избирателей по возрасту

	DV = Явка избирателей
Возраст	0.0103 (0.0001)
Константа	–0.1381 (0.0066)
r^2	0.991
Root-MSE	0.151
Наблюдения	50

Краткая история регрессии

Насколько могут судить историки статистики, регрессия была изобретена (или открыта?) примерно в конце XIX в. Первый опубликованный пример линейной регрессии обнаружен в приложении к небольшой книге французского математика Адриана-Мари Лежандра под названием «Новые методы определения орбит комет». Это была работа, имевшая важные последствия для геодезии – изучения измерения Земли, что было задачей с высокими ставками, учитывая экономическое и военное значение мореплавания в XVIII в.

Статус Лежандра как первооткрывателя регрессии оспаривал современник, великий немецкий математик Карл Фридрих Гаусс. В «Теории движения небесных тел, движущихся вокруг Солнца в конических сечениях» 1809 г. Гаусс заявил свою претензию на авторство, написав: «Наш принцип, которым мы пользуемся с 1795 г., недавно был опубликован Лежандром». Лежандра это примечание не обрадовало, и они продолжали язвить друг друга по этому поводу на протяжении всего начала XIX в.

Ни Гаусс, ни Лежандр не называли метод построения линии наилучшего соответствия путем минимизации суммы квадратов ошибок регрессией. Этот термин был придуман ученым конца XVIII в. Фрэнсисом Гальтоном. Гальтон, двоюродный брат Чарльза Дарвина, который также был женат на племяннице Дарвина, был эрудитом (он пробовал себя и преуспел во многих различных областях). Ему также пришла в голову идея современной системы снятия отпечатков пальцев, и он был первым человеком, количественно задокументировавшим феномен «мудрости толпы»[1]. Более тревожно то, что Гальтон был евгеником – сторонником селективного разведения людей. Чтобы внести яс-

[1] Идея состоит в том, что, если вы спросите достаточное количество людей, даже если они не являются экспертами, возможно, их ошибки взаимно компенсируются, и вы получите хороший ответ. Гальтон показал, что, хотя большинство людей плохо угадывают вес быка, если вы спросите сотни людей и усредните их ответы, вы очень близко приблизитесь к правильному весу. К сожалению, это не всегда работает.

ность: мы не поддерживаем и не одобряем евгенику, но регрессия оказывается полезной и для неевгеников.

Интерес Гальтона к евгенике привел к тому, что он захотел количественно изучить эволюцию и наследственность. Он начал с измерения таких простых вещей, как рост. В одном анализе он собрал данные о росте родителей и их детей. После построения диаграммы распределения он оценил среднее соотношение между этими двумя переменными, используя то, что мы теперь называем линией регрессии.

Анализ Гальтона на самом деле был немного сложнее. Он сравнил рост детей со средним ростом их родителей, предварительно нормировав данные так, чтобы рост женщин и мужчин измерялся в одном масштабе. Мы не будем повторять его путь. Чтобы понять идею исследования, представьте себе анализ, подобный анализу Гальтона, но рассматривающий только рост отцов и сыновей. Единицей анализа является пара отец–сын, а уравнение регрессии выглядит следующим образом:

$$\text{Прогнозируемый рост сына} = \alpha + \beta \cdot \text{Рост отца}.$$

Когда Гальтон получил α и β с помощью метода наименьших квадратов, что, по вашему мнению, он обнаружил? Мы могли бы ожидать, что $\alpha^{OLS} = 0$ и $\beta^{OLS} = 1$. Это означало бы, что в среднем сыновья, как правило, имеют тот же рост, что и их отцы, т. е. мы ожидаем, что у отца ростом пять футов сын, скорее всего, тоже будет ростом пять футов, у отца ростом шесть футов сын тоже будет ростом шесть футов и т. д. Вместо этого Гальтон с удивлением обнаружил, что $\alpha^{OLS} > 0$ и $\beta^{OLS} < 1$. Остановитесь на мгновение и подумайте, почему это может быть так.

Результат исследований Гальтона показан на рис. 5.12. Пунктирная черная линия проходит под углом 45°, т. е. представляет собой линию с $\alpha = 0$ и $\beta = 1$. Сплошная серая линия показывает линию наилучшего соответствия регрессии с $\alpha^{OLS} = 38.2$ и $\beta^{OLS} = 0.448$.

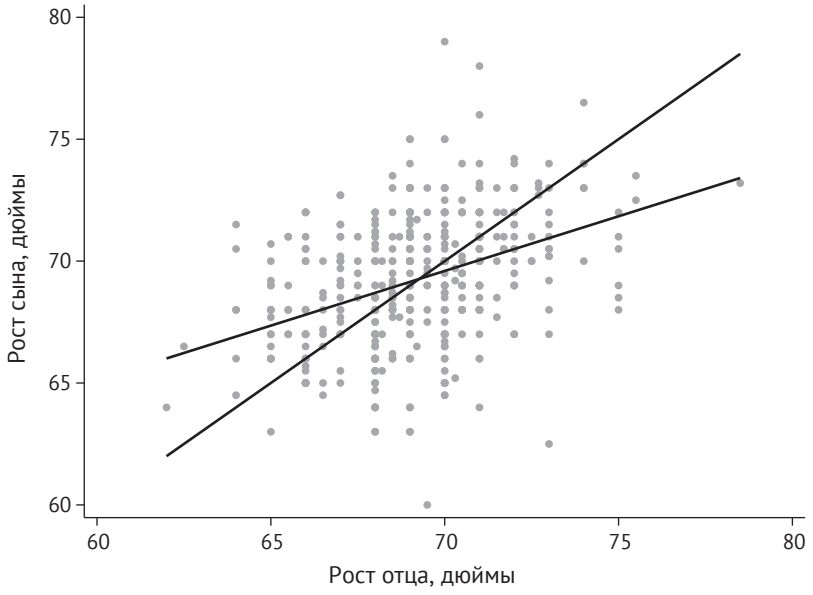


Рис. 5.12. Линия регрессии роста сына относительно роста отца

Давайте начнем с интерпретации этих коэффициентов регрессии. Линия регрессии лежит выше линии 45° для относительно невысоких отцов и ниже для относительно высоких отцов. Это означает, что у высоких отцов, как правило, сыновья выше среднего, но тем не менее ниже их. Точно так же у невысоких отцов, как правило, сыновья ниже среднего, но тем не менее выше их самих. Гальтон назвал это явление «регрессией к посредственности». Сегодня мы обычно называем это явление *регрессией к среднему*, или *возвратом к среднему*, и пониманию этого явления будет посвящена вся глава 8. С тех пор мы используем слово «регрессия» как для обозначения статистической техники Гальтона, так и для явления, которое он открыл с ее помощью. Поэтому неслучайно, что регрессия OLS и регрессия к среднему содержат в своем названии одно и то же слово. У них общая интеллектуальная история.

Подведение итогов

Регрессия – самый важный инструмент, который у нас есть для изучения корреляций. Наклон линии наилучшего соответствия говорит нам о знаке и величине связи между двумя переменными: когда одна увеличивается, в какой степени другая склонна увеличиваться или уменьшаться? Регрессии могут рассказать о многом, но нам следует проявлять бдительность и сохранять критическое мышление. Имея в своем распоряжении мощные компьютерные алгоритмы, вы рискуете расслабиться и впасть в заблуждение. Вы можете избежать некоторых ошибок, нанося данные на график, учитывая потенциальную нелинейность связей и избегая переобучения.

Регрессия показывает взаимосвязь между переменными в наших данных. Если мы просто пытаемся описать данные, это само по себе информативно. Но зачастую мы пытаемся сделать больше. Например, можем попытаться вывести взаимосвязь между переменными в какой-то более крупной совокупности на основе взаимосвязи между этими переменными в имеющихся данных, которые могут представлять собой лишь небольшую выборку из генеральной совокупности. Как узнать, сохранится ли связь, которую мы обнаружили в выборке, в более крупной совокупности? Эти вопросы являются темой главы 6.

Ключевые термины

- **Зависимая переменная:** переменная, связанная с результатом, который мы пытаемся описать, предсказать или объяснить.
- **Независимая или объясняющая переменная:** переменная, которую мы используем, чтобы попытаться описать, предсказать или объяснить зависимую переменную.
- **Уравнение регрессии:** уравнение, линейно связывающее зависимую переменную с некоторыми независимыми переменными.
- **Параметры регрессии:** параметры (точка пересечения и наклоны), которые связывают зависимую переменную с некоторыми независимыми переменными в уравнении регрессии.
- **Ошибка:** разница между значением выходной переменной для отдельной точки данных и прогнозируемым значением для этой же точки данных. Иногда ее также называют остатком.

- **Сумма квадратических ошибок** (SSE): для линии регрессии вычисляют ошибку относительно каждой точки данных, найдя ее вертикальное расстояние от линии. Сумму квадратических ошибок для этой линии находят путем возведения в квадрат каждой отдельной ошибки и их сложения.
- **Регрессия по методу наименьших квадратов** (OLS): метод поиска линии наилучшего соответствия данным, который минимизирует сумму квадратов ошибок.
- **Линия регрессии:** линия наилучшего соответствия данным, полученная в результате регрессии по методу OLS.
- **Пересечение:** в контексте регрессии пересечение сообщает нам прогнозируемое значение результата, когда значения всех независимых переменных равны 0. Это значение также называется постоянным членом, или константой. Иногда пересечение имеет содержательную интерпретацию, но иногда ее нет, потому что не имеет смысла рассматривать ситуацию, когда все объясняющие переменные равны нулю (например, прогнозируемая явка избирателей для людей нулевого возраста). В любом случае мы всегда рассматриваем точку пересечения при выполнении регрессии (за исключением очень необычных обстоятельств, когда мы знаем из теории, что точка пересечения должна быть равна нулю).
- **Функция условного среднего:** функция, которая сообщает вам среднее значение некоторой переменной в зависимости от значения некоторых других переменных.
- **Прогнозирование вне выборки:** использование регрессии (или другого статистического метода) для прогнозирования результатов наблюдений, не входящих в исходные данные, которые вы использовали для создания прогнозов.
- **Переобучение:** попытка спрогнозировать зависимую переменную со слишком большим количеством независимых переменных, так что переменные кажутся предсказывающими зависимую переменную в данных, но не имеют связи с ней в реальном мире.

Упражнения

Загрузите файл SchoolingEarnings.csv и связанный с ним файл README.txt, описывающий переменные в этом наборе данных, по адресу *www.press.princeton.edu/thinking-clearly*. Этот набор данных показывает средний годовой заработок мужчин в возрасте от 41 до 50 лет в Соединенных Штатах в 1980 г. на каждом уровне образования. Одно наблюдение дает средний заработок (в тысячах долларов) для мужчин с восьмилетним образованием, другое – для тех, кто имеет девятилетнее образование, и т. д.

5.1. Найдите регрессию с заработком в качестве зависимой переменной и образованием в качестве единственной независимой переменной. Интерпретируйте коэффициенты.

5.2 Предположим, вам нужен экономичный способ прогнозировать доходы, используя только годы обучения. Что бы вы сделали?

5.3. Давайте разберемся, является ли связь между заработком и образованием приблизительно линейной.

a) Начните с построения точечной диаграммы распределения. Затем нанесите на нее прогнозируемые значения вашей регрессии вместе с точками необработанных данных, как мы это делали в главе 2. Можно ли сказать, что линия регрессии хорошо соответствует данным?
b) Теперь выполните полиномиальную регрессию четвертого порядка (т. е. используйте параметры *образование*, *образование*^{2}, *образование*^{3} и *образование*^{4}). Насколько существенно эти прогнозы отличаются от прогнозов, полученных на основе линейной регрессии?
c) Постройте различные регрессии для нескольких разных уровней обучения. Насколько существенно эти линии отличаются от прогнозов, которые вы получаете на основе одной регрессии, включающей все данные?
d) Можно ли сказать, что простая линейная регрессия подходит для этих данных?

5.4 Аналогично тому, что мы сделали с возрастом и явкой избирателей, проведите несколько тестов вне выборки, чтобы оценить вашу стратегию прогнозирования. Используя только данные об имеющих 12 лет обучения или меньше, проверьте, насколько хорошо ваши различные стратегии из вопроса 3 работают при прогнозировании доходов для тех, кто учится более 12 лет.

Дополнительное чтение и ссылки

Для получения дополнительной информации о ранней истории регрессии см.:

Stephen M. Stigler. 1986. *The History of Statistics: The Measurement of Uncertainty before 1900*. Belknap, Harvard.

Глава 6

Выборки, неопределенность и статистические выводы

О чем эта глава

- Все количественные оценки представляют собой сумму трех слагаемых: истинного значения, систематической ошибки и шума.
- Статистическая проверка гипотез позволяет аналитикам оценить, могла ли оценка возникнуть из-за шума.
- Статистическая значимость и содержательная значимость не одно и то же, и их не следует смешивать.

Введение

В главах 4 и 5 были рассмотрены инструменты, позволяющие нам описывать взаимосвязи между переменными внутри набора данных. Располагая вариациями обеих переменных, мы можем описать корреляцию между этими переменными с помощью регрессии. Но часто нам нужно пойти дальше. Нужно использовать отношения между переменными, найденные в имеющихся данных (ограниченная выборка), чтобы сделать выводы об отношениях, которые существуют между этими переменными в более широком мире (интересующая совокупность). Например, как только мы обнаружим, что преступность выше в теплые дни 2018 г., возникает закономерное желание узнать, сохраняется ли эта взаимосвязь и в другие годы или является просто особенностью данных 2018 г. То есть мы стремимся узнать, отражает ли наблюдаемая закономерность в выборке подлинное явление в совокупности дней, или же она справедлива лишь для выборки данных, которую мы изучили по чистому стечению обстоятельств. В этой главе мы обсудим некоторые инструменты, которые помогут нам сделать вывод о применимости закономерностей.

Оценка

Прежде всего нам нужен общий язык, чтобы говорить о различиях между тем, что мы наблюдаем в нашей выборке, и явлениями в популяции, о которых мы хотели бы узнать. Для этого воспользуемся следующим простым уравнением, которое будет так часто встречаться в оставшейся части книги, что с этого момента мы начнем называть его *нашим любимым уравнением*:

Оценка = Оцениваемая величина + Смещение + Шум.

По ходу дела мы будем тщательно объяснять каждый из этих терминов. Но давайте начнем с некоторых основных определений.

Оценка (estimate) – это число, которое мы получаем в результате нашего анализа. Оценка представляет собой истинное значение переменной в популяции, которое мы пытаемся узнать. Мы надеемся, что наша оценка будет близка к *оцениваемой величине* (estimand). Оценка может не совпадать с оцениваемой величиной по двум причинам: из-за систематической ошибки (смещение, bias) и шума. *Смещение* относится к ошибкам, которые происходят по систематическим причинам, а шум относится к *нефакторным* (несистематическим) ошибкам, возникающим по случайным причинам.

Давайте подготовим почву на простом примере, который позволит нам более четко определить и понять эти термины.

Предположим, мы проводим опрос, чтобы узнать, какой из двух кандидатов (республиканец или демократ) победит на предстоящих выборах. Это исследование можно рассматривать как задачу прогнозирования: мы собираем данные, чтобы спрогнозировать будущего победителя. С другой стороны, это описание в чистом виде: мы хотим знать долю избирателей, которые поддерживают одного кандидата против другого.

В любом случае ключевой проблемой является то, что избирателей слишком много, чтобы мы могли опросить их всех. Практические соображения вынуждают нас провести опрос выборки, составляющей лишь небольшую часть от общей численности избирателей. Таким образом, нам нужно выяснить, какие выводы можно сделать о политических взглядах более обширной группы населения на основе данных, полученных в результате опроса лишь относительно небольшой выборки.

В нашем примере нас интересует доля избирателей среди населения, поддерживающих республиканцев. Обозначим эту долю, представляющую собой число от 0 до 1, буквой q. Поскольку кандидатов всего два, доля сторонников демократов составляет $1 - q$. Итак, q – наша оценка. Пока выборы не состоятся, мы не сможем наблюдать q; мы должны попытаться оценить это значение.

Предположим, мы опрашиваем случайную выборку из 100 избирателей и спрашиваем их, кого они поддержат – республиканца или демократа. Мы могли бы оценить количество избирателей среди всего населения, поддерживающих республиканцев (что мы не можем наблюдать), определив долю людей в нашей выборке, поддерживающих республиканцев (что мы можем наблюдать). Обозначим нашу оценку на основе выборки символом $\hat{q}$. Следуя стандартной практике, мы будем обозначать оцениваемую величину буквой (не обязательно q), а для обозначения оценки этой величины будем использовать ту же букву, но с символом «крышечки» над ней. В данном случае надеемся, что наша оценка $\hat{q}$ будет близка к оцениваемой величине q.

В этом примере *оцениваемой величиной* является истинная доля республиканцев в населении (q) – это ненаблюдаемая величина, которую мы пытаемся узнать с помощью нашего анализа данных. Процесс выборки 100 избирателей и расчета доли сторонников республиканцев называется *оцениванием* – это процедура, которую мы применяем для получения числового результата. Доля

республиканцев в нашей выборке ($\hat{q}$) – т. е. наша оценка – это численный результат, полученный в результате оценивания выборки, и, как мы надеемся, он близок к реальному значению для популяции.

Выяснив разницу между оценкой и оцениваемой величиной, мы сделали первый шаг к пониманию нашего любимого уравнения:

Оценка = Оцениваемая величина + Смещение + Шум.

Величина, которая нас интересует, является оцениваемой величиной. Величина, которую мы наблюдаем в имеющейся выборке данных, является оценкой. В идеальном мире оценка была бы равна значению оцениваемой величины, поскольку процесс оценивания показал бы нам истинное значение интересующего показателя. Но наше любимое уравнение говорит, что это не так. Оценки отличаются от реальности из-за систематической ошибки и шума. Чтобы понять почему, нам нужно больше узнать об этих двух неприятных компонентах.

Почему оценка отличается от оцениваемой величины?

Одинаково важно учитывать и смещение, и шум. Но они различаются, и их отличие часто упускается из виду людьми, что приводит к заблуждениям. Мы рассмотрим их по очереди. В обсуждении смещения и шума нам поможет следующая аналогия.

Энтони любит играть в шотландский керлинг. В керлинге две команды по очереди катают тяжелые гранитные камни по длинной ледяной дорожке. Пока камень скользит, другие члены команды, как сумасшедшие, мечутся перед ним, бегая по льду. Рекомендуем посмотреть видео; это довольно весело. В любом случае очки получает команда, чей камень находится ближе всего к центру мишени (так называемый баттон, button) на дальнем конце ледяной дорожки.

Энтони неплохо играет в керлинг. Он может более или менее точно направлять свои камни туда, куда он хочет. Но, несмотря на его мастерство, иногда его камни не попадают в баттон (в керлинге вы не всегда пытаетесь «дотянуться до баттона», но для целей данного обсуждения будем считать, что это и есть ваша цель). Почему так происходит? Существует множество факторов, неподвластных игроку, но влияющих на то, как скользит камень. Возможно, на дорожке был какой-то мусор, из-за которого метко направленный камень отклонился от курса. Или, может быть, Энтони поскользнулся на льду и случайно изменил направление толчка. В любом случае есть много причин, по которым камни меткого Энтони могут не попасть в баттон.

Итан, напротив, ужасно играет в керлинг. Поэтому, когда он играет в керлинг с Энтони, его камни часто не попадают в цель, обычно отклоняясь влево (не говоря уже о расстоянии). Он хотел бы заявить, что это произошло из-за несистемных факторов, как в случае промахов Энтони. Но если бы это было правдой, он бы с одинаковой вероятностью промахивался как влево, так и вправо. Нет, горькая правда в том, что техника бросков Итана плохая, поэтому его камни систематически направляются левее цели.

Мы считаем, что существует полезная аналогия между керлингом и анализом данных. Отнеситесь к центру мишени в керлинге как к оцениваемой вели-

чине: это истина, к которой вы стремитесь. Процесс оценивания аналогичен скольжению камня по льду. И рассматривайте результат одного броска камня как оценку, полученную в результате одной итерации оценивающего процесса.

Ваш камень (оценка) может не попасть в баттон (оцениваемая величина) по двум причинам. Во-первых, вы, как и Энтони, возможно, хорошо прицелились, но случайные факторы могли сдвинуть камень в ту или иную сторону. Эти случайные факторы подобны шуму. Поскольку эти факторы в среднем случайны, они не заставляют Энтони чаще промахиваться влево или вправо. Действительно, в среднем расположение его камня – на баттоне. Но это не означает, что каждый отдельный камень попадает на баттон; его промахи просто усредняют друг друга. Именно это и делает шум: оценки могут *в среднем* равняться оцениваемой величине, но из-за шума любая конкретная оценка может иметь ошибку.

Во-вторых, как и Итан, вы можете систематически целиться слишком влево. Шум по-прежнему влияет, поэтому иногда можно промахнуться вправо. Но в среднем ваш камень оказывается слева от баттона. Эти регулярные промахи отражают наличие систематического смещения. В отличие от Энтони средний бросок Итана не попадает в цель. Именно так проявляется смещение: когда оно есть, даже средняя оценка не равна оцениваемой величине, не говоря уже о какой-либо конкретной оценке.

Прекрасно, теперь, когда есть аналогия, которая поможет вам понять разницу между смещением и шумом, давайте поговорим о них более подробно.

Смещение

Одна из причин, по которой оценивающий процесс может дать вам оценку, не совпадающую с реальной величиной, заключается в том, что она является смещенной. Представьте себе, что вы применяете свой оценивающий процесс снова и снова бесконечное количество раз всегда к новой, независимой выборке данных. Это привело бы к получению бесконечного числа оценок. Из-за шума некоторые из этих оценок будут больше, чем оцениваемая величина (т. е. в некоторых выборках вы получите большую долю республиканцев, чем в генеральной совокупности), а некоторые из этих оценок будут меньше (т. е. для некоторых выборок вы получите меньшую долю республиканцев, чем в генеральной совокупности). Но вам хотелось бы, чтобы среднее из этого бесконечного числа оценок было равно оцениваемой величине. То есть вы не хотите предсказуемо (или систематически) переоценивать или недооценивать число республиканцев. Вы стремитесь к истине. Мы говорим, что оценивающий процесс является *несмещенным*, если после его применения к бесконечному числу новых независимых выборок среднее значение генерируемых им оценок будет равно оцениваемой величине.

Мы также иногда говорим о среднем значении переменной за бесконечное количество прогонов с точки зрения *ожиданий* (expectation). Мы могли бы сказать, что несмещенная оценка равна оцениваемой величине *в ожидании*. Или мы могли бы сказать, что *ожидаемое значение* несмещенной оценки – это искомое оцениваемое значение.

Есть много причин, по которым политический опрос может быть смещенным. Предположим, избиратели систематически лгут социологам. Возможно, избиратели полагают, что социологи являются демократами, и избиратели

хотят им угодить или ввести в заблуждение, поэтому некоторые избиратели-республиканцы сообщают, что поддерживают демократов. Тогда наша оценка будет смещена в пользу демократов – в среднем мы сообщим о большей поддержке демократов избирателями, чем есть на самом деле. Или предположим, что демократы с большей вероятностью придут на выборы, чем республиканцы, но с одинаковой вероятностью ответят на опросы. Тогда респонденты опроса будут отличаться от избирателей, а оценки опросов будут смещены в пользу республиканцев – в среднем сообщая о большей поддержке республиканцев избирателями, чем есть на самом деле. Наконец, что, если социологи связываются с людьми только по определенному каналу связи (электронная почта, домашний телефон, мессенджер соцсети), а пользователи этого канала систематически отличаются по своим политическим взглядам от населения в целом? Это также приведет к предвзятости. По многим причинам, если бы мы проводили опрос бесконечное количество раз и усредняли оценки, это среднее значение могло бы не соответствовать истинной доле республиканцев в общей численности избирателей. Следовательно, опрос может быть смещенным по любой из этих причин.

В последующих главах мы очень внимательно обсудим источники смещения. Однако в оставшейся части этой главы будем игнорировать источники систематической ошибки, чтобы сосредоточиться на второй потенциальной проблеме – шуме.

Шум

Когда вы берете выборку населения, то неизбежно вносите в свою оценку некоторый шум. Когда вы спрашиваете у 100 случайно выбранных людей из 100 млн их мнение о политических кандидатах, иногда случайно вам доводится поговорить с непропорционально большой долей республиканцев, а иногда – с непропорционально большой долей демократов. В результате даже без смещения любая индивидуальная оценка не обязательно равна оцениваемой величине. Предположим, ваш оценивающий процесс не смещен. Применив его бесконечное число раз, вы получите несмещенную среднюю оценку поддержки республиканцев или демократов. Но каждая отдельная оценка, вероятно, будет несколько отличаться от реального значения из-за шума, т. е. естественной изменчивости, возникающей в результате выборки. Эту естественную изменчивость иногда называют *вариацией выборки* (sample variation); она является распространенным источником шума.

Существуют способы измерения количества шума, связанного с оценивающим процессом. Представьте многократное применение оценивающего процесса к новым независимым выборкам данных бесконечное количество раз. Чем ближе различные оценки будут друг к другу, тем *точнее* процесс оценки. Таким образом, более точная оценка – это оценка с меньшим шумом.

Как получается хороший оцениватель?

В конце концов, мы пытаемся узнать истинное значение оцениваемой величины. Поскольку наша оценка может отличаться от оцениваемой величины из-за

систематической ошибки или шума, нам очень нужна оценка, которая была бы одновременно несмещенной и точной.

Если наш оценивающий процесс не смещен, но недостаточно точен, наши оценки обычно будут отличаться от оцениваемой величины из-за большого количества шума. Например, в нашем примере с опросом, если мы случайным образом побеседуем только с одним избирателем, его мнение отразит объективную оценку среднего мнения электората (если вы проделали это бесконечное количество раз, q из этих раз вы получите республиканцев, и в $1 - q$ случаев вы получите демократов). Но вариация выборки, связанная с оценкой мнения избирателей путем опроса мнения только одного человека, огромна – мы всегда будем оценивать либо 100-процентных республиканцев, либо 100-процентных демократов.

Если наша оценка смещена, но точна, наши оценки обычно будут отличаться от оцениваемой величины, поскольку они очень точно оценивают неправильную величину. Например, если мы выберем десять тысяч избирателей, но сделаем это только в республиканских кварталах, мы получим очень плотно сгруппированные ответы, но они будут систематически переоценивать число республиканцев.

Рисунок 6.1 показывает, что оценки могут быть несмещенными, точными, ни теми, ни другими или и теми, и другими. Черные ромбы представляют собой оцениваемую величину – истинное значение в реальном мире. Серые точки показывают различные оценки, возникающие в результате повторных применений оценивающего процесса, каждый раз с независимой выборкой данных. Если серые точки симметрично распределены вокруг ромба (как керлинговые камни Энтони вокруг баттона), оцениватель является несмещенным. То есть оценки, которые он дает, в среднем верны. Если серые точки плотно сгруппированы вместе, оценка точна. То есть шума очень мало, поэтому оцениватель дает приблизительно одинаковые оценки на каждой итерации. При прочих равных условиях нам, очевидно, хотелось бы, чтобы наша оценка была менее предвзятой и более точной. Однако иногда между этими целями приходится идти на компромисс, и мы вынуждены решать, с каким смещением мы готовы мириться ради определенного выигрыша в точности.

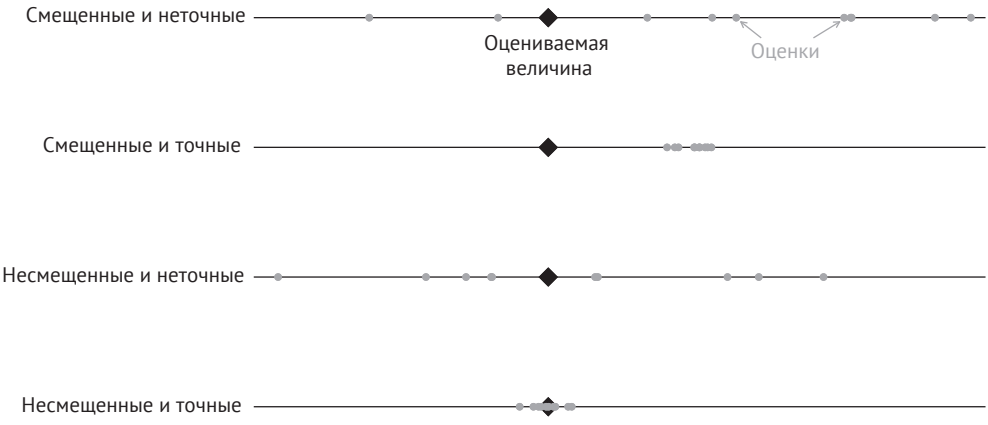


Рис. 6.1. Понимание разницы между отсутствием смещения и точностью

В качестве конкретного примера возможного компромисса между смещением и точностью давайте вернемся к теме опросов. Предположим, у вас есть 2000 долл. и вы хотите провести надежный опрос, чтобы понять, насколько популярен будет политический кандидат, политическое предложение, продукт или потенциальная рекламная кампания. Вы можете опубликовать опрос в интернете, платить людям 20 центов за ответ и получить десять тысяч ответов. Или вы можете заплатить профессиональной фирме, занимающейся опросами, чтобы получить случайную репрезентативную выборку по цене 20 долл. за ответ, т. е. вы сможете позволить себе только сто ответов.

Недорогая онлайновая выборка намного больше, поэтому ваши оценки общественного мнения будут более точными, но они также, вероятно, будут смещенными. Те люди, которые добровольно участвуют в опросах за весьма скромную компенсацию, вряд ли будут репрезентативными для населения в целом. Профессионально проведенный опрос, скорее всего, даст вам менее смещенные оценки, но размер выборки будет меньше, поэтому такие оценки будут менее точными.

Подобный компромисс между смещением и точностью довольно распространен среди аналитиков данных, и мы покажем больше примеров в части III. Правильный способ найти этот компромисс будет зависеть от ваших целей, стоимости различных видов ошибок и конкретного вопроса, на который вы надеетесь ответить.

Если оцениватель не смещен, нам также хотелось бы, чтобы он был как можно более точным. И, как мы уже говорили, мы могли бы даже допустить небольшое смещение в обмен на большой прирост точности. Но если оцениватель сильно смещен, то уже не очевидно, что точность – это хорошо. Во-первых, точная, но смещенная оценка никогда не будет близка к истине, в то время как с меньшей точностью вы иногда можете делать хорошие прогнозы, несмотря на смещение, хотя и случайно. (Итану, вероятно, было бы лучше, если бы в тот момент, когда он толкнул свой камень для керлинга, случилось землетрясение, потому что, по крайней мере, хотя бы тогда его камень остался бы в игре.) Более того, точность может дать вам ложное чувство уверенности. Остерегайтесь точных оценок с неизвестным систематическим смещением.

Количественная оценка точности

Помните главный вопрос этой главы: когда мы оцениваем что-то на основе выборки данных, насколько уверены мы можем быть в выводах о более широкой популяции? Как вы только что видели, если наша оценка смещена, это серьезный повод для беспокойства. Но даже если наша оценка не является предвзятой, нам все равно придется беспокоиться о том, что наши оценки не отражают истинные отношения в более широкой совокупности (оцениваемую величину) из-за шума. Чтобы понять, насколько нам следует беспокоиться по этому поводу, необходимо количественно измерить точность оценивателя. Это делается с помощью статистического показателя, называемого стандартной ошибкой, которую затем можно использовать для построения доверительных интервалов.

Стандартные ошибки

В главе 2 мы говорили о стандартном отклонении как об одном из способов измерения того, насколько широко простирается распределение переменной (или, что аналогично, насколько изменчивой является переменная). Что ж, представьте, что мы запустили наш оцениватель бесконечное количество раз, каждый раз с новой выборкой данных. В этом мысленном эксперименте мы можем рассматривать саму оценку как переменную. Каждый раз, когда мы извлекаем выборку данных и запускаем оцениватель, мы получаем другое значение оценки из-за шума. Следовательно, можно представить распределение оценок, которые мы получим после повторения нашей оценки бесконечное количество раз. Это воображаемое распределение называется *выборочным распределением* (sampling distribution). Стандартное отклонение этого выборочного распределения называется *стандартной ошибкой* (standard error). Стандартная ошибка, если бы мы ее знали, дала бы нам представление о том, насколько далека любая конкретная оценка от средней оценки, поскольку она измеряет, насколько изменчивыми будут наши оценки. Если оценка не смещена, средняя оценка равна оцениваемой величине. Итак, для несмещенной оценки стандартная ошибка приблизительно говорит нам, насколько далека типичная оценка от истинного значения, которое мы пытаемся узнать.

Если стандартная ошибка велика, оценки будут очень разбросаны, а оцениватель будет относительно неточным (т. е. присутствует большая вариация выборки). Если стандартная ошибка невелика, то оценки будут очень близки друг к другу, а оцениватель будет относительно точным (т. е. вариация выборки невелика). Вернемся к рис. 6.1. В третьей строке показан пример некоторых оценок из повторных запусков оценивателя с относительно большой стандартной ошибкой – как следствие, оценки, которые мы видим, сильно разбросаны. (Конечно, мы не видим полного выборочного распределения, поскольку у нас нет бесконечного количества оценок.) В четвертой строке показан пример некоторых оценок из повторных запусков оценивателя с относительно небольшой стандартной ошибкой – как следствие, оценки, которые мы видим, тесно сгруппированы.

Мы можем составить некоторое представление о том, что делает оцениватель точным или неточным (т. е. когда стандартная ошибка будет маленькой или большой). В нашем примере опроса стандартная ошибка примерно равна $\sqrt{\frac{q(1-q)}{N}}$, где N обозначает размер выборки (количество опрошенных людей). Мы не будем здесь показывать вывод формулы (это тема для другой книги), но попытаемся понять, что делает оцениватель более или менее точным, размышляя над формулой.

Начнем с числителя $q(1 - q)$. Обратите внимание, что числитель максимален при $q = 1/2$ и уменьшается по мере того, как q становится больше или меньше. Итак, предположим, что истинная доля республиканцев в населении q либо очень велика (близка к 1), либо очень мала (близка к 0). В том и другом случае значение выражения $q(1 - q)$ очень мало, и, следовательно, стандартная ошибка тоже мала. Почему? Когда q очень велико или очень мало, вероятность ошибки выборки незначительна. Если 99 % избирателей составляют республиканцы, то, когда вы соберете выборку, скажем, в тысячу избирателей, очень маловеро-

ятно, что вы найдете много демократов. Напротив, если q близко к половине, стандартная ошибка велика. Это отражает тот факт, что существует много возможностей для ошибок выборки. Вы можете легко найти фактическое разделение 55-45 или 45-55 в выборке данных, взятой из населения 50-50. Чем ближе q к половине, тем более естественными являются вариации интересующего нас результата, что увеличивает стандартную ошибку.

Теперь рассмотрим знаменатель. Он говорит нам о том, что по мере увеличения размера выборки стандартная ошибка снижается. Это вполне логично. Проблема неточности связана с тем, что наша выборка может неточно отражать всю совокупность. Когда выборка большая, она будет более точно приближаться к генеральной совокупности. Мы можем более точно оценить мнение миллиона человек, опросив 10 000 респондентов, чем после опроса десяти респондентов.

Формула стандартной ошибки на самом деле говорит нам нечто более тонкое, чем простой факт, что малые размеры выборки приводят к неточности. Формула говорит нам о том, что стандартная ошибка уменьшается пропорционально $\sqrt{N}$. Предположим, что истинная доля республиканцев в населении равна q = 0.5. Тогда, если бы мы опросили 1000 избирателей, стандартная ошибка составила бы $\sqrt{\frac{.5 \cdot .5}{1000}} \approx .016$. Предположим, мы провели гораздо более крупный опрос 10 000 человек. Тогда наша стандартная ошибка равна $\sqrt{\frac{.5 \cdot .5}{10\ 000}} = .005$. Таким образом, увеличение размера выборки в 10 раз повышает точность опроса примерно в три раза. Если мы увеличим размер выборки до 100 000, то получим еще примерно трехкратное повышение точности ($\sqrt{\frac{.5 \cdot .5}{100\ 000}} = .0016$). Другими словами, отдача от увеличения размера выборки уменьшается. Стандартная ошибка опроса с участием 10 000 респондентов уже и так очень мала, и добавление большего количества респондентов не приводит к значительному повышению точности.

Вы, возможно, заметили одну непростую вещь: чтобы вычислить стандартную ошибку, нам нужно знать q. Но мы не знаем q, поэтому для начала проводим опрос. На практике мы аппроксимируем стандартную ошибку, подставляя в формулу $\hat{q}$ нашу оценку q. Конечно, это приближение столкнется с проблемами, если у вас действительно маленькое N или значение q, очень близкое к 0 или 1. Предположим, вы поговорили с пятью избирателями и обнаружили, что ни один из них не является сторонником республиканцев. Бездумно применяя описанные выше процедуры, вы можете ошибочно прийти к выводу, что никто не является республиканцем и что ваша стандартная ошибка равна 0. Конечно, это неверно, потому что при небольших выборках и экстремальных значениях q ваше приближение с использованием $\hat{q}$ вводит в заблуждение.

Следует также отметить, что, хотя в нашем примере опроса фигурирует хорошая формула для аппроксимации стандартной ошибки, это не всегда так. К счастью, наши компьютеры часто могут давать достаточно надежные приближения к стандартным ошибкам даже в более сложных обстоятельствах.

Маленькие выборки и экстремальные наблюдения

Стоит сделать паузу и отметить, что тот факт, что небольшие выборки приводят к неточности, объясняет распространенное явление, которое вы, возможно, заметили раньше. Если вы посмотрите данные о городах с самым высоким или самым низким уровнем заболеваемости раком или с самым высоким или самым низким средним доходом, вы найдете список городов с довольно небольшим количеством жителей. Аналогичным образом, если вы посмотрите школы с самыми высокими или самыми низкими средними баллами за тесты, вы найдете список школ с небольшим количеством учеников. Почему так получается?

Мы можем рассматривать средний уровень заболеваемости раком или средний доход горожан как оценку национального уровня заболеваемости раком или среднего дохода, точно так же как средняя поддержка республиканцев в выборке опросов является оценкой средней поддержки республиканцев среди всего населения. Когда количество жителей в городе невелико, это эквивалентно небольшому размеру выборки. Это приводит к меньшей точности (большему количеству шума) в ваших оценках. Значит более вероятно, что ваша оценка будет иметь экстремальное значение в любом направлении. Маленькие города, как правило, доминируют в списке мест с экстремальным уровнем заболеваемости раком или средними доходами не потому, что они обязательно в среднем более или менее предрасположены к раку или более или менее богаты, а потому, что уровень заболеваемости раком и средние доходы в них более изменчивы, чем в местах с более высоким количеством людей для усреднения.

В качестве крайнего примера представьте себе город, в котором живет всего один житель. В этом городе либо 100-процентный уровень заболеваемости раком, либо 0-процентный уровень. Но если в городе 100 000 жителей, уровень заболеваемости раком будет где-то посередине, намного ближе к среднему показателю по стране.

Эту мысль иллюстрирует рис. 6.2, вдохновленный аналогичным графиком из книги Говарда Уэйнера «Познание неопределенного мира». На рисунке представлены данные средних школ Калифорнии за 2012 г. Мы наблюдаем среднюю успеваемость учащихся (индекс академической успеваемости, который в значительной степени определяется результатами стандартизированных тестов) и размер контингента учащихся в каждой школе. Пустые кружки представляют школы с самыми низкими показателями (5 % нижних показателей по академической успеваемости), а закрашенные – школы с самыми высокими показателями (5 % лучших показателей успеваемости). Как видно из линии регрессии, в этих данных на самом деле существует положительная корреляция между размером школы и успеваемостью – в среднем в более крупных школах дети учатся лучше, чем в меньших. Но, что более важно для нас, в обеих группах избыточно представлены маленькие школы.

Понимание того, что небольшие размеры выборки приводят к неточностям, важно по многим причинам. Как указывает Вайнер, неспособность ясно подумать о проблеме может привести к принятию неверных решений. Фонд Билла и Мелинды Гейтс потратил миллиарды долларов на совершенно неэффективную инициативу по созданию малых школ. Доказательством, которое побудило их сделать эти ошибочные инвестиции, стало наблюдение, что школы с неболь-

шим количеством учеников были широко представлены в списках школ с лучшими результатами тестов. Если бы они мыслили немного более критично, они бы также проверили списки школ с худшими результатами тестов и обнаружили, что в этих списках тоже избыточно представлены небольшие школы.

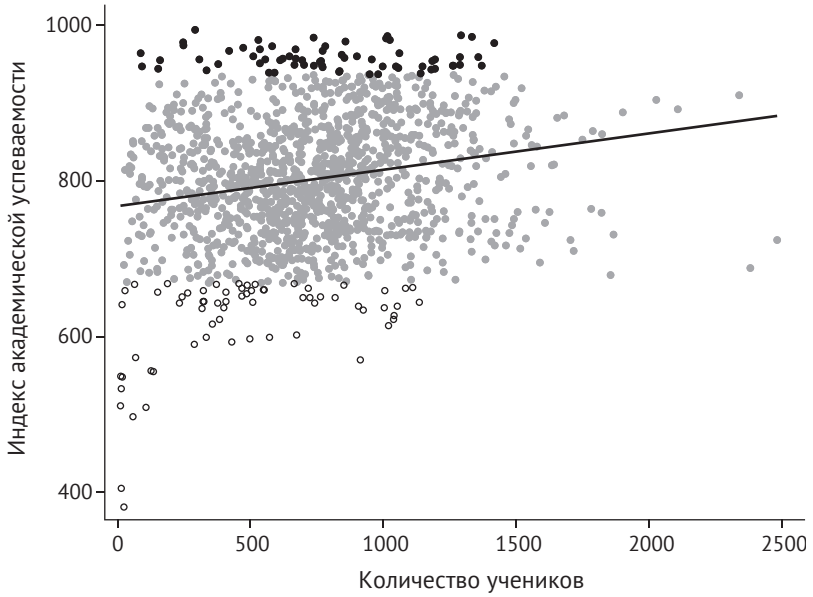


Рис. 6.2. Диаграмма рассеяния и линия регрессии, показывающая небольшую положительную корреляцию между средней успеваемостью и размером школы в средних школах Калифорнии в 2012 г.

Доверительные интервалы

Еще один способ количественной оценки точности – использование доверительного интервала.

Важный математический факт, называемый *законом больших чисел*, говорит нам, что, по мере того как размер нашей выборки становится действительно большим, шум практически исчезает. Но насколько велика должна быть «достаточно большая» выборка?

Другой важный математический факт, называемый *центральной предельной теоремой*, говорит нам, что если опрос действительно объективен, то, если мы будем проводить его повторно, примерно 95 % наших оценок ($\hat{q}$, республиканцы в нашей выборке) окажутся в пределах примерно двух стандартных ошибок оцениваемой величины (q, республиканцы в населении). Поэтому социологи часто говорят о так называемой *статистической погрешности*, которая просто равна удвоенной стандартной ошибке.

Исследователи и социологи также иногда говорят о *доверительном интервале* 95 %. Это интервал, который варьируется от оценки ($\hat{q}$) минус двукратная стандартная ошибка до оценки плюс двукратная стандартная ошибка.

Доверительный интервал 95 % является источником некоторой путаницы. Часто люди ошибочно говорят, что мы на 95 % уверены, что истинное значение

находится в пределах этого интервала. Но это не совсем так. Правильное утверждение намного более неуклюжее. С технической точки зрения мы можем сказать, что, если нет смещения и если мы повторим нашу оценку бесконечное количество раз, истинная оценка будет находиться в пределах 95 % доверительного интервала в 95 % случаев.

Чтобы получить представление о том, как работают доверительные интервалы, давайте вернемся к нашей аналогии с керлингом. Предположим, Энтони бросает бесконечное количество камней на бесконечном количестве ледяных дорожек. В качестве оценки будем рассматривать место на льду, где остановился точный центр его камня. Крайне маловероятно, что эта оценка будет идеально совпадать с баттоном – ваша оценка почти никогда не будет точно равна истинному значению оцениваемой величины. Но камень шире этого места. Поэтому мы можем спросить, как часто баттон будет касаться какой-либо части камня. Это будет зависеть от ширины камня. (Конечно, в керлинге существует строго определенная ширина, но позвольте нам немного поэтической вольности. Мы не шотландцы. И это не Олимпийские игры.) Мы могли бы найти точную ширину камня, чтобы баттон касался какой-либо части камня в 95 % бросков Энтони. Это похоже на 95-процентный доверительный интервал. Дело не в том, что при любом броске мы на 95 % уверены, что баттон закрыт какой-то частью камня, а в том, что в 95 % бросков баттон перекрывается какой-то частью камня.

Эта аналогия также помогает нам рассуждать о доверительных интервалах, отличных от 95 %. Иногда нам хочется уверенности больше, чем 95 %. Поэтому нас может заинтересовать доверительный интервал 99 %. Как вы думаете, шире или уже 99-процентный доверительный интервал, чем 95-процентный? Если Энтони хочет, чтобы 99 % его бросков заканчивались тем, что какая-то часть камня коснулась баттона, ему придется использовать более широкий камень. Следовательно, доверительный интервал 99 % шире, чем доверительный интервал 95 %. Мы должны допустить более широкий круг возможных республиканцев, если хотим быть уверены, что наша оценка находится в пределах этого диапазона в 99 % случаев, а не только в 95.

Статистический вывод и проверка гипотез

Теперь мы можем, наконец, обратиться к главному вопросу этой главы: как делать выводы о генеральных совокупностях, используя оценки на основе выборок? Давайте остановимся на нашем примере опроса. Как мы подчеркивали, проводя опрос, даже если считаем его объективным, мы также хотим, чтобы он был точным, потому что хотим получить оценку, близкую к истине. Как можно это оценить? Насколько хорошо выборка, скажем, из 1000 избирателей оценивает взгляды 140 млн американцев, которые определят, кто победит на предстоящих президентских выборах? Давайте посмотрим.

Проверка гипотез

Часто у нас возникает потребность оценить какую-то конкретную гипотезу. Например, нам нужно удостовериться, разумно ли полагать, что оцениваемая величина больше, меньше или не отличается от некоторой конкретной контрольной точки. Для этого нам нужно прибегнуть к проверке гипотезы.

В примере с предвыборным опросом мы пытаемся узнать, какой кандидат победит на выборах. Предположим, мы провели беспристрастный опрос тысячи избирателей, и это дало оценку доли голосов кандидата от республиканской партии в $\hat{q}$= 0.532, или 53.2 %. Насколько мы должны быть уверены в том, что республиканец действительно победит на выборах, т. е. насколько мы должны быть уверены в том, что более 50 % избирателей проголосуют за республиканца или, другими словами, что $q > 0.5$? Проверка гипотез дает нам возможность ответить на этот вопрос.

Один из вариантов рассуждений по этому поводу заключается в следующем. Результаты нашего опроса свидетельствуют о том, что кандидат от республиканской партии более популярен, чем демократ. Но этого мало. Нам нужно знать, насколько хороши эти свидетельства. То есть нам нужно знать, насколько вероятно, что мы могли бы наблюдать такие свидетельства, *даже если республиканец не более популярен, чем демократ*. Поэтому мы проверяем, насколько вероятно, что мы бы заметили наблюдаемые доказательства, если бы два кандидата были на самом деле одинаково популярны. Этот эталон отсутствия отношений обычно называют *нулевой гипотезой*.

Чтобы понять суть проверки гипотезы, начнем с предположения, что нулевая гипотеза верна, т. е. два кандидата одинаково популярны, поэтому $q = 0.5$. Теперь зададимся вопросом, насколько вероятно, что мы получим результат опроса, по крайней мере, столь же благоприятный для республиканца, как тот, который мы фактически нашли, ($\hat{q} = 0.532$).

У нас уже есть информация, необходимая для ответа на этот вопрос. При истинной доле голосов $q = 0.5$ и опросе одной тысячи избирателей стандартная ошибка нашего оценивателя составляет приблизительно 1.6 процентного пункта ($\sqrt{\frac{.5 \cdot .5}{1000}} \approx .016$). Наша оценка 0.532 соответствует превышению нулевой гипотезы на две стандартные ошибки ($0.5 + 2 \cdot 0.016 = 0.532$). (Мы выбрали эти числа не случайно.)

Как мы говорили ранее, центральная предельная теорема гласит, что 95 % оценок из беспристрастного опроса, который мы проводили, будут находиться в пределах двух стандартных ошибок от истинного значения, а это означает, что только 5 % оценок будут отклоняться более чем на две стандартные ошибки от истинного значения. Более того, в половине этих неудачных случаев оценка будет на две стандартные ошибки ниже истинного значения (что указывает на заметное лидерство демократа). Таким образом, если нулевая гипотеза верна, вероятность того, что мы получим результат опроса, столь же благоприятный для республиканца, как и тот, который мы получили, составляет около 2.5 %, или 1 из 40.

Мы выбрали числа, которые упрощают этот расчет, но ваш компьютер может выполнить его для любого результата опроса. Статистики занимаются разработкой методов проведения подобных расчетов. На языке статистики анализ, который мы только что провели, называется *односторонним z-тестом*. Вам не обязательно знать о z-тестах, чтобы понять остальную часть этой книги, но, если вы хотите узнать о них, можете обратиться практически к любой книге по статистике. («Википедия» также довольно надежна в отношении такого материала.) В более общем плане важно то, что проверка гипотез – это стратегия

оценки вероятности получения столь экстремального результата, как ваш, при условии, что нулевая гипотеза верна.

Статистическая значимость

Мы только что увидели, что, если нулевая гипотеза верна, вероятность того, что мы получили бы результат, столь же благоприятный для кандидата-республиканца, как тот, который мы нашли, составляет всего 0.025. Эта вероятность называется *p*-значением. Если наше значение *p* действительно низкое, мы можем заключить, что нулевая гипотеза вряд ли верна. Таким образом, мы располагаем статистически убедительными свидетельствами того, что избиратели действительно отдают предпочтение республиканцам, – если бы истинная доля голосов была разделена поровну, весьма маловероятно, что результат опроса был бы настолько благоприятным для республиканцев (и еще менее вероятно, что при столь благоприятной доле голосов в пользу республиканцев опрос отдавал бы предпочтение демократам).

Общий подход заключается в том, чтобы заранее указать определенный порог (чаще всего 0.05), и, если значение *p* ниже этого порога, мы говорим, что отвергаем нулевую гипотезу, и делаем вывод, что у нас есть *статистически значимые* доказательства в пользу гипотезы, которую мы проверяли.

Конечно, проверка гипотез не дает определенных выводов. При пороге значимости 0.05 существует 5-процентная вероятность получения статистически значимого результата, даже если нулевая гипотеза верна. Но проверка гипотез дает один из способов количественного рассуждения о том, может ли закономерность или результат, обнаруженный вами в наборе данных, отражать подлинное явление, а не просто быть следствием шума.

Очень часто люди ошибочно полагают, будто значение *p* равно вероятности того, что нулевая гипотеза верна. Это не так. Оно равно вероятности получения такой же экстремальной оценки, как и та, которую вы получили, если нулевая гипотеза верна. Эти две вероятности обычно различаются. Действительно, чтобы вычислить первую величину, вам нужно было бы иметь гораздо больше информации (например, насколько вероятно, что вы считали значение нулевой гипотезы истинным, прежде чем увидели доказательства). Мы обсудим эти вопросы в части 4.

Статистический вывод о взаимосвязях

До сих пор мы развивали наши идеи о смещении, шуме и проверке гипотез в простой ситуации, когда просто пытаемся узнать долю избирателей, поддерживающих кандидата от республиканской партии. Но все эти концепции и инструменты статистического вывода можно применить к гораздо более интересным задачам, включая оценку взаимосвязей, таких как корреляции.

Предположим, мы построили регрессию, чтобы оценить взаимосвязь между результирующей и объясняющей переменной. В предыдущей главе вы узнали, как можно использовать линейную регрессию для поиска коэффициентов, описывающих взаимосвязь между двумя переменными в наборе данных. Но теперь давайте поговорим об этом с точки зрения оценки и статистических выводов.

Предположим, что наш набор данных состоит из информации о доходах и образовании случайной выборки из тысячи работников, но на самом деле нас интересует среднее соотношение между доходом и образованием среди всех работников. Итак, мы пытаемся сделать выводы о корреляции в генеральной совокупности (наш предмет оценивания) на основе корреляции данных (наша оценка). Как нам это сделать?

Начнем со следующего уравнения, которое описывает взаимосвязь между доходом и образованием населения:

$$\text{Доход}_i = \alpha^{OLS} + \beta^{OLS} \cdot \text{Образование}_i + \text{ошибка}_i.$$

Это уравнение похоже на то, которое мы изучали в главе 5. Доход_i – это доход человека i, Образование_i – это продолжительность обучения человека i в годах, а ошибка_i – это разница между доходом человека i и доходом, предсказанным линией регрессии OLS для человека с соответствующим образованием. Параметры α^{OLS} и β^{OLS} принимают любые значения, минимизирующие сумму квадратов ошибок по совокупности. Например, β^{OLS} – это средняя степень увеличения дохода с каждым дополнительным годом обучения работника. Эти параметры α^{OLS} и β^{OLS} являются характеристиками мира. Мы их не знаем. Но поскольку нас интересует, как в среднем меняется доход в зависимости от образования, оцениваемой величиной является β^{OLS}.

Мы не знаем β^{OLS}, потому что не наблюдаем доходы и образование каждого отдельного человека в генеральной совокупности. Но мы можем оценить его, применив линейную регрессию к нашим данным о тысяче работников. Следуя соглашению обозначать оценки «шапочкой», назовем оценку этой регрессии $\hat{\beta}^{OLS}$. Это коэффициент регрессии, и он отражает корреляцию между образованием и доходом в нашей выборке. (Часто опускают верхний индекс OLS и просто используют обозначение $\hat{\beta}$, что вполне приемлемо, если понятно, о чем идет речь.)

К сожалению, β^{OLS} и $\hat{\beta}^{OLS}$ – это не одно и то же. Первое – это оцениваемая величина, а второе – оценка, которая, как мы знаем из нашего любимого уравнения, может отличаться от реальности как из-за систематической ошибки, так и из-за шума. Предположим, что наша выборка работников была извлечена из совокупности строго случайным образом, поэтому систематического смещения нет. (Подробнее о случайной выборке и несмещенности мы поговорим в главе 11). Но шум все равно остался. Если мы хотим знать, насколько близко $\hat{\beta}^{OLS}$ к истинному β^{OLS}, нужно вычислить стандартную ошибку.

Наша оценка связи между доходом и образованием $\hat{\beta}^{OLS}$ имеет стандартную ошибку аналогично тому, как ее имела ранее рассмотренная оценка доли республиканцев в населении q. Стандартная ошибка дает нам представление о том, насколько в среднем наша оценка будет далека от истины, если мы повторим нашу оценку бесконечное количество раз с независимыми выборками данных. Как и в случае с результатом опроса, существуют формулы для расчета этой стандартной ошибки. На данный момент вам не нужно беспокоиться

о формуле, поскольку компьютер рассчитает стандартную ошибку за вас. Более технический подход к стандартным ошибкам – это тема для другой книги.

Оценив стандартную ошибку, связанную с коэффициентом регрессии, вы можете поступить с ней так же, как и со стандартной ошибкой результата опроса. Например, можете построить доверительный интервал 95 %. Вы также можете проводить проверки гипотез и вычислять *p*-значения. Все это поможет оценить, насколько точна ваша оценка истинной взаимосвязи.

Один из распространенных вопросов, который интересует людей, заключается в том, существуют ли убедительные доказательства наличия истинной взаимосвязи. Предположим, что вы нашли положительное значение $\hat{\beta}^{OLS}$; доход и образование в вашей выборке положительно коррелируют. Есть ли у вас основания полагать, что они положительно коррелируют в более широкой популяции? Вы можете начать отвечать на этот вопрос, проверив нулевую гипотезу о том, что истинная связь между доходом и образованием на самом деле равна нулю. Для этого вы спрашиваете, какова вероятность получить такую большую оценку, как $\hat{\beta}^{OLS}$, если на самом деле корреляция между доходом и образованием населения отсутствует (т. е. $\beta^{OLS} = 0$). Если вы получите небольшое значение *p* и отклоните нулевую гипотезу, то у вас есть статистически значимые доказательства существования связи между доходом и образованием населения.

Одна из причин, по которой статистические выводы такого рода очень полезны, заключается в том, что время от времени мы находим в доступных данных взаимосвязи, которые не соответствуют истинным взаимосвязям в более широкой совокупности. Такова природа зашумленных данных. Поэтому всегда следует проверять, не являются ли наши выводы просто результатом шума.

Что, если у нас есть данные для всей совокупности?

Иногда у нас есть данные для всей интересующей совокупности. Например, предположим, что мы хотим узнать корреляцию между участием в университетских занятиях легкой атлетикой и средним баллом студентов Чикагского университета. Университет располагает соответствующими данными по каждому отдельному студенту, и в этом случае нам не нужно оценивать предполагаемую взаимосвязь по выборке. Мы могли бы точно измерить истинную корреляцию между легкой атлетикой и средним баллом для всей совокупности студентов Чикагского университета.

Но возникает сложный вопрос. Имеет ли смысл думать о стандартных ошибках, доверительных интервалах и статистической значимости, когда у нас есть данные для всей совокупности? Один из аргументов заключается в том, что эти инструменты не имеют значения, поскольку не было выборки. Так что шума нет. Оценка совпадает с оцениваемой величиной. Следовательно, нет необходимости заботиться о статистических выводах.

Но мы по-прежнему считаем, что есть веские причины обратить внимание на концепцию шума и связанные с ней меры неопределенности, даже если у нас есть данные для всей совокупности. Сейчас мы поясним почему.

Предположим, мы обнаружили небольшую положительную корреляцию между занятиями университетскими видами спорта и средним баллом. По-прежнему имеет смысл задаться вопросом, возникла ли эта разница по какой-то причине или просто по совпадению.

Что будет означать случайная корреляция? Предположим, что нет веских оснований ожидать заметную разницу между средним баллом спортсменов и неспортсменов: процедура поступления одинакова для обоих типов студентов, участие в спортивных соревнованиях не влияет на средний балл, успеваемость не влияет на участие в спортивных соревнованиях и т. д. Тем не менее между студентами существуют всевозможные несистематические различия, из-за которых их средние баллы отличаются друг от друга. И при любом конечном (тем более ограниченном) числе студентов обязательно будет хотя бы небольшая разница между успеваемостью спортсменов и неспортсменов, даже если для этой разницы нет веской причины.

Чтобы оценить, возникла ли наблюдаемая корреляция случайно или по какой-то причине, мы не можем собрать больше данных. У нас уже есть все данные о студентах Чикагского университета. Однако оказывается, что инструменты статистических выводов и проверки гипотез по-прежнему предоставляют полезный способ проверки, была ли наблюдаемая закономерность совпадением или нет.

Один из способов решить эту проблему состоит в том, что, хотя у нас есть данные обо всех реальных студентах университета, эти реальные студенты представляют собой лишь небольшую выборку гораздо более крупной гипотетической совокупности студентов, которые могли бы учиться в университете. Мы можем начать с нулевой гипотезы о том, что истинная корреляция в этой более крупной гипотетической популяции равна нулю, и задаться вопросом, насколько вероятно, что мы случайно наблюдали бы корреляцию аналогичную той, которую мы наблюдали среди реальных студентов. Конечно, для этого требуется метафизический переход от реальной популяции к гипотетической. Но немного заняться метафизикой – это цена, которую, вероятно, стоит заплатить, чтобы сохранить нашу способность анализировать, отражают ли некоторые наблюдаемые отношения подлинную, предсказуемую закономерность или были просто случайностью.

Содержательная и статистическая значимость

Статистическая проверка гипотез часто бывает полезной и информативной, поскольку нам нужно знать, не возникло ли наблюдаемое явление чисто случайно (например, из-за вариаций выборки). Однако *статистическая* значимость (низкое значение p, указывающее на то, что результат вряд ли возник случайно) – это не то же самое, что *содержательная* значимость, и мы должны быть осторожны, чтобы не смешивать эти два понятия. Часто нас интересует не доказательство существования какого-то явления (вопрос статистической значимости), а то, насколько велико и важно это явление, и это уже вопрос содержательной значимости. Например, руководители корпорации Coca-Cola наверняка знают, что их реклама оказывает положительное влияние на продажи. Но это знание ничего не говорит о том, сколько тратить на рекламу. Для

этого нужно знать, насколько велико влияние рекламы на продажи. Позвольте нам привести примеры, иллюстрирующие два сценария, когда количественные аналитики могут сбиться с пути, делая упор на статистическую, а не на содержательную значимость.

Социальные сети и голосование

В 2012 г. шесть исследователей опубликовали в журнале Nature исследование, показавшее, что люди с большей вероятностью проголосуют на промежуточных выборах в США в 2010 г., если на их страницах в Facebook будет отображаться баннер с указанием, кто из их друзей проголосовал. Исследование было примечательным по нескольким причинам.

Facebook позволил исследователям рандомизировать опыт 61 млн пользователей соцсети избирательного возраста в США в день выборов. И действительно, экспериментальное воздействие, судя по всему, привело к увеличению явки – предполагаемый эффект от знания того факта, что близкий друг уже проголосовал, имеет высокую статистическую значимость ($p = 0.02$). Исследователи пришли к выводу, что «сильные связи играют важную роль в распространении как онлайн-, так и реального поведения в социальных сетях людей». Исследование получило широкое освещение в прессе, поскольку продемонстрировало, насколько сильно социальное окружение влияет на голосование.

Большинство наблюдателей не заметили, что предполагаемое влияние баннеров Facebook на явку избирателей составило менее 0.4 процентных пункта. Это очень маленький эффект, возможно, не имеющий большого значения для предвыборной кампании или понимания выборов. Тот факт, что 0.4 % избирателей, имеющих право голоса, можно убедить проголосовать с помощью баннера в Facebook, сообщающего о голосовании друзей, не говорит нам о том, что сильные социальные связи играют важную роль в распространении поведения. Конечно, при размере выборки в 61 млн практически любая ненулевая оценка будет статистически значимой. Это неплохо. Большие размеры выборки означают, что наши оценки достаточно точны, поэтому мы сможем более надежно обнаружить подлинные связи. Однако мы не можем слепо предполагать, что любой статистически значимый результат также является *существенно* значимым.

Вы убедились, что статистически значимые результаты могут быть существенно незначительными. Теперь рассмотрим пример, когда верно обратное.

Второй закон о реформе

В статье 2011 г. в Quarterly Journal of Political Science Сэмюэл Берлински и Торун Деван оценивают влияние Второго закона о реформе избирательного права 1867 г. на выборы в Соединенном Королевстве. Несмотря на то что Второй закон о реформе примерно удвоил численность имеющих право голоса избирателей и впервые привел на избирательные участки избирателей из рабочего класса, авторы сообщают, что он мало повлиял на результаты выборов: «Нет никаких доказательств того, что либералы [одна из крупнейших британских партий] получили электоральную поддержку благодаря изменениям в электоральном праве».

Но правильная ли это интерпретация? Когда авторы говорят об отсутствии доказательств, они имеют в виду, что их оценки влияния закона о реформе не являются статистически значимыми. Поэтому формально у них нет основа-

ний утверждать, что наблюдаемые результаты возникли не случайно. Однако данные исследования, хотя и не являются статистически значимыми, на самом деле свидетельствуют о том, что Второй закон о реформе имел важные последствия. Численные оценки показывают, что удвоение электората в результате закона о реформе увеличило долю голосов либеральной партии на 8 процентных пунктов, что является существенно значимым эффектом, подразумевающим, что новые избиратели из рабочего класса, получившие избирательные права в результате реформы, с гораздо большей вероятностью поддержали либеральную партию, чем более богатые избиратели, имевшие избирательные права и раньше. Однако, хотя эта оценка существенно значима, она также неточна и, следовательно, не является статистически значимой. Сосредоточив внимание на этой статистической незначительности, Берлински и Деван приходят к выводу, что Второй закон о реформе не возымел большого эффекта. Но факты на самом деле заставляют нас предположить, что он оказал большое влияние на политическую обстановку. Вот только само это предположение является статистически ненадежным.

Хотя статистическая значимость полезна и информативна, ее часто неправильно используют и неправильно понимают. На протяжении всей этой книги мы пытаемся доказать, что критическое мышление и данные дополняют, а не заменяют друг друга. Тот факт, что мы занимаемся статистикой, не означает, что мы должны перестать предметно думать о вопросах, на которые стремимся ответить. Мы должны использовать статистические выводы, когда это возможно. Но также нужно всегда напоминать себе о необходимости делать существенные выводы на основе имеющихся фактов.

Подведение итогов

Оценки могут отличаться от истинного значения оцениваемой величины по двум причинам: из-за систематической ошибки (смещение) и шума. Смещению посвящена глава 9. В этой главе мы сосредоточились на шуме – различиях между оценкой и оцениваемой величиной, возникающих из-за несистематических особенностей нашей выборки. Поскольку шум уникален, его среднее значение будет равно нулю, если мы будем повторять нашу процедуру оценки снова и снова бесконечное количество раз, всегда на независимой выборке данных. Но в любой конкретно выборке влияние шума может быть весьма значительным.

Наличие шума означает, что у нас всегда остаются определенные сомнения в том, отражает ли взаимосвязь в выборке данных (оценка) реальную взаимосвязь в более крупной интересующей совокупности (оцениваемая величина). Мы обсудили методы количественной оценки этой неопределенности и проверки гипотезы о том, что предполагаемая взаимосвязь реальна, в сравнении с нулевой гипотезой о том, что предполагаемая взаимосвязь была результатом наличия шума.

Наличие шума создает проблемы, выходящие за рамки неопределенности. Например, в главе 7 мы рассмотрим проблему, заключающуюся в том, что, если одно и то же исследование проводится снова и снова, некоторые итерации дадут статистически значимые результаты из-за шума, даже если исследуемой взаимосвязи не существует. Если рассматривать только эти статистически зна-

чимые результаты, то научная деятельность может привести к систематически неверным выводам. В главе 8 мы рассмотрим, как присутствие шума создает загадочный феномен возврата к среднему значению (за экстремальными наблюдениями обычно следуют менее экстремальные), – ситуацию, когда отсутствие критического мышления может привести к различным видам ошибочной интерпретации числовых данных.

Ключевые термины

- **Популяция (генеральная совокупность):** объекты мира, о которых мы пытаемся получить знания.
- **Выборка:** часть популяции, для которой у нас есть данные.
- **Оцениваемая величина:** не наблюдаемая напрямую величина, которую мы пытаемся узнать (оценить) с помощью анализа данных.
- **Оцениватель:** процедура, которую мы применяем к данным для получения числового результата оценки.
- **Оценка:** числовой результат, полученный в результате применения оценивателя к определенному набору данных.
- **Смещение (систематическая ошибка):** различия между нашей оценкой и оцениваемой величиной, возникающие по систематическим причинам, т. е. по причинам, которые в среднем сохраняются на многих различных выборках данных.
- **Шум:** различия между нашей оценкой и истинным значением оцениваемой величины, возникающие из-за *несистематических* факторов, относящихся к текущей выборке.
- **Несмещенность:** оценка/оцениватель не имеет смещения, если при повторении нашей процедуры оценивания бесконечное число раз среднее значение наших оценок будет равняться истинному значению.
- **Ожидание или ожидаемое значение:** среднее значение бесконечного числа выборок переменной называется ожидаемым значением переменной или ее ожиданием.
- **Точность:** оценка/оцениватель является точным, если при многократном повторении процедуры оценивания различные оценки будут близки друг к другу. Чем более похожи гипотетические оценки, полученные в результате многократного запуска оценивателя, тем точнее оценка.
- **Выборочное распределение:** распределение оценок, которое мы получили бы, если бы повторили оценивание бесконечное число раз, каждый раз с новой выборкой данных.
- **Стандартная ошибка:** стандартное отклонение выборочного распределения. Если оцениватель не смещен, стандартная ошибка дает нам представление о том, насколько в среднем наша оценка будет далека от истины, если мы многократно повторим нашу процедуру с независимыми выборками данных.
- **Погрешность:** социологи часто умножают стандартную ошибку на 2 и называют это погрешностью.
- **95-процентный доверительный интервал:** если бы мы применяли оцениватель бесконечное количество раз всегда к новой выборке дан-

ных, оценочное значение входило бы в 95-процентный доверительный интервал (каждый раз вычисляемый заново) в 95 % случаев. Неправильно говорить, что мы на 95 % уверены в том, что истинная оценка находится в 95-процентном доверительном интервале.

- **Проверка гипотезы:** статистические методы оценки того, насколько мы можем быть уверены в том, что некоторые особенности данных отражают реальные свойства мира, а не являются результатом шума.
- **Нулевая гипотеза:** гипотеза о том, что некоторые наблюдаемые свойства данных или явления полностью являются следствием шума.
- **Статистическая значимость:** мы говорим, что у нас есть статистически значимые доказательства для некоторой гипотезы, когда мы можем отвергнуть нулевую гипотезу с некоторым заранее заданным уровнем уверенности (обычно 95 %).
- ***p*-значение:** вероятность обнаружения взаимосвязи, столь же сильной или более сильной, чем связь, обнаруженная в данных, если нулевая гипотеза верна. Мы используем *p*-значения для оценки статистической значимости. Например, если *p*-значение меньше 0.05, то у нас есть статистически значимые доказательства (при уровне достоверности 95 %) того, что взаимосвязь реальна. Важно отметить, что *p*-значение не равно вероятности того, что нулевая гипотеза верна.

Упражнения

6.1. Рассмотрите следующие стратегии проведения политического опроса для предсказания доли голосов на предстоящих выборах. Обсудите вероятную степень смещения и точности для каждого из них.

a) Fox News просит своих зрителей позвонить и сообщить, кого они поддерживают на выборах. Они получают более 100 000 звонков.

b) Nailbiter Polling (новая фирма на рынке консалтинга) проводит опросы, а затем, независимо от ответов, всегда сообщает о ничейном результате: 50 % в пользу кандидата A и 50 % в пользу кандидата B.

c) Surpriseing News Polls (еще один новый игрок) проводит крупные репрезентативные опросы, вычисляет среднюю поддержку каждого кандидата, а затем подбрасывает монетку. Если монета выпала орлом, они добавляют 10 % к поддержке кандидата A, а если решкой, то вычитают 10 % из поддержки кандидата A.

d) Опросный центр Средней Америки изготавливает распечатку списка зарегистрированных избирателей, переходит на среднюю страницу списка и берет интервью по телефону у десяти человек в середине этой средней страницы.

6.2. Отец Энтони, Пит, недавно купил колесо рулетки, чтобы устроить подпольное казино в своем гараже. Если вы не знакомы с правилами игры в рулетку: колесо вращается, и шарик случайным образом попадает в одну из 38 ячеек на колесе, каждая из которых имеет номер и цвет. На этом колесе 18 красных, 18 черных и две зеленые ячейки. Игрок может сделать ставку на красное, и в этом случае он удвоит свои деньги, если шарик попадет в красную ячейку, но в противном случае потеряет свои деньги.

Если колесо действительно работает честно, а это означает, что шарик с одинаковой вероятностью попадет в любую ячейку, Пит рассчитывает заработать на этих ставках, поскольку игрок выигрывает 18 раз из 38, а Пит выигрывает остальные 20 раз из 38. Конечно, если колесо окажется нечестным в пользу игрока, Пит быстро разорится. Чтобы проверить колесо, Пит провел три пробных вращения, и, к его большому разочарованию, шарик все три раза попадал в красную ячейку. Учитывая доступную нам на данный момент информацию, что мы можем сказать со статистической точки зрения о смещенности исходов рулетки в пользу красных ячеек?

a) Какова нулевая гипотеза?
b) Каково *p*-значение?
c) Дайте содержательную интерпретацию *p*-значения и, что важно, объясните, чем оно не является.
d) Не обращая внимания на законность гаражного казино, какой дополнительный совет вы бы дали Питу, чтобы помочь выяснить, честна ли его рулетка?

6.3 Вернемся к анализу зависимости доходов от образования из упражнений предыдущей главы. Когда вы строите регрессию заработка от обучения, ваш компьютер, вероятно, выдает вам не только приблизительные коэффициенты, но и некоторые другие числа, которые вы не понимали, пока не прочитали эту главу. Для коэффициента, связанного с количеством лет обучения, вы должны были получить оценку 1.16, т. е. каждый дополнительный год обучения соответствует увеличению заработка примерно на 1160 долл. Каковы предполагаемая стандартная ошибка, *p*-значение и 95-процентный доверительный интервал, связанные с этим коэффициентом? Дайте содержательную интерпретацию каждому из этих чисел.

Дополнительное чтение и ссылки

Если вас интересует удивительно интересная история о том, как ученые и статистики пришли к общепринятому 5-процентному порогу статистической значимости, прочтите статью:

Michael Cowles and Caroline Davis. 1982. *On the Origins of the .05 Level of Statistical Significance.* American Psychologist 37 (5): 553–58.

Если вы вообще интересуетесь историей вероятностей и статистикой (а это вам должно быть интересно), прочтите книги:

Ian Hacking. 2006. T*he Emergence of Probability: A Philosophical Study of Early Ideas about Probability and Statistical Inference*, 2nd Edition. Cambridge University Press;

Ian Hacking. 1990. *The Taming of Chance.* Cambridge University Press.

Для знакомства с увлекательной историей проверки статистических гипотез и проблем, которые возникают, когда ученые смешивают статистическую и содержательную значимость, мы рекомендуем книгу:

Stephen T. Ziliak and Deirdre N.McCloskey. 2008. *The Cult of Statistical Significance: How the Standard Errors Cost Us Jobs, Justice, and Lives.* University ofMichigan Press.

Наше обсуждение выборок небольшого размера, маленьких школ и Фонда Гейтса опиралось на материалы из работы:

Howard Wainer. 2009. *Picturing the Uncertain World: How to Understand, Communicate, and Control Uncertainty through Graphical Display.* Princeton University Press.

Данные об успеваемости и зачислении в школы Калифорнии взяты с сайта *https://www.cde.ca.gov/re/pr/reclayout12b.asp*.

Исследование о Facebook и явке избирателей, на которое мы ссылаемся:

Robert M. Bond, Christopher J. Fariss, Jason J. Jones, Adam D. I. Kramer, Cameron Marlow, Jaime E. Settle, and JamesH. Fowler. 2012. *A 61-Million-Person Experiment in Social Influence and Political Mobilization.* Nature 489 (7415): 295–98.

Статья с анализом Второго закона о реформе избирательного права:

Samuel Berlinski and Torun Dewan. 2011. *The Political Consequences of Franchise Extension: Evidence fromthe Second ReformAct.* Quarterly Journal of Political Science 6 (34): 329–76.

Глава 7

Завышение значимости и занижение отчетности

О чем эта глава

- Если аналитики проводят множество сравнений, но принимают во внимание только статистически значимые результаты, у них будет много ложноположительных результатов и завышенных оценок.
- Эти ложноположительные результаты могут быть результатом недобросовестного поведения исследователя (*p*-хакинг). Но они могут возникнуть и в сообществе вполне честных исследователей (*p*-скрининг).
- Простого решения не существует, но у аналитиков и потребителей есть инструменты, позволяющие если не исключить, то хотя бы снизить риск впасть в заблуждение.

Введение

Хотя статистическая проверка гипотез – полезный инструмент, он далеко не так надежен, как кажется. Чтобы понять, почему научные исследования и количественный анализ данных так часто дают обманчивые или ненадежные результаты, мы начнем с неожиданного примера – истории, казалось бы, удивительного морского существа.

Может ли осьминог быть футбольным экспертом?

В 2008 и 2010 гг. осьминог Пауль попал в заголовки газет благодаря своему очевидному мастерству в предсказании исходов футбольных матчей. Перед матчем между двумя национальными сборными владельцы Пауля давали ему два лотка с едой, каждый из которых был отмечен флагом одной из стран-участников. Пауль выбирал один из лотков, а владельцы смотрели, флаг какой страны прикреплен к лотку. Считалось, что сборной этой страны предсказана победа в предстоящем матче. Пауль был удивительно точен, и журналисты и игроки с нетерпением ждали его предсказаний.

Пауль был предметом большого восхищения и даже негодования. По словам Ника Коллинза из The Telegraph, аргентинский шеф-повар был так разгневан после того, как осьминог правильно предсказал победу Германии над Аргентиной, что «пригрозил в отместку приготовить из Пауля блюдо». Игроки делали

ставку на точность предсказаний Пауля еще до того, как он их сделал. Коллинз сообщил, что «букмекерская контора Уильяма Хилла приняла так много ставок на то, что Пауль правильно предскажет результат финальной игры между Испанией и Голландией, что ей пришлось снизить коэффициенты с равных до 10/11».

Скептик мог бы возразить, что, хотя осьминоги впечатляюще умны, Пауль никак не мог обладать особым знанием исхода футбольных матчей. Даже экспертам трудно предсказать результаты игр, в которых кто-то кому-то забивает гол. И Пауль, по-видимому, ничего не знал ни об играющих командах, ни даже о футболе в целом. Был ли успех Пауля всего лишь удачей?

Как было сказано в главе 6, у нас есть инструменты для оценки того, является ли наблюдаемая закономерность результатом банального везения или, говоря более научным языком, влияния шума. Мы можем проверить гипотезу и рассчитать *p*-значение.

Как Пауль справится с такой проверкой гипотезы? За свою карьеру осьминог сделал 14 прогнозов и оказался прав в 12 из них. Это очень хороший результат. Предположим, что нулевая гипотеза о том, что это было слепое везение, верна, т. е. Пауль выбирал команды совершенно случайным образом, при этом вероятность выбора каждого лотка была равной. Чтобы выяснить, мог ли рекорд Пауля стать результатом исключительного везения, нужно вычислить вероятность того, что он угадал бы правильно минимум 12 раз, если бы действовал строго наугад.

Эта задача настолько проста, что вы можете вычислить *p*-значение вручную. Основная идея такова. Предположим, что Пауль угадывает наобум. Подсчитайте, насколько вероятно, что он наберет ровно 12 правильных ответов, насколько вероятно, что он наберет ровно 13 правильных ответов, и насколько вероятно, что он наберет ровно 14 правильных ответов. Сумма этих трех чисел дает нам вероятность того, что Пауль продемонстрировал свой выдающийся результат благодаря чистому везению.

Прежде чем продолжить обсуждение осьминога, давайте научимся вычислять вышеупомянутые вероятности. Мы должны очень хорошо понимать, что такое «чистое везение».

Для начала будет полезно упростить задачу. Наша нулевая гипотеза состоит в том, что осьминог Пауль дает прогнозы наугад, т. е. предсказание Пауля аналогично тому, как человек подбрасывает монету и она выпадает орлом. Итак, давайте порассуждаем о подбрасывании монеты. Предположим, вы подбрасываете монету три раза (чуть позже мы доберемся до 14, как у Пауля). В табл. 7.1 показаны все возможные исходы (О – орел, Р – решка).

Таблица 7.1. Возможные исходы трехкратного подбрасывания монеты

	Три орла	Два орла	Один орел	Ноль орлов
		ООР	ОРР	РРР
Исход	ООО	ОРО	РОР	
		РОО	РРО	

Аналогично если Пауль спрогнозировал три игры, то ноль, один, два или три его прогноза могут оказаться верными.

Какова вероятность того, что у вас выпадет, скажем, ровно два орла? Что ж, всего существует восемь исходов, и, если мы подбрасываем честную монету, все они одинаково вероятны. Из этих восьми три предполагают выпадение двух орлов. Таким образом, вероятность того, что при трех подбрасываниях монеты выпадет ровно два орла, равна 3/8. Аналогично, если бы Пауль прогнозировал три игры наугад, вероятность того, что он угадает ровно два раза, равна 3/8.

Но это не совсем та величина, которую мы хотим знать. Мы хотим знать вероятность того, что вам выпадет *как минимум* два орла или что Пауль правильно предскажет *как минимум* две игры.

Помимо выпадения двух орлов, у вас также может выпасть три орла. Это может произойти только одним способом, поэтому вероятность того, что выпадет три орла, равна 1/8. Следовательно, вероятность того, что у вас выпадет как минимум два орла, равна 3/8 + 1/8 = 1/2. Аналогично, если бы Пауль наобум предсказал три игры и просто угадал, вероятность того, что он случайно угадает как минимум два раза, равна 1/2.

Но Пауль не просто предсказал три игры. Он предсказал 14. Составлять таблицу для 14 подбрасываний монеты было бы довольно скучно. Поэтому нужно придумать, как проанализировать эту задачу в более общем виде.

Предположим, вы подбросили монету n раз. Насколько вероятно, что вам выпадет ровно k орлов? Давайте начнем с расчета вероятности того, что каждая из первых k монет выпадет орлом, остальные – решкой. Вероятность того, что первые k выпадут орлом, равна $\frac{1}{2}^k$. Вероятность того, что остальные выпадут решкой, равна $\frac{1}{2}^{n-k}$. Таким образом, вероятность того, что первые k выпадут орлом, а остальные $n - k$ выпадут решкой, равна $\frac{1}{2}^k \times \frac{1}{2}^{n-k}$.

Конечно, это только один из способов получить ровно k орлов. k орлов не обязательно должны выпасть за первые k подбрасываний. Это может быть любая комбинация исходов, содержащая k орлов из n подбрасываний. Существуют $\frac{n!}{k!(n-k)!}$ различных способов получить ровно k орлов при подбрасывании монеты n раз. Таким образом, общая вероятность получить ровно k орлов при подбрасывании монеты n раз равна

$$\frac{1}{2}^k \times \frac{1}{2}^{n-k} \times \frac{n!}{k!(n-k)!}.$$

Восклицательные знаки выше означают *факториал*. Выражение $n!$ называется *факториалом* n, и оно определяется как произведение n и каждого положительного целого числа, меньшего n. Так, например, 3! = 3×2×1 = 6.

Давайте посмотрим, подтверждает ли эта формула наши выводы, сделанные ранее в нашем примере с подбрасыванием трех монет. Если мы подбросим монету три раза, какова вероятность того, что у нас выпадет ровно два орла? Поскольку $n = 3$ и $k = 2$, мы вычисляем вероятность следующим образом:

$$\frac{1}{2}^{2} \times \frac{1}{2}^{3-2} \times \frac{3!}{2!(3-2)!} = \frac{1}{4} \times \frac{1}{2} \times \frac{3 \times 2 \times 1}{2 \times 1 \times 1} = \frac{3}{8}.$$

Теперь мы можем вычислить вероятность того, что осьминог Пауль правильно предскажет 12 или более игр из 14, если он будет выбирать наугад. Вероятность того, что он наберет ровно 12 случайных угадываний, равна

$$\frac{1}{2}^{12} \times \frac{1}{2}^{14-12} \times \frac{14!}{12!(14-12)!} \approx .00555.$$

Вероятность того, что он наберет ровно 13 угадываний, равна

$$\frac{1}{2}^{13} \times \frac{1}{2}^{14-13} \times \frac{14!}{13!(14-13)!} \approx .00085.$$

Вероятность того, что он угадает все 14 ответов, равна

$$\frac{1}{2}^{14} \times \frac{1}{2}^{14-14} \times \frac{14!}{14!(14-14)!} \approx .00006.$$

Таким образом, вероятность того, что Пауль правильно назовет 12 или более игр, составляет примерно 0.00555 + 0.00085 + 0.00006 ≈ 0.0065, примерно 1 из 155. Другими словами, если бы Пауль не был знатоком футбола, крайне маловероятно, что он давал бы настолько точные прогнозы чисто случайно. Конечно, именно поэтому все были одержимы Паулем. И, похоже, они были правы. Используя стандартный подход к проверке статистических гипотез, который мы представили в главе 6, можно отвергнуть нулевую гипотезу, согласно которой Пауль просто действует наугад, и прийти к выводу о наличии статистически значимых доказательств того, что Пауль действительно является опытным футбольным предсказателем.

Приведенный выше анализ очень похож на то, что сделали два математика, Крис Бадд и Дэвид Шпигельхартер, когда давали интервью о Пауле еще в 2010 г. Но если мы присмотримся к прогнозам осьминога более внимательно, то увидим, что выводы математиков слишком щедры по отношению к экстрасенсорным способностям Пауля.

Пауль жил в Германии, и его в первую очередь просили предсказать исход игр, в которых участвовала Германия. Фактически в 13 из 14 игр участвовала Германия. Более того, у Пауля была сильная склонность выбирать Германию. Возможно, ему понравился этот флаг, потому что он видел его много лет на стене бара. Возможно, по неизвестным нам причинам лоток с немецким флагом был его любимым лотком. Возможно, владельцы Пауля неосознанно научили его выбирать Германию. Кто знает? Так уж совпало, что Германия действительно хороша в футболе – она побеждает в большей части матчей. Так что, возможно, успех Пауля не столь удивителен. Давайте повторим приведенный выше анализ, учитывая эту информацию.

Пауль предсказал исход 13 игр с участием Германии и выбрал Германию победителем 11 из этих игр. Фактически Германия выиграла 9 из них. Наша нулевая гипотеза снова состоит в том, что предсказания осьминога были просто

слепой удачей в том смысле, что у него не было особых знаний о футболе. Но на этот раз вместо того, чтобы воображать, что он с равной вероятностью выберет любую коробку, представьте, что он предрасположен выбрать немецкую коробку. Допустим, эта предрасположенность означала, что его вероятность выбрать Германию была равна 11/13 в любой игре, в которую играла Германия, поскольку именно с такой частотой Пауль фактически выбирал Германию. Если Германия выиграла 9 из 13 игр, а Пауль каждый раз выбирал Германию с вероятностью 11/13, насколько вероятно, что он оказался прав 11 или более раз по чистой случайности? Это *p*-значение можно вычислить вручную, но сделать это сложно. Поэтому мы запустили простое моделирование на нашем компьютере, чтобы найти приближенное значение. С учетом этих уточненных предположений вероятность того, что Пауль угадает 11 или более игр подряд из 13, составляет примерно 0.03 или 1 из 33 – все еще маловероятно, но гораздо более вероятно, чем 1 из 155.

Что мы думаем? По-прежнему маловероятно, что успех Пауля можно объяснить исключительно глупой удачей. Даже если бы он был склонен предсказывать победу только сильной немецкой команды, вероятность того, что Пауль добьется такого же успеха, составляла всего 3 %. Таким образом, традиционная проверка гипотезы с порогом 0.05 все равно приведет к отказу от нулевой гипотезы. У нас по-прежнему есть статистически значимые доказательства того, что Пауль хорошо предсказывает футбольные матчи.

Вы не удивитесь, узнав, что мы по-прежнему настроены скептически. Но почему? Пауль не единственный осьминог. Что, если на самом деле по Германии разбросано десять осьминогов, каждый из которых пытается предсказать исход футбольных матчей? Мир, конечно, узнает только о самом успешном из них. Если это так, то мы до сих пор не проверили правильную гипотезу, чтобы выяснить, насколько вероятно, что точность Пауля была обусловлена банальной удачей. Если бы действительно существовало десять осьминогов, пытающихся предсказать футбольные матчи, и если бы Пауль просто оказался тем, кто преуспел и поэтому стал знаменитым, то, вместо того чтобы спрашивать, насколько вероятно, что прогнозы Пауля окажутся точным по счастливой случайности, мы должны спросить, насколько вероятно, что любой из десяти осьминогов продемонстрирует такую точность по счастливой случайности. Потому что, если бы оказалось, что осьминожка Паулина была права в 12 из 14 случаев, а Пауль – нет, то мы бы говорили о Паулине и никогда бы не услышали о Пауле.

Выяснить, насколько вероятно, что какой-то осьминог из десяти окажется столь же точным, как Пауль, относительно несложно. Но чтобы выполнить расчет, нам нужно понять еще один факт о *p*-значениях. Напомним, что *p*-значение – это вероятность наблюдения результата, по крайней мере столь же экстремального, как тот, который вы наблюдаете, если нулевая гипотеза верна. Итак, если нулевая гипотеза верна, как часто вы будете наблюдать столь же экстремальный результат, как результат с p-значением 0.05? Ровно в 5 % случаев. И если нулевая гипотеза верна, как часто вы будете наблюдать столь же экстремальный результат, как с *p*-значением 0.2? Ровно 20 % случаев. И т. д. для каждого *p*. Это всего лишь повторение определения *p*-значения.

Но из этого факта мы узнаем нечто важное. Когда нулевая гипотеза верна, мы должны наблюдать значение *p* меньше или равное 0.05 в 5 % случаев, зна-

чения *p* меньше или равные 0.2 в 20 % случаев, значения *p* меньше или равные 0.5 в 50 % случаев и т. д. Следовательно, должно быть так, что, когда нулевая гипотеза верна, вы с одинаковой вероятностью найдете каждое *p*-значение. (На техническом языке это означает, что значения *p* равномерно распределены в условиях нулевой гипотезы.)

Итак, какова вероятность того, что хотя бы один из наших немецких осьминогов по простой случайности сгенерирует рекорд предсказания со значением *p*, по крайней мере таким же хорошим, как рекорд Пауля 0.03? Мы только что увидели, что вероятность получить значение $p = 0.03$ или ниже только по чистой случайности, равна 0.03. Следовательно, вероятность того, что какой-либо осьминог даст значение *p* выше 0.03, равна 0.97. Если есть два осьминога и они делают свои предположения независимо, вероятность того, что ни один из них не даст значение *p* лучше, чем 0.03, равна 0.97^2. Таким образом, вероятность того, что хотя бы один из них генерирует значение $p = 0.03$ или выше, равна $1 - 0.97^2$ (т. е. единица минус вероятность того, что оба генерируют значение *p* хуже, чем 0.03). Если десять осьминогов делают случайные предположения, вероятность того, что хотя бы один из них выдаст такое же хорошее значение *p*, как у Пауля, равна $1 - 0.97^{10} \approx 0.26$. Другими словами, если бы десять немецких осьминогов прошли через ту же последовательность предсказаний, что и Пауль, вероятность того, что хотя бы один из них достиг бы рекордной точности предсказаний, по крайней мере такой же великолепной, как у Пауля, составляет примерно 1 из 4, даже если ни один из них не является футбольным экспертом. Это должно заставить нас гораздо более скептически относиться к способностям Пауля.

Мы не знаем, сколько немецких осьминогов было занято в бизнесе футбольных прогнозов. Но мы знаем, что в нем участвовало множество других животных. Шутка ли, дикобраз Леон, бегемот Петти, тамарин Антон и попугай Мани предсказывали победителей футбольных матчей примерно в то же время, что и Пауль. И это только те, кто попал в новости. Вероятно, были еще десятки животных, о которых мы никогда не слышали. И это обсуждение касается только футбола. А как насчет всех остальных видов спорта? А как насчет всех неспортивных результатов, которые можно предсказать? Если бы барсучиха Джуди умела предсказывать победителей студенческих футбольных матчей, кот Стив умел предсказывать победителей выборов в конгресс, а выдра Фрэн – изменения на фондовом рынке, они тоже были бы знаменитостями. Но их предсказания были не более чем случайностью, поэтому мы о них так и не услышали.

На это сразу же указали математики Бадд и Шпигельхартер. Шпигельхартер отмечает, что «если кто-то подбрасывает монету и получает один и тот же результат 9 или 10 раз, это само по себе не является чем-то примечательным, но человеку, подбрасывающему монету, это покажется таковым». Другими словами, если достаточное количество людей подбрасывает монеты, один из них обязательно выкинет несколько орлов подряд. И, несмотря на то что у кого-то случайно *должна была* выпасть удачная серия решек, этот человек мог ошибочно заключить, что у него нечестная монета или что он особенно умело подбрасывает монету. К сожалению, как мы увидим, эта проблема применима к гораздо более серьезным ситуациям, чем подбрасывание монеты и прогнозирование футбола, и имеет далеко идущие последствия.

Предвзятость публикации

Статистическая проверка гипотез и *p*-значения явно полезны. Когда мы обнаруживаем закономерности в данных, мы хотим знать, отражают ли они реальные явления или же они могли легко возникнуть в результате случайности.

Но есть проблема, которую подчеркивает история осьминога Пауля. Ни общественность, ни более широкое научное сообщество не видят всех проверок гипотез, которые были (или могли быть) проведены. Обычно принято обнародовать только статистически значимые результаты. Просто не так уж интересно писать об осьминожке Мэри, которая предсказывает футбольные матчи примерно так же хорошо, как подбрасывание монеты. Но если существует 20 разных животных, делающих футбольные прогнозы, мы ожидаем, что у одного из них значение *p* будет меньше 0.05 по чистой случайности, даже если никто из них не разбирается в футболе. Проблема в том, что в новостях напишут только об одном случае. Поэтому, если мы будем основывать свои убеждения исключительно на том, что нам сообщают, у нас будут систематически ошибочные убеждения.

Выбор наилучших результатов из большого количества испытаний, т. е. *завышение значимости* (over-comparing) и избирательное сообщение только об интересных или статистически значимых случаях, т. е. *занижение отчетности* (under-reporting), – опасная комбинация, которая встречается повсеместно. Поэтому, когда мы слышим о новом, захватывающем научном результате, нам трудно понять, насколько достоверно он отражает подлинное явление.

Эта проблема завышения значимости и занижения отчетности влияет не только на уверенность в том, что конкретное открытие является подлинным. Она также влияет на нашу способность накапливать знания в определенной области в течение продолжительного времени. Мы знаем, что любая оценка, даже несмещенная, может быть далека от истинной оценки из-за шума. Есть надежда, что по мере накопления оценок шум усредняется, так что среднее значение большого количества несмещенных оценок становится очень близким к истинной оценке. Завышение значимости и занижение отчетности означает, что этого может не случиться при анализе опубликованных оценок – тревожное явление, называемое *предвзятостью публикации* (publication bias). Чтобы понять, почему оно возникает, давайте вернемся к нашему любимому уравнению:

$$\text{Оценка} = \text{Оцениваемая величина} + \text{Смещение} + \text{Шум}.$$

Предположим, существует большое количество исследований, посвященных одному и тому же вопросу. Каждое исследование хорошо спланировано и дает объективную оценку рассматриваемого явления. Таким образом, единственная причина, по которой оценки отличаются друг от друга или от истинного значения в наблюдаемом мире (оцениваемая величина), заключается в шуме.

Но давайте также предположим, что мы столкнулись с занижением отчетности (т. е. нам не сообщают о каждом результате), и поэтому мы слышим только о результатах исследований, в которых доказательства достаточно сильны, чтобы отвергнуть нулевую гипотезу о том, что истинная оценка равна нулю (т. е. о том, что наша оценка была результатом шума). Чтобы результат был статистически отличим от нуля, предполагаемая взаимосвязь должна быть до-

статочно большой относительно стандартной ошибки. Таким образом, если мы слышим только о статистически значимых результатах, мы слышим только об оценках, которые были достаточно большими по величине.

Это означает, что для любой истинной оцениваемой величины в конечном итоге станут известны только те оценки, когда шум оказался достаточно большим по величине, чтобы отодвинуть величину оценки весьма далеко от нуля и сделать ее статистически значимой. Таким образом, в результате завышения значимости и занижения отчетности не только наши *p*-значения будут неверными, но и набор оценок, о которых мы узнаем из опубликованных исследований, будет систематически завышать истинное значение оцениваемой величины.

К сожалению, хотя мы начали с предположения об отсутствии систематической ошибки в наших оценках, мы узнали, что процесс завышения значимости и занижения отчетности вносит систематическую ошибку не в какую-то одну оценку, а в общее распределение оценок, представленных в научной литературе. Следовательно, усредняя все оценки, мы не приближаемся к истинной оценке, даже если количество оценок очень велико. То есть мы сталкиваемся с явлением, которое называется предвзятостью публикации.

Рисунок 7.1 иллюстрирует, как это происходит. В верхней части рисунка мы видим 50 несмещенных, но неточных оценок. Истинная оценка равна 1, и, поскольку наши оценки не смещены, среднее всех оценок также равно 1.

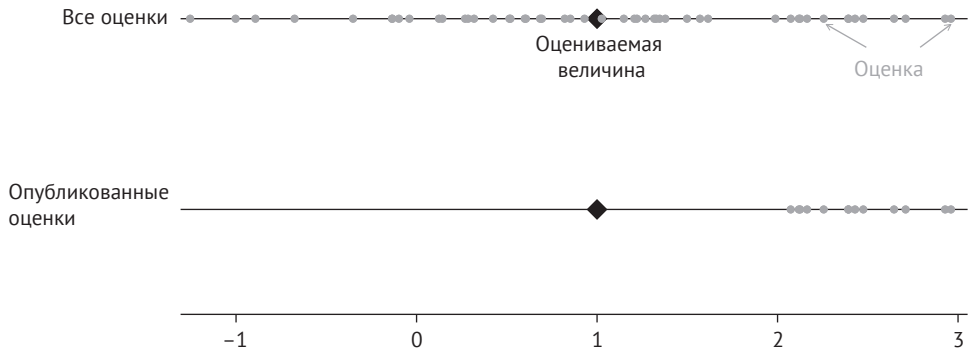


Рис. 7.1. Публикация только статистически значимых оценок создает предвзятость публикации

Мы вычисляем стандартную ошибку и обнаруживаем, что 95-процентный доверительный интервал составляет от –2 до 2. То есть вероятность того, что оценка возникла бы даже при отсутствии связи (т. е. истинная оцениваемая величина была 0), составляет менее 5 %, только если эта оценка превышает 2 по величине. Таким образом, только оценки больше 2 считаются статистически значимыми (у нас нет оценок меньше –2). Статистически значимые оценки приведены в нижней части рисунка.

Предположим, что исследователи публикуют только эти статистически значимые оценки. Разумеется, теперь опубликованные оценки систематически превышают оцениваемую величину. Поэтому, если бы мы основывали свои убеждения об истинном значении оцениваемой величины только на опубликованных оценках, у нас сложилось бы завышенное представление. Это и есть предвзятость публикации.

Завышение значимости и занижение отчетности, приводящие к предвзятости публикаций, могут возникать по-разному. Давайте рассмотрим два примера.

p-хакинг

Одна из причин, по которой мы можем столкнуться с завышением значимости и занижением отчетности, – это неправильное или нечестное поведение отдельных аналитиков. Научное сообщество называет игру с данными или тестами до тех пор, пока значение *p* не станет ниже определенного порога, *p-хакингом*. Допустим, ученый проводит эксперимент и не получает статистически значимых доказательств ожидаемого или желаемого результата. Этот ученый может решить, что, вероятно, что-то было не так с первой попыткой, и поэтому попытается немного подкорректировать эксперимент. Фактически ученый может продолжать проводить подобные эксперименты до тех пор, пока один из них не станет статистически значимым. Если он проведет эксперимент достаточное количество раз, то из-за воздействия шума рано или поздно получит искомый результат, даже если изучаемого явления в реальности не существует. Это проблема завышения значимости отдельного эксперимента. И, конечно, если недобросовестный ученый сообщает только о результатах одного эксперимента, который дал статистически значимый результат, мы сталкиваемся с проблемой занижения отчетности и, следовательно, предвзятости публикации.

Или, допустим, у аналитика есть некоторая гибкость в реализации того или иного статистического теста. Предположим, вас попросили на работе провести анализ взаимосвязи между продуктивностью работников и наличием у них стола для работы стоя. Как организовать анализ – рассмотреть всех работников сразу или построить отдельные регрессии для разных возрастных групп? Стоит ли включать полиномы возраста более высокого порядка? Следует ли разделять данные женщин и мужчин? А как насчет людей с разными должностными обязанностями, разными заболеваниями и т. д.? Как видите, существуют разные подходы к проведению анализа, и все они выглядят достаточно разумно. Если вы будете последовательно пробовать разные методы, вы в конечном итоге получите статистически значимый результат, даже если на самом деле нет никакой реальной связи между продуктивностью и работой стоя. Таким образом, перебор методик анализа является разновидностью завышения значимости.

Еще один способ завышения значимости – это рассмотреть множество разных эффектов воздействия. Предположим, вы хотите оценить эффективность какой-то новой таблетки от сердечно-сосудистых заболеваний. Вы можете провести отличное клиническое исследование, в данных которого вообще отсутствует предвзятость. Но, возможно, вы собрали данные о многих показателях среди подопытных пациентов: смертности, частоте сердечных приступов и инсультов, уровне холестерина, продолжительности госпитализации, способности заниматься физическими упражнениями, субъективном ощущении здоровья и т. д. Затем вы можете проверить, оказывает ли таблетка статистически значимое влияние на каждый из этих результатов. Если вы рассмотрите достаточно много разных показателей, то, скорее всего, обнаружите статистически значимый результат по одному из них просто из-за шума – т. е. люди, получившие лекарство, и люди, получившие плацебо, будут различаться по некоторым результатам, даже если лекарство вообще ни на что не влияет. А если

вам не хватает надлежащей этики, вы можете просто сообщить о результатах этого единственного исхода в надежде убедить врачей прописывать пациентам новую таблетку.

Как видите, *p*-хакинг может принимать самые разные формы, и вам, как количественному аналитику, придется усердно работать, чтобы избежать его, а как потребителю количественных данных – чтобы обнаружить недобросовестность[1].

p-скрининг

Конечно, *p*-хакинг вызывает большую озабоченность. Но не обязательно кто-то должен действовать нечестно или небрежно, чтобы возникла проблема чрезмерной значимости или занижения отчетности. Это может случиться, даже если абсолютно все будут вести себя идеально честно и ответственно!

Представьте себе, что 20 ученых по всей стране пришли к одной и той же научной идее. Предположим, речь идет об эффективности потенциального нового лекарства от рака. На самом деле идея ошибочна: препарат не работает. Но у ученых нет возможности узнать это с самого начала. Как и полагается поступать ученым, они разрабатывают исследования для проверки препарата. Все 20 лабораторий, независимо и не подозревая о других, проводят один и тот же высококачественный эксперимент, но на разных выборках. Каждая из них набирает большую выборку пациентов с соответствующим типом рака. Половине из них случайным образом назначают прием нового препарата. Другая половина получает плацебо. В конце периода исследования ученые смотрят, привел ли прием лекарства к повышению доли пациентов с ремиссией в группе, получавшей лекарство, по сравнению с группой, получавшей плацебо.

Допустим, 19 из 20 лабораторий не нашли статистически значимых доказательств (проценты ремиссии среди тех, кто принимал препарат, и среди тех, кто получал плацебо, неразличимы) и пришли к выводу, что препарат не работает. Такие нулевые результаты никого не заинтересуют. «Еще одно лекарство не лечит рак» не лучший заголовок для газетной статьи. Поэтому научные журналы неохотно принимают к публикации статьи с нулевыми результатами. Как следствие, эти лаборатории вряд ли удосужатся описать свои результаты, вместо этого просто перейдя к более многообещающим направлениям исследований. Иногда это называют *проблемой картотеки* (file drawer problem), поскольку статистически незначимые результаты просто убирают на дальнюю полку. Даже если лаборатории напишут о своих «закрытиях», у них могут возникнуть проблемы с поиском журнала, заинтересованного в публикации. В любом случае мы получаем заниженную отчетность, а научное сообщество и общественность не узнают об этих 19 нулевых результатах.

Одна (не)везучая лаборатория из 20 находит статистически значимые доказательства того, что препарат работает. Мы знаем, что препарат не работает (хотя ученые так не думают), поэтому мы знаем, что это чистая случайность. Так уж получилось, что по причинам, не имеющим ничего общего с препаратом,

[1] Интересный факт: термин «*p*-хакинг» был придуман Джозефом Симмонсом, Лейфом Нельсоном и Ури Симонсоном в глубоком исследовании, в котором они показали среди прочего, что, используя стандартные методы социальных наук, они могут обеспечить статистически значимые доказательства того, что прослушивание песни «When I'm Sixty-Four» группы Beatles делает испытуемых моложе!

у людей, получавших препарат в этом эксперименте, также наблюдались более высокие показатели ремиссии, чем у людей, которым было назначено получать плацебо. Такое иногда случается. Из-за влияния шума оценка может отличаться от оцениваемой величины даже при отсутствии систематической ошибки.

Поскольку другие исследования либо никогда не были описаны, либо никогда не были опубликованы, с точки зрения ученого, работающего в этой лаборатории, все существующие данные указывают на эффективность нового препарата. Итак, эта единственная лаборатория пишет научную статью о своих открытиях. Поскольку результат неожиданный и заслуживает внимания, он, скорее всего, будет охотно опубликован и освещен в научной прессе. И действительно, если вы посмотрите на это исследование, оно выглядит великолепно. Лаборатория провела хороший, беспристрастный эксперимент. Ученые сделали только одно подходящее сравнение по поводу одного подходящего результата. Никакого *p*-хакинга не было. И данные подтверждают их гипотезу. Таким образом, все верят в этот результат, хотя на самом деле он абсолютно ошибочен. Если бы у нас был доступ к данным всех 19 «неудавшихся» экспериментов, мы бы увидели, что преобладающие доказательства указывают на прямо противоположный вывод. То есть мы в итоге получаем предвзятость публикации.

Не существует общепринятого термина, который бы описывал как ученых, которые не утруждают себя описанием результатов, обнаруживающих незначительные эффекты или вообще их отсутствие, потому что их будет трудно опубликовать (проблема картотеки), так и журналы, которые неохотно публикуют такие результаты, даже если они хорошо описаны. Но мы считаем, что эти два явления полезно рассматривать вместе, поскольку они оба порождают предвзятость публикаций, несмотря на то что все участники процесса ведут себя надлежащим образом. По аналогии с *p*-хакингом мы называем эту проблему *p-скринингом*. Проблема здесь не в том, что какой-то отдельный исследователь пытается получить статистически значимый результат. Проблема в том, что научное сообщество посредством своей практики публикаций отсеивает исследования, значения *p* которых превышают определенный порог. При *p*-хакинге мы не видим нулевых результатов, потому что их прячут нечестные исследователи. При *p*-скрининге мы не видим нулевых результатов, потому что честные исследователи не могут публиковать такие результаты. В любом случае исход один и тот же. Результаты, которые мы видим, страдают от предвзятости публикации, поскольку проводится множество сравнений, но публикуются только статистически значимые.

Из-за *p*-скрининга научные данные (и наши знания во многих других областях) могут оказаться ненадежными, даже если все ведут себя правильно. Вы должны задуматься о том, что подобные вещи происходят постоянно. Фактически вы можете спросить себя: каким знаниям, полученным из научных публикаций, я вообще могу доверять? Как только вы начнете критически думать о проблеме, вы увидите ее признаки повсюду.

Являются ли большинство научных «фактов» ложными?

Как мы видели, чрезмерная значимость и занижение отчетности приводят к предвзятости публикаций. И эта практика довольно глубоко укоренились

во многих научных сообществах и культуре. Осознание этого факта привело к чему-то вроде экзистенциального кризиса во многих научных областях, когда критически мыслящие ученые начали задаваться вопросом, не являются ли многие широко признанные научные факты ложными вследствие приписывания чрезмерной значимости и занижения отчетности.

Это серьезная проблема. Многие вещи, которые мы считаем правдой, на самом деле являются ложными из-за предвзятости публикации. Но, разумеется, не все. Аналитики озаботились вопросом о том, как распознать ситуацию, когда научное сообщество страдает от серьезной предвзятости публикаций. Чтобы лучше разобраться в том, как это сделать, рассмотрим две проблемы и различные способы их решения. Мы даже обсудим несколько советов о том, как обнаружить *p*-хакинг.

Экстрасенсорное восприятие

В 2010 г. психолог из Корнелла Дэрил Бем устроил шумиху, опубликовав исследование в Journal of Personality and Social Psychology, престижном академическом журнале по психологии, в котором утверждалось, что люди обладают экстрасенсорным восприятием (ЭСВ). Часто утверждения о паранормальных явлениях развенчивают академические исследователи и количественные аналитики, но в данном случае источником диковинного утверждения стал уважаемый штатный профессор Лиги плюща.

В эксперименте Бема студентов просили предсказать, за какой виртуальной заслонкой на экране их компьютера (слева или справа) скрывается интересующий объект. Бем сообщил о статистически значимых доказательствах того, что его испытуемые явно вышли за рамки случайного угадывания в предсказании будущего и определении правильной заслонки.

Это очень интересное открытие, если вы редактор журнала, который заботится о славе, или научный журналист, который нуждается в читательской аудитории. Результат впечатляющий. Ученый, о котором идет речь, работает в авторитетном университете. Статья опубликована в крупном научном журнале. Нет никаких оснований полагать, что данные являются фальшивыми. Исследование предоставляет научные доказательства по меньшей мере удивительного явления. Какой нормальный журналист сможет устоять перед этой историей?

Несмотря на это исследование и все внимание средств массовой информации, которое оно привлекло, мы вполне уверены, что у людей нет экстрасенсорного восприятия. Так что же произошло на самом деле?

Конечно, существуют обычные проблемы, связанные с проверкой статистических гипотез. Если аналитик использует порог значимости 0.05, существует 5-процентная вероятность найти подтверждение результата (т. е. отклонить нулевую гипотезу), даже если результат ложный (т. е. нулевая гипотеза верна). И, как будет показано в части IV, если у вас уже есть веские основания полагать, что у людей нет экстрасенсорного восприятия, вам не следует сильно менять свои убеждения в ответ на это исследование.

Но у нас есть и другие проблемы, связанные с темами этой главы. Это именно тот случай, когда многие исследователи проводят эксперименты, но публикуются только те, у кого есть статистически значимые доказательства маловероятного явления. Очевидно, никто не собирается публиковать статью, в которой

сообщается, что люди угадывают правильную заслонку не лучше, чем случайно. Мы и так в это верим. Поэтому нам следует сильно беспокоиться о предвзятости публикации из-за *p*-скрининга.

Вероятно, также нельзя исключать возможность *p*-хакинга. Бем сообщил о результатах девяти различных экспериментов, проведенных в течение десяти лет. Эти эксперименты относительно просты и недороги в проведении. Поскольку Бем, судя по всему, был предан изучению экстрасенсорного восприятия, было бы логично предположить, что он на самом деле провел намного больше экспериментов по экстрасенсорному восприятию за этот десятилетний период. И если это так, то девять экспериментов, о которых он сообщил, вполне могли продемонстрировать самые убедительные доказательства существования экстрасенсорного восприятия.

Есть также некоторые признаки завышения значимости и занижения отчетности в самом исследовании. Например, Бем не находит доказательств существования экстрасенсорного восприятия в его обобщенном виде; он находит их только тогда, когда объект за занавеской носит эротический характер. Что касается других видов объектов, он не находит никаких доказательств паранормальных явлений. Ну конечно! Ведь совершенно естественно, что в процессе эволюции люди развили экстрасенсорное восприятие, позволяющее обнаруживать эротические картинки за занавеской, но не другие объекты, наподобие хищника в засаде. Более того, в некоторых тестах он обнаруживает эффект только для женщин, а не для мужчин; в других он находит результаты только для тех, кому просто скучно. Учитывая количество разнообразных экспериментов, проведенных Бемом, было бы удивительно, если бы он случайно не наткнулся на несколько статистически значимых результатов.

Обнадеживает тот факт, что сообщество психологов сохранило скептическое отношение и довольно быстро отреагировало на статью Бема. Несколько последующих исследований пытались, но не смогли воспроизвести полученные результаты. Однако, к нашему разочарованию, Journal of Personality and Social Psychology сначала отказался публиковать повторные исследования, опровергающие утверждение Бема. Редактор обосновал это решение тем, что журнал имеет давнюю политику отказа от публикации простых повторений эксперимента. К счастью, журнал в конечном итоге изменил свое мнение и опубликовал сводный анализ попыток повторить эксперимент, который убедительно свидетельствует о том, что первоначальный результат был ненадежным. Этот случай иллюстрирует одно важное решение проблемы завышения значимости и занижения отчетности: критический анализ результатов в научном сообществе и стремление исследовать, повторяемы ли эти результаты. Более подробно повторяемость результатов мы обсудим позже в этой главе.

Явка избирателей на голосование

Политические кампании включают в себя множество действий, направленных на получение голосов избирателей, – телефонные звонки, прямую почтовую рассылку, обход домов с агитацией и т. д. С 1990-х гг. ученые активно используют избирательные кампании для проведения экспериментов, чтобы узнать об эффективности различных действий. В таких исследованиях некоторым людям случайным образом назначается воздействие (например, прямая рас-

сылка с информацией о дате выборов или местонахождении их избирательного участка), а других людей случайным образом помещают в контрольную группу (например, не получающую никакой дополнительной информации). Мы можем узнать о среднем влиянии усилий по привлечению избирателей на выборы, сравнив показатели явки в двух группах.

Согласно опубликованным данным средний предполагаемый эффект от воздействия в виде прямого приглашения посетить выборы представляет собой увеличение явки избирателей примерно на 3.5 %. Более того, почти ни в одной опубликованной статье не сообщается об эффекте менее 1 %. Таким образом, если организаторы избирательной кампании ознакомятся с опубликованными отчетами, они придут к выводу, что усилия по привлечению избирателей весьма эффективны.

Но экспериментов с голосованием было проведено намного больше, чем опубликовано научных работ по этой теме, а это означает, что некоторые эксперименты не привели к публикации. Интересно, почему?

Если наши опасения по поводу завышения значимости и занижения отчетности верны, то можно предположить, что ответ заключается в *p*-скрининге: эксперименты, которые не дали статистически значимых доказательств эффекта, не привели к публикации статьи. Если это так, то существует предвзятость публикации. Поэтому следует ожидать, что истинный средний эффект от прямого приглашения на голосование будет меньше, чем предполагают опубликованные результаты.

Трое политологов, Дон Грин, Мэри МакГрат и Питер Аронов, решили проверить эти опасения количественно. Им удалось получить данные более чем 200 экспериментов, проведенных различными учеными на протяжении многих лет. По итогам лишь некоторых экспериментов были опубликованы статьи. Другие так и остались «в картотеке». Политологи провели анализ, чтобы определить средний эффект от прямого приглашения на выборы по всем двум сотням экспериментов. Результат: 0.5 % – это значительно меньше, чем средний эффект в 3.5 %, показанный в опубликованных отчетах! Неопубликованные данные гораздо меньше подтверждают эффективность усилий по повышению явки избирателей, чем опубликованные.

Эффективность усилий по повышению явки избирателей – одна из наиболее тщательно изучаемых тем в социологии. Казалось бы, кандидаты или кампании, желающие найти лучший способ распределения ограниченных ресурсов, могут обратиться к опубликованным исследованиям для обоснования своего решения. Однако теперь мы знаем, что это приведет лишь к тому, что они переоценят эффективность усилий по привлечению избирателей в 7 раз, что еще раз продемонстрирует, насколько серьезными могут быть последствия *p*-скрининга.

Выявление *p*-хакинга

Всегда сложно узнать наверняка, применялся ли *p*-хакинг в отдельном исследовании. В каком-то смысле это хорошо – проявлять доброту, предполагая, что большинство людей большую часть времени пытаются вести себя честно и разумно. Тем не менее критическое мышление помогает нам понять, насколько широко распространена проблема *p*-хакинга. Наилучшим доказательством этому служит исследование *p*-значений в опубликованной научной литера-

туре. Мы не будем здесь утверждать, что какие-то конкретные исследования были подтасованы. Но этот обзор поможет нам понять, почему научные публикации буквально пестрят многочисленными признаками *p*-хакинга.

Вот как это работает. Мы начинаем с размышления о том, как будет выглядеть распределение *p*-значений в литературе в четырех различных возможных состояниях мира:

1) если в мире нет реальных взаимосвязей и нет *p*-хакинга;
2) если в мире есть реальные взаимосвязи и нет *p*-хакинга;
3) если в мире нет реальных взаимосвязей и есть *p*-хакинг;
4) есть ли в мире есть реальные взаимосвязи и есть *p*-хакинг.

Затем мы сравниваем фактическое распределение *p*-значений, указанных в научной литературе, с распределениями, которые получились бы в каждом из этих четырех состояний мира, чтобы попытаться выяснить, с каким состоянием мы, скорее всего, столкнулись. Далее будем предполагать, что присутствует *p*-скрининг (поэтому нет значений *p*, превышающих 0.05). Мы просто хотим выяснить, не применялся ли в публикациях *p*-хакинг. Но все, что мы собираемся сказать, верно и без *p*-скрининга.

Логику этих случаев можно понять, обратившись к рис. 7.2, адаптированному из исследования Симонсона, Нельсона и Симмонса 2014 г., которые впервые предложили изучить распределение *p*-значений для оценки *p*-хакинга.

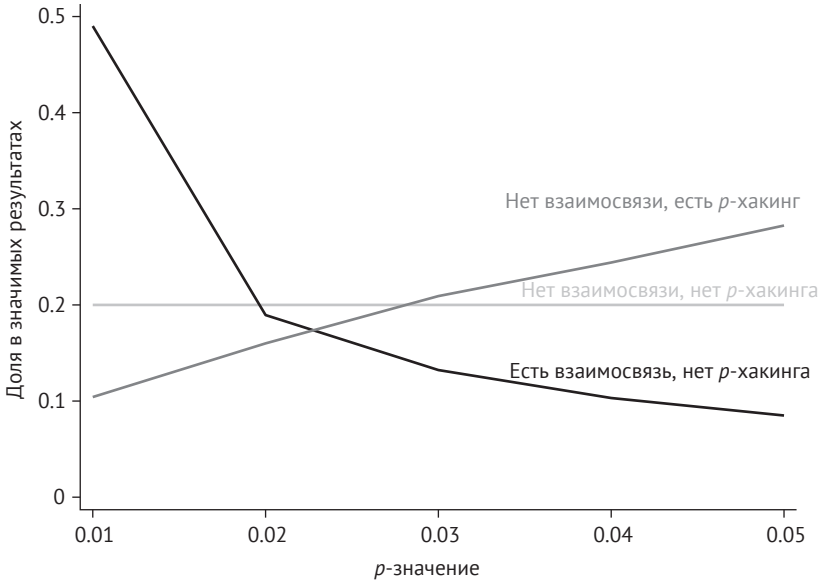


Рис. 7.2. *p*-хакинг искажает распределение *p*-значений в научных публикациях

Начнем со случая 1 – в мире нет реальных взаимосвязей и нет *p*-хакинга. Если в мире нет реальных взаимосвязей между наблюдаемыми переменными, это означает, что нулевая гипотеза верна. И, как мы уже говорили, если нулевая гипотеза верна, то все значения *p* имеют равные шансы встретиться в любом конкретном исследовании. Таким образом, если *p*-хакинг отсутствует, наблю-

даемое распределение значений *p* в опубликованных исследованиях должно выглядеть примерно равномерным – т. е. в диапазоне от 0 до 0.05 разные значения *p* должны появляться примерно с одинаковой частотой. Это показано светло-серой линией на рис. 7.2.

Есть только две причины, по которым мы можем увидеть отклонение от этого единообразия. Первая – если в мире существует настоящая взаимосвязь между переменными. Вторая – если применялся *p*-хакинг.

Это подводит нас к случаю 2: в мире существует реальная взаимосвязь (т. е. нулевая гипотеза ложна) и нет *p*-хакинга. Если мы на самом деле изучаем реальную взаимосвязь между явлениями мира, у нас больше шансов обнаружить статистически значимые отношения, чем если бы в мире не было этой взаимосвязи. Итак, в случае 2, когда существует реальная взаимосвязь и нет *p*-хакинга, мы ожидаем, что распределение *p*-значений в опубликованной записи будет искажено в сторону более низких *p*-значений. То есть, отражая тот факт, что существует реальная взаимосвязь, в случае 2 должно быть больше низких *p*-значений, чем в случае 1. Итак, если мы видим распределение с преобладанием более низких *p*-значений, это наводит на размышления о том, что в публикации представлены реальные взаимосвязи. Этому случаю соответствует черная кривая на рис. 7.2.

Другая причина, по которой мы можем увидеть отклонение от случая 1, связана с *p*-хакингом. Это подводит нас к случаю 3: нет реальной взаимосвязи и применяется *p*-хакинг. Как мы уже обсуждали, когда в мире нет реальных взаимосвязей, каждое *p*-значение одинаково вероятно. Но что происходит при *p*-хакинге? Допустим, исследователи обнаружили значение *p* ниже 0.05. Они могут просто сообщить об этом статистически значимом результате. Но предположим, что они нашли значение *p*, близкое к 0.05, но выше него. У них может возникнуть соблазн заняться *p*-хакингом, манипулируя условиями эксперимента, составом подгрупп и т. д., пока они не найдут значение *p* ниже 0.05, о котором смогут сообщить как о статистически значимом. Последствием этого *p*-хакинга станет целый ряд заявленных значений *p*, близких к 0.05, но чуть ниже. Таким образом, в отличие от случая 2, где мы видели преобладание низких *p*-значений среди статистически значимых результатов, в случае 3 мы будем наблюдать преобладание более высоких *p*-значений. Этот случай иллюстрируется темно-серой кривой на рис. 7.2.

Случай 4 объединяет случаи 2 и 3. Если существует истинная взаимосвязь, это смещает распределение в сторону низких *p*-значений. Если при этом также применялся *p*-хакинг, он снова сместит ситуацию в сторону высоких значений *p*. Поэтому трудно понять, чего ожидать в этом случае. Но тем не менее, просто учитывая различия между случаями 1, 2 и 3, мы можем добиться некоторого прогресса в выявлении *p*-хакинга в научной литературе.

К сожалению, во многих научных публикациях распределение *p*-значений соответствует случаю 3. Симонсон, Нельсон и Симмонс изучили статьи в известном журнале по психологии, чтобы выяснить, нет ли каких-либо тревожных сигналов, указывающих на *p*-хакинг. Они установили определенные слова, которые могут служить признаком завышенной значимости. Одно из таких слов – «исключение», например: «мы исключили эту переменную (или группу, или результат) из своего анализа, потому что она не оказала желаемо-

го влияния на результат». Другое слово – «преобразование», например: «мы преобразовывали возраст в возраст^2, возраст^3, возраст^4 и т. д., пока результаты не подтвердили нашу гипотезу». Более темная кривая на рис. 7.3 показывает распределение *p*-значений для исследований, в которых не используются слова, являющиеся признаками *p*-хакинга. Обнадеживает то, что в этих исследованиях мы видим преобладание низких *p*-значений, т. е. можно рассчитывать, что они выявляют подлинные взаимосвязи (случай 2). Более светлая кривая на рис. 7.3 показывает распределение *p*-значений для исследований, в которых используются слова, являющиеся признаками *p*-хакинга. Вызывает тревогу то, что в этих исследованиях есть основания подозревать *p*-хакинг; мы видим преобладание более высоких *p*-значений (случай 3). Таким образом, хотя этот способ анализа не говорит нам точно, какие статьи основаны на *p*-хакинге, он позволяет нам взглянуть на распределение *p*-значений в литературе и задаться вопросом, насколько нам следует беспокоиться о том, что любой научный консенсус, основанный на этих исследованиях, достигнут путем *p*-хакинга.

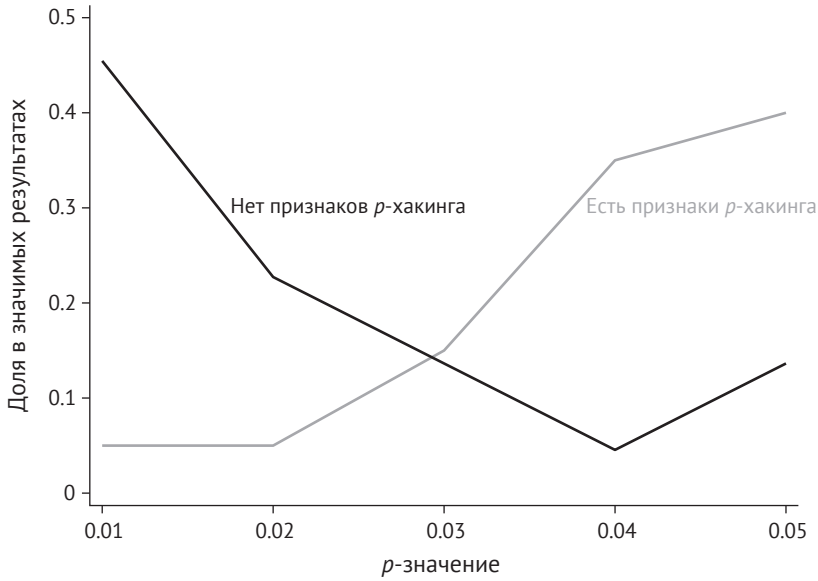


Рис. 7.3. Использование распределения *p*-значений для выявления признаков *p*-хакинга

Возможные решения проблемы

Предвзятость публикаций – коварная проблема для науки. Поэтому ученые задумались о том, как изменить научную практику, чтобы уменьшить проблему завышения значимости и занижения отчетности.

Уменьшение порога статистической значимости

Возможно, мы сможем решить проблему предвзятости публикаций, используя более строгий порог значимости для *p*-значений. Похоже, что общепринятый порог 0.05 упрощает подгонку условий, пока не будет найден статистически значимый результат. В 2017 г. 72 исследователя из различных областей науки

опубликовали в журнале Nature Human Behavior письмо, призывающее научное сообщество принять значительно более низкий порог значимости $p < 0.005$.

С одной стороны, более низкий порог значимости, безусловно, затруднит получение статистически значимого результата путем завышения значимости. С другой стороны, снижение порога значимости, вероятно, усилит стимулы для *p*-хакинга, поскольку сделает статистически значимые результаты более редкими и, следовательно, более ценными. Новый порог может даже усилить самоуспокоенность в отношении этих проблем, заставив всех нас мыслить менее критично. И хотя порог 0.005 означает меньше ложных срабатываний (т. е. отказов от нулевой гипотезы, когда она верна), это происходит за счет большего количества ложноотрицательных результатов (т. е. невозможности отвергнуть нулевую гипотезу, когда она ложна). Поэтому пороговое значение является своего рода компромиссом и, вероятно, должно зависеть от конкретного вопроса.

Корректировка *p*-значения при многократном тестировании

Предполагается, что *p*-значение говорит нам о вероятности получения результата, по крайней мере такого же сильного, как ваш результат, если нулевая гипотеза верна. Как мы видели, если исследователи завышают значимость, *p*-значение перестает отражать эту вероятность. Оно слишком низкое.

Если мы знаем, сколько тестов было проведено, мы можем попытаться исправить *p*-значение. Как мы обсуждали в случае с осьминогом Паулем, если исследователи проведут десять независимых тестов, но сообщат только о самом низком *p*-значении, равном 0.03, истинное *p*-значение будет больше похоже на $1 - (1 - 0.03)^{10} \approx 0.263$. Корректировка *p*-значений таким образом для учета количества проведенных тестов – это хороший способ для исследователей обеспечить прозрачность, а для потребителей количественной информации – лучше оценить состояние доказательств. К сожалению, это тоже не панацея. Простой расчет, приведенный выше, работает только в том случае, если тесты действительно независимы. Если тесты тесно связаны друг с другом (например, если мы проверяем одну и ту же гипотезу с одними и теми же данными, но используем несколько разные переменные в регрессии или фокусируемся на несколько разных подгруппах наблюдений), может быть гораздо менее ясно, как правильно скорректировать *p*-значение.

Не зацикливайтесь на статистической значимости

Порог 0.05 – это просто произвольное число. Существенно важные эффекты могут быть статистически незначимыми, а статистически значимые результаты могут быть существенно неважными. Статистическая проверка гипотез – полезный инструмент количественной оценки неопределенности, но им можно злоупотреблять. Мы должны использовать *p*-значения, когда это уместно. Но они не являются окончательным средством оценки правдоподобности количественных результатов. Вы не можете ограничиться вычислениями; вам нужно продолжать критически мыслить. В части IV мы обсудим, как следует объединять количественные данные с другими знаниями, чтобы критически рассуждать о трансформации наших убеждений после получения новых свидетельств.

Предварительная регистрация

По крайней мере, в некоторых ситуациях – например, когда исследователи сами создают данные с помощью нового опроса или эксперимента – они могут заранее заявить о том, какие именно тесты собираются провести, прежде чем увидят данные. Для этого исследователи предварительно публикуют план своего эксперимента, указывая, что именно они планируют проверять и как они планируют это делать, прежде чем фактически приступить к исследованию. Если они предварительно регистрируют ограниченное количество тестов, это предотвращает подгонку экспериментов под результат. Этот подход также усложняет занижение отчетности.

У читателей возникнут обоснованные подозрения, если в научной статье будут представлены результаты только трех из десяти запланированных тестов. Более того, некоторые научные журналы теперь готовы принимать к публикации научные исследования только на основе предварительно зарегистрированного плана эксперимента и обязательства исследователей сообщать о результатах независимо от того, что они обнаружат, что также помогает избежать занижения отчетности. Предварительная регистрация является полезным инструментом для смягчения проблем завышения значимости и занижения отчетности. Рассмотрим этот подход на примере, чтобы определить его достоинства и границы возможностей.

Предварительная регистрация схемы испытания лекарств

Проблема завышения значимости и занижения отчетности имеет большое значение в клинических испытаниях новых лекарств: компания, которая вложила много средств в новое лекарство, может поддаться искушению пересматривать методику эксперимента, состав подгрупп или результаты, пока не найдет какой-то результат, доказывающий эффективность препарата при лечении заболевания у определенной группы людей. Национальный институт сердца, легких и крови (NHLBI) с 1970 г. профинансировал множество клинических испытаний новых лекарств и пищевых добавок, а в 2000 г. они придумали хитрый способ использования предварительной регистрации для борьбы с проблемой подгонки результатов. Они потребовали от разработчиков лекарства или добавки заранее заявлять об ожидаемом результате применения. Согласно этим новым правилам клиническое испытание лекарства объявляется успешным только в том случае, если исследователи демонстрируют статистически значимое влияние препарата на достижение заранее зарегистрированного результата.

Исследование Роберта Каплана и Вероники Ирвин, проведенное в 2015 г., показывает, что, после того как NHLBI стал требовать предварительную регистрацию, доля успешных испытаний упала с 57 до 8 %. Это говорит о том, что многие из «успешных» испытаний до предварительной регистрации были результатом завышения значимости влияния, а не какого-либо реального эффекта препарата или добавки. Это большой успех подхода с предварительной регистрацией.

Важно отметить, что, даже если предварительная регистрация помогает сократить завышение значимости и занижение отчетности, нам все равно придется беспокоиться обо всех остальных проблемах статистических выводов.

Возьмем, к примеру, 8-процентный показатель успеха препаратов после предварительной регистрации. Каплан и Ирвин используют порог значимости 0.05. Это означает, что, даже если лекарство окажется абсолютно неэффективным, мы ожидаем, что 5 % таких испытаний случайно генерируют статистически значимые доказательства эффективности. Таким образом, 8-процентный показатель успеха лишь ненамного превышает вероятность полностью случайного везения. Это означает, что даже после того, как мы увидим успешное предварительно зарегистрированное исследование, мы все равно не можем быть полностью уверены в эффективности препарата. На самом деле вероятность того, что положительный предварительно зарегистрированный результат на самом деле является ложноположительным, составляет 5 из 8.

Повторяемость

Один из способов оценить, является ли предполагаемый эффект подлинным, – это воспроизвести его на новых, независимо созданных данных. *Повторяемость*, или *репликация*, не является гарантией истинности эффекта. Но, как мы видели на примере исследований экстрасенсорного восприятия, она способна предоставить дополнительные свидетельства о том, является ли предполагаемый эффект подлинным, или это просто результат завышения значимости и занижения отчетности.

Предположим, мы делаем только одно сравнение и используем порог значимости 0.05. Вероятность найти статистически значимые доказательства взаимосвязи явлений, даже если их не существует, составляет 0.05. Но если мы проведем исследование дважды, каждый раз используя независимые данные, вероятность найти статистически значимые доказательства в обоих исследованиях при отсутствии реальной взаимосвязи составит $0.05 \times 0.05 = 0.0025$. Если мы проведем исследование в третий раз, вероятность того, что мы найдем доказательства несуществующей взаимосвязи во всех трех исследованиях, составит $0.05^3 = 0.000125$, что весьма маловероятно. Повторяя исследование, мы уменьшаем вероятность того, что придем к ложноположительному выводу. Это особенно верно, если репликация выполняется независимыми группами исследователей, у которых нет личной заинтересованности в достижении первоначальных результатов.

Конечно, репликация не панацея, и нам нужно продолжать критически мыслить. Неспособность отвергнуть нулевую гипотезу не является доказательством того, что нулевая гипотеза верна. Если доверять только тем результатам, которые удалось независимо получить несколько раз, мы можем иногда ошибочно отвергать реальные эффекты, особенно если мы проводим репликации на разреженных или зашумленных данных, где эффекты трудно обнаружить. В идеале мы должны собрать много данных и повторить эксперимент на больших выборках.

Иногда такая репликация возможна, а иногда нет. Если исследователи, проводящие испытания препарата против холестерина, соберут данные о весе и обнаружат неожиданное влияние препарата на потерю веса, они смогут набрать новую группу субъектов и посмотреть, показывает ли новая группа испытуемых аналогичную потерю веса по сравнению с новой контрольной группой. Но если мы обнаружим феномен, связанный с выборами губернаторов XX в., поведением спутников Венеры или стратегиями лидерства в преддве-

рии Первой мировой войны, это будет единственная выборка. Нет возможности собрать больше данных. В этом случае нам, возможно, стоит рассматривать репликацию менее буквально и более концептуально. Это можно сделать не путем прямого повторения существующих результатов, а задавая вопросы наподобие: «Если это явление истинно, какие еще гипотезы также должны быть верными?» Давайте рассмотрим пример.

Футбол и выборы

Энтони в соавторстве с Пабло Монтаньем написал статью, иллюстрирующую этот подход. В статье анализируется известное исследование Эндрю Хили, Нила Малхотры и Сесилии Мо, опубликованное в журнале Proceedings of the National Academy of Sciences, в котором утверждается, что исход студенческих футбольных матчей влияет на то, кто победит на выборах. В частности, в исходном исследовании сказано, что действующая партия показывает лучшие результаты на выборах в конгресс и губернаторов в округах, где проживают команды, победившие до выборов. Подобные открытия заставляют некоторых людей беспокоиться о судьбе демократии. (Не нас, но это тема для другой дискуссии.)

Во всяком случае, исследование Хили и его соавторов во многом примечательно. Футбольные победы и поражения кажутся довольно случайными, поэтому нет особых причин беспокоиться о предвзятости. Но это именно тот случай, когда стоит беспокоиться о ложноположительном результате в результате завышения значимости или занижения отчетности. Например, почти наверняка существует проблема *p*-скрининга: опубликует ли престижный научный журнал статью, показывающую, что результаты футбольных матчей в колледжах, похоже, не оказывают никакого влияния на выборы в конгресс? Более того, существует множество других видов спорта, которые можно было бы использовать для прогнозирования успеха действующего президента: другие исследовательские группы, изучавшие влияние поражений в баскетболе или керлинге на выборы, возможно, не обнаружили никакой связи и, следовательно, не опубликовали статьи. Поэтому нам не следует делать поспешных выводов только потому, что в единственной опубликованной статье представлены доказательства, подтверждающие гипотезу о том, что поражения в одном конкретном виде спорта связаны с результатами выборов.

Энтони и Пабло не смогли провести чистую независимую репликацию, потому что невозможно повторно провести десятилетия футбольных матчей в колледже и выборов в конгресс. Вместо этого они рассмотрели независимые теоретические прогнозы – дополнительные гипотезы, которые, как ожидается, также должны иметь силу, если исходы футбольных матчей действительно влияют на выборы. Например, если результаты футбольных матчей среди студентов влияют на выборы, можно ожидать, что эта связь будет особенно сильной в тех местах, где большое внимание уделяется студенческому футболу. Если избиратели обвиняют действующих политиков в плохих футбольных результатах, можно ожидать, что влияние футбольного поражения на партию действующего президента будет сильнее, когда действующий президент добивается переизбрания, по сравнению с тем, когда баллотируется какой-то новый кандидат от той же партии. И т. д. Проверка таких гипотез, которые говорят о лежащем в их основе механизме, – это способ проверить, может ли не-

который предполагаемый эффект отражать реальные отношения в мире или является результатом шума (т. е. ложного срабатывания).

Вот несколько примеров того, что нашли Энтони и Пабло. Оказывается, что предполагаемый эффект футбольных игр, наоборот, меньше в округах, где больше людей уделяет внимания студенческому футболу, чем в округах, где люди меньше интересуются играми студентов, и не больше, когда действующий президент действительно баллотируется на переизбрание, и одинаково сильны как за пределами округа проживания команды, как и внутри округа проживания. Более того, они не обнаружили никаких доказательств связи между результатами футбольных игр команд НФЛ и результатами выборов, несмотря на то что команды НФЛ имеют такую же региональную поддержку, как и команды колледжей по футболу, и игры НФЛ примерно в десять раз более популярны, чем игры в американский футбол.

Энтони и Пабло проверили множество независимых теоретических прогнозов, которые потенциально могли бы сбыться, если бы связь между футболом и выборами была реальной, но ни одно из них не получило подтверждения в данных. Из этого они пришли к выводу, что влияние футбольных матчей в колледжах на результаты выборов очень маловероятно. Это не классическая репликация. Но этот пример показывает, как изучение доказательств дополнительных гипотез, связанных с механизмом, лежащим в основе исходной гипотезы, может помочь пролить свет на силу доказательств неожиданных результатов.

С идеей независимой репликации связано использование резервных выборок, которое мы обсуждали в главе 5, когда говорили о переобучении. Предположим, у вас есть большая выборка данных, и вы хотите изучить ее на предмет взаимосвязей. Возможно, имеет смысл выделить случайно выбранную часть этих данных из исследования. Например, вы можете случайным образом выбрать половину наблюдений и использовать их в качестве исследовательского набора данных. Затем, после того как вы обнаружите несколько интересных взаимосвязей, сможете проверить, появляются ли эти взаимосвязи в отложенной выборке данных, которую вы еще не проанализировали. Если завышение значимости привело к ложноположительному результату в вашем первоначальном анализе, такая же связь вряд ли проявится в контрольной выборке. Но если вы обнаружили реальное явление, то следует ожидать, что оно сохранится в отложенных данных.

Проверка важных и правдоподобных гипотез

Если вы встретили исследование, которое никогда бы не было опубликовано, если бы исследователи обнаружили противоположное явление (например, не смогли отвергнуть нулевую гипотезу), вам следует заподозрить завышение значимости или занижение отчетности. Но если исследование дает ответ на вопрос, который для нас важен независимо от того, каким окажется этот ответ, опасения по поводу завышения значимости и занижения отчетности практически исчезают. В частности, если результаты могут быть опубликованы независимо от результата, мы можем меньше беспокоиться о *p*-скрининге и можем полагать, что у исследователя нет причин заниматься *p*-хакингом.

К счастью, многие важные научные исследования попадают в эту последнюю категорию. Если исследование проверяет серьезную гипотезу, тестирует

перспективную методику лечения, которая может сработать, или оценивает реальное воздействие на политическую ситуацию, ответ интересен, каким бы он ни оказался.

Напротив, многие забавно звучащие вопросы с неожиданными ответами попадают в первую категорию. И к сожалению, такие исследования выглядят чрезвычайно привлекательными для большей части научной прессы. Вспомните исследование экстрасенсорного восприятия. Никому не интересна статья, в которой не обнаружено доказательств существования экстрасенсорного восприятия. Отсюда возникают опасения как по поводу *p*-хакинга, так и по поводу «эффекта картотеки». Следующий пример проиллюстрирует эту мысль.

Поза власти

Знаменитое исследование Эми Кадди, Даны Карни и Энди Япа якобы доказывает замечательную эффективность принятия *позы власти*. Авторы утверждали, что, хотя обычно внутренние установки являются причиной нашего поведения, небольшие изменения в нашем поведении могут, наоборот, изменить наши внутренние установки. В частности, приняв позу, которая внутренне ассоциируется у вас с властью, вы пробудите в себе чувство напористости и будете вести себя соответственно.

Хотя лежащие в основе этой гипотезы научные соображения вызывают серьезные сомнения, ее сторонники продолжают утверждать, что принятие правильной позы заставляет людей испытывать чувство силы и приводит к физиологическим изменениям, включая повышение уровня тестостерона и кортизола. Ранее не было веских причин полагать, что это может быть правдой. И трудно представить, чтобы крупный журнал опубликовал исследование, показывающее, что принятие позы власти ни на что не влияет. Так что читатели должны были быть настроены скептически с самого начала. Тем не менее, поскольку результаты оказались забавными, удивительными и оптимистичными, исследование привлекло огромное внимание. Оно было опубликовано в престижном научном журнале, и о нем написали в крупных средствах массовой информации, а Кадди пригласили с выступлением на TED, которое оказалось чрезвычайно популярным.

Неудивительно, что результат оказался ошибочным. Многочисленные попытки репликации не смогли обнаружить заявленные эффекты. А один из соавторов, Дана Карни, в конце концов дезавуировала работу, задокументировав множество случаев, когда открытие стало результатом *p*-хакинга.

За пределами науки

Мы сосредоточились на том, как завышение значимости и занижение отчетности создают серьезные проблемы для научного сообщества. Но, как показывает история осьминога Пауля, проблема шире. Действительно, вы сталкиваетесь с этими явлениями регулярно, часто даже не замечая того.

Предположим, кто-то пытается вам что-то продать – например, машину, финансовый совет или подписку на приложение для знакомств. Продавец может сказать: «Этот автомобиль занял первое место по степени удовлетворенности клиентов через пять лет после покупки!» Звучит здорово. Но вы можете спросить себя, сколько показателей удовлетворенности они проанализировали, прежде

чем найти тот, по которому был оценен этот автомобиль номер один. Учитывали ли они надежность, историю ремонтов, безопасность, долговечность, расход бензина и стоимость при перепродаже в дополнение к удовлетворенности клиентов? На каком отрезке времени рассматривали удовлетворенность клиента – один, два или три года с момента покупки? Если это так, то они провели множество сравнений и рассказали вам лишь о том показателе, который представляет машину в наилучшем свете. Это не беспристрастная оценка качества автомобиля; это эквивалент *p*-хакинга в исполнении продавца.

Точно так же ваш финансовый консультант может сказать вам: «Этот фонд взаимного кредитования входил в S&P 500 семь из последних восьми лет». Звучит неплохо. Но как он вел себя по сравнению с индексом Доу-Джонса или индексом широкого рынка? Как обстояли дела за последние девять, десять, пятнадцать лет? Выбрал ли консультант сравнение с индексом S&P за последние восемь лет, потому что это был естественный показатель или потому что именно он сделал фонд лучшим?

В общем, вам нужно привыкнуть постоянно думать о проблемах завышенной значимости и заниженной отчетности, а не только о формальной проверке гипотез и статистической значимости. Всякий раз, когда вам предлагают какое-либо доказательство, вы должны спросить себя, является ли оно естественным критерием или первым, на которое вам самому пришло в голову взглянуть. В противном случае вы можете взять паузу, чтобы поразмыслить над тем, существуют ли в этой области другие правдоподобные критерии сравнения, и не станете ли вы очередной жертвой преднамеренного подбора наилучших показателей.

В качестве подтверждения того факта, что эта проблема действительно встречается повсюду, позвольте привести пример еще одной ситуации, когда нам подсовывают завышенную значимость и заниженную отчетность, – феномен суперзвезд.

Суперзвезды

Мы склонны восхищаться людьми, которые действительно успешны. Нам нравится изучать их жизнь. Вы уже знаете одну причину, по которой это может привести к ошибочным выводам: корреляция требует вариаций. Еще одна причина, по которой нам не следует так спешить восхищаться и изучать суперзвезд, заключается в том, что, возможно, в них нет ничего особенного, кроме удачи.

Билл Миллер изучал экономику в колледже, служил офицером военной разведки, пробовал себя в докторской программе по философии и работал казначеем в сталелитейной и цементной компаниях, прежде чем в 1981 г. (в 31 год) занял должность директора по исследованиям в Legg Mason Capital Management. Миллер явно был умным парнем, и его ждала многообещающая карьера. В следующем году он начал управлять паевым фондом Legg Mason Value Trust. В течение первого десятилетия или около того эффективность фонда была посредственной, немного отставая от средней по рынку. Но в конце концов Миллер добился своего, достигнув больших доходов в конце 1990-х и начале 2000-х гг. К 2006 г. коллеги-инвесторы и репортеры заметили, что Legg Mason Value Trust 15 лет подряд опережал рынок – беспрецедентная полоса успеха, которая вывела Билла Миллера в высшие эшелоны финансовой славы.

Разумеется, все хотели узнать секрет Миллера. Что сделало его таким успешным инвестором? Возможно, это удивительно, но Миллер добился успеха не за счет глубоких знаний нишевых отраслей или технических торговых алгоритмов. Его фонд в основном инвестировал в небольшое количество уже известных компаний, таких как Google, Amazon, eBay, J.P. Morgan и Aetna. Описывая свою инвестиционную философию в письме инвесторам в 2006 г., Миллер сообщил, что он просто ищет «наибольшую ценность». Далее он рассуждал о том, что отличает его фонд от многих конкурентов: «Мы отличаемся от многих инвесторов тем, что готовы анализировать акции, которые выглядят перспективными, чтобы увидеть, так ли это на самом деле. На самом деле большинство из них перспективны, и лишь редкие не оправдывают ожиданий».

Миллер говорит, что все очень просто. Он инвестирует в компании, которые недооценены, и получает доход с роста акций. Но, прежде чем мы придем к выводу, что Билл Миллер – гениальный инвестор, давайте рассмотрим возможность того, что Миллеру просто повезло, как осьминогу Паулю.

В финансах есть идея, называемая *гипотезой эффективного рынка*. По сути, она говорит о том, что ни один фонд или инвестиционная стратегия не могут систематически превосходить средние рыночные показатели в долгосрочной перспективе. В общих чертах логика выглядит так. Если бы какой-нибудь гениальный инвестор придумал инвестиционную стратегию, которая предсказуемо превзошла бы рынок, другие инвесторы подражали бы этой стратегии. Это изменит цены на активы, торгуемые в рамках этой стратегии. Инвесторы будут продолжать делать это до тех пор, пока эта стратегия не перестанет превосходить рынок. Например, если цена акций компании полностью отражает всю имеющуюся информацию о стоимости этой компании, чего и следовало ожидать на большом рынке, где множество людей торгуют на основе наилучшей доступной информации, не должно быть никакого способа систематически предсказывать изменения цены акций без ведома инсайдеров.

Если гипотеза эффективного рынка верна, то Миллер и другие управляющие фондами и финансовые аналитики просто делают то же самое, что подбрасывание монеты. И мы знаем, что, если достаточное количество людей подбрасывает монеты, некоторые из них просто по счастливой случайности получат длинную цепочку орлов. Итак, чтобы оценить, действительно ли Миллер гений, нам нужно задаться вопросом, насколько вероятно, что он просто окажется тем парнем, которому по счастливой случайности попала длинная череда орлов.

Для начала давайте представим, что обыграть рынок – это все равно, что выбросить орел честной монетой. Это наша нулевая гипотеза. Теперь мы должны спросить: если наша нулевая гипотеза верна, насколько вероятно, что кому-то выпадет 15 орлов подряд?

Шансы на то, что конкретный инвестор в течение 15-летнего периода случайно обыграет рынок каждый год, очень малы. Вероятность того, что какой-то инвестор по счастливой случайности обыграет рынок в течение одного года, равна 1/2. Вероятность того, что инвестор обыграет рынок благодаря удаче два года подряд, равна $1/2 \times 1/2$, а вероятность того, что инвестор обыграет рынок 15 лет подряд благодаря чистой удаче, равна $(1/2)^{15}$, или примерно 1 из 30 000. Так что, возможно, Миллер – гений; если бы он просто подбрасывал монеты, вероятность того, что он добьется такого успеха, составляет всего 1 из 30 000.

А может, и нет. Давайте убедимся, что мы критически воспринимаем некоторые вещи.

На рынке очень много инвесторов, и, если бы кто-то из них обыгрывал рынок 15 лет подряд, он был бы так же знаменит, как Миллер, и мы бы обсуждали именно его. Следовательно, актуальный вопрос не в том, насколько вероятно, что один конкретный управляющий фондом, Билл Миллер, случайно обыграет рынок 15 лет подряд. Актуальный вопрос заключается в том, насколько вероятно, что один *из множества* управляющих фондом случайно обыграет рынок 15 лет подряд.

Обратите внимание: это похоже на предвзятость публикации или проблему осьминога Пауля. Из-за предвзятости публикаций мы слышим только о немногих исследованиях со статистически значимыми результатами, а остальные остаются в тени. Что касается осьминога, мы слышали только об одном животном из множества, которое правильно спрогнозировало несколько футбольных матчей подряд. Точно так же мы слышим только о немногих инвесторах, у которых действительно долгая полоса везения. Во всех трех случаях, если мы рассматриваем только исследования, животных или инвесторов, о которых нам приходится слышать, мы переоцениваем вероятность того, что их успех отражает реальное явление в мире.

В любой год на рынке торгуют как минимум 24 000 профессиональных фондов, и, предположительно, каждый из них продолжит торговать, если обыграет рынок. Итак, давайте предположим (в качестве нашей нулевой гипотезы), что существует 24 000 управляющих фондами, ни один из которых не обладает каким-либо особым пониманием. Это означает, что для каждого из них вероятность обыграть рынок в определенный год составляет 50 на 50. Таким образом, выяснить, насколько вероятно, что один из них обыграет рынок 15 лет подряд, – это все равно, что выяснить, насколько вероятно, что кто-то выбросит 15 орлов подряд, если каждый из 24 000 человек подбросит по 15 честных монет.

Если провести тот же расчет, что и для осьминога Пауля, то ответ будет примерно 0.52, или 1 к 2.[1] Очень маловероятно, что какой-либо конкретный инвестор сможет обыграть рынок 15 лет подряд благодаря чистой удаче. Но если принять во внимание тысячи инвесторов, то на самом деле вполне вероятно, что один из них будет обыгрывать рынок 15 лет подряд, даже если никто из них не обладает выдающейся проницательностью, и все они просто подбрасывают монеты.

Эти расчеты выглядят для Миллера еще хуже, если учесть, что 15-летняя полоса выглядела бы столь же впечатляюще, если бы она началась в любом другом году. Если мы рассмотрим все фонды и все возможные 15-летние периоды, то представляется весьма вероятным, что у какого-либо управляющего фондом в какой-то момент случайно возникнет такая полоса. Эти расчеты в сочетании с нашим знанием гипотезы эффективного рынка должны заставить нас скептически относиться к любому, кто утверждает, что знает секреты победы на рынке. Огромное количество трейдеров и фондов означает, что среди них

[1] Вероятность того, что конкретный инвестор будет в выигрыше 15 лет подряд, равна 0.5^{15}. Таким образом, вероятность того, что конкретный инвестор будет проигрывать 15 лет подряд, равна $1 - 0.5^{15}$. Таким образом, вероятность того, что ни один из 24 000 не ответит правильно 15 лет подряд, равна примерно $(1 - 0.5^{15})^{24000}$. Отсюда вероятность того, что хотя бы один из инвесторов будет в выигрыше 15 лет подряд, составляет $1 - ((1 - 0.5^{15})^{24000})$, или приблизительно 0.52.

обязательно найдется тот, у кого сложится исключительно хороший послужной список. Разумеется, именно о нем напишут в новостях. Итак, прежде чем передать свои сбережения инвестиционному менеджеру, спросите себя, вложили бы вы те же деньги, делая ставки на футбольные прогнозы осьминога Пауля? Если нет, позвольте порекомендовать вам рассмотреть недорогие индексные фонды.

Как вы думаете, что случилось с Биллом Миллером после шквала публикаций в прессе в середине 2000-х? Полоса удач закончилась в 2006 г. Его фонд потерял 55 % своей стоимости во время финансового кризиса 2008 г., после чего фонд барахтался на рынке еще несколько лет, и в конце концов Миллер ушел со своего поста в 2012 г. В тот период, когда Миллер управлял Legg Mason Value Trust, фонд фактически *отставал от рынка*. Увы, его историческая победная серия по-прежнему позволяет ему регулярно появляться в программах новостей, где он рассуждает о рыночных условиях и выборе акций. В 2017 г. его новый фонд Miller Opportunity Trust снова попал в новости о впечатляющих доходах. В чем секрет? Большая ставка на Apple, самую ценную компанию в мире.

Подведение итогов

Завышение значимости или занижение отчетности может произойти из-за нечестного поведения исследователя (*p*-хакинг) или из-за особенностей поведения сообщества полностью честных исследователей (*p*-скрининг). В любом случае это приводит к предвзятости публикаций – явлению, при котором опубликованные результаты систематически вводят в заблуждение, поскольку существует смещение в сторону публикации статистически значимых результатов. У этих проблем нет простого решения. Но, научившись критически мыслить о них, вы сможете предсказывать, когда это может произойти, и придумать методы, которые хотя бы смягчат последствия.

В главе 8 мы обратимся к еще одной проблеме, создаваемой присутствием шума: возврату к среднему значению. Как только мы разберемся в причинах возврата к среднему значению, вы увидите, что этот эффект в сочетании с завышением значимости и занижением отчетности помогает объяснить то, что кажется поистине загадочным явлением, – тенденцию научных оценок сокращаться с течением времени.

Ключевые термины

- **Предвзятость публикации:** явление, при котором значимость публикуемых результатов систематически преувеличивается, поскольку существует выраженная склонность к публикации статистически значимых результатов.
- ***p*-хакинг**: поиск множества различных способов проведения эксперимента, сравнения или определения статистической модели до тех пор, пока не будет найден тот, который дает статистически значимый результат, а затем обнародование только этого результата.
- ***p*-скрининг**: социальный процесс, при котором сообщество исследователей, используя свои стандарты публикаций, отсеивает исследования с *p*-значениями, превышающими определенный порог, что приводит к предвзятости публикаций.

Упражнения

7.1. Вернемся к вопросу из главы 6 о рулетке Пита. Можете ли вы пересмотреть какие-либо свои советы или выводы в свете уроков, извлеченных из этой главы?

7.2. В конце апреля 2020 г. Национальный институт здравоохранения объявил результаты исследования по использованию препарата «Ремдесивир» для лечения COVID-19. Некоторым пациентам с COVID-19 случайным образом давали «Ремдесивир»; другим давали плацебо. Исследование выявило статистически значимые доказательства того, что лечение «Ремдесивиром» сокращает время выздоровления, измеряемое количеством дней, которые потребовались пациенту для выписки из больницы после приема препарата. Исследование было двойным слепым (ни пациенты, ни врачи не знали, принимал ли испытуемый настоящий препарат или плацебо). Размер выборки был достаточно большим (сотни пациентов). Назначение лечения было чисто случайным.

   a) На основе знаний, полученных в этой главе, определите еще две части информации, которые помогут вам оценить, насколько вы должны быть уверены в эффективности «Ремдесивира».
   b) Оказывается, исследование было предварительно зарегистрировано. План предварительной регистрации определил 28 показателей, которые ученые собирались измерить. Как это меняет ваши представления о том, подтверждают ли результаты реальный эффект? Почему?
   c) В плане предварительной регистрации также был указан один показатель в качестве основного представляющего интерес результата. Основным интересующим результатом был тот, о котором сообщалось в отчете: сколько времени потребовалось, чтобы пациента выписали из больницы. Влияет ли это на ваш ответ на предыдущий вопрос? Обоснуйте свое мнение.
   d) Но постойте, в этой истории есть еще один последний поворот. План предварительной регистрации фактически был пересмотрен в ходе исследования. Оказывается, продолжительность госпитализации не указывалась в качестве основного показателя до пересмотра 16 апреля 2020 г. До этого основным показателем значилась оценка пациента по восьмибалльной шкале, измеряющей тяжесть заболевания. Это отражено в версии плана от 2 апреля 2020 г. В своем заявлении исследователи подчеркнули, что они не видели данных, полученных в результате исследования, до изменения основного показателя. Подумайте, как все это влияет на ваше мнение о результатах исследования.

7.3. Загрузите набор данных VoterSurveyData2016.csv и связанный с ним README.txt, описывающий переменные в этом наборе данных, по адресу *http://www.press.princeton.edu/thinking-clearly*. Предположим, мы хотим знать, повлияло ли предыдущее знакомство с Дональдом Трампом до того, как он стал политиком, на поведение избирателей на президентских выборах в США в 2016 г. Чтобы продемонстрировать влияние Трампа, в опросе людей спрашивали, смотрели ли они телешоу «Ученик», которое вел Трамп, и смотрели ли они фильм «Один дома 2» с Трампом в эпизодической роли.

a) Используя доступные данные, постарайтесь найти как минимум три интересные, статистически значимые взаимосвязи, позволяющие предположить, что предыдущее знакомство с Трампом повлияло на поведение избирателей на президентских выборах 2016 г.

   Если вам сложно найти статистически значимые взаимосвязи, подумайте обо всех фактах, которые вы можете проверить. Можете использовать просмотр шоу «Ученик», фильма «Один дома 2» или обоих в качестве меры предыдущего воздействия Трампа. В качестве интересующего вас результата вы можете использовать поддержку Трампа, поддержку Хиллари Клинтон или явку избирателей в 2016 г. Вы можете группировать данные, чтобы избирательно рассмотреть интересующие подгруппы избирателей (например, женщины, чернокожие, южане, богатые, молодые и т. д.).

b) Как только вы обнаружите три статистически значимые взаимосвязи, интерпретируйте их по существу и подумайте, что они означают. Узнали ли вы что-нибудь интересное об электоральном поведении американцев?

c) Данные, которые вы только что проанализировали, представляют собой реальные данные опроса, проведенного в рамках Совместного исследования выборов в конгресс 2016 г. Мы случайным образом выбрали тысячу респондентов и поделились с вами частью их ответов. Однако выше мы солгали, когда сказали, что респондентов спрашивали, смотрели ли они «Ученик» или «Один дома 2». Мы выдумали эти переменные. (Извините, мы больше не будем вам врать.) Более того, значения этих переменных были сгенерированы совершенно случайным образом. Объясните, почему вам тем не менее удалось обнаружить связь между этими переменными и политическим поведением. Ожидаете ли вы, что эта связь сохранится и дальше, если мы предоставим данные еще о тысяче респондентов и снова случайным образом сгенерируем данные о воздействии?

7.4. Найдите недавно опубликованное научное исследование, в отношении которого вас беспокоят проблемы завышения значимости и занижения отчетности. Объясните свои опасения.

   Могут ли авторы сделать что-нибудь, чтобы устранить или смягчить ваши опасения, не собирая дополнительных данных? Есть ли дополнительная информация, которую вы хотели бы получить от авторов? Какие дополнительные анализы, по вашему мнению, должны провести авторы?

Дополнительное чтение и ссылки

Подробнее о *p*-хакинге см.:

Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn. 2011. *False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant.* Psychological Science 22 (11): 1359–66;

Uri Simonsohn, Joseph P. Simmons, and Leif D. Nelson. 2014. *P-Curve: A Key to the File-Drawer.* Journal of Experimental Psychology 143 (2): 534–47.

Чтобы просмотреть исследование ESP и прочитать о некоторых неудачных репликациях, см.:

Daryl J. Bem. 2011. *Feeling the Future: Experimental Evidence for Anomalous Retroactive Influence on Cognition and Affect.* Journal of Personality and Social Psychology 100 (3): 407–25;

Jeff Galak, Robyn A. LeBoeuf, Leif D. Nelson, and Joseph P. Simmons. 2012. *Correcting the Past: Failures to Replicate Psi.* Journal of Personality and Social Psychology 130 (6): 933–48.

Дополнительная информация о *p*-скрининге в контексте исследований посещаемости выборов представлена здесь:

Donald P. Green,Mary C.McGrath, and Peter M. Aronow. 2013. *Field Experiments and the Study of Voter Turnout.* Journal of Elections, PublicOpinion and Parties 23 (1): 27–48.

Письмо 72 исследователей о необходимости снижении порога статистической значимости:

Benjamin, Daniel J., et al. 2017. *Redefine Statistical Significance.* Nature Human Behavior 2: 6–10.

Исследование частоты нулевых результатов в исследованиях NHLBI до и после предварительной регистрации:

Robert M. Kaplan and Veronica L. Irvin. 2015. *Likelihood of Null Effects of Large NHLBI Clinical Trials Has Increased over Time.* PLoS One 10 (8).

Чтобы узнать больше о том, влияют ли результаты студенческого футбола на выборы, см.:

Andrew J. Healy, Neil Malhotra, and Cecilia Hyunjung Mo. 2010. *Irrelevant Events Affect Voters' Evaluations of Government Performance.* Proceedings of the National Academy of Sciences 107 (29): 12804–09;

Anthony Fowler and B. Pablo Montagnes. 2015. *College Football, Elections, and False-Positive Results in Observational Research.* Proceedings of the National Academy of Sciences 112 (45): 13800–04.

В блоге Эндрю Гельмана есть хорошее обсуждение всего эпизода с позой власти, где часто освещаются вопросы, связанные с предвзятостью публикаций: *https://statmodeling.stat.columbia.edu/2017/10/18/beyond-power-pose-using-replication-failures-better-understanding-data-collection-analysis-better-science/*.

Оригинальное исследование «позы власти» и первая попытка репликации:

Dana R. Carney, Amy J. C. Cuddy, and Andy J. Yap. 2010. *Power Posing: Brief Nonverbal Displays Affect Neuroendocrine Levels and Risk Tolerance.* Psychological Science 21 (10): 1363–68;

Eva Ranehill, Anna Dreber, Magnus Johannesson, Susanne Leiberg, Sunhae Sul, and Roberto A. Weber. 2015. *Assessing the Robustness of Power Posing: No Effect on Hormones and Risk Tolerance in a Large Sample of Men andWomen.* Psychological Science 26 (5): 653–56.

Опровержение результатов исследования со стороны Карни можно найти по адресу *http://faculty.haas.berkeley.edu/dana_carney/pdf_my%20position%20on%20power%20poses.pdf*.

Если вы хотите просмотреть полную историю изменений плана предварительного анализа для исследования «Ремдесивира», упомянутого в упражнении 2, можете найти ее по адресу *https://clinicaltrials.gov/ct2/history/NCT04280705*.

Глава 8

Возврат к среднему значению

О чем эта глава

- Многие показатели имеют тенденцию возвращаться к среднему значению, а это означает, что за экстремальными наблюдениями часто следуют вполне заурядные.
- Это явление свойственно практически любому эффекту, который является функцией как сигнала (т. е. чего-то реального в мире), так и шума.
- Если вы не учитываете возможность возврата к среднему значению, легко неверно истолковать данные.
- Не следует ожидать возврата к среднему значению показателей, которые отражают наши убеждения о будущем, таких как прогнозы выборов или цены на акции.

Введение

Как подчеркивается в нашем любимом уравнении, мир зашумлен, и большинство количественных измерений отражают как измеряемую величину, так и шум. Это имеет множество последствий для того, как мы изучаем и понимаем мир.

Одним из наиболее распространенных, но наименее понятных последствий жизни в зашумленном мире является *возврат к среднему значению*. Грубо говоря, за необычно большими или маленькими измерениями обычно следуют (и предшествуют) измерения, которые ближе к среднему значению.

Хотя возврат к среднему значению не является стандартной темой в книгах по количественному анализу данных, его повсеместное распространение означает, что количественная информация часто будет вводить вас в заблуждение, если вы не понимаете это явление. Поэтому мы считаем, что важно разобраться в его природе.

Исчезает ли истина?

В главе 7 мы говорили о том, что из-за завышения значимости и занижения отчетности нам часто приходится скептически относиться к новым, неожиданным научным открытиям. Поэтому, когда исследователи впервые сообщают о выдающемся достижении, часто первый вопрос, который им задают, звучит так: «Повторяются ли наблюдения?» То есть, если мы проведем новое независимое исследование, разработанное аналогично первоначальному исследованию, обнаружим ли мы аналогичный эффект? В этом случае наши сомнения

понятны: мы обеспокоены тем, что опубликованные результаты отражают превратности случая, а не реальные явления в мире. Прежде чем рукоплескать выдающимся достижениям, мы хотим убедиться, что они подтверждаются многочисленными исследованиями. И в самом деле, ученым частенько не удается воспроизвести разрекламированные результаты в последующих исследованиях. В некоторых областях науки эта ситуация встречается настолько часто, что ученые начали говорить о *кризисе повторяемости*, подрывающем доверие к результатам научных исследований.

Джонатан Скулер, известный психолог из Калифорнийского университета в Санта-Барбаре, заметил любопытную закономерность сбоев повторяемости в некоторых из своих самых влиятельных исследований. Дело в том, что эффекты, обнаруженные Скулером в одном из исследований, обычно не исчезали полностью при воспроизведении, но они систематически уменьшались. Скулер обнаружил, что многие коллеги испытали то же самое; повторные результаты часто были хуже первоначальных.

Одним из возможных объяснений этой закономерности является то, что после проведения исследования испытуемые узнают о результатах и меняют свое поведение. Этот феномен, когда испытуемые меняют свое поведение, потому что они знают, что их изучают, иногда называют *эффектом Хоторна*[1]. Другой термин – *эффект опроса* – относится к ситуациям, когда испытуемые ведут себя по-другому, потому что они знают, чего ждут экспериментаторы, и пытаются им угодить.

Скулер и его коллеги быстро исключили эффект Хоторна и другие подобные объяснения, потому что они обнаружили ту же закономерность в исследованиях птиц, которые, по-видимому, понятия не имеют, что их изучают, и не заботятся о том, чтобы угодить исследователям-людям. Так чем же еще можно объяснить своеобразную картину исчезновения эффектов?

Скулер (возможно, в шутку) начал называть это явление *космическим привыканием*. Он задался вопросом, не существует ли во Вселенной какая-то неизвестная сила, которая заставляет эффекты уменьшаться каждый раз, когда их изучают. Одна из аналогий, которую он приводит, связана с привыканием человеческого чувственного восприятия. Когда что-то впервые касается вашей руки, вы остро это ощущаете. Однако со временем вы привыкаете, и ощущение прикосновения уменьшается. Возможно, Вселенная устроена именно так. Когда мы впервые наблюдаем какое-то явление, возникает острый эффект. Но со временем Вселенная привыкает к нашим исследованиям, и мы наблюдаем этот эффект все меньше и меньше. Другими словами, ученые фактически изменяют реальность каждый раз, когда изучают ее. Звучит пугающе.

Теория Скулера о космическом привыкании привлекла большое внимание средств массовой информации, включая эпизод популярного радиошоу и под-

[1] Интересный факт: термин «эффект Хоторна» появился в результате исследования взаимосвязи между условиями труда и производительностью на фабрике Hawthorne Works под Чикаго. Но впоследствии оказалось, что данные были проанализированы плохо. Экономисты Стивен Левитт и Джон Лист повторно проанализировали исходные данные и показали, что явление, которое первоначальным исследователям казалось эффектом Хоторна, скорее всего, связано с другими факторами, такими как различия поведения в разные дни недели, а не с тем, что испытуемые меняют свое поведение в ответ на изучение.

каста Radio Lab, а также статью в журнале New Yorker под названием «Истина исчезает». Но, прежде чем вслед за Скулером выдвигать гипотезы о новых космических силах, давайте посмотрим, может ли критическое мышление помочь нам решить загадку уменьшения эффекта без привлечения мистических сил.

Фрэнсис Гальтон и возврат к среднему

Как мы упоминали в главе 5, Фрэнсис Гальтон сделал столь же странное открытие в 1860-х гг. Он собрал данные о росте родителей и их детей и проделал то же самое с растениями, собрав данные о размере и весе семян душистого горошка у родителей и потомков.

Гальтон нарисовал диаграммы рассеяния этих данных, поместив размеры родителей на горизонтальную ось, а потомков – на вертикальную. Затем он построил линию регрессии по данным. Вы можете увидеть один из графиков Гальтона для роста родителей (с поправкой на биологический пол) и их детей на рис. 8.1.

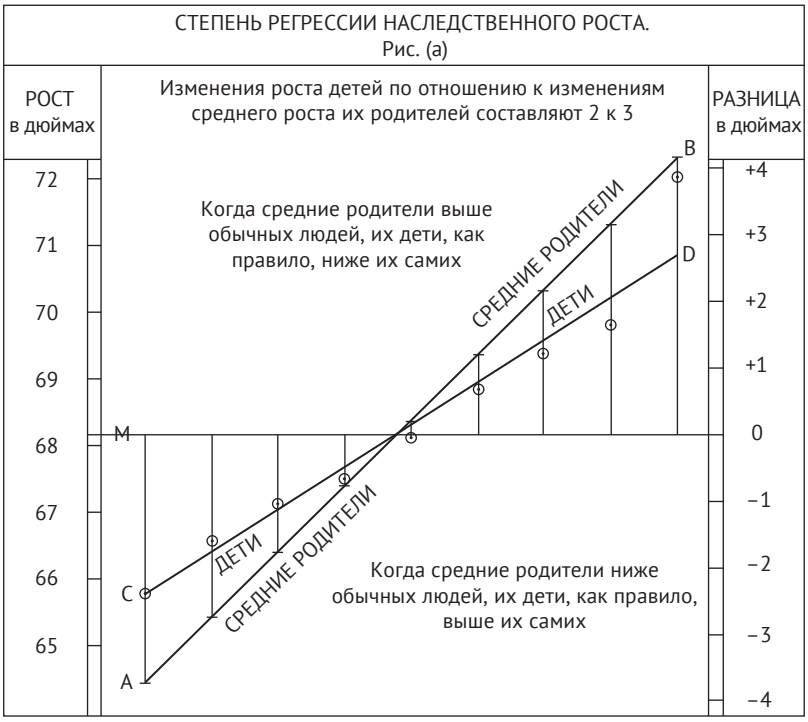


Рис. 8.1. Репродукция иллюстрации Гальтона о возврате к среднему значению

Первоначально Гальтон ожидал, что линия регрессии будет иметь угол 45°, т. е. ее точка пересечения будет равна 0, а наклон 1. Это кажется разумным предположением. Оно было бы верным, если бы в среднем дети были того же роста, что и их родители (опять же, с поправкой на биологический пол).

Однако оказывается, что это предположение неверно ни для людей, ни для душистого горошка. На рисунке Гальтона линия под углом 45° обозначена надписью «Средние родители». Это измеренный Гальтоном средний рост роди-

телей ребенка после предварительной корректировки, чтобы поместить рост женщин и мужчин на одну и ту же шкалу. Горизонтальная ось (без обозначения) соответствует росту родителей. Тогда линия, обозначающая рост родителей, должна проходить под углом 45°. Линия с надписью «Дети» – это линия регрессии, проходящая через данные, которые характеризуются двумя значениями: *x* – рост родителей, а *y* – рост детей.

Как вы можете видеть на этом рисунке, регрессия Гальтона пересекается с осью *y* в области положительных значений, т. е. при самом низком росте родителей линия регрессии лежит выше линии, проведенной под углом 45°. Кроме того, линия регрессии Гальтона имеет наклон, который положителен, но явно меньше 1 – она растет медленнее, чем линия под углом 45°.

Что говорит эта регрессия о взаимосвязи между ростом родителей и ростом детей? Тот факт, что наклон положительный, означает, что в среднем чем выше родители, тем выше их дети. Как мы видим на рисунке, тот факт, что точка пересечения оси *y* положительна, означает, что у особенно низких родителей дети, как правило, выше их самих. На левой стороне горизонтальной оси (где родители невысокого роста) линия регрессии лежит выше линии под углом 45°. Но, поскольку ее наклон меньше 1, линия регрессии увеличивается медленнее, чем линия под углом 45°. И действительно, две линии пересекаются посередине. Таким образом, на правой стороне горизонтальной оси (где родители высокие) линия регрессии лежит ниже линии под углом 45°. У самых высоких родителей дети обычно ниже их самих.

Как мы упоминали в главе 5, Гальтон называл это явление «регрессией к посредственности». Благодаря Гальтону мы теперь используем слово «регрессия» для обозначения практики сопоставления линий с данными. Именно поэтому некоторые люди называют явление, при котором показатели склонны возвращаться к среднему значению (что является предметом этой главы), *регрессией к среднему значению* (regression to the mean). Однако, чтобы не путать эти два понятия, мы будем называть последнее явление иначе – *возвратом* к среднему значению (reversion to the mean).

Открытия Гальтона во многом напоминают «космическое привыкание» Скулера. Возможно, во Вселенной существует какая-то невидимая сила, которая приближает рост членов семьи к среднему значению. Возможно, когда Вселенная видит слишком высоких родителей или слишком большие семена душистого горошка, она восстанавливает порядок, уменьшая их потомство. Или, возможно, когда Гальтон измерил высоту родителей и диаметр семян душистого горошка, он каким-то образом заставил их потомство уменьшиться! Гальтон, вероятно, был озадачен своими неожиданными открытиями. Но он не спешил делать сверхъестественные выводы.

В конце концов Гальтон понял, что происходит. Размер определяется многими факторами. Чтобы как можно яснее и проще представить идею Гальтона, давайте подумаем о семенах душистого горошка. И давайте представим, что на размер семени душистого горошка влияют всего две вещи: (1) гены, которые оно унаследовало от своего родителя, и (2) количество прямого солнечного света, которое оно получало во время плодоношения. Наследование генов размера делает семя крупнее при прочих равных условиях. Аналогично получение большего количества солнечного света делает семя крупнее при прочих равных условиях.

В рамках этой простой модели давайте поразмыслим о том, как семя может оказаться особенно большим или особенно маленьким. Предположим, вы нашли очень большое семя душистого горошка. Оно могло стать большим, потому что унаследовало гены большого размера от своих предков. Оно также могло стать большим, потому что выросло в необычайно солнечном году. Или это может быть комбинация того и другого. Скорее всего, если семя *действительно* большое, в его пользу сработали оба фактора: родитель с генами большого размера и отличное солнце.

Так чего же нам ожидать, если эти крупные семена будут посажены и дадут собственное потомство? Они передадут свои гены размера, превышающего средний. Но, скорее всего, растение-потомок не получит такое же количество солнечного света, как его предок. В среднем растения будут расти на солнце средней интенсивности. Таким образом, у второго поколения семена будут крупнее среднего из-за унаследованных генов. Но, скорее всего, они будут меньше родительских, потому что предку особенно повезло с пребыванием на солнце, а потомку – нет. То же самое справедливо и для очень маленьких семян. Они получают гены своих родителей, отвечающие за небольшой размер. Но следующее поколение горошка, скорее всего, получит больше солнечного света, чем предок, и, следовательно, семена будут крупнее родительских.

Итак, если эта простая модель верна, то мы будем наблюдать явление, открытое Гальтоном. Более крупные растения, как правило, имеют более крупных потомков (линия регрессии имеет положительный наклон). Но размер возвращается к среднему значению: у действительно низких родителей, как правило, дети ниже среднего, но выше родителей, а у действительно высоких родителей дети, как правило, выше среднего, но ниже родителей (наклон линии регрессии меньше 1).

Очевидно, что на самом деле ситуация с размером семян сложнее. На него влияют многие вещи, помимо генов и солнца. Но пример показывает суть. Нам следует ожидать возврат к среднему значению, если размер частично определяется генами, которые систематически передаются от родителя к потомку, а частично определяются несистематическими или случайными факторами, которые не коррелируют между поколениями (например, воздействием солнца). То же самое касается человеческого роста, как мы видим на графике Гальтона.

В более общем смысле у любого показателя, который частично является функцией систематических факторов (иногда их называют сигналом), а частично функцией случайных или несистематических факторов (которые мы иногда называем шумом), произойдет возврат к среднему значению. Представьте себе многократно наблюдаемый исход, где при каждом наблюдении результат отражает комбинацию систематического сигнала (например, генов) и случайного шума (например, солнечного света). Экстремальные результаты обычно возникают из-за совпадения экстремальных значений как сигнала, так и шума. В других итерациях, пока сигнал остается фиксированным, шум принимает новое случайное значение. И ожидаемо значение шума будет средним. Таким образом, ожидается, что экстремальные значения в одной итерации вернутся к среднему значению в других итерациях.

Многие явления в мире имеют подобную структуру «сигнал–шум». Поэтому явление возврата к среднему значению встречается повсеместно. Следовательно, критическое отношение к доказательствам заставляет нас помнить о возможности возврата к среднему значению. Далее мы сначала углубимся в природу возвра-

та к среднему значению, чтобы убедиться в точном понимании происходящего. Затем рассмотрим различные примеры из реального мира, чтобы понять, когда нам следует и не следует ожидать возникновения возврата к среднему значению.

Возврат к среднему значению не является силой притяжения

Одно из распространенных заблуждений относительно возврата к среднему значению заключается в том, что его считают чем-то наподобие земного притяжения, т. е. что мир полон отклонений и что со временем все показатели неизбежно притягиваются к среднему значению. Это неправильно.

Чтобы проверить себя, попробуйте ответить на следующие вопросы:

1) Джон-младший исключительно высокий. Если бы вам пришлось угадывать, вы бы предположили, что сын Джона-младшего, Джон 3-й:
   a) ниже Джона-младшего?
   b) того же роста, что и Джон-младший?
   c) выше Джона-младшего?

2) Джон-младший исключительно высокий. Если бы вам нужно было угадать, вы бы предположили, что отец Джона-младшего, Джон-старший:
   a) ниже Джона-младшего?
   b) того же роста, что и Джон-младший?
   c) выше Джона-младшего?

Прежде чем мы перейдем к ответам, давайте начнем с того, что на оба вопроса вы должны были дать один и тот же ответ. Важно понять почему.

Для многих людей, как только они узнают о возврате к среднему значению, вопрос 1 становится интуитивно понятным. Джон 3-й, вероятно, ниже Джона-младшего. Джон-младший особенно высок. Так что у него, вероятно, есть гены высокого роста (сигнал). И еще, вероятно, с ним произошли какие-то особые вещи, из-за которых он стал особенно высоким (шум). Его сын, Джон 3-й, вероятно, унаследует его гены высокого роста. Но вы можете предположить, что прочие специфические факторы будут менее экстремальными. Это логика возврата к среднему значению, как объяснил Гальтон.

Но, по нашему опыту, вопрос 2 часто вызывает некоторые затруднения. Вы можете начать рассуждать следующим образом. Джон-младший очень высокий. В нашем мире происходит возврат к среднему значению. Так что рост Джона-младшего, вероятно, ближе к среднему, чем у его отца.

Если воспроизвести возврат к среднему значению в обратную сторону, раз Джон-младший был очень высоким, его отец, должно быть, был настоящим гигантом! Следовательно, можно предположить, что, хотя ответом на вопрос 1 является (a), ответом на вопрос 2 является (c).

Ничего страшного, если вы рассуждали в этом направлении. Аргумент имеет определенную привлекательность. Но он ошибочный, и важно понять почему. На вопросы 1 и 2 следует дать ответ (a). И логика для отца Джона-младшего идентична логике для сына Джона-младшего: это прежняя логика возврата к среднему значению. Вот как это происходит, следите внимательно.

Предположим, вы наблюдаете некоторый результат, состоящий из сигнала и шума, и этот результат на удивление велик. (Этот аргумент, конечно, работает и для удивительно малых значений.) Затем предположим, что вы соби-

раетесь наблюдать другой результат, где сигнал такой же, как и при первом наблюдении, но будет новый, независимый выброс шума. Поскольку первое наблюдение очень велико, оно, вероятно, отражает большую величину сигнала и большую величину шума. Новое наблюдение снова имеет большое значение сигнала, но значение шума, вероятно, будет меньше. Так что новое наблюдение, скорее всего, будет меньше.

Важно отметить, что, приводя этот аргумент, мы ничего не сказали о том, какой результат возник первым во времени. Мы говорили только о порядке, в котором вы их наблюдали. Возврат к среднему значению – это не сила гравитации, которая с течением времени притягивает вещи к среднему значению. Для логики возврата к среднему значению абсолютно не важно, что во времени произошло первым. Итак, если Джон-младший очень высокий, а у его сына такой же сигнал (гены) и независимый шум, то его сын, вероятно, ниже его ростом. И если Джон-младший очень высокий, а у его отца такой же сигнал (гены) и независимый шум, то его отец, вероятно, тоже ниже его ростом.

Проще всего это увидеть в реальном мире на данных спортивных соревнований, где мы наблюдаем, как один и тот же спортсмен снова и снова выполняет одно и то же упражнение. В этих данных хорошо видно, что возвращение к среднему значению характеризует данные, перемещающиеся вперед и назад во времени.

На рис. 8.2 представлена диаграмма распределения результатов первых двух раундов Открытого чемпионата США по гольфу среди женщин 2019 г. Очки игроков в первом раунде отложены по горизонтальной оси, а во втором раунде – по вертикальной. Что вы здесь видите?

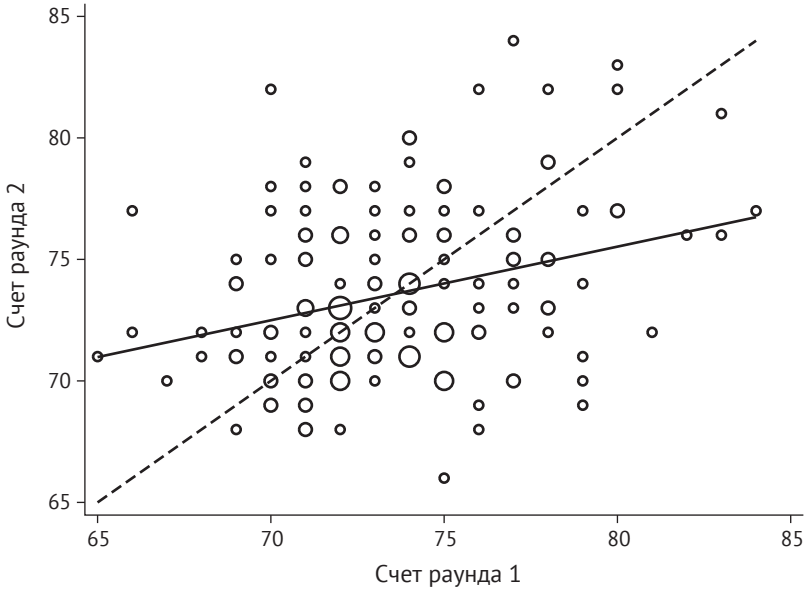


Рис. 8.2. Результаты в раундах Открытого чемпионата США по гольфу среди женщин 2019 г. В тех случаях, когда несколько игроков набрали одинаковые очки, размер круга увеличивается, чтобы отразить количество игроков. Линия под углом 45° пунктирная, а фактическая линия регрессии OLS – сплошная

В среднем существует положительная корреляция между результатами в двух раундах; линия регрессии наклонена вверх. Игроки, показавшие лучшие результаты (в гольфе чем меньше очков, тем лучше) в первом раунде, также имели тенденцию добиваться лучших результатов и во втором раунде. Это логично: некоторые игроки играют лучше других (сигнал). Но наклон линии регрессии намного меньше 1; линия регрессии ниже, чем линия под углом 45°. Если результат игрока в 1-м раунде был хуже среднего, его результат во 2-м раунде, как правило, был лучше, чем в 1-м раунде. И если результат игрока в 1-м раунде был лучше среднего, его результат во 2-м раунде, как правило, был хуже, чем в 1-м раунде. Это явление из той же серии, что обнаруженная Гальтоном закономерность с ростом родителей и детей или размером семян горошка. И мы можем заверить вас, что не выбрали этот пример специально. Вы увидите ту же самую картину на любом турнире по гольфу.

Комментатор гольфа может посмотреть на эти данные и рассказать историю, объясняющую результаты. Возможно, игроки, которые действительно хорошо провели первый раунд, устали от напряжения. Они выдохлись. И это объясняет, почему во втором раунде они выступили хуже. И возможно, игроки, у которых был плохой первый раунд, поняли, что им нужно менять стратегию или по-настоящему сосредоточиться. И, соответственно, улучшили свои показатели.

Такое возможно. Но это также может быть просто возврат к среднему значению. Результаты в гольфе зависят как от навыков (сигнала), так и от удачи (шума). Игроки, набравшие лучший результат в определенном раунде, вероятно, лучше, чем средний игрок на поле. Но, скорее всего, им также сопутствовало везение. Было сделано несколько удачных ударов, которые с такой же легкостью могли бы оказаться неудачными. Их несколько плохих бросков были удачно скомпенсированы, что уберегло их от неприятностей. И т. д. В других раундах они по-прежнему играют лучше среднего, поэтому мы ожидаем, что их результаты будут выше среднего. Но их везение вряд ли продолжится, поэтому их результат, вероятно, будет хуже, чем в предыдущем экстремальном раунде.

Обратите внимание: в логике предыдущего абзаца ничто не зависело от того, какой раунд наступит первым. Это потому, что возврат к среднему значению – это не какая-то сила, притягивающая вещи к среднему значению с течением времени. Осознание этого позволяет нам выяснить, какая история – рассказ комментатора или возвращение к среднему – с большей вероятностью будет правдой.

Представьте игрока, который показал особенно хороший результат во втором раунде. Какой результат вы должны предположить для первого раунда – лучше или хуже второго? Прежняя логика состояла в том, чтобы предположить, что из-за возврата к среднему значению для того, чтобы игрок мог получить хороший результат во втором раунде, он должен был показать еще более хороший результат в первом раунде, что позволяло ему сохранить хороший результат во втором раунде, несмотря на возврат к среднему значению. Но теперь вы более искушенный аналитик. Логика возврата к среднему значению не имеет ничего общего со временем. Счет в каждом раунде игры в гольф представляет собой комбинацию сигнала и шума. Если у игрока в какой-то момент выдался особенно хороший раунд, мы должны ожидать, что *другой* раунд этого игрока (с тем же сигналом, но другим шумом) будет хуже независимо от того, какой раунд был первым. И если в какой-то момент у игрока выдался особенно пло-

хой раунд, мы должны ожидать, что *другой* раунд этого игрока будет лучше независимо от того, какой раунд был первым.

На рис. 8.3 показан тот же график, который вы видели раньше, но оси перевернуты так, что раунд 2 находится на горизонтальной оси, а раунд 1 – на вертикальной. Общая картина практически не изменилась. Люди, которые имели особенно хорошие результаты во втором раунде, были лучше среднего в раунде 1, но все же хуже в раунде 1, чем в раунде 2. Как и в случае с Джоном-младшим и Джоном-старшим, возврат к среднему значению работает назад во времени так же хорошо, как и вперед.

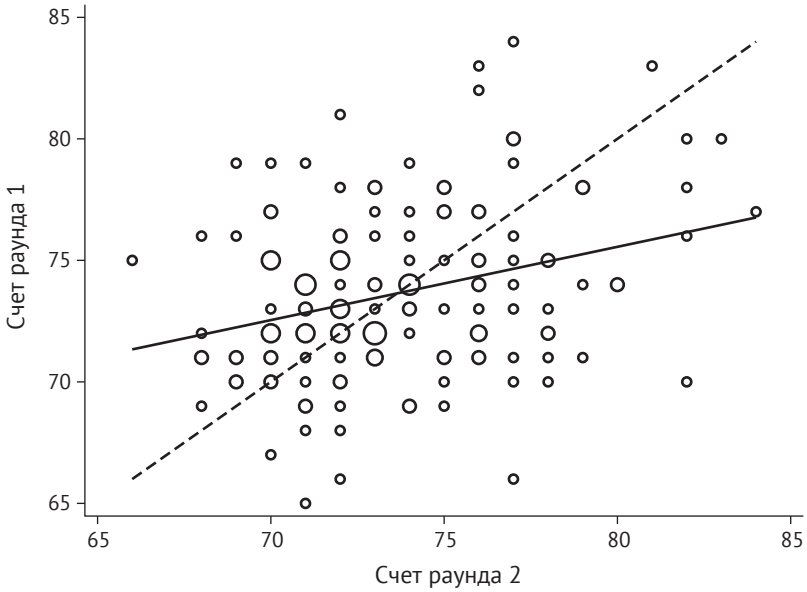


Рис. 8.3. Результаты в раундах Открытого чемпионата США по гольфу среди женщин 2019 г. с перевернутыми осями

Анализ обеих версий графика показывает, почему объяснение комментатора вряд ли верно. Трудно поверить, что хороший результат во втором раунде заставляет игроков чувствовать дополнительное влияние, которое каким-то образом распространяется назад во времени, ухудшая их результат в первом раунде. И тем не менее мы видим одну и ту же картину независимо от того, смотрим ли мы вперед или назад во времени. Объяснение – возврат к среднему значению.

Поиск помощи

Возврат к среднему значению особенно трудно поддается критическому мышлению в условиях, когда мы обращаемся за помощью, если что-то неожиданно пошло неправильно (например, вы внезапно заболели или плохо сдали экзамен). Почему?

Если что-то неожиданно пошло неправильно, это говорит о том, что у нас есть какое-то основополагающее ожидание того, как все должно идти – воз-

можно, сформированное на основе длительного прошлого опыта, – и что мы отклонились от этого ожидания в плохом направлении. Ожидание того, каким должен быть нормальный ход событий, представляет собой сигнал. А отклонения от этих ожиданий представляют собой шум.

Давайте рассмотрим пару конкретных ситуаций.

Предположим, вы относительно здоровый человек. Вы считаете, что в вашем случае хорошее здоровье отражает истинный сигнал. Даже самые здоровые люди время от времени чувствуют себя плохо – из-за гриппа, боли в спине или чего-то еще – по несистематическим причинам, которые, возможно, не отражают каких-либо фундаментальных изменений в их базовом состоянии здоровья. И у большинства здоровых людей бывают дни, когда они чувствуют себя особенно бодрыми и здоровыми. Таким образом, мы можем считать день, когда вы чувствуете себя хорошо, отражением вашего истинного сигнала с очень небольшим шумом. Дни, когда вы чувствуете себя плохо, можно рассматривать как дни с особенно большими отрицательными значениями шума. А дни, когда вы чувствуете, что можете выйти и покорить мир, можно рассматривать как дни с особенно большими положительными значениями шума.

Или, возможно, нашим читателям будет более близок пример с учебой. У вас есть некоторый базовый уровень академических знаний, который отражает, насколько вы сильный ученик в той или иной области. Это ваш базовый сигнал. Но в некоторые дни вы справляетесь с тестом намного лучше, чем обычно, – возможно, этому способствовал удачный набор вопросов или особенно хороший ночной сон. А в некоторые дни вы справляетесь с тестом намного хуже, чем обычно, возможно, из-за неудачного набора вопросов или компьютерных игр до поздней ночи. Эти несистематические факторы представляют собой шум.

Но как эти примеры связаны с возвратом к среднему значению? Спросите себя: в какие дни человек может обратиться за помощью, скажем, к мануальному терапевту? Вероятно, в те дни, когда он просыпается с болью в спине, которая ощущается сильнее, чем обычно, т. е. в те дни, когда шум особенно неприятен. И спросите себя: какие студенты обращаются за помощью к репетиторам по подготовке к экзаменам? Вероятно, студенты, чьи результаты по важному тесту оказались хуже, чем они ожидали, учитывая их понимание своих базовых способностей. Если это так, то принцип возвращения к среднему значению говорит нам, что, даже если массаж спины или занятия с репетитором на самом деле вообще не помогают, мы все равно должны ожидать, что люди, которые обращаются за такой помощью, увидят улучшения. И если они не задумываются о возвращении к среднему значению, они, скорее всего, будут слишком признательны мануальному терапевту или репетитору. Возвращение к среднему значению может стать хорошей бизнес-моделью.

Проблема заблуждений такого рода широко распространена. Мы уже видели пример в главе 1, где обсуждали теорию разбитых окон. Напомним, что в Нью-Йорке в рамках новой стратегии полиция сосредоточила внимание на мелких правонарушениях на участках с самым высоким уровнем преступности и обнаружила, что после применения стратегии преступность на этих участках снизилась. Но это именно то, чего мы ожидаем от возврата к среднему значению, даже если теория разбитых окон вообще не работает. На участках с самым высоким уровнем преступности ситуация будет улучшаться, а на участках

с наименьшим уровнем преступности ситуация будет ухудшаться независимо от каких-либо политических изменений. Из-за этого возврата к среднему значению действия полиции, нацеленные на участки с самым высоким уровнем преступности, будут выглядеть для наивного наблюдателя так, будто они действительно эффективны, даже если это не так.

Еще один возврат к среднему значению скрывался под другим примером из главы 1. Помните, сыну Итана врачи рекомендовали попробовать безглютеновую диету, потому что у него был недостаточный вес? Их идея заключалась в том, что, если ребенок начнет набирать вес после перехода на безглютеновую диету, это будет свидетельством непереносимости глютена. Но возврат к среднему значению говорит о том, что мы могли бы ожидать увеличение веса Эйба даже без диеты. Из месяца в месяц вес ребенка является функцией как сигналов (например, здоровья, генетики), так и шума (например, случайных особенностей окружающей среды, особенностей траектории роста). Если у ребенка за месяц наблюдался слишком низкий вес, вероятно, в этом месяце на него повлияли чрезвычайно отрицательные факторы шума. Со временем значения шума изменятся в противоположную сторону, поэтому вес начнет расти. Было бы ошибкой интерпретировать это как свидетельство того, что некоторые изменения в поведении (например, отказ от глютена) объясняют увеличение веса.

Рассматривая события с этой точки зрения, вы увидите, что многие виды воздействия будут выглядеть так, будто они работают, даже если они ничего не дают. Обычно люди обращаются за помощью, когда дела идут хуже всего. Можно ожидать, что со временем ситуация улучшится даже без помощи со стороны из-за возврата к среднему значению. Давайте рассмотрим один особенно яркий пример, который был тщательно исследован учеными.

Работает ли операция на колене?

Существует множество дорогостоящих методов лечения, доказательства эффективности которых мало отличаются от доказательств эффективности работы полиции в теории разбитых окон. Например, не существует рандомизированных исследований, подтверждающих эффективность различных видов хирургического вмешательства. Представьте себе пациента, который обращается к хирургу с болью в суставе. Врач рекомендует операцию. По окончании восстановительного периода пациент говорит, что чувствует себя лучше. Врач может опираться на обширные знания об анатомии и физиологии тела, которые дают веские основания полагать, что операция действительно помогла. Но мы должны, по крайней мере теоретически, допустить возможность того, что здесь параллельно сработал возврат к среднему значению, т. е. что многие пациенты испытали бы ощутимое улучшение без хирургического вмешательства.

Действительно, после проведения рандомизированного исследования исследователи иногда обнаруживают, что обычная операция на самом деле не дает предполагаемого положительного эффекта. Например, в исследовании 2002 г. артроскопической хирургии остеоартрита коленного сустава исследователи обнаружили, что обычно назначаемые операции не оказали заметного влияния на боль в колене. Да, пациенты сообщали об уменьшении боли в коленях через две недели после операции, но другие пациенты, которым просто сделали разрезы на коже и сказали, что им сделали операцию, сообщили

о таком же уменьшении боли в коленях. Врачи на самом деле делали ложные операции некоторым своим пациентам, и эти обманутые пациенты чувствовали себя как минимум не хуже, чем тем, кому сделали настоящую операцию. Почему все пациенты почувствовали себя лучше? Можно предположить, что вы соглашаетесь на операцию, только когда испытываете особенно сильную боль, поэтому у большинства пациентов наступает облегчение через несколько недель даже без операции.

Возвращение к среднему, эффект плацебо и космическое привыкание

Вы только что убедились, что без ясного понимания эффекта возврата к среднему значению очень трудно правильно истолковать степень, в которой внешние воздействия, включая медицинские манипуляции, на самом деле способствуют улучшению результатов. Но оказывается, что возврат к среднему значению может создать проблемы, даже когда мы проводим тщательные научные исследования. Давайте посмотрим, как это происходит в нескольких разных ситуациях. Сначала мы рассмотрим широко известный эффект плацебо. Затем вернемся к проблеме «космического привыкания», с которой мы начали эту главу.

Эффект плацебо

Немногие явления в медицине упоминаются чаще, чем так называемый *эффект плацебо*. Многие люди подозревают, что вера в лечебное воздействие каким-то образом активирует собственные целебные силы организма, независимо от прямого эффекта самого лечения. По этой причине медицинские исследователи стараются сравнивать эффективность новых лекарств или методов лечения с плацебо. Они хотят учесть возможность того, что вера в лечение исцелит вас сама по себе. Поэтому они используют такие вещи, как сахарные таблетки или фальшивые операции, чтобы подопытные не знали, получают ли они настоящее лечение или нет.

Почему исследователи-медики и другие ученые считают, что существует эффект плацебо? Одним из источников доказательств являются сами медицинские испытания. В таких экспериментах есть подопытная группа, получающая препарат, и контрольная группа, получающая таблетку плацебо. И часто в таких исследованиях улучшается здоровье обеих групп. Улучшение в контрольной группе рассматривается как доказательство эффекта плацебо.

Но теперь вы можете догадаться, что такого рода доказательства эффекта плацебо неубедительны. Люди из контрольной группы (и подопытной группы) приняли участие в медицинском исследовании, потому что они плохо себя чувствовали. Можно ожидать, что им в какой-то степени полегчает даже при отсутствии лечения, и не обязательно потому, что их разум исцеляет их тела. Это может быть просто возврат к среднему значению.

Если вы действительно хотите проверить эффект плацебо, нужно разделить группу подопытных на подгруппу, которая получала таблетку плацебо, и подгруппу, которая не получала вообще ничего. (Конечно, у вас также может быть группа, которая получила настоящее лекарство. Давайте не будем забывать,

ради чего мы этим занимаемся.) Лишь немногие исследования непосредственно проверяют эффект плацебо по сравнению с отсутствием воздействия. Те, кто это делает, обычно не обнаруживают никаких доказательств эффекта плацебо. Более того, исследования, которые действительно обнаруживают доказательства эффекта плацебо, обычно сфокусированы на чисто субъективных показателях. Люди могут чувствовать себя лучше после приема плацебо, даже если они не стали объективно здоровее.

Например, в 2011 г. группа исследователей из Гарвардской медицинской школы опубликовала в New England Journal of Medicine статью, в которой сравнивались эффекты настоящего лечения, приема плацебо и отсутствия лечения у пациентов с астмой. Интересно, что как настоящее лечение (ингаляция альбутерола), так и лечение плацебо (ингалятор плацебо или иглоукалывание) привели к тому, что пациенты стали чувствовать себя лучше. Но когда ученые измерили объем легких испытуемых, эффект был обнаружен только после настоящего лечения; прием плацебо был не полезнее, чем полное отсутствие воздействия. Если и существует доказательство эффекта плацебо, то это свидетельство самообмана разума, а не победы разума над материей.

Короче говоря, стоит критически подумать о возврате к среднему значению, и мы увидим, что в медицине мало убедительных доказательств эффекта плацебо. Тем не менее почему-то почти все верят в эффект плацебо – наверное, потому что плохо умеют распознавать и правильно интерпретировать возврат к среднему значению. Даже некоторые из величайших исследователей-медиков стали жертвами этой путаницы. Несмотря на скудость доказательств, широко распространено мнение, что витамин С очень полезен для здоровья. (Чтобы внести ясность, небольшое количество витамина С необходимо, чтобы избежать цинги, но почти каждый житель развитого мира естественным образом получает нужное количество этого витамина с повседневной пищей. Существует мало доказательств полезности приема дополнительных доз витамина С.) Важным источником этого широко распространенного суеверия является Лайнус Полинг, всемирно известный химик и двукратный лауреат Нобелевской премии, который активно защищал витамин С. Некоторые ученые утверждают, что Полинг знал об ограниченной полезности витамина С, но он верил в эффект плацебо и считал, что убедить людей в исцеляющей силе таблеток витамина С или стакана апельсинового сока – дешевый и простой способ заставить работать эффект плацебо, побуждая человеческое тело каким-то образом вылечить себя.

Объяснение космического привыкания

Возможно, вы заметили связь между разговорами о сигнале и шуме в этой главе и нашим любимым уравнением. Напомним, наше любимое уравнение гласит, что оценка на основе данных состоит из трех компонентов – истинного значения (оцениваемой величины), смещения и шума:

$$\text{Оценка} = \text{Оцениваемая величина} + \text{Смещение} + \text{Шум}.$$

Представьте себе исследование, которое действительно хорошо спроектировано для изучения оценки и исключает смещение. Оценка, полученная в результате этого исследования, будет состоять из истинного значения оцениваe-

мой величины и шума, вызванного такими факторами, как вариация выборки. Истинное значение остается постоянным в различных исследованиях одного и того же явления. Но полученные оценки тем не менее варьируются от исследования к исследованию из-за шума. Итак, если мы представим себе несколько исследований одного и того же явления, можно рассматривать истинное значение как сигнал. А шум – он всегда шум.

Таким образом, можно ожидать, что повторные научные исследования одного и того же явления продемонстрируют возврат к среднему значению. Если первое исследование обнаружило особенно выдающуюся взаимосвязь, то, вероятно, истинная взаимосвязь действительно велика, но также вероятно, что вариация выборки в этом исследовании создала шум в положительном направлении. В следующем исследовании было бы логично ожидать получения меньшей оценки, поскольку, хотя истинная связь (т. е. оценка) по-прежнему велика, шум от вариации выборки на этот раз будет не таким большим.

Осознав это, мы теперь готовы вернуться к идее «космического привыкания» и понять, почему его вряд ли нужно объяснять мистической силой, посредством которой Вселенная приспосабливается к деятельности ученых на Земле.

Если это не мистические силы, то почему предполагаемые эффекты имеют тенденцию уменьшаться при воспроизведении? Отчасти это возвращение к среднему значению. Но это еще не весь ответ. Чтобы по-настоящему понять, что происходит с «космическим привыканием», вам нужно объединить свое понимание возврата к среднему значению с обсуждением предвзятости публикаций из главы 7. Давайте разберемся почему.

Представьте себе нескольких ученых, независимо друг от друга изучающих какое-то явление – скажем, помогает ли дополнительное время на размышление принимать более правильные решения, что является одной из тем исследований Джонатана Скулера. Каждый ученый проводит исследование. Один обнаружил, что люди, которым дано время подумать, принимают значительно худшие решения. Другой обнаружил, что люди, которым дано время подумать, принимают немного худшие решения. Третий не находит никакой связи. А четвертый считает, что люди, которым дано время подумать, принимают немного лучшие решения. Поскольку размер выборки во всех этих исследованиях не очень большой, статистически значимыми являются только крупные результаты.

Почему возникают разные оценки? Вполне возможно, что дополнительное время на обдумывание оказывает какое-то влияние на решения. Это оцениваемая величина нашего любимого уравнения. Мы также можем рассматривать это влияние как сигнал, который является общим для всех упомянутых исследований. Но есть множество несистематических факторов, «шумов», которые влияют на наблюдаемую взаимосвязь (оценку) в любом конкретном исследовании. Например, даже несмотря на то, что это эксперименты, в любом из них по стечению обстоятельств может получиться так, что люди, которым дано время подумать, по своей природе будут гораздо хуже принимать решения, чем люди, которым не дано время на раздумья. Этот большой компонент отрицательного шума приведет к тому, что исследование покажет особенно большой отрицательный эффект. В другом исследовании люди, которым дали время подумать, могли оказаться немного лучше в принятии решений, чем люди, которым не дали времени подумать. Этот положительный шумовой

компонент приведет к тому, что исследование обнаружит более благоприятную связь между временем обдумывания и принятием решений. Итак, как мы видим, результаты этих исследований состоят как из сигнала, так и из шума. Таким образом, мы должны ожидать, что повторные исследования будут возвращаться к среднему значению.

Теперь давайте подумаем, какие открытия будут повторены с наибольшей вероятностью. Исходя из нашего предыдущего обсуждения *p*-скрининга и предвзятости публикаций мы бы предположили, что лишь одно из этих исследований привлечет достаточное внимание научного сообщества и заслужит независимое воспроизведение – то, которое имеет большую статистически значимую отрицательную связь между временем на размышление и качеством принятия решений. Это исследование имеет два преимущества перед другими исследованиями. Во-первых, в нем есть статистически значимые результаты, поэтому вероятность его публикации выше. Во-вторых, его вывод весьма удивителен: кто бы мог подумать, что тщательное обдумывание приводит к принятию худших решений?

Смотрите, что происходит. Опубликовано большое удивительное открытие. Оно было сделано в ходе хорошо спланированного исследования, поэтому люди думают, что это, вероятно, правда. Но поскольку это важный результат, ученые также захотят посмотреть, повторится ли он. Чего нам следует ожидать от них? Что ж, мы только что увидели, что открытие с большей вероятностью будет опубликовано и потребует повторения (как из-за его неожиданности, так и из-за того, что оно с большей вероятностью преодолеет порог статистической значимости), когда предполагаемый размер эффекта особенно велик по величине. Но мы также знаем, что выдающиеся оценки, вероятно, являются результатом как больших значений сигнала, так и больших значений шума. Таким образом, из-за возврата к среднему значению, когда мы приступим к повторению этого исследования, мы должны ожидать обнаружения меньшего (по величине) предполагаемого размера эффекта (как, собственно, и было обнаружено в трех других неопубликованных исследованиях). То есть из-за сочетания предвзятости публикации и возврата к среднему значению мы обречены наблюдать космическое привыкание!

Космическое привыкание и генетика

Рисунок 8.4 – наша любимая иллюстрация феномена космического привыкания, возникающего в результате предвзятости публикаций и возврата к среднему значению. На рисунке показаны меняющиеся данные о связи между конкретными генами и заболеваниями. Различные кривые представляют различные гипотетические связи между генами и заболеваниями. Каждая точка данных показывает знак и размер предполагаемой взаимосвязи с учетом всех доступных данных в произвольный момент времени. На этом конкретном графике значение 1 на вертикальной оси означает, что данные не показывают никакой связи между геном и заболеванием. Значение ниже 1 означает, что данные показывают отрицательную связь между геном и заболеванием. А значение выше 1 означает, что данные свидетельствуют о положительной связи между геном и заболеванием. Чем дальше от 1, тем больше (по абсолютной величине) предполагаемая взаимосвязь.

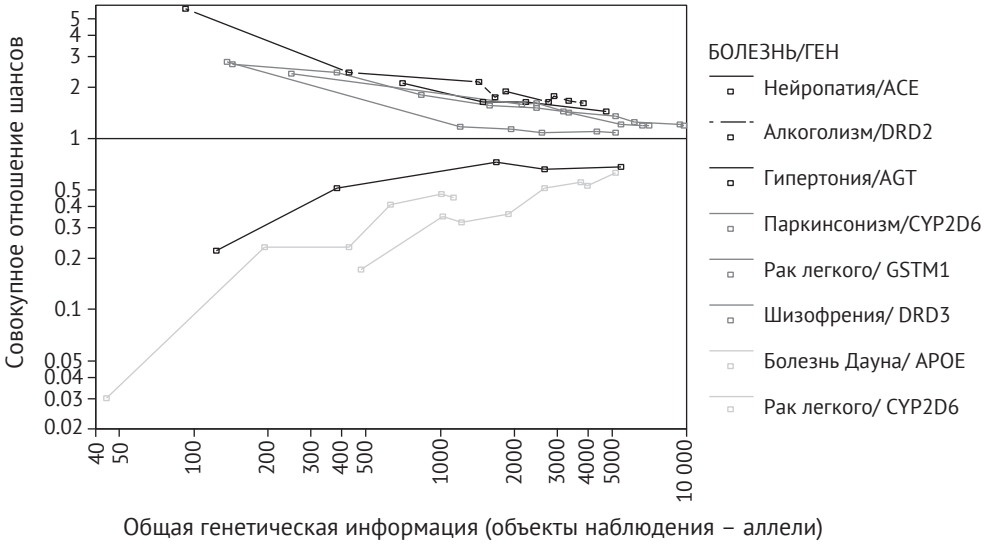


Рис. 8.4. Предполагаемый размер эффекта уменьшается по мере накопления большего количества данных в генетических исследованиях

Точки данных в крайнем левом углу графика – это предполагаемые связи между генами и заболеваниями из самого первого исследования, опубликованного по этой теме. Следующая точка данных по направлению вправо показывает предполагаемую взаимосвязь с учетом данных как первого, так и второго опубликованных исследований. Этот принцип работает по мере продвижения вправо, пока мы не доберемся до самого последнего опубликованного исследования.

То, что вы видите на рис. 8.4, похоже на космическое привыкание. Первое опубликованное исследование обнаруживает тесную связь между геном и заболеванием. Это исследование, которое публикуется в престижном журнале и освещается в прессе. Но по мере того, как ученые начинают воспроизводить результаты, происходит возврат к среднему значению. В последующих исследованиях масштаб предполагаемых эффектов обычно меньше. По мере того как мы добавляем все больше и больше информации, мы приближаемся к истине, которая довольно далека от завышенной оценки, указанной в первоначальном исследовании. Как мы видим, в конечном итоге данные свидетельствуют по крайней мере об очень слабой связи между геном и рассматриваемым заболеванием. Но об этом не напишут в новостях.

Убеждения не возвращаются к среднему значению

Понимание возврата к среднему значению, скорее всего, станет для вас источником постоянного беспокойства. Комментаторы, аналитики и случайные наблюдатели из раза в раз неправильно понимают природу этого феномена, придумывая сложные теории для объяснения закономерностей, отражающих простое статистическое явление. Это звучит так, будто мы должны подозревать возврат к среднему значению почти везде, и в грубом приближении так и есть. Мы должны наблюдать возврат к среднему значению любой перемен-

ной, на которую влияют сигнал и шум. Поэтому, вместо того чтобы перечислять все случаи возврата к среднему значению, давайте подумаем о ситуациях, в которых этого ожидать не следует.

Во-первых, мы не ожидаем увидеть значительный возврат к среднему значению, если сигнал намного превышает шум. Предположим, мы повторили наш предыдущий анализ игры в гольф, но вместо того, чтобы отображать результаты двух разных раундов в рамках одного турнира, мы построили график средних результатов за два разных сезона LPGA Tour. Средний результат за весь сезон содержит гораздо больше информации о способностях игрока. Большая часть удач и неудач, составляющих шум между раундами, усредняется, поэтому возврат к среднему значению в этой картине будет меньше (но все же будет).

Но есть ситуации, когда нам вообще не следует ожидать возврата к среднему значению. Возьмем, к примеру, фондовый рынок. Стоит ли нам ожидать возврата к среднему значению цен на акции? Возможно, мы могли бы победить рынок и стать миллиардерами, воспользовавшись возвратом к среднему значению.

Из логики возврата к среднему значению следует, что в среднем компании с низкими ценами акций должны подрасти в будущем, а компании с высокими ценами должны упасть. Судя по всему, возврат к среднему также предсказывает, что за повышением должно последовать снижение, и наоборот. Можем ли мы использовать эту информацию, покупая акции, которые только что упали, и продавая акции, которые только что выросли?

Ответ на этот вопрос почти наверняка – нет. Предположим, что произошло возвращение к среднему значению цен на акции. Умные инвесторы поймут это, будут следовать описанной выше стратегии и заработают много денег. Но если достаточное количество инвесторов будет следовать этой стратегии, возврат к среднему значению исчезнет, потому что акции с низкими ценами будут расти, а цены акций с высокими ценами будут снижаться в ответ на решения рынка о покупке и продаже. Гипотеза эффективного рынка, кратко обсуждавшаяся в главе 7, гласит, что, поскольку большое количество трейдеров ищет такого рода возможности, мы не сможем предсказать изменения цен на акции, и, следовательно, нам не следует ожидать возврата к среднему значению.

Возврат к среднему значению довольно распространен в деловом мире. Мы видим это в отношении корпоративных доходов и прибылей, несмотря на желание стартапов и венчурных капиталистов прогнозировать будущие доходы, делая линейные, а иногда и экспоненциальные прогнозы на основе прошлых доходов. Так почему же мы не видим возврата к среднему значению цен на акции? Основная причина в том, что цены на акции отражают представления о будущем, а доходы – нет. Цена акций определяется убеждениями инвесторов в долгосрочной стоимости компании. И если бы происходило возвращение к среднему значению, это означало бы, что инвесторы совершают систематические ошибки при формировании своих убеждений. Если бы существовали акции, цена которых должна в будущем гарантированно вырасти, все инвесторы кинулись бы их скупать, моментально подняв цену акций и аннулируя наши ожидания.

Фондовый рынок – это лишь один пример общего явления. Когда дело касается убеждений, понятие возврата к среднему значению неприменимо. Было бы бессмысленно говорить что-то вроде «Сегодня я считаю, что у республи-

канцев есть 60-процентный шанс победить на следующих выборах, но в день голосования моя уверенность будет ниже». Это не имеет смысла, потому что ваше убеждение – это всего лишь ваша вера в будущее и ничего больше. Если вы ожидаете, что в день выборов ваше убеждение составит 55 %, а не 60, то ваше убеждение должно составлять 55 % уже сегодня.

Подведение итогов

Во второй части вы узнали, как количественно оценивать корреляции и как определить, отражают ли корреляции, обнаруженные в данных, реальные явления или просто шум. Затем мы обратились к другим проблемам, возникающим из-за присутствия шума: завышению значимости, занижению отчетности и возврату к среднему значению.

Наше любимое уравнение подсказывает нам, что шум не единственная причина, по которой оценка может не совпадать с истинным значением оцениваемой величины. Нам также следует беспокоиться о систематическом смещении. По причинам, которые мы начали изучать в главе 3, смещение представляет собой особенно важную проблему, когда мы пытаемся изучить причинно-следственные связи: когда мы говорим, что корреляция не подразумевает причинно-следственную связь, мы имеем в виду, что корреляция между двумя явлениями может быть смещенной оценкой причинно-следственной связи между ними. В части III мы сосредоточимся на причинно-следственных связях, сначала более подробно изучив источники смещения, а затем изучая стратегии объективной оценки причинно-следственных связей.

Ключевые термины

- **Эффект Хоторна:** явление, при котором субъекты меняют свое поведение, потому что знают, что их изучают.
- **Эффект опроса:** конкретный пример эффекта Хоторна, при котором субъекты исследования меняют свое поведение, чтобы попытаться угодить исследователю.
- **Сигнал:** систематический компонент результата, который сохраняется во всех наблюдениях.
- **Шум:** случайные компоненты результата, которые меняются от наблюдения к наблюдению.
- **Возврат к среднему:** явление, при котором, если одно наблюдение величины, состоящей из сигнала и шума, особенно велико (соответственно, мало), другие наблюдения обычно будут меньше (соответственно, больше).

Упражнения

8.1. В начале каждого бейсбольного сезона многие обещают побить рекорд хоум-рана, но почти никто этого не делает. Давайте подумаем почему.

Предположим, вы совершили феноменальное количество хоум-ранов в первых 20 играх сезона и готовы побить рекорд.

a) Как вы думаете, в следующих 20 играх вы, скорее всего, сделаете количество хоум-ранов выше среднего или ниже среднего? Почему?
b) Какова вероятность того, что в следующих 20 играх вы совершите меньше хоум-ранов, такое же количество хоум-ранов или больше хоум-ранов, чем в первых 20 играх? Почему?
c) Комментатор замечает, что игроки, достигшие высокого результата после первых 20 игр, почти никогда не побивают рекорд. Комментатор утверждает, что игроки теряют самообладание, когда начинают думать о близком рекорде. Какие данные вы могли бы собрать и как их проанализировать, чтобы убедиться, верна ли эта интерпретация?

8.2. Энтони однажды прошел курс обучения у известного эконометриста, который выдвинул следующий аргумент: сын Пола Джон имеет IQ на уровне гения. Следовательно, из-за возврата к среднему сам Пол должен был иметь IQ сверхгениального уровня.

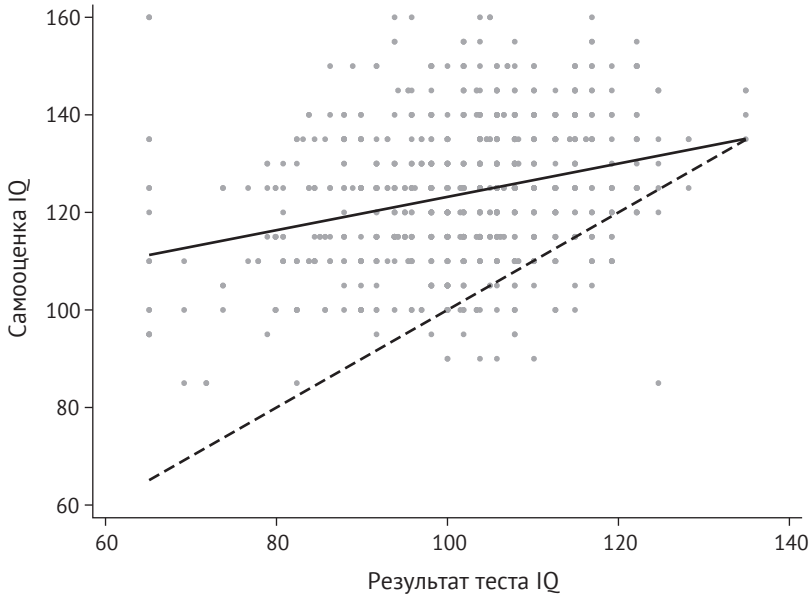
a) Что не так с рассуждениями эконометриста? Как вы думаете, IQ Пола ниже или выше среднего? Он ниже или выше, чем у Джона?
b) Как бы изменился ваш ответ, если бы мы сказали вам, что Полом в этом примере является Пол Самуэльсон, лауреат Нобелевской премии, которого многие считают выдающимся ученым-экономистом XX в.?

8.3. На момент написания этой книги цена акций Zoom (компании, специализирующейся на онлайн-видеоконференциях, с которой многие из нас слишком хорошо познакомились в 2020 году) только что упала примерно на 18 % в ответ на победу Джо Байдена на президентских выборах в США в 2020 г. и публикации многообещающих результатов по вакцинам против COVID-19. Ваш друг утверждает, что из-за возврата к среднему сейчас самое подходящее время для покупки акций Zoom.

a) Объясните своему другу простым языком, что не так в его рассуждениях.
b) Не зная каких-либо дополнительных подробностей, за исключением того, что акции Zoom недавно упали, что вы предполагаете – ваш друг заработает деньги, потеряет деньги или просто вернет инвестиции?

8.4. Психологи утверждают, что степень, в которой человек может точно оценить свои способности в какой-то области, зависит от его реальных способностей в этой области. Возможно, люди, у которых нет способностей в определенной области, даже не обладают нужными знаниями, чтобы понять, насколько они на самом деле несовершенны. Это явление назвали *эффектом Даннинга–Крюгера* в честь исследователей, которые первыми разработали эту гипотезу.

Типичные доказательства, предлагаемые в пользу гипотезы Даннинга–Крюгера, показаны на рисунке ниже. Сначала испытуемых просили предсказать свой собственный IQ, а затем они проходили тест на IQ. На рисунке показана диаграмма распределения этих двух баллов по каждому предмету. Линия регрессии показана серым цветом, а пунктирная линия под углом 45° – черным. Люди с низким IQ (по данным теста) были склонны переоценивать свои способности, а люди с высоким IQ (по данным теста) в среднем были правы.



a) Можете ли вы придумать другое объяснение этого эмпирического феномена, которое не обязательно подразумевало бы, что люди с высоким IQ лучше оценивают свой собственный IQ? Вам захочется подумать о возврате к среднему, но это само по себе не поможет, поскольку люди с высокими результатами тестов не недооценивали свой IQ. Возможно, вам также захочется подумать о смещении. Вспомните наше любимое уравнение.

b) Чтобы проверить свою догадку, смоделируйте на компьютере данные, которые дадут аналогичный результат, хотя оценки людей с высокими способностями столь же зашумлены, как и оценки людей с низкими способностями. (Подсказка: стоит помнить, что IQ, оцениваемый в ходе теста, не является идеальным показателем истинного интеллекта.)

c) Загрузите файл IQdata.csv и связанный с ним файл README.txt, описывающий переменные в этом наборе данных, по адресу *http://www.press.princeton.edu/thinking-clearly*. Это данные, использованные для построения графика. (Мы получили данные из исследования Жиньяка и Заженковски, проведенного в 2020 г.) Давайте подумаем, как можно определить, действительно ли люди с высоким IQ лучше оценивают свой собственный IQ.
   i) Сначала вычислите абсолютное значение ошибки для каждого испытуемого (т. е. насколько далеко была их самооценка IQ от результата тестирования IQ?).
   ii) Теперь регрессируйте эту абсолютную ошибку на IQ, оцениваемый тестами, и интерпретируйте результаты.

d) Как вы можете видеть на рисунке, люди склонны переоценивать свой IQ.
   i) Оцените среднюю степень систематической ошибки в этих данных.
   ii) Вычтите эту оценку систематической ошибки из самооценки IQ каждого участника, чтобы получить самооценку с поправкой на смещение.

iii) Используя эту самооценку с поправкой на смещение, пересчитайте абсолютную величину ошибок, т. е. подсчитайте, насколько в среднем самооценка человека с поправкой на смещение отличается от его IQ, оцененного тестом.

iv) Наконец, регрессируйте эту новую меру ошибки на IQ, оцениваемый тестами, и интерпретируйте результаты.

е) Дайте окончательную оценку гипотезы Даннинга–Крюгера на основе этих данных. Действительно ли люди с высоким интеллектом лучше оценивают свой интеллект?

8.5. Найдите подходящий пример, когда аналитик не предусмотрел возвращение к среднему значению, хотя следовало бы это сделать. В частности, ищите доказательства, представленные в пользу конкретной теории или явления, которые также можно легко объяснить с помощью возврата к среднему. Ваш пример может быть взят из газетной статьи, научного исследования, политического доклада или заявления политика, бизнес-лидера или спортивного комментатора. Кратко изложите утверждение, сделанное аналитиком, и доказательства, которые предположительно подтверждают это утверждение. Объясните, почему данные в равной степени согласуются с возвратом к среднему значению. В качестве бонуса подумайте о том, как вы потенциально могли бы сделать выбор между утверждением аналитика и возвратом к среднему значению.

Дополнительное чтение и ссылки

Статья в New Yorker о космическом привыкании:

Jonah Lehrer. 2010. *The Truth Wears Off: Is There Something Wrong with the Scientific Method?* The New Yorker. December 13.

Упомянутое нами исследование, в котором проводится повторный анализ доказательств эффекта Хоторна:

Steven D. Levitt and John A. List. 2011. *Was There Really a Hawthorne Effect at the Hawthorne Plant?: An Analysis of the Original Illumination Experiments.* American Economic Journal: Applied Economics 3 (1): 224–238.

Источником рис. 8.1 является оригинальная статья Гальтона на эту тему:

Francis Galton. 1886. *Regression towards Mediocrity in Hereditary Stature.* Journal of the Anthropological Institute of Great Britain and Ireland 15: 246–263.

Исследование, сравнивающее настоящие операции на колене с ложными операциями:

J. Bruce Moseley, Kimberly O'Malley, Nancy J. Petersen, Terri J. Menke, Baruch A. Brody, David H. Kuykendall, John C. Hollingsworth, Carol M. Ashton, and Nelda P. Wray. 2002. *A Controlled Trial of Arthroscopic Surgery for Osteoarthritis of the Knee.* New England Journal of Medicine 347 (2): 81–88.

Статья, показывающая, что среди пациентов с астмой эффект плацебо, по-видимому, справедлив для субъективных показателей, но не влияет на объективные показатели:

Michael E. Wechsler, John M. Kelley, Ingrid O. E. Boyd, Stefanie Dutile, Gautham Marigowda, Irving Kirsch, Elliot Israel, and Ted J. Kaptchuk. 2011. *Active Albuterol or Placebo, Sham Acupuncture, or No Intervention in Asthma.* New England Journal of Medicine 365 (2): 119–126.

Рисунок 8.4 взят из статьи:

John P. A. Ioannidis, Evangelia E. Ntzani, Thomas A. Trikalinos, and Despina G. Contopoulos-Ioannidis. 2001. *Replication Validity of Genetic Association Tests.* Nature Genetics 29: 306–309.

Задание по эффекту Даннинга–Крюгера составлено на основе статьи:

Gilles E. Gignac and Marcin Zajenkowski. 2020. *The Dunning-Kruger Effect Is (Mostly) a Statistical Artefact: Valid Approaches to Testing the Hypothesis with Individual Differences Data.* Intelligence 80: 101449.

ЧАСТЬ III

Является ли связь причинно-следственной?

Глава 9

Почему корреляция и причинно-следственная связь не одно и то же

О чем эта глава

- Корреляция не обязательно подразумевает причинно-следственную связь.
- Есть две ключевые причины, по которым наблюдаемая корреляция может быть смещенной оценкой причинно-следственной связи: искажающие факторы и обратная причинно-следственная связь.
- Критическое мышление часто помогает нам распознать искажающие факторы.
- Существует важное различие между искажающими факторами и механизмами.

Введение

Как мы обсуждали в главах 2 и 3, информация о корреляциях и причинно-следственных связях полезна для разных целей. Знание корреляций само по себе помогает нам описать мир и спрогнозировать наличие определенных свойств мира исходя из наблюдения других его свойств. Знание причинно-следственных связей особенно ценно для принятия решений, поскольку оно говорит нам, как предпринимаемые действия повлияют на мир. Вспомните наше определение причинности из главы 3. Причинно-следственная связь – это изменение некоторой характеристики мира, которое происходит в результате изменения какой-либо другой характеристики мира.

Поэтому, когда мы говорим, что какое-то действие оказывает причинное влияние на какой-то результат, мы утверждаем, что результат был бы другим в контрфактическом мире, в котором действие было бы другим. Опережающее знание последствий наших действий позволяет нам предвидеть и взвешивать их затраты и выгоды.

С прагматической точки зрения именно это различие объясняет, почему принцип «корреляция не то же самое, что причинно-следственная связь» так важен. Если мы ошибочно примем корреляцию за причинно-следственную связь, мы можем в конечном итоге совершить большие ошибки, исходя из ложных представ-

лений о том, как эти действия повлияют на результаты. В этой главе мы научимся критически мыслить о разнице между корреляцией и причинно-следственной связью, обсудим источники систематических ошибок, которые могут сделать корреляции ненадежными оценками причинно-следственных связей, и начнем рассматривать, как это влияет на изучение причинно-следственных связей.

Чтобы наглядно показать, насколько важна эта тема, давайте рассмотрим пример, в котором принимаются важные решения о том, как использовать ресурсы, и где мы можем с некоторой уверенностью отделить корреляцию от причинно-следственной связи. Пример касается темы так называемых чартерных школ[1] (Charter School) в США.

Чартерные школы

В своем душераздирающем фильме «В ожидании Супермена» Дэвид Гуггенхайм рассказывает историю нескольких маленьких детей из бедных семей. В каждом случае ребенка зачисляют в некачественную государственную школу. И в каждом случае родители прилагают все усилия, чтобы отдать своего ребенка в чартерную школу (или, в некоторых случаях, в специальную школу).

Чартерные школы тоже работают за государственный счет, но независимо от системы государственных школ. Некоторые чартерные школы управляются некоммерческими организациями, а другие – коммерческими корпорациями. Идея движения чартерных школ заключается в поощрении инноваций и выбора. Чартерные школы свободны от некоторых ограничений (например, профсоюзных договоров, устаревших учебных программ), с которыми сталкиваются государственные школы. У них есть возможность по своему усмотрению вводить новшества в учебные программы, стимулы для учителей и т. д., чего не могут сделать обычные государственные школы. А поскольку им приходится конкурировать за учеников, у них будет мотивация предлагать новые и потенциально лучшие подходы к образованию. Действительно ли чартерным школам удается улучшить результаты обучения – это предмет горячих споров.

Во многих районах количество детей, подающих заявки на поступление в чартерные школы, превышает возможности зачисления. По закону, когда число желающих учиться в чартерной школе превышает количество учащихся, она должна принимать учеников посредством лотереи. Семьи подают заявки в школу, и после этого удача определяет, какие дети получат заветные места. Как ярко иллюстрирует фильм, шансы складываются не в пользу детей. В некоторых чартерных школах есть сотни претендентов на пару десятков мест.

По ходу фильма нам рассказывают множество фактов о работе чартерных школ, в которые стремятся поступить учащиеся. По сравнению с государственными школами с социально-экономически схожим контингентом учащихся, эти чартерные школы имеют лучшие результаты тестов, более высокий процент выпуска, меньший уровень преступности и т. д. Действительно, чартерные школы, показанные в фильме, работают намного лучше, чем государственные школы, практически по всем измеримым показателям.

В конце фильма мы обнаруживаем, что лишь немногие из учеников, за которыми мы следили, были приняты в школу по своему выбору. Остальные будут

[1] Школа с собственным уставом (вид муниципальной спецшколы для особой группы детей, часто для этнической группы). – *Прим. перев.*

зачислены в государственные «фабрики неудач», где, как нам остается полагать, их потенциал будет потрачен впустую.

Но правильный ли это вывод? Действительно ли попадание в чартерную школу гарантирует лучшие результаты обучения ребенка? На карту поставлено нечто большее, чем наши чувства к детям в фильме. За последние несколько десятилетий чартерные школы стали одним из доминирующих подходов к школьной реформе в Соединенных Штатах. Развитие чартерных школ как альтернативы традиционным государственным школам получило поддержку обеих партий – например, его агрессивно продвигали администрации Буша и Обамы. Доля учащихся государственных школ, посещающих чартерные школы, выросла с менее 1 % в 1999 г. до более 6 % в 2021 г. Но критики выражают обеспокоенность по поводу того, что это расширение могло привести к сокращению ресурсов, доступных традиционным государственным школам, и тем самым причинить вред учащимся в этих школах. Поэтому нам бы очень хотелось знать, оказывают ли чартерные школы положительное влияние на академические результаты учащихся.

Вот что мы знаем. Определенно, существует корреляция между посещением чартерной школы и успеваемостью. В среднем по городу учащиеся из семей с низкими доходами, посещающие чартерные школы, демонстрируют лучшие показатели обучения, чем учащиеся из таких же семей, посещающие традиционные государственные школы.

В качестве примера рассмотрим школу Пройсса (Preuss School) – чартерную школу, созданную Калифорнийским университетом в Сан-Диего. Школа Пройсса обслуживает учащихся средних и старших классов с низким доходом со всего округа Сан-Диего. По общему мнению, это замечательная школа, отправляющая почти 100 % своих учеников, чьи родители обычно не заканчивали даже колледж, получать высшее образование в университете. Школа получила похвалы из многих источников, в том числе от журнала Newsweek, который неоднократно называл школу Пройсса «лучшей революционной средней школой» в стране.

И действительно, дети из школы Пройсса учатся намного лучше, чем их сверстники в государственных школах Сан-Диего. Например, взгляните на данные на рис. 9.1, демонстрирующие разницу в проценте учащихся, сдавших стандартизированные тесты по математике и английскому языку в школе Пройсса и в городских школах Сан-Диего (SDCS).

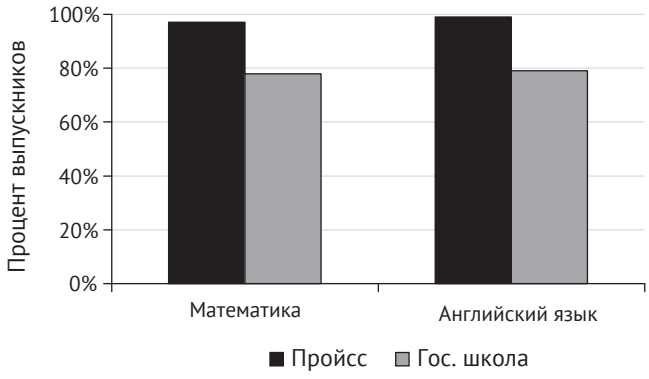


Рис. 9.1. Результаты стандартизированных тестов в школе Пройсса и городских школах Сан-Диего (SDCS)

Как и в историях из фильма «В ожидании Супермена», судя по этим данным, школа Пройсса оказывает огромное влияние на успеваемость своих учеников. Но чтобы не делать поспешные выводы, давайте подвергнем эти данные критическому анализу. Данные показывают положительную корреляцию между посещением школы Пройсса и успеваемостью. Но означает ли тот факт, что учащиеся школы Пройсса (или других чартерных школ) учатся лучше, чем учащиеся государственных школ, что посещение чартерной школы приводит к улучшению успеваемости? Иными словами, если мы рассмотрим контрфактический мир, в котором дети из школы Пройсса и дети из государственной школы поменялись местами, начнут ли бывшие ученики государственной школы учиться лучше, а бывшие ученики Пройсса – хуже? Подразумевает ли корреляция причинно-следственную связь? Это именно то, что нам нужно знать, если мы собираемся выяснить, является ли расходование ресурсов на чартерные школы хорошим решением.

Конечно, вполне возможно, что чартерные школы действительно способствуют более высокой успеваемости учащихся, а это означает, что в контрфактическом мире, где учащиеся чартерных школ посещали обычные государственные школы, их успеваемость была бы хуже, а у бывших учеников государственной школы – лучше. Но другое возможное объяснение, которое так красноречиво иллюстрирует фильм «В ожидании Супермена», заключается в том, что учащиеся и семьи, выбирающие чартерные школы, изначально во многом отличаются от обычной семьи среднего ученика государственной школы. То есть, возможно, объяснение корреляции заключается не в лучшей школе, а в лучших учениках. Если это так, можем ли мы быть уверены, что они все равно не превзойдут своих сверстников в контрфактическом мире?

Спросите себя: при каких обстоятельствах родители из экономически неблагополучных семей запишут своего ребенка на участие в лотерее, чтобы он попал в чартерную школу? Нам приходят на ум два обстоятельства. Во-первых, если родители считают, что их ребенок особенно талантлив, у них может быть особая мотивация отдать ребенка в школу с хорошей репутацией. Во-вторых, если родители сами уделяют особое внимание образованию своего ребенка, они с большей вероятностью проделают необходимую работу, чтобы обеспечить ребенку участие в лотерее.

Природный талант и участие родителей сами по себе являются весьма важными факторами, определяющими успеваемость учащихся. Предположим, что контингент учащихся, участвующих в лотереях чартерных школ (и, следовательно, попадающих в эти школы), в среднем более талантлив и происходит из семей, более заинтересованных в образовании, чем население в целом. Тогда, даже если сами чартерные школы не окажут никакого выдающегося влияния на успеваемость своих учеников, эти ученики тем не менее превзойдут население в целом просто благодаря своим талантам и большей поддержке со стороны семьи. Иными словами, если бы все дети ходили в одну и ту же школу, дети, которые сейчас учатся в чартерных школах, все равно демонстрировали бы результаты выше среднего, потому что они более талантливы и у них более вовлеченные и ответственные родители.

Помните вопрос, который нас волнует: можно ли ожидать, что отправка ребенка в чартерную школу улучшит успеваемость этого ребенка по сравнению

с тем, чего он достиг бы в местной государственной школе? Приведенное выше обсуждение показывает, что мы не можем узнать ответ на этот важный вопрос, сравнивая успеваемость учащихся в чартерных школах с традиционными государственными школами. Мы бы сравнили группу особенно талантливых и амбициозных детей из очень преданных родительскому делу семей с населением в целом. Как мы узнаем, возникли ли различия между этими группами из-за влияния чартерных школ, из-за глубинных социальных различий или из-за того и другого? Проще говоря, мы бы сравнивали яблоки с апельсинами.

Чтобы определить, действительно ли чартерные школы являются причиной отличных результатов своих учеников, нам нужно провести сравнение, которое лучше всего отражает контрфактическую природу причинности. Для этого нам нужен способ сравнивать яблоки с яблоками. Идеальный вопрос, на который мы хотели бы ответить, звучит примерно так: если бы все остальные условия у двух детей были одинаковыми, добился бы ребенок, который пошел в чартерную школу, большего успеха, чем ребенок, поступивший в государственную школу? Мы, очевидно, не можем ответить на этот вопрос. Но мы можем подойти ближе, попытавшись спросить примерно так: если бы все остальное в двух группах детей было в среднем одинаковым, дети, которые ходят в чартерные школы, в среднем учились бы лучше, чем дети, которые ходят в государственные школы, или нет?

Чтобы ответить на этот последний вопрос, нам нужно выйти за рамки простого сравнения детей из чартерных школ с детьми из государственных школ. Мы делаем это, сужая наше внимание только до детей, которые пытались поступить в чартерные школы. Все эти дети были достаточно талантливыми или имели достаточно ответственные семьи, чтобы подать заявление в чартерную школу. Но в результате приемной лотереи некоторые счастливчики попали в чартерную школу, а другие – нет. Поскольку лотерея была случайной, пул победителей и пул проигравших должны иметь в среднем одни и те же характеристики (т. е. если бы мы проводили лотерею снова и снова, дети, выигравшие в лотерею, были бы столь же мотивированы или талантливы, как и те, кто проиграл). Таким образом, мы можем узнать гораздо больше о фактическом эффекте посещения чартерной школы, сравнивая академическую успеваемость тех, кто поступал и выиграл вступительную лотерею, с теми, кто поступал, но проиграл в лотерее. Если в этом более узком сравнении положительная корреляция между посещением чартерной школы и успеваемостью сохранится, мы будем чувствовать себя гораздо увереннее, давая ей причинную интерпретацию, потому что теперь мы сравниваем яблоки с яблоками.

Это сравнение действительно было сделано для многих чартерных школ. Давайте начнем с того, что проведем такое сравнение со школой Пройсса. У нас нет данных для этого сравнения для того же стандартизированного теста, что и на рис. 9.1, но есть данные для другого важного стандартизированного теста, показанного на рис. 9.2.

Это сравнение демонстрирует важность сравнения яблок с яблоками. Да, учащиеся школы Пройсса, как правило, превосходят учащихся городских школ Сан-Диего в целом. Но когда учеников, выигравших в лотерею, сравнивают с учениками, проигравшими в лотерею, корреляция исчезает – разницы в успеваемости практически нет.

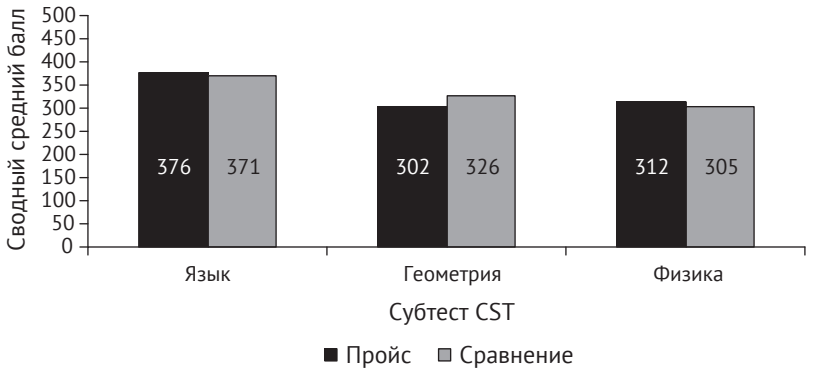


Рис. 9.2. Сравнение результатов стандартизированных тестов детей, выигравших и не выигравших вступительную лотерею в школу Пройсса

Аналогичные результаты получены в результате исследований многих других чартерных школ. Конечно, разные исследования обнаруживают разные вещи. Исследователи из Бостона обнаружили, что дети, поступившие в чартерные школы по программе «Знание – сила», учатся лучше, чем дети, которые подали заявку, но проиграли лотерею. Но мы считаем, что следующий вывод из другого исследования программ выбора школ в Сан-Диего более типичен для научных публикаций по этому поводу:

> «В подавляющем большинстве случаев мы не обнаружили никаких доказательств того, что победители и проигравшие в той или иной лотерее показали разные результаты в этих тестах на успеваемость через 1–3 года после проведения вступительной лотереи. Мы интерпретируем это так, что выигрыш в лотерею не помогает и не вредит росту достижений».

Подумайте, что это значит. Когда мы сравниваем яблоки с яблоками, оказывается, что высокоэффективные чартерные школы практически не влияют на успеваемость учащихся. Большинство очевидных успехов этих школ связано с тем фактом, что дети, участвующие в лотереях чартерных школ, уже отличаются от среднестатистического ученика в плане обучения. В любом случае их успехи в учебе будут выше среднего. Такое отделение корреляции от причинно-следственной связи может изменить ваши взгляды на то, как нам следует тратить ресурсы на образование.

Критический анализ потенциальных исходов

Когда корреляцию между двумя переменными можно правдоподобно интерпретировать как убедительное свидетельство причинно-следственной связи? Мы только что увидели пример того, как легко можно сделать ошибочный вывод, что корреляция указывает на причинно-следственную связь, и какое значение этот вывод может иметь для принятия решений. Но давайте попробуем более систематично подойти к вопросу о том, почему корреляция не всегда является свидетельством причинно-следственной связи, чтобы научиться более

четко различать, когда у нас есть, а когда нет достоверной оценки причинно-следственной связи.

Не забывайте, что мы определяем причинно-следственные связи контрфактически. В главе 3 мы представили понятие потенциальных исходов, которое помогает нам более ясно мыслить о таких контрфактах.

Предположим, мы пытаемся оценить влияние посещения чартерной школы на успеваемость, измеряемую результатами стандартизированных тестов. Таким образом, интересующим нас *исходом* являются результаты стандартизированных тестов, а интересующим *воздействием* – посещение чартерной школы. Обозначим исход – результаты стандартизированных тестов – буквой Y. Обозначим воздействие – посещение чартерной школы – бинарной переменной T. Если для какого-то ученика $T = 1$, это означает, что он посещал чартерную школу. Если $T = 0$, это означает, что он посещал государственную школу. Иногда мы говорим, что объект с $T = 1$ *подвергался воздействию*, а объект с $T = 0$ *не подвергался*, хотя зачастую совершенно не имеет значения, какую группу называть подвергнутой воздействию (например, мы могли бы назвать воздействием посещение государственной школы).

В метафизическом смысле для каждого человека существует некий стандартизированный балл по тестам, который он получил бы, если бы пошел в чартерную школу, и некий стандартизированный балл, который он получил бы, если бы не пошел в чартерную школу. Однако мы можем наблюдать только один из них. Тем не менее наличие обозначений для каждого из этих потенциальных исходов помогает нам определить контрфакты:

$$Y_{1i} = \text{результат для объекта } i, \text{ если } T = 1,$$

$$Y_{0i} = \text{результат для объекта } i, \text{ если } T = 0.$$

Используя эти обозначения, влияние посещения чартерной школы на результаты тестов человека i можно записать так:

$$\text{Влияние}_i = Y_{1i} - Y_{0i}.$$

Мы говорим, что причинность связана с контрфактическими сравнениями, потому что для любого человека в определенный момент времени мы можем наблюдать только одну из двух величин – Y_{1i} или Y_{0i}. Это означает, что мы не можем напрямую измерить влияние посещения чартерной школы на отдельного человека.

Вместо этого мы можем надеяться оценить средний эффект от посещения чартерной школы группой людей из некоторой интересующей нас популяции. Для каждой интересующей нас группы населения определим обозначение среднего балла за тест, если все ходили в чартерную школу, и среднего балла за тест, если все ходили в государственную школу, следующим образом:

$$\bar{Y}_1 = \text{средний результат, если бы все объекты имели } T = 1,$$

$$\bar{Y}_0 = \text{средний результат, если бы все объекты имели } T = 0.$$

Используя эти обозначения, мы теперь можем подумать о *среднем эффекте воздействия* (average treatment effect, ATE):

$$\text{ATE} = \bar{Y}_1 - \bar{Y}_0.$$

Конечно, мы не можем непосредственно наблюдать этот средний эффект, так же как мы не можем наблюдать влияние чартерных школ на человека. Невозможно сделать так, чтобы все объекты одновременно подвергались и не подвергались воздействию. Действительно, в любой момент времени каждый объект находится или в том, или в другом состоянии. Но мы можем попытаться оценить ATE.

Первое, что мы могли бы сделать, чтобы попытаться оценить средний эффект воздействия, – это просто посмотреть на корреляцию, т. е. сравнить средние результаты тестов учащихся, посещающих чартерную школу (испытавших воздействие), со средними результатами тестов учащихся, посещающих государственные школы (без воздействия).

Начнем с рассуждения о том, что наша популяция разделена на две группы: тех, кто ходил в чартерные школы ($\mathcal{T}$) и тех, кто ходил в государственные школы ($\mathcal{U}$). Обозначим средние результаты тестов в каждой из этих групп через:

$$\overline{Y}_{1\mathcal{T}} = \text{средний результат среди объектов с } T = 1,$$

$$\overline{Y}_{0\mathcal{U}} = \text{средний результат среди объектов с } T = 0.$$

Мы будем называть разницу в средних результатах тестов между этими двумя группами *разностью средних значений популяции*. Она записывается очень просто:

$$\text{Разность средних значений популяции} = \overline{Y}_{1\mathcal{T}} - \overline{Y}_{0\mathcal{U}}.$$

Конечно, мы можем наблюдать не всю популяцию, а лишь выборку из нее. Допустим, мы наблюдаем только за учениками одной конкретной чартерной школы. Таким образом, разница в средних результатах тестов, которую мы наблюдаем в нашей выборке, равна разнице в средних результатах тестов среди учащихся чартерных и государственных школ среди всего населения плюс некоторый шум. Итак, мы имеем

$$\text{Выборочная разность средних значений} = \underbrace{\overline{Y}_{1T} - \overline{Y}_{0U}}_{\text{Разность средних значений популяции}} + \text{ шум},$$

что является лишь мерой корреляции между результатами стандартизированных тестов и посещением чартерной школы в нашей выборке.

Конечно, нам нужно знать средний *эффект* от посещения чартерной школы, а не просто корреляцию. Чтобы понять, в чем разница, полезно ввести еще два понятия – *средний эффект воздействия на подвергшихся воздействию* (average treatment effect on the treated, ATT) и *средний эффект воздействия на не подвергавшихся воздействию* (average treatment effect on the untreated, ATU). ATT – это средний эффект от посещения чартерной школы среди тех учащихся, которые на самом деле посещали чартерную школу. То есть:

$$\text{ATT} = \overline{Y}_{1\mathcal{T}} - \overline{Y}_{0\mathcal{T}}.$$

ATU – это средний эффект от посещения чартерной школы среди тех учащихся, которые на самом деле посещали государственные школы. То есть:

$$\text{ATU} = \overline{Y}_{1U} - \overline{Y}_{0U}.$$

Обратите внимание на две вещи. Во-первых, ATE – это просто средневзвешенное значение ATT и ATU, где веса зависят от количества детей в каждой группе[1]. Во-вторых, как и ATE, ATT и ATU принципиально не наблюдаемы. Мы не наблюдаем результаты тестов, которые получили бы учащиеся чартерной школы, если бы они ходили в государственные школы ($\overline{Y}_{0T}$). И мы не наблюдаем, какие результаты тестов получили бы учащиеся государственных школ, если бы они пошли в чартерную школу ($\overline{Y}_{1U}$).

Теперь, когда у нас есть все эти обозначения, мы сможем предметно рассуждать о разнице между корреляцией и причинно-следственной связью. Для начала давайте сравним разницу в средних значениях, которую мы фактически наблюдаем (что является нашей мерой корреляции), с ATT – эффектом посещения чартерных школ среди учащихся, которые посещали чартерные школы. Это поможет нам сформировать определенное представление, которое мы затем сможем применить для сравнения разницы в средних значениях с ATU и в конечном счете с ATE.

Нам нужно вернуться к работе с нашим любимым уравнением:

$$\text{Оценка} = \text{Оцениваемая величина} + \text{Смещение} + \text{Шум}.$$

То есть мы хотим найти способ записать уравнение

$$\text{Выборочная разность средних значений} = \text{ATT} + \text{смещение} + \text{шум}.$$

Как нам это сделать?

Прежде всего вспомним из вышесказанного, что

$$\text{Выборочная разность средних значений} = \underbrace{\overline{Y}_{1T} - \overline{Y}_{0U}}_{\text{Разность средних значений популяции}} + \text{шум}.$$

Теперь мы собираемся хитрым образом переписать разность средних значений популяции, прибавляя и вычитая из него $\overline{Y}_{0T}$. Мы знаем, что это кажется странным. Но поверьте нам, это поможет. По крайней мере, ясно, что мы не причиняем никакого вреда, поскольку, прибавляя и вычитая один и тот же член уравнения, мы на самом деле просто прибавляем ноль. В любом случае, когда мы это сделаем, получим

$$\text{Выборочная разность средних значений} = \overbrace{\overline{Y}_{1T} - \overline{Y}_{0T}}^{\text{ATT}} + \overbrace{\overline{Y}_{0T} - \overline{Y}_{0U}}^{\text{Смещение}_{\text{ATT}}} + \text{шум},$$

[1] Средневзвешенное значение – это просто среднее значение, при котором мы придаем разный вес разным элементам. Например, предположим, что 75 % популяции подвергаются воздействию, а 25 % – нет; тогда ATE – это средневзвешенное значение ATT и ATU с 75 % веса ATT. То есть $\text{ATE} = \frac{75 \cdot \text{ATT} + 25 \cdot \text{ATU}}{75 + 25} = .75 \cdot \text{ATT} + .25 \cdot \text{ATU}$

где мы добавили к смещению индекс ATT, чтобы указать, что это смещение, которое мы получаем при использовании разницы в средних значениях для оценки ATT.

Наш алгебраический трюк и в самом деле был очень крутым, не так ли? Добавляя и вычитая один и тот же член, мы смогли прийти к нашему любимому уравнению. Выборочная разность средних значений (оценка) равна ATT (оцениваемая величина) плюс смещение, плюс шум!

Но что именно в данном случае означает этот член «смещения»? Если мы пытаемся оценить эффект от посещения чартерной школы, наше сравнение результатов тестов среди учащихся, которые посещали и не посещали чартерную школу, является необъективным, если мы ожидаем, что эти две группы учащихся получили бы разные средние баллы по результатам стандартизированных тестов даже в контрфактическом мире, где все они ходили в государственные школы (т. е. $\overline{Y}_{0T} - \overline{Y}_{0U} \neq 0$). В этом случае мы говорим, что две группы имеют *базовые различия*.

До сих пор мы видели, как разница в средних результатах тестов между учениками чартерных и государственных школ может быть необъективной оценкой истинного влияния посещения чартерных школ на тех учащихся, которые посещают чартерные школы (ATT). Мы могли бы провести аналогичный анализ влияния посещения чартерных школ на учащихся государственных школ (ATU):

$$\text{Выборочная разность средних значений} = \overbrace{\overline{Y}_{1U} - \overline{Y}_{0U}}^{\text{ATU}} + \overbrace{\overline{Y}_{1T} - \overline{Y}_{1U}}^{\text{Смещение}_{\text{ATU}}} + \text{шум}$$

Здесь мы обнаруживаем аналогичное смещение, но теперь базовые различия, которые нас беспокоят, связаны с различиями в результатах между группами, подвергшимися и не подвергшимися воздействию, если бы обе группы подверглись воздействию (т. е. $\overline{Y}_{1T} - \overline{Y}_{1U} \neq 0$). Поскольку общий средний эффект воздействия (ATE) представляет собой всего лишь средневзвешенное значение ATT и ATU, смещение, связанное с использованием разницы в средних значениях для оценки ATE, возникает из-за обоих этих типов базовых различий.

Вспомните разницу в успеваемости между учащимися школы Пройсса (есть воздействие) и учениками городских школ Сан-Диего (нет воздействия). Мы были обеспокоены тем, что эта связь не является причинно-следственной, поскольку, скажем, учащиеся Пройсса в среднем были более талантливы в учебе, чем учащиеся городских школ Сан-Диего. Если учащиеся Пройсса на самом деле более талантливы в учебе, то между двумя группами учащихся существуют базовые различия – разница в успеваемости между учащимися будет существовать, даже если все учащиеся в обеих группах пойдут в одну и ту же школу (т. е. $\overline{Y}_{0T} - \overline{Y}_{0U}$ и $\overline{Y}_{1T} - \overline{Y}_{1U} > 0$). Поскольку это явно не сравнение яблок с яблоками», мы не можем быть уверены, что разница в средних результатах между двумя группами является свидетельством влияния школы Пройсса. Даже если бы ATE был равен нулю, мы все равно наблюдали бы положительную разницу в средних значениях. Именно это и означает утверждение, что корреляция не то же самое, что причинно-следственная связь.

Лотерея стала убедительным доказательством именно потому, что она рандомизировала учащихся на подвергавшихся и не подвергавшихся воз-

действию. Рандомизация гарантирует, что в среднем обе группы одинаковы в отношении потенциальных результатов. То есть, если бы мы проводили рандомизацию снова и снова, в среднем две группы имели бы одинаковые исходные результаты. (Конечно, для любого отдельного этапа рандомизации все же могут существовать непричинные различия в успеваемости между двумя группами просто из-за различий в выборке или других видов шума.) Следовательно, разница в средних результатах между победителями лотереи и проигравшими в лотерее – это объективная оценка причинного эффекта школы.

Говоря о причинно-следственной связи, мы часто используем язык, напоминающий об экспериментах, как мы это сделали здесь, обсуждая воздействие. Мы делаем это, потому что эксперименты дают отличный способ критически подойти к выводу причинно-следственной связи из корреляции. Если проводится экспериментальная рандомизация по группам, то не существует систематических исходных различий между группами, подвергавшимися и не подвергавшимися воздействию.

Однако важно отметить, что во многих случаях, когда нас интересует причинно-следственная связь, нам фактически не удается провести эксперимент. Вместо этого некоторые люди испытывают воздействие, а другие нет по причинам, которые от нас не зависят. В таких случаях мы должны быть очень осторожны при интерпретации корреляции между эффектом и воздействием как оценки причинно-следственной связи. Как вы видели, положительная корреляция между результатами тестов (эффект) и посещением школы Пройсса (воздействие) среди популяции в целом на самом деле не свидетельствовала о причинно-следственной связи. Причина в том, что социальные процессы привели к возникновению базовых различий между группами, испытавшими и не испытавшими воздействие.

Источники смещения

Чтобы правильно интерпретировать корреляции, нам необходимо отчетливо понимать, когда возникают систематические базовые различия, поскольку именно они приводят к смещению. Есть два основных источника таких различий: искажающие факторы и обратная причинно-следственная связь. Понимание природы этих источников – большой шаг к умению различать, когда вы можете, а когда не можете узнать что-то достоверное о причинно-следственной связи на основе корреляции.

Искажающие факторы

Искажающий фактор (confounder) – это свойство мира, которое удовлетворяет двум условиям:

1) оно оказывает влияние на статус воздействия;
2) оно оказывает влияние на результат сверх того влияния, которое оно оказывает на статус воздействия.

Искажающие факторы создают базовые различия и, следовательно, систематическое смещение. Предположим, какое-то свойство мира увеличивает вероятность того, что люди подвергнутся воздействию. И предположим, что

это свойство также повышает вероятность того, что люди добьются определенного результата. Тогда из-за фактора, искажающего результат, возникнет корреляция между этим результатом и воздействием по причинам, не связанным с каким-либо фактическим эффектом воздействия. Следовательно, если такие искажающие факторы существуют (и мы ничего не сделали для их объяснения, о чем мы поговорим в следующих главах), то было бы ошибкой интерпретировать корреляцию между эффектом и воздействием как несмещенную оценку причинного эффекта воздействия.

Чтобы добавить немного конкретики, вспомните нашу обеспокоенность тем, что корреляция между посещением школы Пройсса и академической успеваемостью не была убедительным доказательством причинно-следственной связи. Это беспокойство было связано с базовыми различиями, возникшими в результате того, что более талантливые дети с большей вероятностью будут стремиться (или иметь семьи, которые стремятся) в школу Пройсса. Другой способ выразить ту же обеспокоенность заключается в том, что изначальный талант ребенка к учебе является искажающим фактором. Этот талант влияет на статус воздействия: дети, которые более талантливы, с большей вероятностью попадут в группу воздействия (т. е. пойдут в школу Пройсса). Талант к учебе влияет на результаты помимо того, что он влияет на статус воздействия: более одаренные дети будут лучше сдавать тесты просто в силу своих способностей, помимо того факта, что они стремятся в лучшие школы. Изучение победителей и проигравших в лотерею помогло опровергнуть причинно-следственную связь, поскольку разорвало связь между талантом (искажающий фактор) и посещением школы Пройсса (воздействие).

Рассмотрим еще один пример. Многие исследования показывают сильную отрицательную корреляцию между состоянием экономики страны и тем, переживает ли она гражданскую войну. Есть основания полагать, что в основе этой корреляции может лежать причинно-следственная связь – например, возможно, когда дела в экономике идут лучше, люди тоже живут лучше и, следовательно, с меньшей вероятностью будут готовы участвовать в серьезных беспорядках. Но, прежде чем интерпретировать корреляцию как причинно-следственную, необходимо подумать о том, есть ли потенциальные искажающие факторы. Один из факторов, о котором стоит подумать, связан с однородностью общества. В самом деле однородное общество, сохраняющее эту однородность на протяжении длительного исторического периода, намного меньше склонно к гражданской войне в период экономических неурядиц, чем общество, которое раздирают культурные или социальные противоречия. Следовательно, неоднородность общества является искажающим фактором. Таким образом, мы не имеем права интерпретировать корреляцию между экономическим процветанием и риском гражданской войны как объективную оценку причинно-следственной связи.

Итак, первый шаг в оценке того, является ли корреляция свидетельством причинно-следственной связи, – это спросить себя, существуют ли какие-либо искажающие факторы. Схематическое изображение на рис. 9.3 поможет вам не забыть это сделать. Вопрос, который задает рисунок, заключается в следующем: существуют ли для рассматриваемой пары воздействие–эффект какие-либо факторы, которые, по вашему мнению, относятся к группе искажающих факто-

ров? Для утвердительного ответа на этот вопрос должны соблюдаться два условия. Во-первых, стрелка от искажающего фактора к воздействию должна иметь смысл – т. е. вы должны быть уверены, что искажающий фактор оказывает влияние на воздействие. И во-вторых, должна иметь смысл стрелка от искажающего фактора к результату – т. е. вы должны быть уверены, что искажающий фактор может оказать влияние на результат воздействия. Если вы можете назвать хотя бы один фактор, удовлетворяющий обоим этим условиям, то у вас есть разумные опасения по поводу искажающих факторов, и следует опасаться причинно-следственной интерпретации корреляции между воздействием и эффектом.

Рис. 9.3. Искажающий фактор оказывает влияние на воздействие и параллельно влияет на эффект

Обратная причинно-следственная связь

Второй источник предвзятости, о котором нам следует беспокоиться, – это *обратная причинность* (reverse causality). Обратная причинность существует, если результат каким-то образом меняет воздействие. Обратная причинно-следственная связь создает базовые различия, потому что, если результат влияет на то, испытает ли группа воздействие или нет, в результатах между группами, испытавшими и не испытавшими воздействие, возникнут систематические различия, которые не связаны с эффектом воздействия.

Рассмотрим еще раз наш пример отрицательной корреляции между состоянием экономики страны и гражданской войной. Мы уже видели, что в основе этих отношений могут быть факторы, способные сбить с толку. Но может быть и обратная причинно-следственная связь. Например, в ходе гражданской войны разрушается инфраструктура, нарушается производство, гибнут люди. Все эти последствия гражданской войны напрямую снижают экономическое процветание. Следовательно, отрицательная корреляция между показателем экономического процветания и гражданской войной может отражать влияние войны на экономику, а не влияние экономики на войну. Возможность такой обратной причинно-следственной связи является еще одной причиной того, что причинная интерпретация этой корреляции не оправдана.

Схематическое изображение на рис. 9.4 – это способ напомнить себе о необходимости проверки обратной причинно-следственной связи, прежде чем интерпретировать корреляцию как прямую причинно-следственную связь. Вопрос, который задает рисунок, заключается в следующем: существуют ли для любого данного воздействия и эффекта потенциальные источники обратной причинно-следственной связи? То есть можем ли мы придумать причины, по которым стрелка может идти от результата к воздействию? Если да, то

ваше беспокойство по поводу обратной причинно-следственной связи вполне обосновано, и следует с осторожностью относиться к причинно-следственной интерпретации корреляции между воздействием и результатом.

Рис. 9.4. Обратная причинно-следственная связь возникает, когда результат влияет на воздействие

В общем, если кто-то покажет вам корреляцию между результатом и воздействием, без дополнительной информации и исследований вам будет сложно узнать, каким образом возникла эта корреляция: воздействие влияет на результат или результат меняет воздействие, или искажающие факторы влияют как на воздействие, так и на результат, или наблюдается некая комбинация всех этих возможностей.

На рис. 9.5 представлена общая схема размышлений о двух источниках смещения, которые мы обсуждали. Теперь, когда вы прочно овладели понятиями, представленными на рис. 9.5, давайте потренируемся критически размышлять о корреляции и причинно-следственной связи, искажающих факторах и обратной причинно-следственной связи, подробно рассмотрев пару примеров.

Рис. 9.5. Искажающие факторы и обратная причинно-следственная связь – два ключевых источника систематической ошибки при оценке причинно-следственных связей

Новый взгляд на правило 10 000 часов

Наверное, вам не знакомо имя Дэна МакГлафлина, но в 2010 г. о нем довольно много писали в прессе. В апреле того же года МакГлафлин оставил работу фотографа, чтобы осуществить мечту – профессионально играть в гольф. Он плани-

ровал заниматься гольфом не менее 30 часов каждую неделю в течение более шести лет, пока не наберет 10 000 часов целенаправленных тренировок. Он верил, что к концу этих 10 000 часов он станет опытным игроком в гольф, готовым пройти квалификацию на турнир PGA (Ассоциация профессионалов гольфа). Все получилось не совсем так, как планировал МакГлафлин. Он не прошел в PGA, но открыл, судя по всему, довольно крутую компанию по производству газированных напитков.

Донкихотский план МакГлафлина должен показаться знакомым. Он довел правило 10 000 часов Малкольма Гладуэлла, о котором мы говорили еще в главе 4, до его (не)логической крайности. Дескать, талант вторичен; большой успех зависит от того, как вы потратите эти 10 000 часов. Таким образом, любой из нас может добиться практически всего, даже карьеры в профессиональном спорте, если просто посвятит 10 000 часов своей жизни действительно серьезным тренировкам.

Как мы уже обсуждали, доказательства Гладуэлла в пользу правила 10 000 часов – даже просто как утверждение о корреляции – сомнительны из-за отсутствия вариаций. Но МакГлафлин полностью полагался не только на эти доказательства. Его также вдохновили исследования психолога Андерса Эрикссона из Университета штата Флорида.

Эрикссон утверждает, что ключом к сверхвысоким результатам в любом деле является целенаправленная практика. По его словам, после достижения определенного уровня знаний в какой-то задаче профессиональный уровень людей имеет тенденцию к стабилизации, даже если они продолжают набираться опыта или общей практики. Единственный способ продолжать совершенствоваться на этом этапе – это целенаправленная практика – работа над упражнениями, специально предназначенными для определенных аспектов производительности. Чем более осознанной будет практика, тем лучше будет результат. Что отличает настоящих мастеров своего дела от хороших, но не великих экспертов, так это количество времени, посвященное осознанной практике.

Идея 10 000 часов возникла в результате плодотворного исследования опытных музыкантов, проведенного Эрикссоном и его коллегами. В отличие от Гладуэлла у Эрикссона есть вариации. Он исследовал учащихся курса скрипки элитной музыкальной школы в Берлине. Все студенты, участвовавшие в исследовании, были опытными скрипачами. Но тем не менее их игра различалась по уровню. Эрикссон попросил преподавателей выделить три группы – самых лучших скрипачей, которые, скорее всего, сделают карьеру солистов или в крупных оркестрах, хороших скрипачей, у которых меньше шансов сделать успешную исполнительскую карьеру, и самую слабую группу скрипачей, которые, скорее всего, сделают карьеру преподавателей. (Стараемся не обижаться.)

Скрипачей спрашивали об их истории практики: о возрасте, в котором они начали, количестве часов в неделю, типе деятельности, которой они занимались во время практики, уровне концентрации и т. д. Их также попросили вести дневники, в которых записывались их практические привычки. Вооружившись этой информацией, исследователи смогли сравнить практическое поведение разных групп скрипачей. Вывод: лучшие скрипачи к 18 годам практиковались не менее 10 000 часов, в то время как менее опытные скрипачи из второй группы практиковались только около 7 500 часов, а будущие преподаватели из третьей группы практиковались только около 5 000 часов. Более того, лучшие

скрипачи отличались тем, что сознательно тратили большую часть своего времени на целенаправленные занятия. Например, они тратили больше времени на сложные задачи, направленные на повышение отдельных нюансов исполнительского мастерства, а не на простое повторение удобных произведений, которые они уже освоили. Подобные исследования сообщают об аналогичных результатах для шахматистов, гимнастов и других людей. Таким образом, данные показывают положительную корреляцию между сознательной практикой и высокими достижениями.

Учитывая эти доказательства, похоже, что и правило 10 000 часов, и упор на целенаправленные занятия не могут быть такими уж надуманными. Наиболее эффективные эксперты во многих областях, похоже, не отличаются измеримыми физическими характеристиками. Ключевым моментом, по-видимому, является тот факт, что лучшие результаты получают профессионалы, которые тренируются наибольшее количество часов и наиболее целенаправленно. Так что, возможно, целенаправленные тренировки в течение 10 000 часов действительно помогут вам выйти на мировой уровень. Возможно, у Дэна МакГлафлина были неплохие шансы стать следующим Тайгером Вудсом.

Но, прежде чем вы отложите книгу и пойдете играть в гольф, давайте подумаем еще немного. Эрикссон не повторил ошибки Гладуэлла. У него были вариации, и поэтому он установил корреляцию между сознательной практикой и достижениями. Но это не означает, что корреляция отражает причинный эффект. Чтобы прийти к такому выводу, нам нужно подумать об искажающих факторах и обратной причинно-следственной связи.

Вот одна из возможных проблем. Предположим, врожденный природный талант действительно важен. Представьте себе двух девочек, обе любят играть на скрипке. Одна из них более талантлива от природы, чем другая. Они обе усердно занимаются, проводя по многу часов со скрипкой. На первом этапе тяжелая работа окупается, и обе быстро прогрессируют. Но со временем начинает проявляться разница в талантах. Более талантливая девочка быстрее и точнее осваивает сложные музыкальные произведения. Она получает больше похвал и возможностей для выступлений, чем менее талантливый ребенок.

Время идет, и двое детей становятся подростками. Появляются новые возможности и развлечения – свидания, спорт. Каждый подросток должен решить, сколько времени и энергии ему продолжать уделять занятиям игрой на скрипке. Более талантливая из двух девочек-подростков обнаруживает, что каждый раз, когда она посвящает день игре на скрипке, она осваивает новые навыки и репертуар. Этот прогресс и достижения вдохновляют. Возникает петля положительной обратной связи, благодаря которой практика ведет к успеху, что вдохновляет на дальнейшую практику и большую концентрацию. Поэтому она продолжает посвящать себя целенаправленным занятиям игрой на скрипке, достигнув этих волшебных 10 000 часов к 18 годам.

Менее талантливая из двух подростков также прогрессирует каждый раз, когда посвящает день игре на скрипке, но делает это медленнее и с меньшим мастерством. Произведение, на освоение которого у более талантливого подростка уходит неделя, может занять у нее месяц. И даже тогда она играет с меньшей точностью и артистичностью. Ее достижения приходят медленнее и получают меньше похвал. Отсутствие прогресса разочаровывает. Получив менее положи-

тельные отзывы, она убеждается, что практика не так уж полезна. В результате она по-прежнему любит играть на скрипке и продолжает усердно заниматься, но по мере появления новых возможностей она, скорее всего, возьмет несколько часов или день перерыва в занятиях по игре на скрипке, чтобы реализовать их. И даже во время занятий, возможно, она менее сосредоточена, потому что у нее появились новые интересы. К 18 годам она тратит на тренировки на три четверти меньше времени, чем ее более талантливая подруга, и делает это намеренно.

Двое молодых людей, подобных тем, кого мы только что описали, легко могли бы стать одноклассниками музыкальной школы в исследовании Эрикссона. Более талантливого из них преподаватели считали бы одним из лучших скрипачей, а менее талантливый был бы признан хорошим, но не великим. Сравнивая их, Эрикссон обнаружил бы, как и случилось, что более сильный из двух скрипачей потратил 10 000 часов на целенаправленные занятия, в то время как более слабый потратил всего около 7500 часов на менее осознанную практику.

Из этого сравнения Эрикссон приходит к выводу, что разница в их успехе обусловлена практикой. Но, как мы видели, такая причинная интерпретация корреляции ошибочна. Как показано на рис. 9.6, врожденный талант вполне может быть фактором, влияющим на количество целенаправленной практики (воздействие) и оказывающим прямое влияние на достижения (результат), помимо эффекта от практики. Различия в талантах вызывают базовые различия в достижениях.

Рис. 9.6. Является ли корреляция между осознанной практикой и достижениями объективной оценкой причинно-следственной связи?

Конечно, из истории, которую мы рассказали, вовсе не следует, что практика не имеет никакого эффекта. На успех, несомненно, влияют талант, практика и сочетание того и другого (самый одаренный от природы человек в мире не смог бы стать великим скрипачом без практики). Но в нашем гипотетическом примере более талантливая ученица, скорее всего, все равно будет лучше играть на скрипке, чем ее одноклассница, даже если она занималась всего 7500 часов, а менее талантливая ученица все равно будет играть хуже, даже если она заставит себя заниматься 10 000 часов.

Здесь важна степень, в которой корреляция отражает причинно-следственную связь с учетом смещения на разницу в талантах. Чтобы понять почему, вспомните Дэна МакГлафлина. До ухода с работы МакГлафлин не проявлял никаких признаков игрока в гольф мирового класса. Этот факт поможет объяснить, почему он раньше не тратил тысячи часов на целенаправленную практику. Если значительная часть корреляции между практикой и успехом обусловлена врожденным талантом, а не причинным следствием практики,

то усилия МакГлафлина вряд ли приведут к желаемому результату. Глубоко талантливые люди много тренируются и добиваются больших успехов. Это не значит, что человек, которому не хватает глубокого таланта, но который заставляет себя много тренироваться, добьется такого же успеха. И поэтому, возможно, ему следовало предвидеть, что его приключение закончится именно так, в компании по производству газированных напитков, а не на поле PGA. (Как бы то ни было, Итан считает, что управлять частной компанией по производству газированных напитков намного интереснее, чем профессионально играть в гольф. Энтони с этим не согласен.)

Диетическая газировка

К слову о газированных напитках, на момент написания этой книги эксперты по питанию почти пришли к единому мнению, что диетическая газировка вредна для здоровья. Исследования экспертов в уважаемых научных журналах связали потребление диетических газированных напитков с рядом проблем со здоровьем, включая ожирение, диабет и сердечные приступы.

Любопытно, что, несмотря на все якобы веские доказательства опасности диетической газировки, у ученых пока нет убедительного объяснения (кроме неблагоприятных последствий употребления диетической газировки для зубов – кислота вредна для зубной эмали). Они ломали голову, пытаясь объяснить, почему напитки, практически не содержащие калорий, каким-то образом вызывают у людей лишний вес.

Было выдвинуто несколько теорий. Одно из объяснений состоит в том, что в диетической газировке содержатся химические вещества, которые могут быть вредны для нас. Но все, что мы потребляем, состоит из химических веществ, так что это не является исчерпывающим объяснением. Другое объяснение состоит в том, что диетическая газировка сбивает с толку ваше тело и заставляет его хотеть больше калорий. Гипотеза гласит, что после употребления диетической газировки ваш мозг ожидает получить калории от сладкого напитка, а когда этого не происходит, мозг заставляет вас совершить набег на кухню в поисках печенья и чипсов. В этом смысле ваш мозг похож на ребенка, которому пообещали дать конфету, но в последний момент вручили шпинат. Третье объяснение заключается в том, что диетическая газировка (и, предположительно, все сладкое) снижает чувствительность ваших вкусовых рецепторов, а это означает, что вам нужно есть все больше и больше сладких продуктов, чтобы получить желаемые ощущения.

Мы не диетологи, но ни одно из этих объяснений не звучит убедительно для наших неподготовленных ушей. Действительно, мы могли бы придумать столь же убедительные истории о том, почему эффект должен быть направлен в противоположную сторону. Диетическая газировка позволяет сладкоежкам насладиться освежающим лакомством, не потребляя дополнительных калорий. Диетическая газировка может даже заставить ваш мозг думать, что вы употребили много калорий, и, следовательно, ускорить метаболизм, что полезно для здоровья и снижения веса. Как мы говорим, мы не эксперты, но кажется не менее правдоподобным и то, что диетическая газировка полезна для здоровья, особенно в качестве заменителя сладких напитков. Так почему же эксперты поголовно согласны с тем, что диетическая газировка вредит здоровью?

Мы познакомились с их исследованиями и выделили основные доказательства. Существует отрицательная корреляция между употреблением диетической газировки и последствиями для здоровья. Люди, которые пьют диетическую газировку, чаще страдают ожирением, диабетом и целым рядом других проблем со здоровьем, чем люди, которые не пьют никаких сладких напитков. Прежде чем согласиться с диетологами в том, что эта корреляция отражает истинное причинное влияние диетической газировки на здоровье, нам следует задуматься о том, нет ли здесь каких-то искажающих факторов или обратной причинно-следственной связи.

Что, если частые перекусы заставляют людей чаще пить газировку (потому что газировка хорошо сочетается с едой) и приводят людей к ожирению по причинам, не связанным с газировкой? Тогда перекус будет искажающим фактором. Или, возможно, это обратная причинно-следственная связь: что, если ожирение или диабет заставляют людей чаще пить диетическую газировку? Если вы любите газированные напитки и обнаружили у себя диабет, вы перейдете на диетические газированные напитки, не так ли? Точно так же можно представить себе переход от диетических газированных напитков к обычным сладким напиткам, если со здоровьем все в порядке. Очевидно, что искажающие факторы и обратная причинно-следственная связь вызывают серьезную озабоченность, и мы не должны рассматривать корреляцию между диетическими газированными напитками и последствиями для здоровья как надежную оценку причинного эффекта.

Насколько похожи искажающие факторы и обратная причинность?

Углубившись в рассуждения об искажающих факторах и обратной причинно-следственной связи, стоит остановиться и подумать о том, как они связаны друг с другом. Часто проблему, которая, по-видимому, связана с обратной причинно-следственной связью, можно также рассматривать с точки зрения искажающих факторов, где соответствующим искажающим фактором является просто ожидаемый результат.

Чтобы понять, что мы имеем в виду, вернемся еще раз к нашему примеру отрицательной корреляции между экономикой и риском гражданской войны. Мы видели, что существуют как искажающие факторы, так и обратная причинность, которые делают недостоверной причинную интерпретацию этой корреляции. Рассмотрим еще одну проблему. Предположим, что по разным причинам (например, этнические и религиозные разногласия, гражданские войны в соседних странах) люди считают, что какая-то страна подвержена высокому риску гражданской войны. Этот риск гражданской войны может отпугнуть инвестиции в страну, привести к оттоку капитала, вызвать утечку мозгов и т. д. Таким образом, ожидание *будущей* гражданской войны может привести к ослаблению экономики страны уже сейчас. Это явление можно рассматривать как случай обратной причинно-следственной связи: риск гражданской войны приводит к экономической слабости. Но, возможно, ситуация будет понятнее, если рассматривать ее как разновидность искажения, когда искажающими являются любые факторы, которые заставляют людей поверить в то, что страна подвержена высокому риску гражданской войны. Эти факторы ведут к осла-

блению экономики, сдерживая инвестиции и вызывая бегство специалистов из страны. И, по-видимому, они заставляют людей поверить в то, что в стране существует высокий риск гражданской войны именно потому, что они оказывают независимое влияние на возможность начала гражданской войны.

Давайте рассмотрим другой пример – расходы на предвыборную кампанию.

Расходы на предвыборную кампанию

Политические кандидаты тратят огромное количество времени на сбор денег для своих предвыборных кампаний. Члены конгресса, например, часто проводят несколько часов в день в колл-центре, обзванивая богатых избирателей и прося их помочь финансировать следующее переизбрание. (Оказывается, работа в конгрессе не такая уж гламурная.)

Конечно, политики делают это, потому что уверены, что деньги на предвыборную кампанию необходимы для победы на выборах. А консультанты предвыборной кампании постоянно советуют кандидатам, сколько им следует потратить на телевизионную рекламу, цифровую рекламу, прямую почтовую рассылку и личную работу с избирателями. Избирательные кампании, очевидно, представляют собой большой бизнес, основанный на идее, что кандидаты могут повысить свои шансы на успех, собирая и тратя больше денег.

Учитывая масштабы расходов на предвыборную кампанию, политологи посвятили много времени и труда оценке отдачи от этих усилий. Могут ли расходы на рекламу действительно повлиять на результаты выборов? И достаточно ли велики эти последствия, чтобы оправдать миллионы долларов, пожертвованные на финансирование кампаний, и многие тысячи часов, потраченные на сбор этих долларов?

Одно из самых ранних и наиболее влиятельных исследований расходов на предвыборную кампанию было проведено Гэри Джейкобсоном в 1978 г. Джейкобсон заключает, что расходы на предвыборную кампанию, похоже, значительно улучшают электоральные перспективы претендентов, но приносят мало пользы действующим президентам. Действительно, расходы на предвыборную кампанию действующих политиков могут даже оказаться контрпродуктивными, нанося ущерб их успехам на выборах!

Каковы доказательства Джейкобсона в пользу утверждения о том, что расходы на предвыборную кампанию помогают претендентам, но не действующим политикам? Судя по данным, наблюдается очень сильная положительная корреляция между расходами претендентов и долей голосов, отданных за этих претендентов. Однако расходы действующих политиков отрицательно коррелируют с долей голосов, набранных этими политиками.

По мнению Джейкобсона, одно из объяснений этой корреляции заключается в том, что действующие игроки обычно собирают и тратят больше денег, чем претенденты. Возможно, некоторая начальная сумма расходов на уровне, который мы обычно видим для претендентов, помогает кандидату добиться известности и убедить избирателей. Но, возможно, слишком большие расходы со стороны уже известного действующего политика раздражают и отпугивают потенциальных сторонников. В связи с этим действующие политики совершают систематические ошибки, как теряя время на сбор денег, так и расходуя эти деньги после того, как они их собрали.

Конечно, сравнения, лежащие в основе этих корреляций, не могут быть прямыми. Нам нужно подумать об искажающих факторах и обратной причинно-следственной связи.

Одна из серьезных проблем в этом отношении связана с электоральным потенциалом. Какие претенденты, как правило, способны собрать и потратить много денег? По-видимому, популярные политики с реальными шансами на победу. Доноры, скорее всего, захотят инвестировать именно в электорально сильных или особо талантливых кандидатов. Поэтому было бы ошибкой интерпретировать положительную корреляцию между расходами претендентов и результатами выборов как чисто причинную. Она, по крайней мере частично, отражает базовые различия в электоральном потенциале между претендентами, которые могут и не могут собрать много денег.

В этом примере мы хотим, чтобы вы обратили внимание, что влияние электорального потенциала можно рассматривать как проблему обратной причинно-следственной связи или как проблему искажения. С точки зрения обратной причинно-следственной связи вы можете описать ситуацию следующим образом: «Когда претендент известен, он может собрать и потратить больше на свою кампанию». Если рассматривать это как искажающий фактор, вы можете описать его следующим образом: «Когда у претендента есть характеристики, которые делают его конкурентоспособным, они влияют как на его способность собирать и тратить деньги, так и на его способность победить на выборах». Оба предложения описывают одну и ту же проблему, просто сформулированную несколько по-разному.

Аналогичный аргумент справедлив и для действующих политиков. В целом, хотя они тратят и собирают довольно много денег, большинство действующих политиков на выборах в США находятся в достаточной электоральной безопасности. Те, кому действительно нужно приложить много усилий для сбора и расходования денег, – это политики, уязвимые с электоральной точки зрения, например замешанные в коррупционном скандале. Следовательно, действующие политики тратят огромные деньги не тогда, когда они сильны, а когда они слабы. И опять же, эту проблему можно рассмотреть с точки зрения обратной причинно-следственной связи – «Электорально слабые политики тратят больше денег» – или с точки зрения искажающих факторов – «Характеристики, которые ослабляют потенциал политиков, ухудшают их конкурентоспособность, заставляют тратить больше денег и приводят к худшим, чем в среднем, результатам выборов».

Дальнейшие исследования с использованием рандомизированных экспериментов и других умных методов, направленных на выявление причинно-следственной связи, в целом показали, что расходы на предвыборные кампании действительно оказывают положительное влияние как на претендентов, так и на действующих политиков, хотя размер этого эффекта обычно невелик. Кандидатам, возможно, придется потратить сотни долларов, чтобы повлиять на один голос, а это означает, что прямое существенное влияние на исход крупных выборов посредством пожертвований на кампанию, как правило, невозможно. Например, рассмотрим губернаторскую или сенаторскую гонку в большом штате США. Даже в гонке, которая считается очень конкурентной, исход, скорее всего, будет определяться разницей в несколько сотен тысяч голосов. Это означает, что, если бы спонсоры захотели напрямую повлиять на исход вы-

боров, им пришлось бы потратить десятки миллионов долларов и надеяться, что их расходы не вызовут компенсирующей реакции со стороны сторонников оппонента. По этой причине даже самые крупные спонсоры, вероятно, повлияли на весьма небольшое количество выборов.

Как видите, не так уж важно, считаем ли мы такие случаи обратной причинно-следственной связью или следствием влияния мешающих факторов. Что действительно важно, так это проверка корреляции на предмет возможных базовых различий, будь то из-за искажающих факторов или обратной причинно-следственной связи. Если базовые различия существуют, следует проявить должную осторожность, прежде чем интерпретировать корреляцию как свидетельство наличия причинно-следственной связи.

Признаки смещения

При наличии искажающих факторов или обратной причинно-следственной связи корреляция между воздействием и результатом не является объективной оценкой истинной причинно-следственной связи (будь то ATE, ATT, ATU или другие показатели причинности, которые мы обсудим в последующих главах). Но иногда мы можем добиться некоторого прогресса в изучении причинности, задав вопрос, переоценивает или недооценивает корреляция причинный эффект.

Давайте вспомним наше любимое уравнение, на этот раз записанное в терминах причинно-следственной связи:

Наблюдаемая корреляция (оценка) = Истинный причинный эффект (оцениваемая величина) + смещение + шум.

Предположим, что мы наблюдаем положительную корреляцию между назначением определенного лечения и показателями выживаемости после инсульта. Но предположим также, что есть факторы, которые вы не учли, поэтому в вашей оценке истинного причинного эффекта имеется смещение. Если у вас есть основания полагать, что систематическая ошибка положительна, то наблюдаемая корреляция является переоценкой истинного причинного эффекта лечения. Это означает, что на основании только наблюдаемой положительной корреляции вы не можете утверждать, что лечение вообще хоть что-то дает. Даже если истинный причинный эффект равен нулю, в среднем вы будете наблюдать положительную корреляцию, которая обусловлена факторами, создающими положительное смещение.

Теперь предположим, что смещение является отрицательным, а не положительным. В этом случае наблюдаемая корреляция является недооценкой истинного причинного эффекта лечения. Если вы уверены, что наблюдаемая корреляция положительна, вы должны быть еще более уверены в том, что истинный причинный эффект положителен. И это может быть полезно знать. Например, предположим, что проведение лечения было бы хорошей идеей (учитывая различные издержки), даже если истинный эффект равен наблюдаемой корреляции. Тогда тот факт, что наблюдаемая корреляция является недооценкой истинного эффекта, предполагает, что вам следует назначить лечение.

Обратите внимание, что вы все равно можете ошибиться из-за шума. Даже если вы в среднем недооцениваете истинный причинный эффект, это не означает, что конкретная оценка всегда будет ниже истинного эффекта. Это просто означает, что ваши оценки будут ниже, чем реальный эффект в среднем.

Поскольку такого рода размышления о знаке смещения в оценке иногда могут повлиять на принимаемое решение, полезно потратить немного времени на концептуальное размышление о том, когда искажающие факторы ведут к переоценке истинного эффекта на основе корреляции, а когда – к недооценке.

Начнем с обсуждения взаимосвязи между голосами и расходами претендентов на избирательную кампанию, где мы беспокоились, что электоральный потенциал может вносить искажения причинности. Но в какую сторону направлены эти искажения? Кажется логичным, что высокий потенциал кандидата положительно влияет как на сбор средств, так и на количество голосов за претендентов. Следовательно, некоторые дополнительные голоса в пользу претендентов с высокими расходами на самом деле получены благодаря личным качествам кандидатов, а не вследствие больших расходов. Значит корреляция между расходами и количеством голосов будет переоценкой истинного эффекта.

Чтобы увидеть еще один пример, давайте вернемся к обсуждению положительной корреляции между посещением чартерной школы и результатами стандартизированных тестов. Здесь, как мы сказали, одним из возможных искажающих факторов является наличие выраженных способностей к учебе у детей, поступающих в чартерную школу. И тот факт, что эти ученики более талантливы или мотивированы, может напрямую повлиять на их результаты тестов.

Если эта гипотеза верна, то в какую сторону смещена оценка корреляции между посещением чартерной школы и результатами стандартизированных тестов? Давайте подумаем об этом. Высокая одаренность положительно влияет на вероятность того, что ребенок пойдет в чартерную школу, а также положительно влияет на результаты тестов. Это означает, что наблюдаемая положительная связь между посещением чартерной школы и результатами тестов отчасти является результатом различий в природных способностях. Следовательно, этот искажающий фактор приводит к переоценке истинного эффекта. То есть смещение в нашем любимом уравнении положительное.

Совершенно очевидно, что аналогичные рассуждения верны, если бы у нас был фактор, который негативно влияет как на посещение чартерной школы, так и на результаты тестов. Действительно, это тот же самый случай, но с противоположным знаком. Если мы говорим о таком факторе, как «отсутствие способностей к учебе», а не о «способности к учебе», то этот фактор оказывает негативное влияние на воздействие и результат, но по-прежнему приводит к тому, что наблюдаемая корреляция является переоценкой истинного эффект. Таким образом, как показано на рис. 9.7, если у вас есть фактор, который с одинаковым знаком влияет как на воздействие, так и на результат (независимо от того, отрицательный или положительный), то неспособность учесть этот фактор приведет к возникновению положительной систематической ошибки. В таких обстоятельствах наблюдаемая корреляция будет склонна завышать истинный эффект.

Теперь рассмотрим фактор, который оказывает взаимно противоположное влияние на воздействие и результат. Например, предположим, что учащиеся из бедных районов более мотивированы подавать заявления в чартерные шко-

лы (возможно, потому что их местные государственные школы недостаточно финансируются), но будут хуже учиться из-за проблем в их среде обитания. Это искажающий фактор – он влияет как на воздействие (посещает ли ученик чартерную школу), так и на результат (оценки за стандартизированные тесты). Но, в отличие от таланта (который положительно влияет на воздействие и потенциальные результаты), этот фактор создает негативное смещение. Следовательно, наблюдаемая корреляция между посещаемостью чартерных школ и результатами стандартизированных тестов является скорее недооценкой истинного эффекта.

Почему это так? В нашей новой гипотезе жизнь в бедном районе положительно влияет на вероятность того, что ученик пойдет в чартерную школу, но отрицательно влияет на результаты тестов. Это означает, что наблюдаемая корреляция между посещением чартерной школы и результатами тестов отражает тот факт, что среди учащихся чартерных школ преобладают дети из бедных районов. Этот факт способствует снижению результатов тестов учащихся чартерных школ по причинам, не имеющим ничего общего с обучением в чартерной школе. Если бы в чартерных и государственных школах обучалось одинаковое количество детей из более богатых и бедных районов, положительная корреляция между посещаемостью чартерных школ и результатами тестов была бы еще более положительной. Следовательно, этот искажающий фактор приводит к недооценке истинного эффекта.

Очевидно, то же самое было бы верно, если бы у нас был фактор, который отрицательно влияет на вероятность посещения чартерной школы и положительно – на результат. Таким образом, как показано на рис. 9.7, если у вас есть фактор, который влияет на воздействие с одним знаком и на результат с противоположным знаком, этот фактор создает отрицательное смещение. В таких обстоятельствах наблюдаемая корреляция будет меньше истинного эффекта.

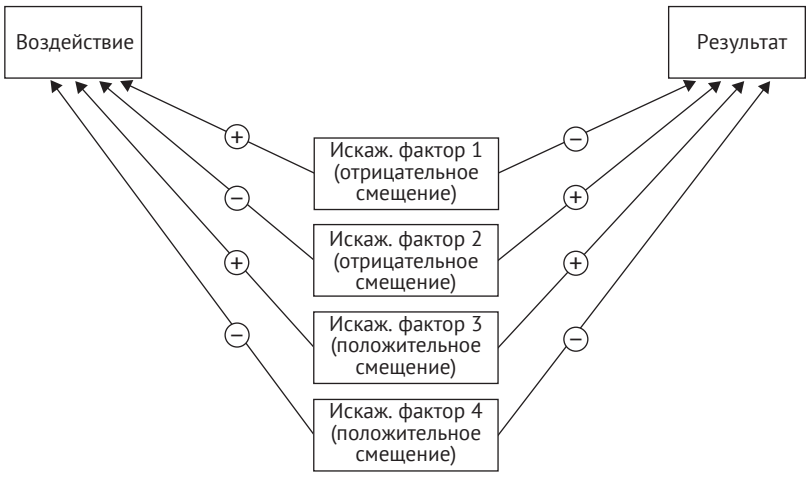


Рис. 9.7. Зависимость от знака воздействия искажающего фактора

Влияние знака еще очевиднее в случае обратной причинно-следственной связи. Результат влияет на воздействие через петлю обратной связи. Таким об-

разом, если результат оказывает положительное влияние на воздействие, смещение является положительным. Это означает, что наблюдаемая корреляция является завышенной оценкой истинного причинного эффекта. А если результат оказывает отрицательное влияние на воздействие, смещение оказывается отрицательным, поэтому наблюдаемая корреляция является заниженной.

Если бы кроме знака у нас была и другая информация, мы могли бы не только обозначить смещение, но и сказать что-то о его величине. При некоторых допущениях смещение, вызванное искажающим фактором, представляет собой просто влияние искажающего фактора на результат, умноженное на меру корреляции между искажающим фактором и воздействием (измеряемую коэффициентом, который вы получили бы в результате регрессии искажающего фактора на воздействие).

Как мы увидим в главе 10, если у нас есть данные об этом факторе, мы можем попытаться устранить смещение с помощью методики учета и ограничения искажающего фактора. Но даже если у нас нет этих данных, можно все равно сделать некоторые предположения о том, в какой степени искажающий фактор влияет на результат и коррелирует с воздействием, чтобы оценить величину систематической ошибки.

Из предыдущих рассуждений следует, что мы можем кое-что узнать о причинных эффектах даже на основе смещенных оценок. Значит не нужно спешить выбрасывать свой анализ в мусорную корзину только потому, что обнаружились искажающие факторы: если у нас есть хорошие предположения о направлении и величине смещения, то мы все равно сможем многому научиться. Но зачастую трудно понять, в какой степени наблюдаемая корреляция является результатом смещения, поэтому мы стараемся не использовать простые корреляции для изучения причинно-следственных связей. Менее наивные и более информативные подходы к причинному выводу будут в центре внимания последующих глав.

Разумный подход к изучению причинных эффектов на основе потенциально смещенных корреляций основан на обратном порядке действий. Вместо того чтобы делать выводы о величине эффекта на основе догадок о величине систематической ошибки, мы можем начать с предположения, что истинный эффект равен нулю, а затем задаться вопросом, насколько большой должна быть систематическая ошибка, чтобы объяснить наблюдаемую корреляцию. Если оценка этой ошибки неправдоподобно велика, то мы можем заключить, что эффект, скорее всего, не равен нулю. Этот вид анализа часто называют *анализом чувствительности к смещениям*. В этой книге мы не будем касаться подробностей, но, как правило, полезно подумать об источниках систематической ошибки, их вероятных признаках, их вероятных величинах и о том, что это означает для эффекта, который вы пытаетесь оценить.

Зная различные источники систематических ошибок и их вероятные признаки, вы сможете более глубоко понять, почему корреляция не обязательно является свидетельством причинно-следственной связи. Истинный эффект может быть нулевым, но наблюдаемая корреляция могла возникнуть из-за искажения или обратной причинно-следственной связи. Точно так же, как мы обсуждали в главе 3, причинно-следственная связь не обязательно подразумевает корреляцию. Даже если какое-то воздействие имеет большой положительный эффект,

искажающий фактор или обратная причинно-следственная связь способны создать большое отрицательное смещение. Это может привести к небольшой, нулевой или даже отрицательной (как в случае с расходами на предвыборную кампанию и голосами за действующих политиков) наблюдаемой корреляции, несмотря на положительный эффект воздействия. Короче говоря, корреляция не обязательно означает причинно-следственную связь, но верно и обратное: причинно-следственная связь не обязательно означает корреляцию.

Опираясь на достигнутое нами углубленное понимание взаимосвязи корреляции и причинности, перейдем к более сложному и показательному примеру.

Контрацепция и ВИЧ

Одним из величайших бедствий общественного здравоохранения нашего времени является распространение вируса иммунодефицита человека (ВИЧ) и вызываемого им синдрома приобретенного иммунодефицита (СПИД) в Африке. Исследователи усердно работали, чтобы определить, почему эти заболевания распространяются так быстро, и попытаться остановить эту волну. Одна из гипотез, которая привлекла внимание как ученых, так и представителей общественного здравоохранения, заключается в том, что использование женщинами препаратов гормональной контрацепции может увеличить риск передачи ВИЧ, вызывая изменения в иммунной системе или тканях тела.

В исследовании 2012 г., опубликованном в журнале The Lancet Infectious Diseases, исследователи представили доказательства, подтверждающие эту гипотезу. Исследователи проанализировали данные более чем 3500 пар, в которых один партнер был инфицирован ВИЧ, а другой – нет. У них были данные о различном поведении, о котором сообщали сами женщины (например, об использовании презервативов, наличии других сексуальных партнеров), а также о том, получала ли женщина гормональные контрацептивы в клинике, проводившей исследование. В данных также сообщалось, заразился ли неинфицированный партнер ВИЧ в течение года или двух. Наконец, для тех партнеров, которые заразились ВИЧ, генетический скрининг предоставил информацию о том, передался ли он от партнера к партнеру или из какого-то третьего источника.

Было сделано два важных открытия. Во-первых, ВИЧ-отрицательные женщины, использовавшие гормональную контрацепцию, в два раза чаще заражались ВИЧ от своих инфицированных партнеров-мужчин, чем ВИЧ-отрицательные женщины, не использовавшие гормональную контрацепцию. Во-вторых, ВИЧ-инфицированные женщины, использовавшие гормональную контрацепцию, в два раза чаще передавали ВИЧ своим ВИЧ-отрицательным партнерам-мужчинам, чем ВИЧ-инфицированные женщины, не использовавшие гормональную контрацепцию. Эти результаты подтверждались исследованием об использовании презервативов с ограничением искажающего фактора. (Мы подробнее поговорим о том, что означает учет и ограничение факторов, в следующей главе.) На основе этих выводов авторы, New York Times, Национальное общественное радио и многие другие СМИ сообщили, что гормональная контрацепция, вероятно, увеличивает риск передачи ВИЧ.

Это исследование шагнуло далеко вперед по сравнению с предыдущими исследованиями по такому критически важному вопросу. Но до сравнения яблок с яблоками все-таки далеко. Что могло пойти не так?

Больше всего беспокоит возможность влияния искажающих факторов: женщины, принимающие гормональные контрацептивы, отличаются от женщин, которые их не принимают, по многим неизмеренным параметрам, потенциально способным влиять на передачу ВИЧ. Если это так, то наблюдаемая корреляция между использованием гормональных контрацептивов и передачей ВИЧ может являться ошибочной оценкой истинной причинно-следственной связи.

Одна из проблем заключается в том, что женщины, более активные в сексуальном плане, также могут с большей вероятностью использовать гормональную контрацепцию. Авторы исследования, опубликованного в журнале Lancet, не смогли случайным образом распределить женщин на принимающих и не принимающих гормональные контрацептивы. Женщины принимали гормональные препараты, если они этого хотели. Сексуальная активность является фактором риска передачи ВИЧ. Таким образом, независимо от чего-либо еще, более сексуально активные женщины подвергаются большему риску передачи ВИЧ. Если женщины, принимающие гормональные контрацептивы, склонны систематически проявлять более высокую сексуальную активность, у них будет более высокий уровень передачи инфекции, даже если сами противозачаточные средства не играют прямой биологической роли.

В каком направлении этот фактор сместит оценки? Как показано на рис. 9.8, идея заключается в том, что сексуальная активность увеличивает использование гормональных контрацептивов, а также увеличивает передачу ВИЧ по причинам, не связанным с контрацепцией. Следовательно, этот фактор вызывает положительное смещение, и наблюдаемая корреляция переоценивает истинную причинно-следственную связь между гормональной контрацепцией и передачей ВИЧ.

Рис. 9.8. Базовый уровень сексуальной активности приводит к тому, что корреляция между использованием гормональных контрацептивов и передачей ВИЧ дает завышенную оценку причинно-следственной связи

Авторы журнала Lancet знают об этих проблемах и предпринимают некоторые попытки их решения. В частности, женщин спрашивали о прошлом сексуальном поведении и использовании презервативов. Но самооценка поведения, как известно, ненадежна, особенно в отношении таких деликатных тем, как сексуальная активность и использование презервативов.

Механизмы или факторы?

Легко запутаться в том, что является искажающим фактором, а что нет. Одной из наиболее распространенных ошибок является ошибочное представление

о механизмах, с помощью которых воздействие влияет на результат. *Механизм* (иногда называемый также *посредником* или *медиатором*) – это некоторое свойство мира, на которое влияет воздействие, что, в свою очередь, влияет на результат. Иными словами, механизм, а не искажающий фактор является частью того, как воздействие влияет на результат.

Например, один из механизмов, с помощью которого чартерная школа помогает учащимся получить более высокие результаты тестов по сравнению с государственной школой, заключается в наличии более продвинутых классов, которые лучше готовят учащихся к тестам. Глядя на корреляцию, которая показывает, что учащиеся чартерных школ сдают стандартизированные тесты лучше, чем учащиеся государственных школ, возникает соблазн сказать: «Да, но картину искажает тот факт, что эти учащиеся чартерных школ могли посещать более продвинутые классы». Но это неправильно.

Помните, что искажающий фактор – это не просто свойство мира, которое коррелирует с воздействием и результатом. Это свойство мира, которое влияет как на воздействие, так и на результат. Но в нашей истории доступ к продвинутым классам не влияет на то, попадет ли ученик в чартерную школу (воздействие). На самом деле возможность посещать чартерную школу влияет на возможность посещать продвинутый класс, а тот, в свою очередь, влияет на успеваемость учащегося по стандартизированным тестам. Таким образом, доступ к продвинутым классам не является искажающим фактором; это один из механизмов, с помощью которого чартерные школы улучшают результаты тестов. Иногда мы описываем искажающие факторы как *ковариаты до воздействия*, т. е. переменные, которые коррелировали с воздействием и результатом до того, как объект подвергся воздействию, и описываем механизмы как *ковариаты после воздействия*, т. е. переменные, которые коррелируют с воздействием и результатом после того, как объект подвергся воздействию. Рисунок 9.9 иллюстрирует это различие (обратите внимание на направление стрелок).

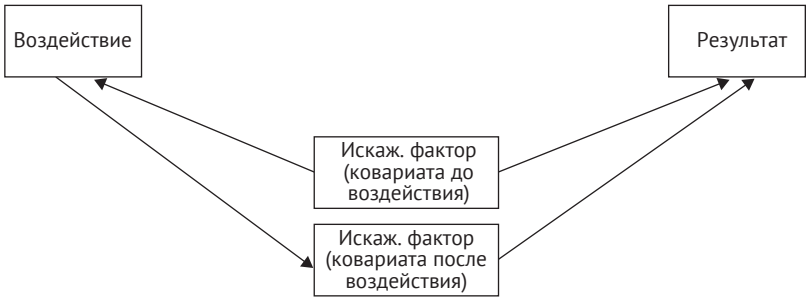


Рис. 9.9. Различие между факторами и механизмами

Как мы уже говорили, в этих вопросах легко запутаться. Итак, давайте рассмотрим пару примеров.

Предположим, медицинское исследование мужчин среднего возраста показало, что те, кто принимает статины, реже умирают от сердечных приступов. Вы заметили также, что те мужчины, которые принимают статины, в среднем богаче и имеют более низкий уровень холестерина. Что из этого является ис-

кажающим фактором, а что может быть механизмом? Подумайте об этом на мгновение, прежде чем мы скажем вам ответ.

Начнем с богатства. Помните: чтобы понять, является ли какое-то свойство мира потенциальным искажающим фактором, вам необходимо задаться вопросом, может ли оно повлиять как на воздействие, так и на результат. Итак, мы задаем два вопроса.

1. Может ли богатство мужчины влиять на то, принимает ли он статины? Конечно, ответ – да. Более состоятельные мужчины, по-видимому, имеют больше возможностей позволить себе лекарства, а также, вероятно, с большей вероятностью обратятся к врачу, который пропишет им эти лекарства.
2. Может ли богатство человека повлиять на риск смерти от сердечно-сосудистых заболеваний? Опять же, ответ – да. Более состоятельные мужчины могут позволить себе вести здоровый образ жизни (например, посещать тренажерный зал) и с большей вероятностью получат быстрый доступ к медицинской помощи в случае сердечного приступа.

Таким образом, у нас есть все основания полагать, что в данном случае богатство является искажающим фактором.

А как насчет снижения уровня холестерина? Медицинские данные свидетельствуют о том, что более высокий уровень холестерина может повлиять на вероятность сердечного приступа (хотя трудно выявить причинный эффект). Но влияет ли холестерин на то, принимает ли человек статины? Здесь нам может понадобиться немного больше информации, в частности когда именно измерялся уровень холестерина.

Если уровень холестерина измерялся до того, как человек начал принимать статины, то это хороший кандидат на роль искажающего фактора. В конце концов, люди обычно предпочитают принимать статины, когда у них высокий уровень холестерина. (С учетом ваших навыков из предыдущего раздела о «выявлении смещения», заставляет ли это вас думать, что исследование занижает или завышает эффективность статинов?)

Но если уровень холестерина измеряли после того, как человек начал принимать статины, то это механизм. Мы полагаем, что одним из способов, с помощью которого статины могут снизить риск сердечно-сосудистых заболеваний, является снижение уровня холестерина. Если это правда и если бы мы случайным образом выбрали некоторых людей для приема статинов, а других нет, мы могли бы ожидать, что у тех, кто принимал статины, будет более низкий уровень холестерина (и меньший риск сердечных заболеваний). Эта разница в уровнях холестерина не является проблемой для вывода об эффективности статинов; скорее, это механизм, с помощью которого достигается такая эффективность.

Вот еще один пример. Предположим, нас интересует, помогает ли хорошее состояние экономики снизить риск гражданской войны. Мы обнаружили, что действительно существует отрицательная корреляция между доходом на душу населения и частотой, с которой в стране происходят гражданские войны. Но мы также отмечаем, что развитая демократия положительно коррелирует с доходом на душу населения и отрицательно коррелирует с риском гражданской войны. Должны ли мы в этом случае думать о демократии как о помехе или механизме?

Это непростая задача. Несложно увидеть, как демократия может сбивать с толку. Наличие демократической формы правления может улучшить качество управления государством. А хорошее управление может привести к росту экономики страны. Более того, демократия может дать людям ненасильственные способы разрешения политических споров, тем самым напрямую снижая риск гражданской войны. В этой истории демократия является искажающим фактором, поскольку она оказывает прямое причинное влияние как на воздействие (доход на душу населения), так и на результат (гражданская война). С другой стороны, демократию можно рассматривать как механизм. Возможно, по мере того как страны становятся богаче, граждане становятся более развитыми, более образованными, более склонными действовать ради выгоды общества и т. д. Таким образом, более высокий доход на душу населения может увеличить вероятность того, что государство станет демократическим (хотя опять же бывают и процветающие монархии). И тогда по уже изложенным выше причинам демократия может снизить риск гражданской войны. С этой точки зрения демократия является не помехой, а частью механизма, с помощью которого более высокий доход на душу населения снижает риск гражданской войны.

Как показывает этот пример, различие между искажающим фактором и механизмом важно, но не всегда однозначно. На данный момент важно понимать различие на концептуальном уровне, даже если во многих реальных сценариях вы не всегда уверены, к чему относится наблюдаемое явление – к механизму, или к мешающему фактору. Мы вернемся к этой теме в следующей главе, когда будем говорить о преимуществах и недостатках методов ограничения влияния искажающих факторов.

Критические размышления о смещении и шуме

Мы хотели бы сделать паузу, чтобы убедиться, что вы не забыли уроки из части 2 – об оценке существования связи – только потому, что теперь мы переключили свое внимание на размышления о нюансах причинно-следственных явлений. Давайте подумаем о вопросах, которые вы должны задать себе, когда кто-то демонстрирует вам корреляцию и интерпретирует ее как оценку причинно-следственной связи.

Во-первых, действительно ли мы наблюдаем корреляцию? Как было сказано в главе 4, люди часто думают, что они измерили корреляцию, хотя на самом деле это не так, потому что они не собирали данные с вариациями одной из ключевых переменных. Так, например, вам нужно убедиться, что они не ограничились только теми случаями, когда наблюдался ожидаемый результат или присутствовало интересующее воздействие. Если они допустили подобную ошибку, вы даже не сможете узнать по представленным данным, коррелируют ли переменные, не говоря уже о причинно-следственной связи.

Во-вторых, отражает ли предполагаемая корреляция подлинные отношения в мире? Допустим, кто-то предположил, что потребление арахисового масла коррелирует с аппендицитом в выборке из 100 человек – в этой выборке люди, которые ели больше арахисового масла, с большей вероятностью заболели аппендицитом. Вы можете задать себе ряд вопросов. Отличается ли корреляция

статистически от нулевой гипотезы об отсутствии корреляции? Почему рассмотрены данные только по 100 субъектам? Собирались ли данные с целью измерения этой конкретной корреляции? Рассказали бы вам об этом исследовании, если бы не обнаружили никакой корреляции? Если вас беспокоит *p*-хакинг или *p*-скрининг, вы можете усомниться в том, что корреляция между арахисовым маслом и аппендицитом будет наблюдаться среди более обширной популяции, и вам наверняка захочется собрать независимую выборку данных, чтобы проверить, существует ли корреляция между употреблением арахисового масла и аппендицитом в более широкой выборке. Если это не так, есть все основания полагать, что истинная оценка (корреляция в генеральной совокупности) равна нулю и что исследователи обнаружили положительную корреляцию в своей выборке из 100 человек из-за шума.

Чтобы сформировать общую картину, давайте вернемся к частному случаю нашего любимого уравнения, когда мы делаем причинный вывод:

Наблюдаемая корреляция (оценка) = Истинный причинный эффект (оцениваемая величина) + смещение + шум.

Существует два типа случаев, когда предполагаемая корреляция может отклоняться от интересующего причинного эффекта. Во-первых, может присутствовать шум. Шум здесь относится к несистематическим факторам, которые влияют на нашу оценку. Это может произойти из-за вариаций выборки в тех случаях, когда вас интересует совокупность, но у вас есть данные только по ограниченной выборке. Иногда шум может возникнуть из-за других несистематических изменений интересующих вас переменных, которые не зависят от какой-либо причинно-следственной связи (например, вы можете случайно измерить переменные с ошибкой). Можно подумать, что, поскольку шум в среднем равен нулю, мы можем просто игнорировать его. Но тот факт, что шум в среднем равен нулю, не означает, что он равен нулю в какой-либо конкретной выборке. Более того, при наличии *p*-хакинга и *p*-скрининга даже средний шум не будет равен нулю. Этому была посвящена глава 7. Во-вторых, помимо шума, может присутствовать систематическая ошибка, т. е. искажающие факторы или обратная причинно-следственная связь, из-за которой текущая оценка отличается от средней, чему посвящена данная глава.

Столкнувшись с корреляцией, которая, как ожидается, служит доказательством и мерой причинно-следственной связи, полезно рассмотреть все три компонента – истинный причинный эффект, смещение и шум – и попытаться осмыслить роль, которую каждый из них играет в объяснении корреляции. Конечно, зачастую оценка отражает комбинацию всех трех факторов.

В некоторых случаях сложно разделить смещение и шум или даже дать им четкие определения. Давайте посмотрим несколько примеров. В книге Тайлера Вигена «Ложные корреляции» представлены пары тенденций, которые совпадают в течение длительного времени, хотя нет веских оснований полагать, что эти две тенденции каким-либо образом связаны причинно или логически. Мы склонны избегать очевидного термина «ложная корреляция», поскольку не ясно, что имеет в виду человек, который его использует, – смещение или шум.

Рисунок 9.10 иллюстрирует один из примеров Вигена. Он показывает долгосрочную корреляцию между самоубийствами через повешение и государственными расходами на науку в Соединенных Штатах. Несмотря на нестандартную форму представления, эти данные демонстрируют выраженную положительную корреляцию. Если рассматривать каждый год как объект наблюдения, то становится ясно, что в годы с большим количеством самоубийств, чем обычно, расходы на науку также выше среднего. Фактически коэффициент корреляции (*r*) равен 0.992 – это, по сути, самая сильная корреляция, которую можно найти без подтасовки данных.

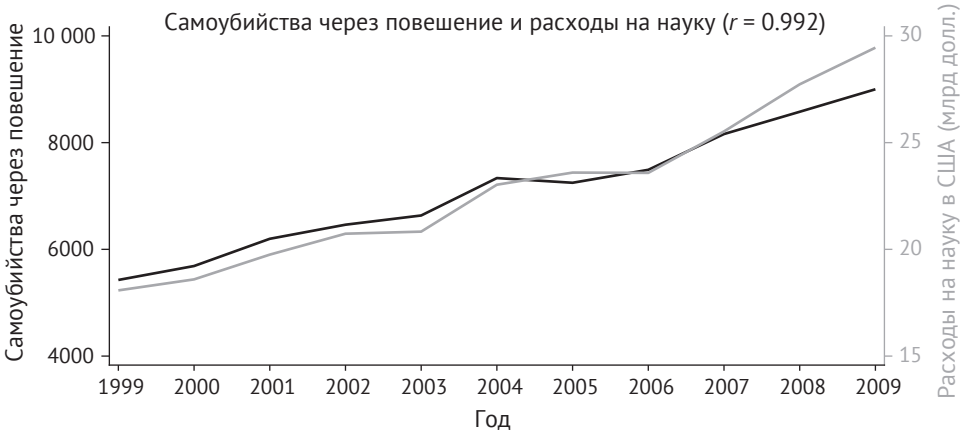


Рис. 9.10. Сильная долгосрочная корреляция между самоубийствами через повешение и государственными расходами на науку

Что тут происходит? Объясняется ли эта корреляция истинным причинным воздействием инвестиций в науку на самоубийства, смещением или шумом? Теоретически возможно, но очень маловероятно, что расходы на науку окажут прямое положительное влияние на количество самоубийств через повешение (или наоборот). Шум, безусловно, кажется правдоподобным объяснением. Если вы изучите достаточно большое количество переменных, вы обязательно обнаружите две из них, которые случайно совпадут, и мы знаем, что именно это и сделал Виген. Он проверял наличие корреляций во времени для многих переменных и выборочно сообщал о корреляциях, которые были статистически значимыми.

Но, возможно, это также смещение. Каким здесь может быть искажающий фактор? Может ли существовать переменная, влияющая как на число самоубийств через повешение, так и на расходы на науку? Одним из потенциальных искажающих факторов является численность населения. За этот период (1999–2009 гг.) население США неуклонно росло примерно с 279 до 307 млн. А рост населения может привести к увеличению как самоубийств, так и расходов на науку.

Чтобы выяснить, что является более важным объяснением наблюдаемой корреляции – смещение или шум, – возможно, стоит подумать о том, ожида-

ете ли вы, что эта корреляция будет сохраняться и в течение многих лет до 1999 г. и после 2009 г. Если вы подозреваете, что корреляция будет сохраняться в более общем плане за пределами известной выборки, то это не может быть просто шум. С другой стороны, если вы думаете, что эта корреляция – всего лишь случайность, вряд ли сохраняющаяся за пределами короткого периода, изученного Вигеном, то это просто шум, не имеющий отношения ни к причинно-следственной связи, ни к смещению.

Рассмотрим еще два примера. На рис. 9.11 показана корреляция между докторскими степенями по социологии в США и количеством запусков некоммерческих космических аппаратов по всему миру. Опять же, наблюдается сильная корреляция. Более того, не так-то просто связать результат с ростом населения (или чем-то еще, меняющимся с течением времени), поскольку корреляция не обусловлена тем, что эти две переменные стабильно увеличиваются с течением времени. В среднем количество космических запусков и докторских степеней по социологии не увеличивается и не уменьшается, но годы с большим количеством космических запусков также, как правило, являются годами с большим количеством докторских степеней по социологии.

Рис. 9.11. Причудливая корреляция между докторскими степенями по социологии и космическими запусками

Мы вполне можем списать это на шум. В космических запусках и докторских диссертациях по социологии из года в год наблюдаются определенные колебания, и за этот период они выстроились в совпадающую последовательность. Но мы подозреваем, что, если рассмотреть данные за следующие 13 лет, корреляция будет близка к нулю.

Наконец, на рис. 9.12 показана корреляция между количеством фильмов, в которых снялся Николас Кейдж, и количеством людей, утонувших в бассейне. Это похоже на еще одно проявление шума. Здесь определенно нет причинно-следственной связи и, вероятно, также нет убедительного искажающего фактора. И, как и в случае с докторскими степенями и космическими запусками, мы готовы поспорить, что эта корреляция не сохранится в будущем.

Рис. 9.12. Мистическая корреляция между фильмами Николаса Кейджа и количеством людей, утонувших в бассейне

Однако корреляция между фильмами с участием Кейджа и утопленниками представляет собой другую концептуальную загадку. Предположим, что этот анализ включал все годы, в течение которых играл Николас Кейдж, и все годы, в течение которых у людей были бассейны (это, очевидно, не так, но постарайтесь представить). Если бы Николас Кейдж перестал сниматься в фильмах и у людей исчезли бы бассейны, мы не смогли бы оценить корреляцию между этими двумя переменными в какой-то будущий период. Так как же мы можем рассуждать о том, является ли эта корреляция результатом шума? Более того, как мы можем говорить о том, что эта корреляция является результатом шума, если мы располагаем всеми возможными данными о фильмах Николаса Кейджа и утоплениях в бассейне? Ведь если мы наблюдаем за полной совокупностью (в данном случае, за фильмами Николаса Кейджа), шума выборки не будет.

Один из способов решить эту загадку – совершить метафизический переход, который мы обсуждали еще в главе 6, когда говорили о статистических выводах при наличии данных для всей совокупности. Конечно, существует наблюдаемая корреляция между фильмами Николаса Кейджа и утоплениями в бассейне в этом мире, но это лишь небольшая выборка из более широкой совокупности альтернативных, гипотетических миров, которые могли бы существовать. Эти миры очень похожи на наш, но все уникальные, несвязанные факторы действуют по-разному. Есть ли у нас основания ожидать, что фильмы Николаса Кейджа будут связаны с утоплениями в бассейнах в других мирах? Если ответ отрицательный, мы можем утверждать, что наблюдаемая нами корреляция – всего лишь шум, хотя у нас есть все данные о Николасе Кейдже и людях, утонувших в бассейне.

Подведение итогов

Мы показали на примерах, что корреляция часто представляет собой смещенную оценку причинно-следственной связи из-за искажающих факторов или

обратной причинно-следственной связи. Именно это мы имеем в виду, когда говорим, что корреляция не то же самое, что причинно-следственная связь.

Если мы знаем, на какие факторы следует обращать внимание, и если мы можем их измерить, сможем ли мы исправить смещение и получить лучшую оценку причинно-следственной связи? Как это сделать – тема главы 10.

Ключевые термины

- **Причинный эффект:** изменение какого-либо свойства мира, которое может произойти в результате изменения какого-либо другого свойства мира.
- **Средний эффект воздействия** (ATE): разница в средних результатах при сравнении двух противоположных сценариев – одного, когда все члены группы испытывают воздействие, и другого, когда все члены группы не испытывают воздействия.
- **Средний эффект воздействия на подвергшихся воздействию (ATT)**: разница в среднем результате при сравнении сценария, в котором вся подгруппа, избранная для воздействия, испытывает воздействие, и контрфактического сценария, когда вся эта группа не испытывает воздействия.
- **Средний эффект воздействия на не подвергшихся воздействию (ATU)**: разница в среднем результате при сравнении контрфактического сценария, в котором все в подгруппе людей, не выбранных для воздействия, испытывают воздействие, и сценария, когда вся эта группа не испытывает воздействия.
- **Разница в средних значениях:** разница в средних результатах при сравнении подгруппы людей, которые фактически испытали воздействие, и подгруппы людей, которые фактически не испытали воздействия.
- **Исходные различия:** различия в среднем потенциальном результате между двумя группами (например, группами, испытавшими и не испытавшими воздействие), даже если эти две группы имеют одинаковый статус воздействия.
- **Искажающий фактор:** свойство мира, которое (1) влияет на статус воздействия и (2) влияет на потенциальный результат сверх того эффекта, который оно оказывает через влияние на статус воздействия.
- **Обратная причинно-следственная связь:** когда результат влияет на статус воздействия.
- **Завышенная оценка:** когда смещение положительное, так что оценка эффекта превышает истинный эффект.
- **Заниженная оценка:** когда смещение отрицательное, так что оценка эффекта меньше истинного эффекта.
- **Механизм (или посредник):** свойство мира, на которое влияет воздействие и которое, в свою очередь, влияет на результат.
- **Ковариата до воздействия:** переменная, которая коррелирует с воздействием и результатом до начала воздействия.
- **Ковариата после воздействия:** переменная, которая начинает коррелировать с воздействием и результатом после воздействия.

Упражнения

9.1. В конце обсуждения примера про насильственное и ненасильственное сопротивление в главе 1 мы спросили вас о следующем:

Почему тот факт, что после насильственных протестов правительство чаще принимает жесткие ответные меры, не обязательно означает, что переход от насилия к миролюбию снизит риск жестких ответных мер?

Мы обещали, что к концу этой главы вы сможете дать убедительный ответ. Итак, укажите хотя бы одну причину, почему тот факт, что насильственные протесты чаще встречают жесткий ответ со стороны правительства, не является убедительным доказательством того, что использование тактики насильственного протеста чаще приводит к жестким мерам.

9.2. Подумайте о завышенных и заниженных оценках в двух наших примерах.

a) В нашем обсуждении игры на скрипке мы отметили, что более талантливый музыкант может больше практиковаться и лучше играть на инструменте по причинам, не имеющим ничего общего с тем, как много он занимается. Означает ли это, что корреляция между практикой и качеством игры является переоценкой или недооценкой истинного влияния практики на игру?

b) В нашем обсуждении расходов на предвыборные кампании мы утверждали, что действующие политики, скорее всего, будут тратить большие средства на свои кампании, когда у них слабая электоральная поддержка. Означает ли это, что наблюдаемое отсутствие корреляции (или даже отрицательная корреляция) между расходами на предвыборную кампанию и результатами выборов является переоценкой или недооценкой истинного влияния расходов на голоса?

9.3. Итан однажды был на встрече, где его проинформировали о том, как анализ данных может улучшить работу университетов. Пример, который больше всего взволновал докладчика, был получен от команды аналитиков данных в отделе развития крупного исследовательского университета (что на жаргоне менеджеров означает сбор денег). Анализируя многолетние данные, команда аналитиков данных обнаружила следующую корреляцию: выпускники, которые делают пожертвования университету шесть лет подряд, с гораздо большей вероятностью будут жертвовать на протяжении всей жизни, чем выпускники, которые делают пожертвования только пять лет подряд.

Докладчик был взволнован, поскольку, по его мнению, этот вывод аналитической группы предполагает четкую стратегию по улучшению сбора средств и вовлечения выпускников. В частности, на основе этого анализа они решили приложить серьезные усилия, чтобы побудить выпускников, которые уже пять лет подряд делали взносы, внести пожертвование и шестой раз – идея заключалась в том, что доказательства корреляции между пожертвованиями в течение шести лет подряд и пожертвованиями в будущем предполагают, будто пожертвования на шестом году имели большое причинное влияние на будущие пожертвования, поэтому ресурсы, потраченные на поощрение еще одного пожертвования на шестой год, использовались с максимальной пользой.

Используя навыки критического мышления, которые вы приобрели в этой главе, приведите два аргумента, свидетельствующие об ошибочности этой идеи.

9.4. Вскоре после выхода книги гарвардского психолога Дэниела Гилберта «Спотыкаясь о счастье» он выступил по телевидению и сообщил Стивену Колберту, что «брак – это одна из лучших инвестиций, которые вы можете сделать в счастье». Этот совет неявно основан на причинном утверждении: брак приносит счастье.

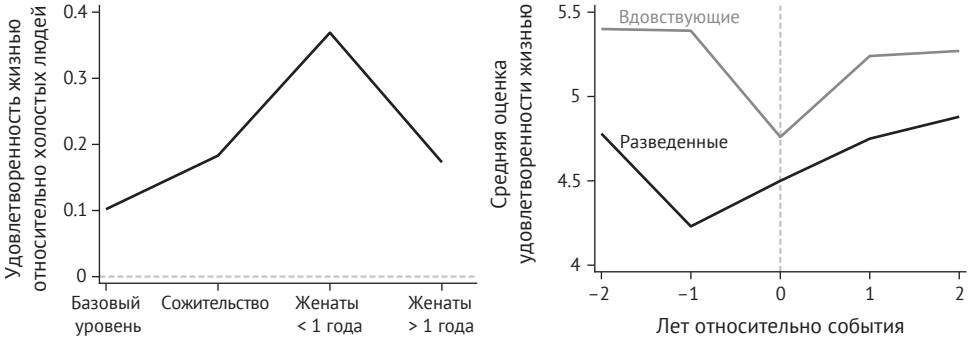
Многие недавние исследования подтверждают положительную корреляцию между браком и счастьем. Но является ли эта связь причинно-следственной?

a) Приведите аргументы в пользу того, почему корреляция между браком и счастьем может быть результатом обратной причинно-следственной связи (счастье приводит к браку, а не наоборот).

b) Определите два фактора, которые, по вашему мнению, могут затруднить причинно-следственную интерпретацию корреляции между браком и счастьем. По каждому из них объясните, почему вы считаете, что искажающий фактор может повлиять как на воздействие (вступление в брак), так и на результат (счастье).

c) Определите знак смещения для каждого из выявленных вами искажающих факторов. Сделав это, объясните, приводит ли каждый из них к переоценке или недооценке истинного причинного эффекта.

d) В исследовании Анке Циммерманн и Ричарда Истерлина изучались люди в период от четырех лет до первого брака и в течение нескольких лет после вступления в брак. Основной вывод проиллюстрирован в левой части рисунка ниже, где показана удовлетворенность жизнью людей, вступивших в брак в течение периода исследования, по сравнению с теми, кто никогда не женился в течение периода исследования. Идя слева направо, мы видим, как удовлетворенность жизнью человека меняется с течением времени по мере того, как они сначала сожительствуют с партнером, затем женятся и продолжают этот брак более года.



Зависимость между браком и удовлетворенностью жизнью

i) Сравните удовлетворенность жизнью людей, состоящих в браке какое-то время, с удовлетворенностью людей, которые не женаты, но живут со своим партнером. Как вы думаете этот график подтверждает выводы Гилберта или противоречит им?

ii) Определите искажающий фактор, который, как предполагает это сравнение, мог существовать в исходной корреляции.

e) Исследование Джонатана Гарднера и Эндрю Освальда также отслеживает динамику отдельных людей, но ставит другой вопрос. Они исследуют, что происходит со счастьем людей, когда браки распадаются. В исследовании рассматриваются два варианта прекращения брака: развод или смерть супруга. Результаты суммированы в правой части рисунка.

Горизонтальная ось показывает годы относительно важного события (развод или смерть супруга) в момент 0. Вертикальная ось показывает удовлетворенность жизнью. Удовлетворенность жизнью показана черным цветом для тех, кто развелся, и серым для тех, кто овдовел.

i) Обратите внимание на первоначальную разницу в удовлетворенности жизнью между теми, кто овдовел, и теми, кто развелся, еще до того, как это событие произошло. Как вы думаете, это подтверждает или опровергает причинную интерпретацию Гилберта? Почему?

ii) Теперь рассмотрим вдов и вдовцов (серая линия). Как изменилось их счастье до, во время и спустя год, после того как скончались их супруги? Как вы думаете, это подтверждает или опровергает причинную интерпретацию Гилберта? Что, по вашему мнению, заставляет эту зависимость проявляться в первоначальной корреляции Гилберта?

9.5. Загрузите файл HouseElectionsSpending2018.csv и связанный с ним файл README.txt, описывающий переменные в этом наборе данных, по адресу *http://www.press.princeton.edu/thinking-clearly*.

a) Постройте линейную регрессию, которая определит взаимосвязь между долей голосов действующего политика и его расходами. (Примечание: возможно, вам придется перекодировать некоторые переменные в наборе данных или создать свои собственные переменные, которые лучше соответствуют вашей цели.)

i) Корреляция положительная или отрицательная?

ii) Согласно этим данным кампании, которые тратят больше, добиваются лучших или худших результатов?

iii) Интерпретируйте величину и направление корреляции между действующими расходами и долей голосов за действующих политиков.

b) Сделайте то же самое, что и выше, для новых претендентов.

c) Как вы думаете, являются ли построенные вами регрессии убедительным доказательством влияния расходов во время избирательной кампании на долю голосов?

i) Назовите три искажающих фактора.

ii) Есть ли в этом наборе данных какие-либо переменные, которые измеряют эти факторы? Если да, определите переменную, которая могла бы достоверно измерить искажающий фактор, присутствующий в наборе данных.

iii) Используя линейную регрессию, оцените, действительно ли расходы действующего политика и расходы претендентов (воздействие) коррелируют с одним из потенциальных искажающих факторов, обнаруженных в наборе данных.

9.6. Приведите свой пример исследователя, журналиста, политика или аналитика, который, по вашему мнению, допустил ошибку, неправильно интерпретировав корреляцию как достоверное свидетельство причинно-следственной связи. Ваш пример не должен быть тесно связан с каким-либо примером, обсуждаемым в этой книге. Проанализируйте представленные доказательства и объясните, почему вы считаете, что эта корреляция не является убедительным доказательством предполагаемой причинно-следственной связи. Обсудите вероятное направление смещения. В качестве дополнительного упражнения продолжайте думать о своем примере, читая следующие четыре главы. Можете ли вы предложить лучший способ более достоверно оценить интересующую причинно-следственную связь?

Дополнительное чтение и ссылки

Обучение в школе Пройсса:

Larry McClure, Betsy Strick, Rachel Jacob-Almeida, and Christopher Reichher. 2005. *The Preuss School at UCSD*. Research report of The Center for Research on Educational Equity, Assessment and Teaching Excellence. *http://www.create.ucsd.edu/_files/publications/PreussReportDecember2005.pdf*.

Исследование по программе «Знание – сила»:

JoshuaD. Angrist, SusanM.Dynarski, Thomas J. Kane, Parag A. Pathak, and Christopher R. Walters. 2012. *Who Benefits from KIPP?* Journal of Policy Analysis and Management 31 (4): 837–60.

Цитата о нулевых гипотезах в литературе, посвященной изучению влияния чартерных школ, взята из работы:

Julian R. Betts, Lorien A. Rice, Andrew C. Zau, Y. Emily Tang, Cory R. Koedel. 2006. *Does School Choice Work?: Effects on Student Integration and Achievement.* Public Policy Institute of California.

Исследование связи между практикой и мастерством скрипачей:

K. Anders Ericsson, Ralf T. Krampe, and Clemens Tesch-Römer. 1993. *The Role of Deliberate Practice in the Acquisition of Expert Performance.* Psychological Science 100 (3): 363–406.

Исследование связи между гормональной контрацепцией и ВИЧ-инфекцией:

ReneeHeffron,DeborahDonnell,Helen Rees, and Connie Celum. 2012. *Use of Hormonal Contraceptives and Risk of HIV-1 Transmission: A Prospective Cohort Study.* The Lancet Infectious Diseases 12 (1): 19–26.

Исследование, изучающее корреляцию между успехом на выборах и расходами на предвыборную кампанию для действующих политиков и претендентов:

Gary C. Jacobson. 1978. *The Effects of Campaign Spending in Congressional Elections.* American Political Science Review 72 (2): 469–91.

Мы обсудили несколько примеров, взятых из книги:

Tyler Vigen. 2015. *Spurious Correlations: Correlation Does Not Equal Causation.* Hachette Books.

В упражнении 4 мы обсудили три исследования ощущаемого благополучия. Общее обсуждение исследований счастья можно найти в книге:

Daniel Gilbert. 2007. *Stumbling on Happiness.* Vintage.

Исследование ощущения благополучия до и после брака:

Anke C. Zimmermann and Richard A. Easterlin. 2006. *Happily Ever After? Cohabitation, Marriage, Divorce, and Happiness in Germany.* Population and Development Review 32 (3): 511–28.

Исследование ощущения благополучия до и после расторжения брака:

Jonathan Gardner and Andrew J. Oswald. 2006. *Do Divorcing Couples Become Happier by Breaking Up?* Statistics in Society 169 (2): 319–36.

Глава 10

Выявление и ограничение искажающих факторов

О чем эта глава

- Если мы сможем выявить искажающий фактор, то сможем ограничить его влияние и смягчить возникающее из-за него смещение.
- Самый распространенный способ ограничения влияния искажающего фактора – включение его в регрессию, хотя существуют и другие подходы.
- Ознакомившись с графиками и простыми примерами, вы получите интуитивное понимание того, как это работает.
- Выявление и ограничение искажающих факторов – не волшебное средство. Этот подход не устраняет смещение, возникающее из-за ненаблюдаемых факторов или обратной причинно-следственной связи.
- Обычно следует выявлять и ограничивать искажающие факторы, а не механизмы.

Введение

В главе 9 вы узнали, что искажающие факторы представляют собой большую проблему, мешающую изучить причинно-следственные связи на основе корреляций. Здесь мы поговорим о первой линии защиты от искажающих факторов – их *учете и ограничении* (controlling)[1].

Вероятно, вы уже слышали, как исследователи говорят о контроле искажающих факторов, но что это на самом деле означает? Контроль предполагает использование статистических методов для поиска корреляции между двумя переменными, сохраняя постоянными значения других переменных. Самый простой способ понять эту идею – это рассмотреть несколько примеров.

Влияние партии на голосование в конгрессе

Не такой уж и удивительный факт в отношении Конгресса США заключается в том, что республиканцы с большей вероятностью будут голосовать консервативно, чем демократы. Одним из способов количественного измерения это-

[1] Далее для краткости мы будем говорить просто «контроль», подразумевая выявление искажающих факторов и ограничение их влияния. – *Прим. перев.*

го явления служит рейтинг, присвоенный каждому представителю конгресса Американским консервативным союзом (American Conservative Union, ACU). Каждый год ACU выбирает 25 важных законопроектов и присваивает каждому конгрессмену рейтинг от 0 до 100 в зависимости от того, как он проголосовал по этим законопроектам. Поскольку ACU представляет сторонников правых (консервативных) политических убеждений, более высокий балл указывает на более консервативный результат голосования.

Мы можем подтвердить утверждение о том, что принадлежность к республиканцам коррелирует с результатами консервативного голосования, проверив, имеют ли республиканцы в среднем более высокие баллы ACU, чем демократы. Таблица 10.1, основанная на данных Палаты представителей за 1997 г., показывает, что так и есть. Представители демократической партии в конгрессе имеют средний балл ACU 19, а республиканцы имеют средний балл ACU 83. В среднем республиканцы голосуют на 64 пункта ACU более консервативно, чем демократы.

Таблица 10.1. Сравнение результатов голосования республиканцев и демократов в Конгрессе США.

	Средний балл ACU
Республиканцы	83
Демократы	19
Разница	64

Эти данные указывают на то, что представители республиканцев и демократов в конгрессе голосуют совершенно по-разному. Чем можно объяснить такую поляризацию?

Одна из идей, выдвигаемых многими политологами, заключается в том, что внутрипартийное давление приводит к расхождениям в поведении законодателей при голосовании. В распоряжении партий имеется множество инструментов, позволяющих оказать давление на рядовых членов, чтобы те проголосовали за линию партии. Возможно, самым важным среди этих инструментов является помощь в сборе средств для кампаний по переизбранию.

Но, прежде чем интерпретировать корреляцию между членством в партии и результатами голосования как свидетельство влияния партийной дисциплины, нам следует рассмотреть возможные искажающие факторы. В данном случае искажающим фактором является какое-то другое свойство мира, которое влияет как на партийную принадлежность представителей конгресса, так и на их результаты голосования.

Как показано на рис. 10.1, очевидным кандидатом на роль искажающего фактора являются личные убеждения. Республиканская партия имеет репутацию консерваторов. Демократическая партия – репутацию либералов. Следовательно, убежденный консерватор с большей вероятностью будет баллотироваться как республиканец, а либерал – как демократ. Более того, личные идеологические пристрастия политика вполне могут повлиять на то, как он будет голосовать по законодательству в конгрессе. Если таким образом рассор-

тировать партии по убеждениям, есть основания полагать, что представители республиканцев будут голосовать более консервативно, а представители демократов – более либерально, даже если партии не будут требовать соблюдения дисциплины. Таким образом, личные политические убеждения, скорее всего, служат искажающим фактором. В свете этого понимания было бы ошибкой интерпретировать корреляцию между членством в партии и результатами голосования как объективную оценку причинного воздействия партийной дисциплины на избирательное поведение представителей.

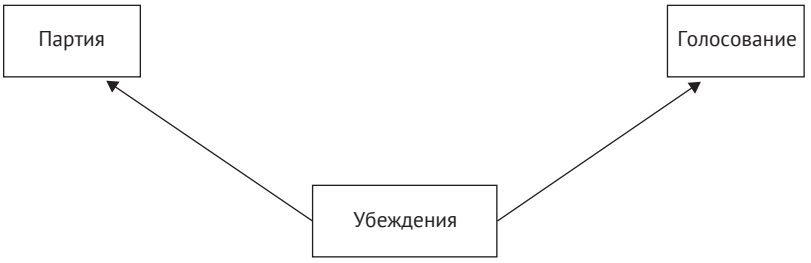


Рис. 10.1. Личные убеждения влияют на то, к какой партии присоединяется политик и как он голосует, находясь у власти. Следовательно, это искажающий фактор

Чтобы устранить эту потенциальную помеху, нужно подвергнуть ее процедуре контроля. В своей простейшей форме контроль фактора личных убеждений просто означает наблюдение за корреляцией между членством в партии и результатами голосования с сохранением постоянства личных убеждений. Для этого нам сначала нужна мера личных политических убеждений. К счастью, у нас есть подходящий вариант.

В 1996 г. внепартийная организация Project Vote Smart провела опрос под названием «Национальный тест на политические убеждения» (National Political Awareness Test, NPAT) среди кандидатов в конгресс. В ходе опроса кандидатам предлагалось высказать свое мнение по широкому кругу вопросов. На основе их ответов Project Vote Smart затем составила рейтинг убеждений по шкале от либералов до консерваторов. На опрос ответили 76 % кандидатов, поэтому мы можем оценить политическую ориентацию большей части представителей конгресса[1].

Чтобы учесть влияние личных убеждений в нашем анализе взаимосвязи между членством в партии и результатами голосования, мы просто сравниваем результаты голосования демократов и республиканцев с аналогичными показателями NPAT. Если NPAT правильно отражает личные убеждения, это сравнение расскажет нам о разнице в результатах голосования республиканцев и демократов с учетом идеологических убеждений (контроль фактора личных убеждений).

В табл. 10.2 представители конгресса распределены по пяти ячейкам на основе их баллов NPAT. В крайней левой ячейке находятся представители с наиболее либеральными убеждениями согласно их ответам NPAT. По мере продвижения вправо ячейки становятся все более консервативными.

[1] На последующих опросах уровень откликов на опрос значительно снизился, что объясняет, почему мы показываем данные за конец 1990-х гг., хотя на тот момент многие наши читатели даже не родились.

Таблица 10.2. Данные о голосовании с учетом баллов NPAT (контроль)

		Либералы ← Процентиль NPAT → Консерваторы				
		1–20	21–40	41–60	61–80	81–100
Республиканцы	Средний балл ACU	----	44	68	86	94
	Кол-во человек	0	4	45	69	69
Демократы	Средний балл ACU	10	18	41	96	84
	Кол-во человек	70	66	24	1	1
Разница средних баллов ACU		**----**	**26**	**27**	**–10**	**10**

Если взглянуть на данные, распределенные таким образом, сразу бросается в глаза несколько вещей. Во-первых, и это самое главное, ни в одной колонке разница между результатами голосования за республиканцев и демократов не приближается к разнице в 64 пункта, которую мы обнаружили до учета личной идеологии. Это говорит о том, что личные убеждения являются важным фактором, искажающим эту корреляцию: большая часть различий в результатах голосования между демократами и республиканцами была связана с тем, что у членов этих двух партий были разные базовые личные предпочтения в отношении политики, а не с партийным давлением. Причина, конечно же, та, о которой мы говорили ранее. Более консервативные люди, как правило, становятся республиканцами, а более либеральные, как правило, – демократами. Этот факт отражается в наблюдении, что с ростом критериев консерватизма в каждой ячейке увеличивается количество республиканцев и уменьшается количество демократов (т. е. почти все республиканцы находятся в 41–100-м процентиле NPAT, а почти все демократы – в 1–60-м процентиле NPAT).

Во-вторых, внутри партии по мере продвижения по категориям убеждений средние баллы ACU по большей части растут. Есть одно исключение – демократы в 61–80-м процентиле голосуют более консервативно, чем демократы в 81–100-м процентиле, – но это сравнение не особенно информативно, поскольку предполагает сравнение только двух человек (идеологически консервативных демократов очень мало).

В-третьих, разница между средними результатами голосования республиканцев и демократов варьируется в зависимости от столбца. То есть соотношение членства в партии и голосования в конгрессе зависит от личных убеждений. Прекрасно, что мы это знаем. Но ведь нам нужна единая, общая мера корреляции между партийной принадлежностью и результатами голосования с учетом личной идеологии, а не измерение убеждений каждого человека. Чтобы получить это единственное число, нам нужно будет взять своего рода средневзвешенное значение различий из различных столбцов. Но как нам решить, какой вес присвоить каждому столбцу?

Когда мы начинаем думать о правильных весах, обратите внимание, что явно есть один столбец, который более информативен, чем другие, о различном поведении демократов и республиканцев со схожими личными убеждени-

ями – столбец для 41–60-го процентиля NPAT. В каждом из остальных столбцов либо очень мало республиканцев, либо очень мало демократов. Но в процентильном столбце 41–60 большое количество представителей обеих партий. И неудивительно: это своего рода центр, где совпадают личные убеждения представителей двух партий. Поэтому, вероятно, имеет смысл сделать так, чтобы наше средневзвешенное значение придавало этому столбцу большой вес.

В более общем плане полезно вспомнить главу 5, где мы узнали о том, что обычная регрессия по методу наименьших квадратов (OLS) представляет собой подгонку линии к данным с целью минимизировать сумму квадратичных ошибок. OLS – это один из принципиальных способов выбора весов для пяти столбцов. (Когда мы говорим о регрессии в этой главе, мы всегда будем иметь в виду регрессию OLS.) Итак, рассмотрим следующую регрессию:

$$\text{Рейтинг ACU} = \alpha + \beta_1 \cdot \text{Республиканец} + \beta_2 \cdot \text{NPAT}_{21-40} + \beta_3 \cdot \text{NPAT}_{41-60} + \beta_4 \cdot \text{NPAT}_{61-80} + \beta_5 \cdot \text{NPAT}_{81-100} + \varepsilon.$$

В этой регрессии предметом анализа является отдельный представитель. Переменная «Рейтинг ACU» представляет собой оценку ACU отдельного представителя. Переменная «Республиканец» представляет собой так называемую *фиктивную переменную*: она принимает значение 1, если представитель является членом республиканской партии, и значение 0, если представитель является членом демократической партии. Различные переменные NPAT также являются фиктивными переменными, принимающими значение 1, если представитель находится в соответствующем процентильном диапазоне, и значение 0 в противном случае[1]. Греческая буква ε (эпсилон) обозначает ошибку.

Коэффициент β_1 в этой регрессии дает нам средневзвешенное значение, о котором мы говорили, т. е. β_1 – это корреляция между баллом ACU и принадлежностью к республиканцам с учетом личных убеждений (измеряемых процентилем NPAT). Мы также получим оценки коэффициентов по четырем задействованным категориям NPAT и пересечение (α). Они тоже имеют свои интерпретации. Однако нас интересует корреляция между оценкой ACU и принадлежностью к республиканцам с учетом личных убеждений, поэтому сосредоточимся на β_1.

Проведение этой регрессии на наших данных дает оценку β_1, обозначенную как $\widehat{\beta}_1$ и равную 24. (Это оценка, а не истинное значение, поскольку наши данные представляют собой выборку, полученную из совокупности всех представителей конгресса, следовательно, наблюдаемая корреляция также отражает шум.) Неудивительно, что это значение очень близко к разнице между средним баллом ACU республиканцев и средним баллом ACU демократов в столбце,

[1] Поскольку все относятся к одной из пяти категорий NPAT, одну из них необходимо опустить. Здесь мы опустили процентиль 1–20. Это аналогично тому факту, что мы не можем включить в регрессию одновременно переменные «Республиканец» и «Демократ», потому что каждый политик является либо тем, либо другим. Но мы не можем отдельно определить эффект от принадлежности к демократам и эффект от принадлежности к республиканцам, поэтому мы просто включаем переменную «Республиканец» и интерпретируем коэффициент как эффект от принадлежности к республиканской, а не демократической партии.

соответствующем процентилю 41–60, где, как мы уже говорили, находится почти вся информация. Регрессия, конечно, немного увеличивает вес других столбцов, снижая оценку с 27 до 24. Но этот столбец, по сути, дает нам ответ.

Даже с учетом убеждений у нас до сих пор нет убедительной оценки причинного воздействия партийной дисциплины на результаты голосования представителей конгресса. Дело в том, что, помимо личных убеждений, может быть много других искажающих факторов. То есть внутри процентильного интервала NPAT может существовать множество других факторов, которые заставляют одних людей становиться демократами, а других – республиканцами и которые также оказывают независимое влияние на их поведение при голосовании в конгрессе. Например, даже при наличии фиксированных личных убеждений демократы могут иметь тенденцию представлять округа с более либеральными избирателями, а республиканцы – округа с более консервативными избирателями. Если политики при голосовании по законопроектам учитывают мнение избирателей, то эти различия в округах станут еще одной проблемой. Мы уверены, что вы можете обнаружить и другие факторы.

По мере роста перечня искажающих факторов создание таблицы, в которой данные разбиваются на все возможные ячейки, становится все более сложным и громоздким. Но пока вы можете измерить потенциальные искажающие факторы, можно контролировать их с помощью регрессии. Это всегда даст вам оценку β_1, отражающую средневзвешенное значение различных ячеек в этой (воображаемой) большой таблице, которая минимизирует сумму квадратов ошибок. Регрессия будет нашим самым важным инструментом контроля искажающих факторов. Поэтому нужно тщательно разобраться, как именно работает контроль искажающих факторов с помощью регрессии.

Примечание о гетерогенных эффектах воздействия

Как мы говорили в главе 3, почти во всех интересных примерах причинно-следственных связей интересующие эффекты неоднородны, т. е. они не одинаковы для каждого объекта наблюдения. Это было хорошо видно в нашем примере, где прививка от гриппа предотвратила заражение гриппом некоторых людей, которые в противном случае могли бы заболеть, но не предотвратила заражение других людей либо потому, что они не привились вовремя, либо потому что прививка от гриппа на них не подействовала. Вероятно, это справедливо и для приведенного выше примера о влиянии партии на голосование. Если говорить о влиянии партии на поименное голосование членов конгресса, этот эффект, вероятно, не одинаков для каждого члена конгресса. Возможно, некоторые члены конгресса имеют прочные личные убеждения и будут голосовать одинаково независимо от какого-либо партийного давления, так что никакого наблюдаемого эффекта не будет. Возможно, другие сильно зависят от поддержки своей партии при переизбрании и будут делать все, что попросят партийные лидеры, что обеспечит наличие выраженного эффекта. А остальные, скорее всего, находятся где-то посередине.

При контроле важно тщательно учитывать подобную неоднородность, поскольку, как показало обсуждение табл. 10.2, как только мы начинаем контролировать искажающие факторы, мы больше не оцениваем средний эффект воздействия по всем объектам. В нашем примере, чтобы оценить взаимосвязь

между партией и голосованием с учетом личных убеждений, мы придаем больший вес членам конгресса с умеренными убеждениями. Это связано с тем, что среди членов конгресса с экстремальными убеждениями редко встречаются вариации партийной принадлежности – по сути, все выраженные консерваторы являются республиканцами, а все выраженные либералы – демократами. Если влияние членства в партии на людей с умеренными убеждениями отличается от влияния на людей с экстремальными убеждениями, придется признать, что мы фокусируем свое внимание на эффекте от партийной принадлежности людей с умеренными убеждениями.

Это признание поднимает острую проблему. Если контроль потенциального искажающего фактора существенно меняет нашу причинную оценку, это может быть признаком того, что оценка без контроля была смещенной и что контроль уменьшил это смещение. Превосходно. Но это также может быть признаком того, что эффекты воздействия гетерогенны, и мы существенно ограничили подмножество объектов, для которых оцениваем средний эффект. Если нас интересует средний эффект воздействия на все наблюдаемые объекты, это может быть очень плохо.

Эти проблемы возникнут и при использовании других методов, помимо контроля, которые мы обсудим позже в книге. Вернемся к этому вопросу в свое время. Иногда вместо того, чтобы оценивать средний эффект воздействия (ATE), мы можем оценить только *локальный* средний эффект (local average treatment effect, LATE), где «локальный» относится к подмножеству объектов, для которых мы можем сгенерировать достоверную оценку. Если эффекты воздействия неоднородны, оценка LATE не обязательно должна совпадать с ATE. Итак, если ATE – это оценка, которая нас действительно интересует, нужно хорошо подумать о том, в какой степени оценки LATE могут или не могут заменить ATE. Но, как говорит экономист Гвидо Имбенс о ситуациях, когда мы можем достоверно оценить лишь локальный средний эффект воздействия, «лучше LATE, чем ничего»[1].

Анатомия регрессии

Ключевыми ингредиентами любой регрессии для причинного вывода являются:

- *зависимая переменная* (dependent variable, также называемая *выходной переменной*);
- *переменная воздействия* (treatment variable);
- набор *переменных контроля* (control variable).

Зависимая переменная – это результат (эффект, исход), который вы пытаетесь понять. Переменная воздействия – это свойство мира, влияние которого на зависимую переменную вы пытаетесь оценить. А переменные контроля являются потенциальными искажающими факторами, которые вы включаете в регрессию, чтобы уменьшить систематическую ошибку.

В простейшем случае, когда имеется только одна переменная контроля, мы запишем уравнение регрессии как

$$Y = \alpha + \beta \cdot T + \gamma \cdot X + \varepsilon, \tag{10.1}$$

[1] Игра слов. Late (англ.) означает «поздно». – *Прим. перев.*

где Y – зависимая переменная, T – переменная воздействия, а X – переменная контроля. Параметры регрессии (т. е. величины, которые мы хотели бы оценить) – это точка пересечения α, эффект воздействия β и «эффект» переменной контроля γ. Существует также ошибка ε, отражающая тот факт, что отклики объектов наблюдения на воздействие отличаются от прогнозируемого результата по особым причинам.

В уравнении регрессии нет ничего, что отличало бы переменную воздействия от переменной контроля. Это различие носит концептуальный характер и обусловлено вопросом, на который вы пытаетесь ответить. Если вы хотите узнать влияние партии на голосование с учетом личных убеждений, переменная партии – это воздействие, а NPAT – ваш контроль. Но если бы вы хотели узнать влияние личных убеждений на голосование, то роли переменных принадлежности к партии и NPAT поменялись бы местами.

Именно поэтому слово «эффект» выше было взято в кавычки при упоминании переменной контроля. Часто нас не волнует параметр регрессии, связанный с переменной контроля (здесь γ). Важно то, что β представляет собой интересующий эффект, и мы попытаемся оценить его объективно.

Один из способов прочитать уравнение 10.1 – это понимать его буквально. Мы можем притвориться, что знаем процесс генерации данных. Каждый отдельный i-й исход (Y_i) равен общему пересечению (α) плюс $\beta \cdot T_i$ плюс $\gamma \cdot X_i$ плюс несистематические факторы (ϵ_i). Другой способ прочитать уравнение – признать, что мы не знаем процесса генерации данных, но тем не менее мы хотели бы оценить β – среднюю линейную зависимость между Y и T с учетом X.

Как мы отмечали в главе 5 (хотя и не совсем так), независимо от процесса генерации данных регрессия OLS всегда дает нам *наилучшее линейное приближение к функции условного ожидания* (best linear approximation to the conditional expectation function, BLACEF). Поэтому нам не обязательно притворяться, что мы знаем процесс генерации данных, чтобы выполнить регрессию. Если нет базовых различий между значениями T с учетом контролируемого фактора X, то BLACEF соответствует среднему эффекту воздействия T на Y. В этом случае знание β очень ценно.

Как и в нашем обсуждении в главе 5, когда мы строим эту регрессию, то получаем оценки $\widehat{\alpha}$, $\widehat{\beta}$ и $\widehat{\gamma}$, находя значения α, β и γ, которые минимизируют сумму квадратов ошибок. Давайте посмотрим, что это значит.

Для любых произвольных значений параметров регрессии, скажем α', β' и γ', соответствующий прогноз Y_i для индивидуального объекта i равен

$$\alpha' + \beta' \cdot T_i + \gamma' \cdot X_i.$$

Характерные ошибки, связанные с этой регрессией, обозначим как ε'. Для каждого наблюдения i они представляют собой фактический эффект минус прогнозируемый эффект:

$$\varepsilon_i' = Y_i - (\alpha' + \beta' \cdot T_i + \gamma' \cdot X_i).$$

Оценки OLS – $\widehat{\alpha}$, $\widehat{\beta}$ и $\widehat{\gamma}$ – представляют собой конкретные значения параметров регрессии, которые минимизируют сумму квадратов этих ошибок. Ваш компьютер может вычислить их очень быстро.

Допустим, мы знаем, что достаточно учесть влияние X, и других искажающих факторов нет. Таким образом, регрессия Y по T и X дает несмещенную оценку влияния T на Y. Вопрос в том, насколько смещенными были бы наши результаты, если бы мы не смогли контролировать X.

Оказывается, мы можем ответить на этот вопрос. Назовем уравнение 10.1 *длинной регрессией*, поскольку оно включает X. Теперь предположим, что вместо него мы построили следующую короткую регрессию:

$$Y = \alpha^S + \beta^S \cdot T + \varepsilon^S. \tag{10.2}$$

Верхний индекс S здесь указывает на то, что речь идет о короткой (short) регрессии. Важно отметить, что нет никакой гарантии, что β^S из короткой регрессии будет таким же, как β из длинной регрессии. На самом деле они и не должны быть одинаковыми, если X является искажающим фактором.

Мы можем количественно оценить смещение, связанное с отсутствием X в регрессии. Рассмотрим регрессию, которая рассматривает переменную контроля (X) как зависимую переменную и регрессирует ее по воздействию (T):

$$X = \tau + \pi \cdot T + \xi.$$

Обратите внимание, что здесь мы использовали другие греческие буквы для обозначения параметров регрессии. Теперь мы называем точку пересечения τ (греческая буква тау), коэффициент воздействия π (греческая буква пи) и ошибку ξ (греческая буква кси). Мы сделали это по нескольким причинам.

Во-первых, и это самое главное, мы не хотели использовать здесь одни и те же буквы, которые имели другое значение выше. Параметр π здесь описывает корреляцию между T и X – это наклон, связывающий изменения T с изменениями X. Мы не хотели, чтобы вы путали его с двумя β, которые мы видели в этом разделе (β и β^S, каждый из которых описывает некоторую версию взаимосвязи между воздействием и результатом Y). Во-вторых, мы также не хотим, чтобы вы думали, что в некоторых греческих буквах есть что-то особенное. Это не тот случай, когда α всегда должна обозначать точку пересечения, β – коэффициент воздействия и т. д. Ведь это всего лишь символы. В разных книгах часто используют разные обозначения. Нам бы хотелось, чтобы вы смогли взглянуть на уравнение и сразу понять, где константа, что такое коэффициент воздействия, какова ошибка и т. д., даже если кто-то использует совершенно другие символы.

Смещение, возникшее по причине исключения X из регрессии результата воздействия, составляет $\beta^S - \beta$. Оказывается, это смещение равно $\pi \cdot \gamma$. То есть:

$$\text{Смещение} = \beta^S - \beta = \pi \cdot \gamma.$$

Иногда мы называем это уравнение формулой *смещения отсутствующей переменной*.

Эта формула говорит нам о том, что короткая регрессия дает смещенную оценку зависимости результата от воздействия, если переменная контроля коррелирует с переменной воздействия (так что $\pi \neq 0$) и влияет на переменную результата (так что $\gamma \neq 0$).

Если мы не можем наблюдать фактор X, то не можем контролировать его, включая в регрессию. Но формула смещения отсутствующей переменной дает

нам возможность задуматься о направлении и степени смещения. Действительно, эта формула воплощает наши идеи из главы 9 о том, как обозначить смещение, как показано на рис. 10.2, повторяющем рис. 9.7, но указывает на то, что параметры регрессии π и γ непосредственно измеряют отношения, важные для определения знака смещения.

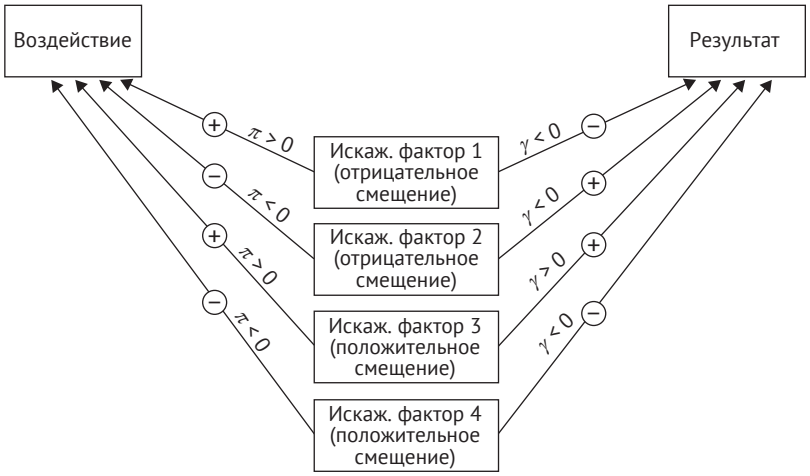


Рис. 10.2. Формула смещения отсутствующей переменной говорит нам, как определить знак смещения от пропущенного искажающего фактора

Если существует ненаблюдаемый искажающий фактор, который, как мы подозреваем, положительно связан как с T (поэтому $\pi > 0$), так и с Y (поэтому $\gamma > 0$), то формула смещения отсутствующей переменной говорит нам, что $\beta^S - \beta > 0$, поэтому мы завышаем эффект T. То же самое верно, если искажающий фактор отрицательно связан как с T, так и с Y (так что π и γ оба отрицательны), – опять же смещение положительное, и мы получаем завышенную оценку эффекта. Если искажающий фактор положительно связан с T, но отрицательно связан с Y (поэтому $\pi > 0$ и $\gamma < 0$) или наоборот (поэтому $\pi < 0$ и $\gamma > 0$), смещение отрицательное, и мы недооцениваем влияние T. Эти соображения сведены в табл. 10.3.

Таблица 10.3. Формула смещения отсутствующей переменной помогает нам подумать о том, приводит ли неспособность контролировать искажающий фактор к переоценке или недооценке причинного эффекта

	Пропущенная переменная положительно коррелирует с воздействием $\pi > 0$	**Пропущенная переменная отрицательно коррелирует с воздействием $\pi < 0$**
Пропущенная переменная положительно коррелирует с результатом $\gamma > 0$	Положительное смещение $\pi \cdot \gamma > 0$	Отрицательное смещение $\pi \cdot \gamma < 0$
Пропущенная переменная отрицательно коррелирует с результатом $\gamma < 0$	Отрицательное смещение $\pi \cdot \gamma < 0$	Положительное смещение $\pi \cdot \gamma > 0$

Как регрессия ограничивает влияние искажающего фактора?

Мы видели, что контроль одной переменной (X) может изменить коэффициент, описывающий взаимосвязь между другой интересующей нас переменной (T) и переменной результата (Y). В частности, контроль X изменит предполагаемую связь между T и Y, если X коррелирует с T и имеет независимую связь с Y. Вот один из способов графически представить, что делает регрессия, когда мы контролируем переменную.

Предположим, мы хотим узнать, как рост влияет на доход, и в этом случае интересующие нас переменные результата и воздействия являются непрерывными (в принципе, они могут принимать бесконечное и несчетное число возможных значений). На рис. 10.3 показаны некоторые данные о доходах и росте, полученные в ходе национального опроса населения, проведенного Бюро статистики труда США. Репрезентативную выборку жителей США, родившихся между 1980 и 1984 гг., спросили об их росте и доходах в 2014 г., когда им было от 34 до 38 лет.

Чтобы облегчить визуализацию, мы сгруппировали респондентов по росту и полу, поэтому каждая точка на рис. 10.3 соответствует группе из 15 или более человек одного пола и роста (измеряется в дюймах). На рисунке показаны средний доход каждой группы, измеряемый в тысячах долларов выше 20 000 долл., и средний рост, измеряемый в футах выше 5 футов. (Сейчас вы поймете, почему мы масштабировали наши переменные таким необычным способом.) Пустые кружки соответствуют группам мужчин, а закрашенные – группам женщин.

Визуально мы наблюдаем сильную положительную корреляцию между ростом и доходом. Что бы мы получили, если бы построили регрессию доходов по росту на основе этих данных, игнорируя пол?

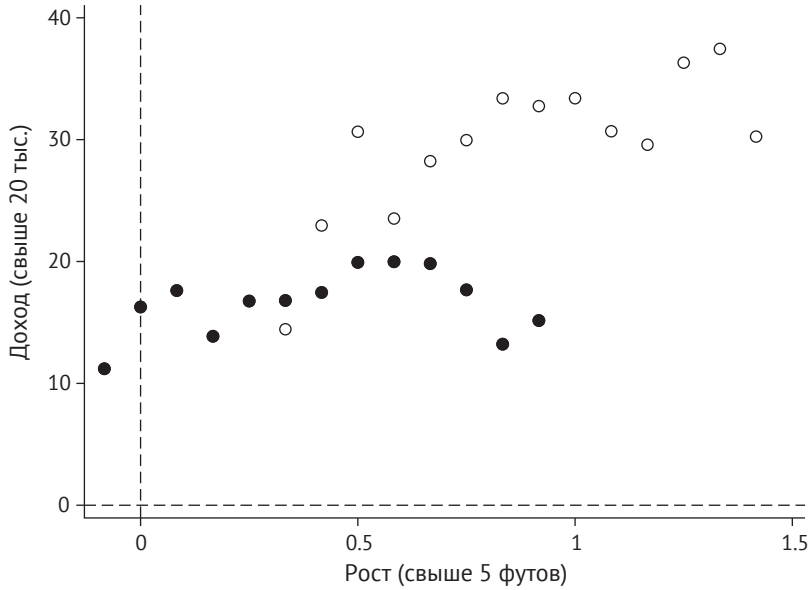


Рис. 10.3. Доход и рост американцев в возрасте от 34 до 38 лет в 2014 г.

Как мы видели ранее, для этого просто нужно найти линию, которая лучше всего соответствует данным. Эта линия изображена на рис. 10.4. Действительно, линия наилучшего соответствия имеет сильный положительный наклон, указывая на то, что в среднем более высокие люди зарабатывают более высокие доходы.

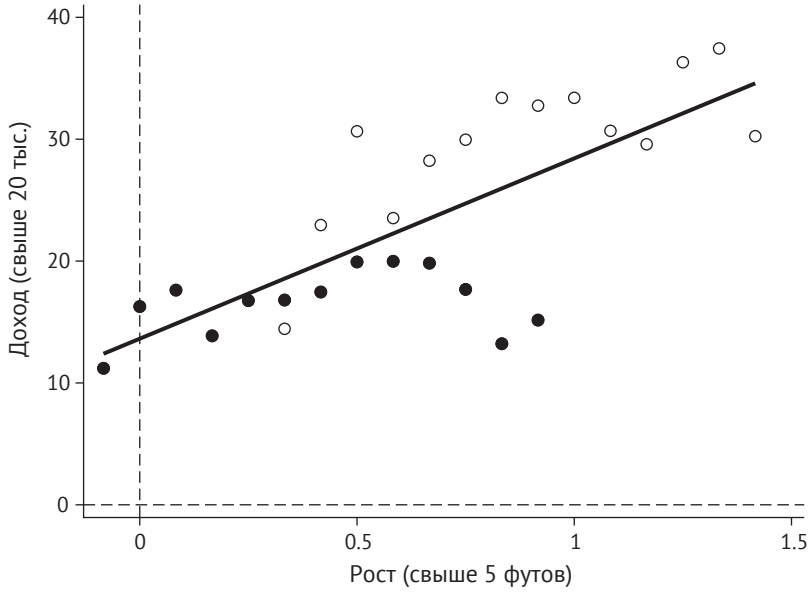


Рис. 10.4. Линия регрессии дохода от роста

Если точнее, регрессия находит линию наилучшего соответствия путем определения значений α и β, которые минимизируют сумму квадратов ошибок в следующем уравнении:

$$\text{Доход} = \alpha + \beta \cdot \text{Рост} + \varepsilon.$$

Эти два значения проиллюстрированы на рис. 10.5. Высота линии при значении Рост = 0 (т. е. когда рост человека составляет 5 футов) равна $\widehat{\alpha}$, а наклон линии равен $\widehat{\beta}$. Для этого конкретного набора данных мы оцениваем наклон примерно в 14.8. В среднем люди ростом на один фут выше зарабатывают дополнительно 14 800 долл. в год!

Конечно, прежде чем выводить причинно-следственную интерпретацию этого коэффициента регрессии, нам следует подумать о факторах, искажающих результат. Один из вариантов – это пол. Мужчины в среднем выше женщин. И мы подозреваем, что мужчины в среднем зарабатывают более высокие доходы, чем женщины, по причинам, не связанным с ростом. (Это может быть результатом гендерной дискриминации на рынках труда или других социальных факторов. Хотя причины, конечно, очень важны, нам не нужно их знать, чтобы контролировать пол респондента как искажающий фактор.) Действительно, на графике видно, что женщины в среднем имеют более низкий рост и более низкие доходы. Таким образом, пол – это фактор, который нам, скорее всего, придется учитывать в этой регрессии.

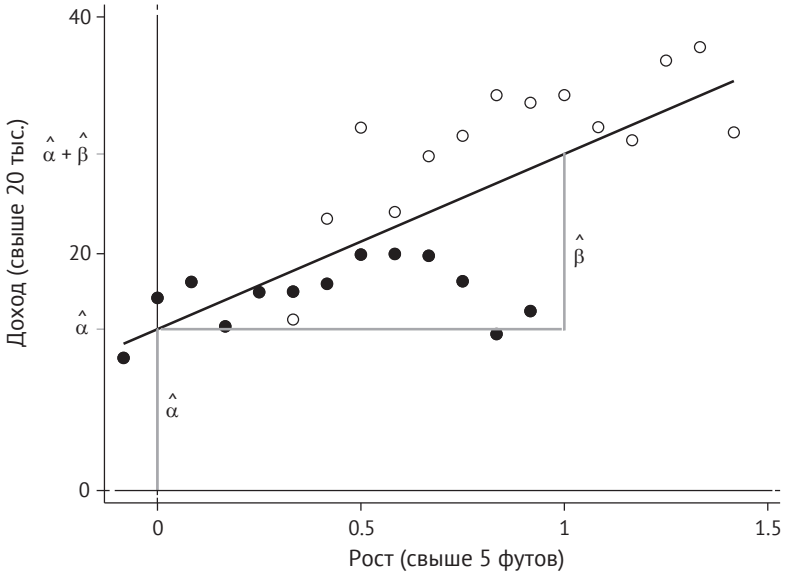


Рис. 10.5. Графическая интерпретация коэффициентов регрессии

Один из способов решить эту проблему – построить отдельные регрессии для мужчин (*M*) и женщин (*W*):

$$\text{Доход} = \alpha^M + \beta^M \cdot \text{Рост} + \varepsilon^M,$$

$$\text{Доход} = \alpha^W + \beta^W \cdot \text{Рост} + \varepsilon^W.$$

Если мы это сделаем, то получим две линии регрессии, как показано на рис. 10.6.

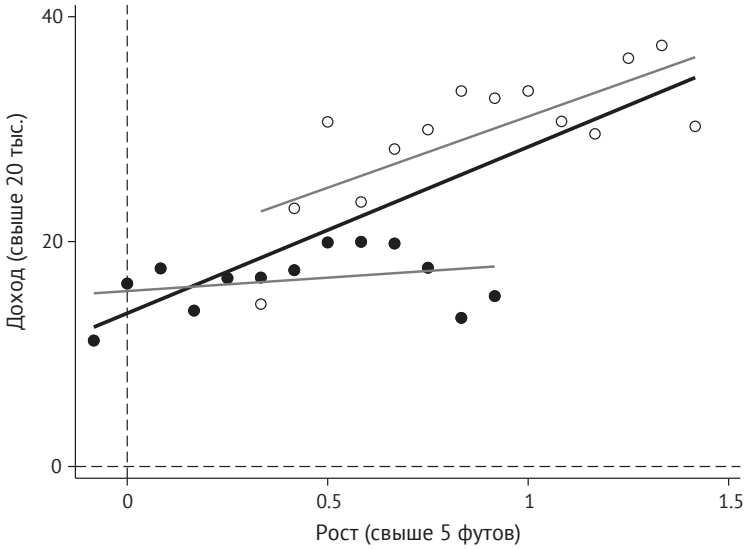


Рис. 10.6. Линия регрессии для объединенных данных (черная) и отдельные линии регрессии для мужчин и женщин (серые)

Отдельные линии регрессии для мужчин и женщин показаны серым цветом, в то время как предыдущая общая линия регрессии по-прежнему показана черным. Интересно, что корреляция между ростом и доходом внутри каждого пола меньше, чем среди населения в целом. То есть и $\widehat{\beta}^{W}$, и $\widehat{\beta}^{M}$ меньше, чем $\widehat{\beta}$ из нашей предыдущей регрессии. Также обратите внимание, что наклон линии для мужчин больше, чем для женщин: $\widehat{\beta}^{M} > \widehat{\beta}^{W}$.

Эта процедура разделения данных и построения отдельных регрессий показывает нам корреляцию между доходом и ростом отдельно для мужчин и женщин. Возвращаясь к нашему примеру с политиками конгресса, мы видим, что это аналогично ячейкам в нижней части табл. 10.2, которые показывают разницу в среднем балле ACU между республиканцами и демократами для каждой группы баллов NPAT.

Хотя знать отдельные корреляции полезно, как и в примере с конгрессменами, обычно от нас ждут единственную сводную оценку корреляции между доходом и ростом с учетом пола. Это число будет представлять собой средневзвешенное значение наклонов двух серых линий на рис. 10.6 (так же как в примере с политикой конгресса, где единственное число представляло собой средневзвешенное значение индивидуальных различий в нижней части табл. 10.2). Но нам нужно знать, как назначать весовые коэффициенты.

Самый простой способ сделать это – построить регрессию дохода в зависимости от роста и пола. Уравнение регрессии будет выглядеть так:

$$\text{Доход} = \alpha + \beta \cdot \text{Рост} + \gamma \cdot \text{Мужчина} + \varepsilon.$$

Как эта регрессия будет отдельно оценивать α, β и γ с графической точки зрения? Вместо того чтобы искать одну линию, которая лучше всего соответствует данным, мы можем искать две строки, которые лучше всего соответствуют данным: одну для мужчин и одну для женщин. Но, в отличие от варианта с двумя отдельными регрессиями, теперь мы ограничиваем эти две линии одинаковым наклоном ($\widehat{\beta}$). На рис. 10.7 показано, как будут выглядеть эти две линии, и приведено сравнение с линиями, которые мы получили, когда проводили отдельные регрессии для мужчин и женщин.

Обратите внимание, что наклон двух черных линий идентичен, а величина наклона находится где-то посередине относительно наклонов двух серых линий, представляющих отдельные регрессии. То есть это средневзвешенное значение двух показателей. На рис. 10.8 показано, что, оценив эти две параллельные линии наилучшего соответствия, мы получим параметры регрессии.

Точка пересечения линии для женщин (переменная Мужчина = 0) равна $\widehat{\alpha}$. Расстояние между двумя линиями равно $\widehat{\gamma}$. А наклон двух линий равен $\widehat{\beta}$. Другими словами, $\widehat{\alpha}$ – это прогнозируемый доход для женщин ростом 5 футов; $\widehat{\gamma}$ – прогнозируемая разница в доходах между мужчинами и женщинами одинакового роста; $\widehat{\beta}$ – среднее соотношение между ростом и доходом с учетом пола.

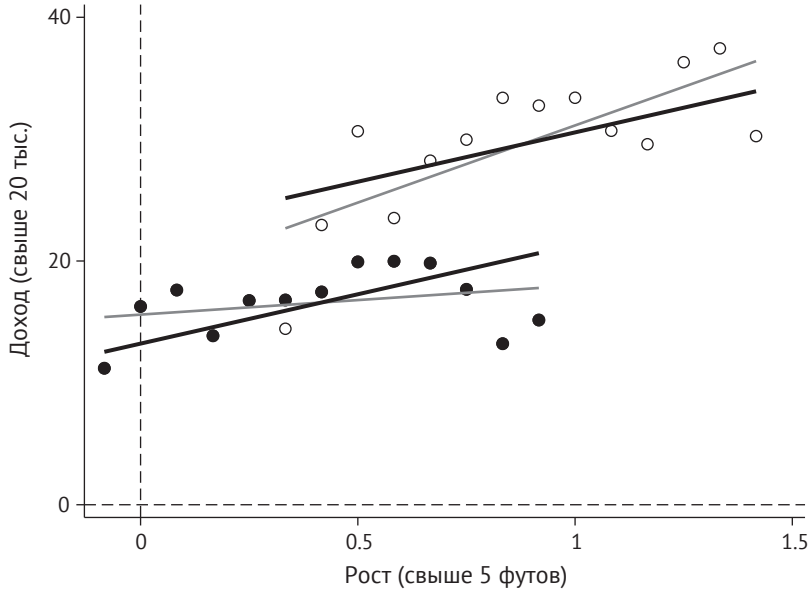


Рис. 10.7. Представление регрессии, в которой мы контролируем фактор пола, включая его в регрессию дохода по росту (черные линии), а также отдельные линии регрессии для мужчин и женщин (серые линии)

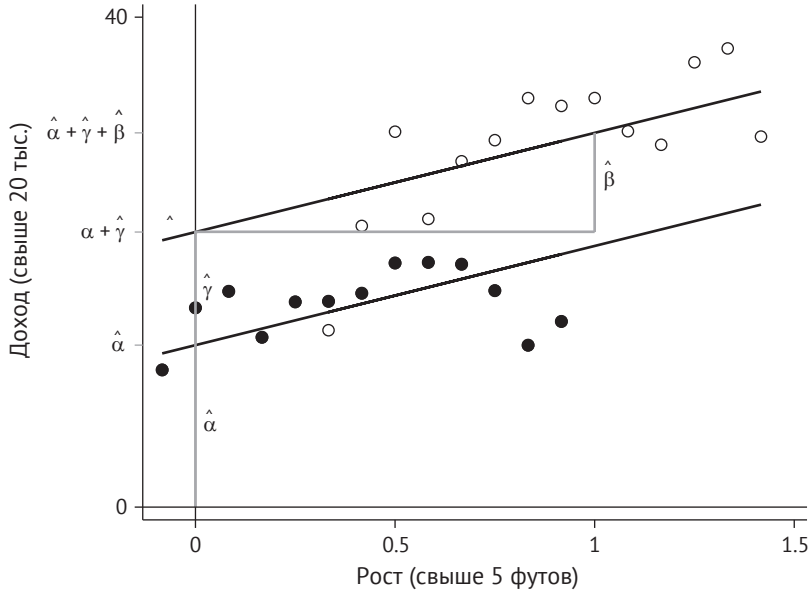


Рис. 10.8. Коэффициенты регрессии при учете фактора пола в регрессии дохода по росту

Неудивительно, что фактор пола имеет существенное значение для оценки взаимосвязи между ростом и доходом. Вместо 14.8 наша новая оценка накло-

на составляет около 8.1. Изменение связано с тем, что пол является фактором, влияющим как на рост, так и на доход. Из главы 9 вы знаете, что, если искажающий фактор положительно коррелирует как с воздействием, так и с результатом, как в данном случае, он создает положительное смещение. Поскольку значение 14.8 было завышенной оценкой истинного влияния роста на доход, с учетом фактора пола мы получаем меньшую оценку.

Стоит отметить, что контроль фактора пола влияет не только на нашу оценку связи между доходом и ростом, но и на точность этой оценки, хотя направление этого эффекта теоретически неоднозначно. С одной стороны, добавление контроля, коррелирующего с результатом, уменьшает остаточную вариацию этого результата, что повышает точность. С другой стороны, добавление контроля, коррелирующего с воздействием, уменьшает остаточную вариацию воздействия, что увеличивает неопределенность наших оценок. Улучшит ли контроль искажающего фактора точность или ухудшит ее, будет зависеть от соотношения этих двух сил.

Учитывая вышеизложенное, может возникнуть соблазн добавить в вашу регрессию дополнительные контрольные переменные (т. е. контролируемые факторы) не с целью уменьшения систематической ошибки, а с целью повышения точности. Это правильно, если вы можете найти переменные, которые сильно коррелируют с результатом, но не с воздействием; включение их в регрессию повысит точность ваших оценок. Однако если вы продолжаете экспериментировать с контрольными переменными до тех пор, пока не получите статистически значимую оценку, это *p*-хакерство и плохая идея.

Поскольку мы говорили об аналогии между тем, что мы только что сделали, и нашим примером с политиками конгресса, давайте вернемся к этому примеру в рамках регрессии. Обратите внимание, что в этом случае воздействие (республиканец или демократ) является бинарным, но потенциальный искажающий фактор (убеждения) непрерывно измеряется по шкале NPAT.

Опять же, начнем с точечной диаграммы, на этот раз рейтинга Американского союза консерваторов (ACU) на вертикальной оси и рейтинга консервативности NPAT на горизонтальной оси. На рис. 10.9 пустые кружки обозначают демократов, а закрашенные – республиканцев.

Так как воздействие является бинарным, мы можем начать с простого сравнения среднего рейтинга ACU для республиканцев и демократов. Рассмотрим следующее уравнение регрессии:

$$\text{Рейтинг ACU} = \alpha + \beta \cdot \text{Республиканец} + \varepsilon.$$

При минимальной сумме квадратов ошибок коэффициент $\widehat{\alpha}$ равен среднему баллу ACU для демократа (Республиканец = 0), а коэффициент $\widehat{\beta}$ представляет собой разницу между средним рейтингом ACU для республиканца и демократа. Таким образом, как мы уже видели, $\widehat{\alpha} = 20$ и $\widehat{\beta} = 84 - 20 = 64$. Это показано на рис. 10.10, где горизонтальные линии соответствуют средним рейтингам ACU для демократов и республиканцев.

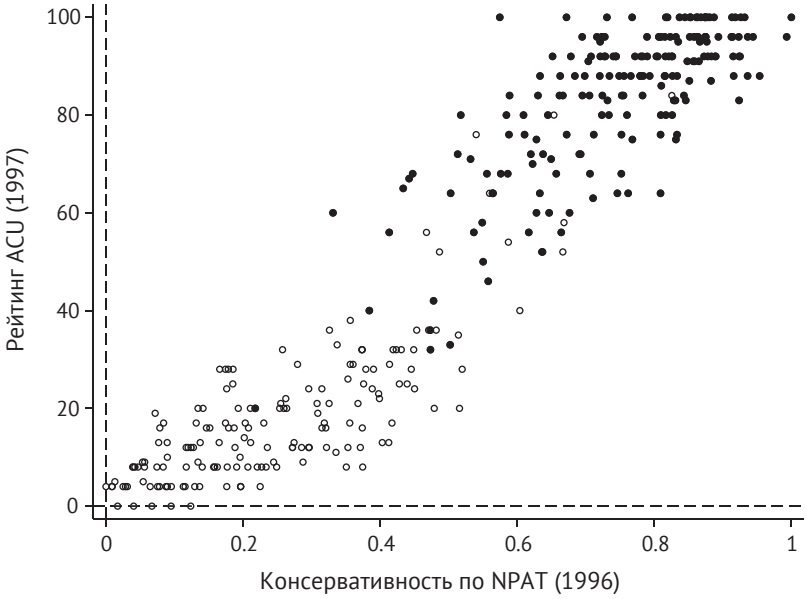


Рис. 10.9. Рейтинг ACU и рейтинг консервативности NPAT

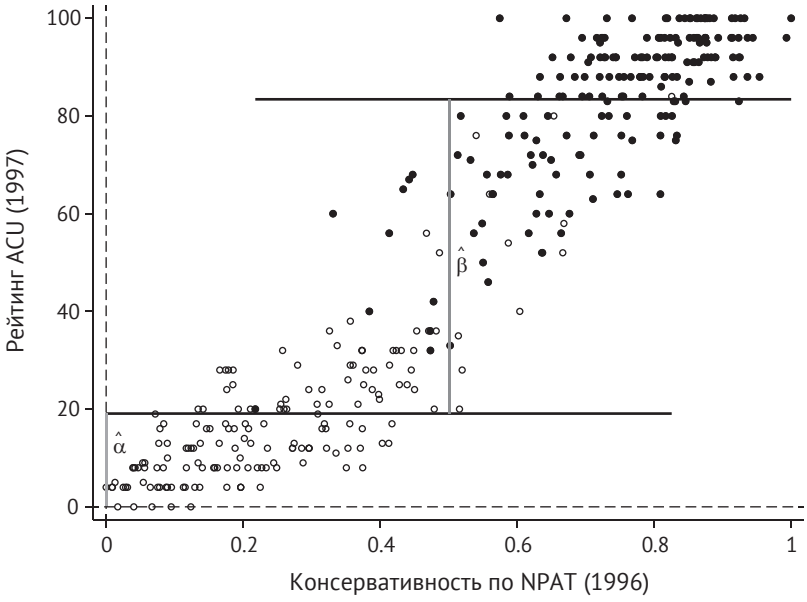


Рис. 10.10. Коэффициенты регрессии рейтинга ACU по партиям

Нас, конечно, беспокоит тот факт, что в этой регрессии личные политические убеждения являются искажающим фактором и $\widehat{\beta}$ не оценивает истинное влияние принадлежности к партии на поведение при голосовании в конгрессе. Поэтому нам следует контролировать фактор личных убеждений. Мы сде-

лаем это, используя рейтинг консервативности NPAT, измеряемый по горизонтальной оси.

В нашем примере с зависимостью дохода искажающим фактором был пол, представленный в виде бинарной переменной. Нашим первым интуитивным шагом к контролю искажающего фактора было построение исходной регрессии (дохода от роста) отдельно для каждого значения искажающего фактора. Затем мы увидели, что окончательный коэффициент регрессии с учетом пола представляет собой средневзвешенное значение наклонов этих двух отдельно построенных линий регрессии.

В примере с конгрессменами, поскольку наш искажающий фактор непрерывен, мы не можем построить отдельную регрессию для каждого значения искажающего фактора. Но мы можем сделать нечто подобное: построить регрессию рейтинга ACU по консервативности NPAT отдельно для демократов и республиканцев (верхние индексы P у коэффициентов регрессии указывают на то, что это регрессия для партии P):

$$\text{Рейтинг ACU} = \alpha^P + \gamma^P \cdot \text{Консервативность NPAT} + \varepsilon^P.$$

Это дает нам две линии регрессии: одну для республиканцев и одну для демократов, как показано на рис. 10.11.

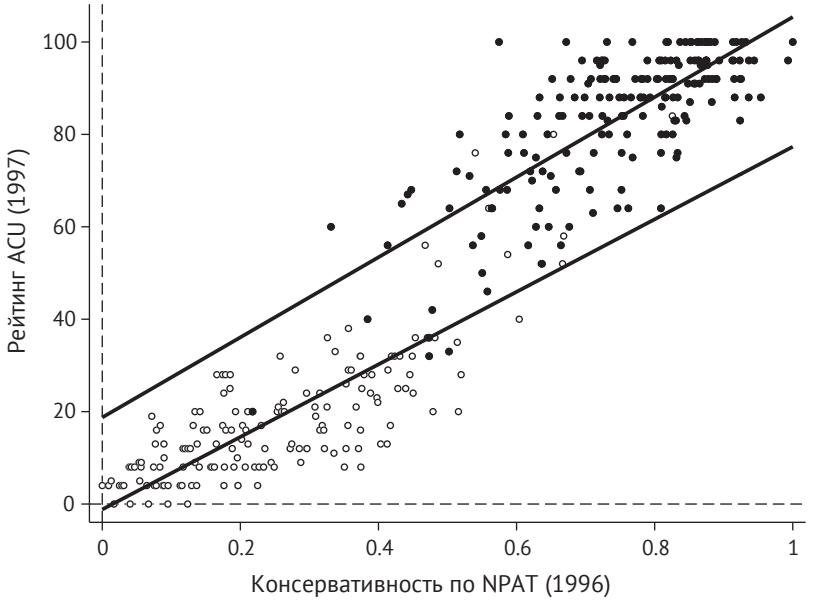


Рис. 10.11. Контроль оценки NPAT (убеждений) во взаимосвязи между оценкой ACU и принадлежностью к партии путем построения двух отдельных регрессий, по одной для каждой партии

Для каждого значения консервативности NPAT прогнозируемый рейтинг ACU республиканца с этим показателем консервативности NPAT равен:

$$\hat{\alpha}^R + \hat{\gamma}^R \cdot \text{Консервативность NPAT}.$$

И для каждого значения консервативности NPAT прогнозируемый рейтинг ACU демократа с этим показателем консервативности NPAT равен:

$$\hat{\alpha}^{D} + \hat{\gamma}^{D} \cdot \text{Консервативность NPAT.}$$

Это означает, что при любом заданном значении консервативности NPAT разрыв между двумя линиями представляет собой разницу в прогнозируемом рейтинге ACU между республиканцами и демократами с этим показателем NPAT. Следовательно, эта регрессия позволяет нам получить непрерывный аналог нашего предыдущего бинарного сравнения. Для каждого значения консервативности NPAT он сообщает нам, какова прогнозируемая разница в среднем рейтинге ACU между республиканцами и демократами.

Но мы еще не закончили. Как и раньше, цель состоит в том, чтобы получить единый показатель взаимосвязи между рейтингом ACU и членством в партии с учетом консервативности NPAT. В настоящее время у нас есть отдельная мера этой взаимосвязи для каждого значения консервативности NPAT. Таким образом, последним шагом в организации контроля искажающего фактора является использование регрессии для создания средневзвешенного значения этих различий, которое минимизирует сумму квадратов ошибок. Мы делаем это с помощью следующей регрессии:

$$\text{Рейтинг ACU} = \alpha + \beta \cdot \text{Республиканец} + \gamma \cdot \text{Консервативность NPAT} + \varepsilon.$$

Рисунок 10.12 иллюстрирует эту регрессию. Параметр $\hat{\alpha}$ сообщает нам средний рейтинг ACU демократа с консервативностью NPAT, равной 0. Параметр $\hat{\gamma}$ показывает наклон зависимости между рейтингом ACU и консервативностью NPAT. Важно отметить, что, в отличие от двух наших предыдущих регрессий, где $\hat{\gamma}^{R}$ и $\hat{\gamma}^{D}$ были разными, эта регрессия предполагает, что наклон зависимости между рейтингом ACU и консервативностью NPAT будет одинаковым для обеих сторон. Следовательно, этот наклон $\hat{\gamma}$ представляет собой средневзвешенное значение $\hat{\gamma}^{R}$ и $\hat{\gamma}^{D}$. Наконец, коэффициент $\hat{\beta}$ представляет собой зазор между двумя линиями. Это расстояние является постоянным при всех показателях консервативности NPAT, поскольку мы заставили $\hat{\gamma}$ быть одинаковым для обеих сторон, сделав линии параллельными. Следовательно, $\hat{\beta}$ оценивает среднюю разницу в рейтинге ACU между республиканцами и демократами с учетом консервативности NPAT.

Контроль и причинно-следственная связь

Хотя контроль позволяет смягчить или устранить смещения, возникающие из-за конкретных факторов, которые вы можете измерить и включить в свою регрессию, в большинстве случаев мы, как правило, по-прежнему скептически относимся к тому, что сам по себе контроль позволяет нам получить объективные оценки причинно-следственных связей. В главе 9 было сказано, что, если мы хотим интерпретировать корреляцию как несмещенную оценку причинного эффекта, мы должны быть уверены в отсутствии базовых различий между объектами, подвергшимися и не подвергшимися воздействию.

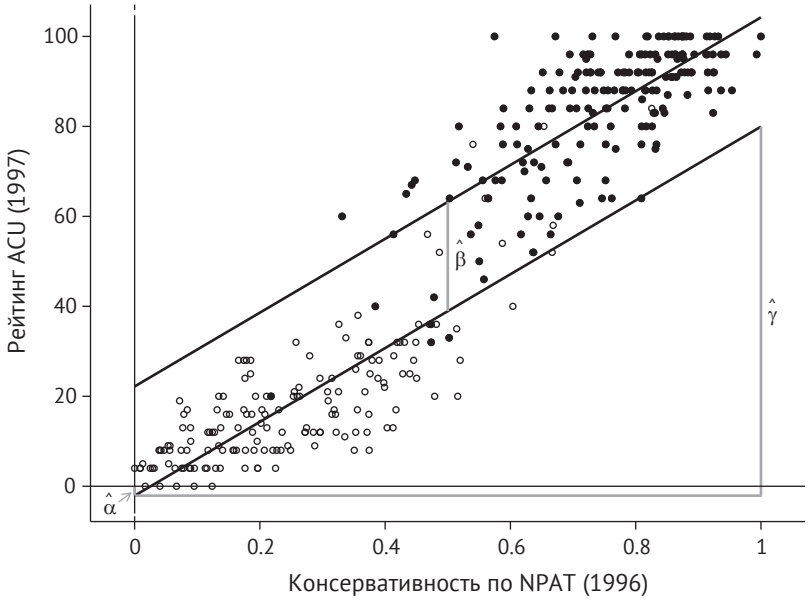


Рис. 10.12. Контроль оценки NPAT (политические убеждения) в регрессии оценки ACU на стороне

Другими словами, нужно сравнивать яблоки с яблоками. Если мы выполняем регрессию *Y* по *T* и *X* (и другим возможным искажающим факторам), мы все равно сделаем аналогичное утверждение, если дадим коэффициенту при *T* причинную интерпретацию. Мы фактически утверждаем, что, за исключением набора переменных, которые мы контролируем, в отношениях *Y* и *T* нет никаких других искажающих факторов, а также нет обратной причинно-следственной связи. Иными словами, чтобы использование контроля искажающих факторов давало объективную оценку причинного эффекта, мы должны учитывать *все* такие факторы.

По нашему опыту, трудно найти ситуации (кроме рандомизированных экспериментов, которые мы обсудим в главе 11), в которых можно поверить в отсутствие упущенных искажающих факторов. Как правило, даже если аналитик контролирует множество факторов, можно предположить наличие других потенциальных факторов, которые либо не наблюдаемы, либо не измерены в данных и, следовательно, не поддаются контролю. Например, спросите себя, можете ли вы вспомнить какие-либо потенциальные факторы, влияющие на взаимосвязь между доходом и ростом, помимо пола. Разумеется, к таким факторам вы можете отнести экономические, биологические, культурные, медицинские и другие различия. Например, богатые родители могут обеспечить своим детям лучшее питание, что способствует более высокому росту, а также содействовать им другими способами, которые позволят в будущем получать более высокие доходы. Трудно представить, что вы сможете измерить и контролировать все возможные искажающие факторы.

Обратная причинно-следственная связь – еще одна причина, по которой мы обычно не надеемся, что контроль искажающих факторов может сам по себе

выявить причинно-следственные связи. В главе 9 мы говорили о том, как искажающие факторы и обратная причинно-следственная связь могут помешать нам провести сравнение однотипных показателей. Идея контроля искажающих факторов состоит в том, чтобы попытаться как можно лучше учесть их влияние, но если существует обратная причинно-следственная связь, т. е. результат влияет на воздействие, то никакой контроль не сможет решить эту проблему.

Позвольте нам привести пример.

Вредят ли нам социальные сети?

Широко распространена обеспокоенность тем, что длительное пребывание в социальных сетях вредит людям. И действительно, многие исследования показывают отрицательную корреляцию между использованием социальных сетей и различными показателями субъективного благополучия и психического здоровья человека.

Разумеется, эта корреляция не обязательно отражает причинное влияние социальных сетей на благополучие. Например, может существовать обратная причинно-следственная связь: возможно, люди, которые грустят, одиноки или расстроены, проводят больше времени в социальных сетях, чем люди, которые более счастливы или имеют больше социальных связей. Или здесь могут проявляться искажающие факторы – возможно, социально-экономический статус, образование или географическое положение влияют как на использование социальных сетей, так и на субъективное благополучие.

Первое, что вы можете сделать, чтобы оценить причинно-следственную связь, – это устранить некоторые искажающие факторы. Насколько хорошо эта стратегия контроля позволит оценить истинный причинный эффект?

Существует исследование, которое может пролить свет на этот вопрос. Группа ученых, интересующихся влиянием социальных сетей, провела эксперимент. Сначала они определили большую группу пользователей Facebook, желающих принять участие в исследовании. (Участники не знали, о чем идет речь.) У каждого из этих людей они запросили показатели субъективного благополучия, использования Facebook и сумму, которую они хотели бы получить за отказ от Facebook на месяц. Затем экспериментаторы случайным образом выбрали некоторых из этих людей и действительно заплатили им за отключение Facebook на месяц (за соблюдением условия они смогли следить). Остальные не отключили Facebook, но продолжали участвовать в исследовании в качестве контрольной группы. Затем в конце эксперимента исследователи снова измерили субъективное благополучие, чтобы увидеть, повлиял ли отказ от Facebook на субъективное ощущение благополучия у участников экспериментальной группы по сравнению с участниками в контрольной группе, которые не отказались от посещения Facebook.

С нашей точки зрения, в этом исследовании отрадно то, что эксперимент путем случайного выбора пользователей Facebook дает объективную оценку влияния социальной сети. В начале эксперимента исследователи также задали вопрос об использовании Facebook и субъективном благополучии, поэтому, сравнивая уровни благополучия людей с разным уровнем использования Facebook в начале исследования, они также могли воспроизвести описанную простую корреляцию в более ранних исследованиях. Более того, исследователи также собрали детальную информацию о людях, участвовавших в их исследовании, и поэтому

могли контролировать некоторые потенциальные искажающие факторы, например доход, возраст, пол, образование, расу, политическую принадлежность.

Если бы они могли контролировать все искажающие факторы, то оценка связи между использованием Facebook и субъективным благополучием в результате рандомизированного эксперимента полностью совпала бы с корреляцией, учитывающей искажения. Сравнивая простую корреляцию, корреляцию с контролем искажающих факторов и экспериментальную оценку, мы можем понять, насколько хорошо в этих условиях контроль помогает выявить истинный причинный эффект.

На рис. 10.13 показаны простая корреляция, корреляция, учитывающая потенциальные искажающие факторы, и экспериментальная оценка (каждая из них окружена 95-процентным доверительным интервалом). Все они измеряются в единицах среднего использования Facebook в день. Как видите, простая корреляция дает наибольшую оценку негативной связи Facebook с субъективным благополучием. Учет потенциальных факторов, искажающих ситуацию, немного снижает эту оценку. Но экспериментальная оценка составляет примерно одну треть размера оценок, полученных на основе простой корреляции, и примерно половину размера оценки с учетом потенциальных искажающих факторов. Это говорит о том, что стратегия контроля в данном случае не избавляет от существенного завышения истинного эффекта.

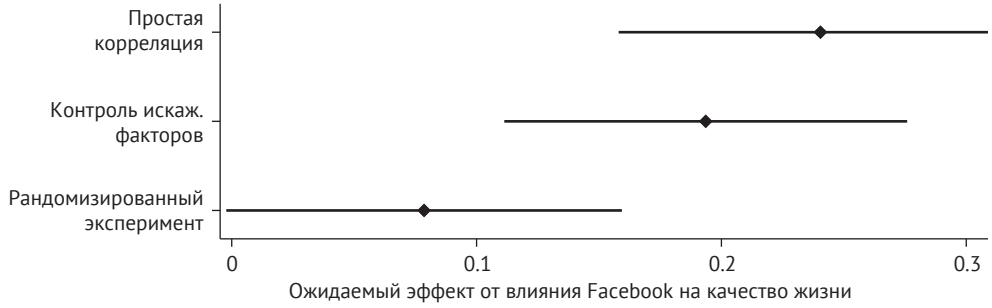


Рис. 10.13. Оценки связи между использованием Facebook и субъективным благополучием

Чтение таблицы регрессии

В этой книге вы много раз встречали графические изображения регрессии. Но когда вы запускаете регрессию на своем компьютере или видите результаты регрессии, обсуждаемые в отчете, они часто бывают представлены в виде таблицы. Поэтому нужно уметь понимать и интерпретировать различные части таблицы регрессии. Впервые мы познакомились с таблицами регрессии в главе 5, но теперь знаем достаточно, чтобы углубиться в детали.

Давайте вернемся к нашему анализу взаимосвязи между протоколами голосования в конгрессе и партийной принадлежностью. Для этой задачи мы построили три регрессии.

В первом случае мы построили регрессию поименного голосования по партиям, не контролируя никакие искажающие факторы:

$$\text{Рейтинг ACU} = \alpha + \beta \cdot \text{Республиканец} + \varepsilon.$$

Во втором случае мы контролировали личные убеждения, включая индикаторы для разных диапазонов баллов NPAT:

$$\text{Рейтинг ACU} = \alpha + \beta_1 \cdot \text{Республиканец} + \beta_2 \cdot \text{NPAT}_{21-40} + \beta_3 \cdot \text{NPAT}_{41-60} + \beta_4 \cdot \text{NPAT}_{61-80} + \beta_5 \cdot \text{NPAT}_{81-100} + \varepsilon.$$

Именно это мы и сделали, чтобы найти правильное средневзвешенное значение в ходе обсуждения табл. 10.2.

В третьем случае мы контролировали личные убеждения, включая непрерывную переменную NPAT:

$$\text{Рейтинг ACU} = \alpha + \beta \cdot \text{Республиканец} + \gamma \cdot \text{Консервативность NPAT} + \varepsilon.$$

В каждом столбце табл. 10.4 представлены результаты одной из этих трех регрессий.

Таблица 10.4. Связь между рейтингом ACU и партийной принадлежностью

Переменные	Рейтинг ACU	Рейтинг ACU	Рейтинг ACU
Республиканец	64.32** (1.71)	23.74** (2.25)	24.28** (1.98)
NPAT_{21-40}		8.01** (1.76)	
NPAT_{41-60}		32.74** (2.29)	
NPAT_{61-80}		52.27** (2.83)	
NPAT_{81-100}		59.77** (2.83)	
Консерватизм NPAT			82.05** (3.44)
Константа	19.09** (1.25)	10.29** (1.24)	−2.10 (1.18)
Наблюдения	349	349	349
r^2	0.80	0.92	0.93

В скобках указана стандартная ошибка. $^{**}p < 0.01$

Давайте разберемся, как читать эту таблицу. Первый столбец просто содержит метки. Во втором столбце показаны результаты нашей первой регрессии: рейтинг ACU по республиканцам без контроля искажающих факторов. В третьем столбце показаны результаты нашей второй регрессии: рейтинг ACU для республиканцев с учетом категории NPAT. В четвертом столбце показаны результаты нашей третьей регрессии: рейтинг ACU для республиканцев с учетом непрерывного показателя консервативности NPAT.

В первой строке мы видим имя нашей зависимой переменной. Для этих трех регрессий это всегда рейтинг ACU. Для каждой регрессии строка с меткой «Республиканец» показывает три фрагмента данных. Верхнее число – это оценка коэффициента перед переменной «Республиканец» в регрессии. Нижнее число в скобках – это стандартная ошибка этой оценки. А звездочки указывают, является ли статистически значимым отличие от нуля (и на каком уровне). Просматривая эту строку, мы видим, что в первой регрессии коэффициент перед переменной «Республиканец» равен 64.32. Но как только мы вводим контроль фактора NPAT, он резко уменьшается. Если учитывать категории NPAT, он падает до 23.74. А если мы будем учитывать непрерывный показатель консервативности NPAT, то он составит 24.28. Два последних значения мало различаются, и это не удивительно – не так уж важно, как именно мы учитываем личные убеждения.

Просматривая таблицу, мы далее получаем оценки коэффициентов, стандартные ошибки и статистическую значимость для каждой из наших контрольных переменных. Вот почему следующие пять строк во втором столбце пусты – мы ничего не контролировали в этой регрессии. В третьем столбце заполнены четыре строки, связанные с категориями NPAT, но строка, связанная с оценкой консервативности NPAT, остается пустой. А в четвертом столбце строки категории NPAT пусты, но заполнена консервативность NPAT. Для всех трех регрессий заполнена строка с меткой «Константа». Это оценка точки пересечения ($\widehat{\alpha}$) для соответствующей регрессии.

Таблица содержит еще два типа данных. Для каждой регрессии таблица показывает, сколько наблюдений было в данных. Здесь ответ – 349, что отражает количество конгрессменов, заполнивших опрос NPAT в 1997 г.

И для каждой регрессии в таблице представлен статистический показатель r^2. В главе 2 было сказано, что это доля вариаций одной переменной, которую можно предсказать с помощью вариаций других переменных. Таким образом, значение 0.93 в нашей итоговой регрессии говорит о том, что в рамках имеющейся у нас выборки данных вы можете предсказать 93 % вариаций рейтинга конгрессмена ACU, используя знание партийной принадлежности и балла консервативности NPAT.

Хотя это звучит довольно привлекательно, мы призываем вас не переоценивать значимость показателя r^2. Фактически, когда мы работаем с регрессией, мы часто даже не смотрим на него. Обычно наша цель не состоит в том, чтобы предсказать или смоделировать изменение нашей зависимой переменной. Цель состоит в том, чтобы выяснить, имеет ли значение ключевая переменная воздействия для нашего результата. Поэтому нас интересует величина коэффициента перед этой переменной. Более того, получение высокого показателя r^2 само по себе не имеет особого смысла. Один из простых способов успешно предсказать большую часть изменений в ваших данных – это просто включить в регрессию множество контрольных переменных. Но это не равносильно пониманию происходящего! Вспомните, что мы говорили о переобучении в главе 5. Тот факт, что вы очень хорошо подогнали регрессию под имеющиеся данные (а высокое значение r^2 означает только это), включив в уравнение множество переменных, не означает, что вы сможете хорошо предсказать, каков будет результат, когда вы столкнетесь с данными не из вашего набора. А в некоторых случаях вы можете получить надежную и объективную оценку интересующего вас параметра даже при низком значении r^2.

Чем искажающий фактор отличается от механизма?

Думать о влиянии искажающих факторов становится сложнее, когда есть какая-то переменная, которая влияет как на воздействие, так и на результат, но на которую, в свою очередь, тоже влияет воздействие. Что мы можем сделать в этой ситуации? Эта переменная является искажающим фактором: она влияет как на воздействие, так и на результат. Значит, нам следует ее контролировать. Но, как мы обсуждали в главе 9, эта переменная также является механизмом: она подвержена воздействию и переносит его на результат. Значит, мы не должны ее контролировать, поскольку это часть пути, по которому воздействие влияет на результат. Что же нам делать?

Чтобы сделать эту головоломку более предметной, давайте вернемся к примеру из главы 9, где нас интересовала взаимосвязь между экономическим благополучием страны и возможностью начала гражданской войны. С одной стороны, развитые страны могут проводить более эффективную политику, которая увеличит доходы, а также предоставит лучшие возможности для ненасильственного выражения политического недовольства, что может напрямую снизить риск гражданской войны. С этой точки зрения вопрос о том, имеет ли страна развитый политический строй или нет, является искажающим фактором (т. е. ковариата до начала воздействия), и поэтому его следует контролировать. С другой стороны, по мере того, как страна становится богаче, ее граждане все больше привыкают решать проблемы мирным путем, да и сами внутренние проблемы теряют свою остроту, что снижает вероятность возникновения гражданской войны. С этой точки зрения политический строй является одним из механизмов, с помощью которых ВВП влияет на вероятность гражданской войны (т. е. ковариата после воздействия), значит, эту переменную не надо контролировать.

Однозначного решения в таких ситуациях действительно нет. Вас проклянут одни, если вы выберете первый вариант, и проклянут другие, если выберете второй – а это означает, что вы не можете узнать много интересного о причинно-следственных связях, которые вас интересуют. Для этого вам понадобится более творческий подход, который станет темой следующих нескольких глав.

Статистика без волшебства

Людям очень хотелось бы верить, что они способны оценить причинно-следственные связи, просто контролируя все искажающие факторы. И они хотят, чтобы вы тоже в это поверили. Но, как мы только что говорили, только полное отсутствие критического мышления может заставить вас поверить, что вы учли все искажающие факторы. Поэтому иногда люди используют математический жаргон, круто звучащие названия статистических методов, сложные компьютерные программы и другие технические чудеса, чтобы попытаться заставить вас думать, как им хочется. Важно не дать себя обмануть. Независимо от того, какие причудливые методы использует аналитик, если фундаментальная стратегия заключается в контроле искажающих факторов и если существуют правдоподобные искажающие факторы, которые либо не наблюдаемы, либо не измерены в данных, то, скорее всего, они так и останутся неучтенными. Компьютеры не творят чудеса. Они могут контролировать наблюдаемые

искажения. Но они не могут сделать ненаблюдаемые искажающие факторы наблюдаемыми.

Чтобы понять, что мы имеем в виду, рассмотрим пример прекрасного и полезного статистического метода, называемого *сопоставлением* (matching). Вот в чем его идея. Предположим, у вас есть непрерывная переменная *X*, которую вы хотите контролировать. Вы можете сопоставить каждый подвергшийся воздействию объект с другим объектом, наиболее близким к значению *X*, но не подвергавшимся воздействию. Затем вы можете вычислить разницу средних значений (или запустить регрессию) для этого сопоставленного набора данных, чтобы оценить влияние *T* на *Y*. Это называется *сопоставлением с ближайшими соседями* (nearest neighbor matching).

Итак, в нашем примере с голосованием в конгрессе вы должны начать с сопоставления каждого конгрессмена-демократа с конгрессменом-республиканцем с наиболее близким показателем NPAT. Затем в этой сопоставленной выборке вы должны вычислить разницу в среднем балле ACU между подобранными парами республиканцев и демократов, что является еще одним способом оценки взаимосвязи между рейтингом ACU и политической партией с учетом балла NPAT.

Если бы у вас было несколько переменных, которые нужно контролировать, вам пришлось бы определить некую суммарную меру того, насколько похожи любые два наблюдения по этим переменным. Для этого существует множество стратегий. Некоторые из них основаны на весьма запутанных вычислениях, из-за чего бывает сложно воспринимать суть этих вычислений. Вы должны стараться сохранять ясность восприятия и не позволять деревьям заслонять лес.

Сопоставление имеет некоторые преимущества перед регрессией как метод контроля искажающих факторов. Одной из приятных особенностей сопоставления является гибкость в том, как управляющая переменная может влиять на интересующий результат. Например, в то время как регрессия предполагает, что связь линейна, сопоставление не делает такого предположения. Сопоставление также имеет недостатки по сравнению с регрессией. Одним из недостатков является меньшая точность, чем у регрессии, поскольку вы используете меньше информации. Еще одним недостатком является то, что оценки соответствия могут быть необъективными, поскольку лучшее совпадение для объекта, подвергшегося воздействию, будет иметь завышенное значение *X*, если, например, *X* положительно коррелирует с *T*. Существуют статистические решения этой проблемы, тоже довольно запутанные и непривычные.

Сопоставление, как и регрессия, является хорошим статистическим методом контроля. Мы не возражаем против него. Мы обеспокоены тем, что аналитикам иногда нравится представлять какой-то очень продвинутый алгоритм сопоставления, а затем говорить что-то вроде «Сопоставление выполняет экспериментальное сравнение объектов, которые различаются по воздействию, но в остальном одинаковы». Подобные утверждения – попытка ослепить вас наукой. Сопоставление – это всего лишь инструмент контроля. Оно не создает более качественное экспериментальное сравнение, чем регрессия, включающая контрольные переменные. Другими словами, оно контролирует переменные, которые наблюдались и сопоставлялись, – не более того. Ваш компьютер, каким бы сложным ни был его статистический алгоритм, не сможет сделать

ненаблюдаемое наблюдаемым. Потому что это было бы волшебство. А в статистике нет места волшебству.

Подведение итогов

Контроль – это способ учесть искажающие факторы и получить более точную и менее предвзятую оценку причинно-следственных связей. Существует много разных способов контроля, но все они, по сути, пытаются сделать одно и то же – получить более достоверные оценки путем сравнения объектов до и после воздействия с аналогичными значениями других наблюдаемых ковариат до воздействия.

Хотя учет искажающих факторов является ценным инструментом, это не панацея. В наиболее интересных случаях по-прежнему будут существовать ненаблюдаемые факторы, которые мы не можем контролировать, обратная причинно-следственная связь или переменные, которые являются частично искажающими факторами, а частично – механизмом воздействия. Таким образом, даже если исследователи учли множество потенциальных искажающих факторов, нам все равно следует беспокоиться о смещенных оценках.

Если контроль предоставляет собой недостаточно убедительную стратегию оценки причинно-следственных связей, то как мы можем сделать ее более убедительной? Один из способов (возможно, единственный) обеспечить объективность оценок – это самостоятельно рандомизировать воздействие. Поэтому следующая глава посвящена так называемому золотому стандарту причинно-следственных связей – рандомизированному эксперименту.

Ключевые термины

- **Контроль:** использование статистического метода для поиска корреляции между двумя переменными, сохраняя значения других переменных постоянными.
- **Фиктивная переменная:** переменная, которая указывает, имеет ли данный объект какое-либо конкретное свойство; принимает значение 1, если объект имеет такое свойство, и 0 в противном случае.
- **Зависимая переменная (результат):** переменная, отражающая свойство мира, которое вы пытаетесь понять или объяснить с помощью регрессии.
- **Переменная воздействия:** переменная, отражающая свойство мира, влияние которого на зависимую переменную вы пытаетесь оценить.
- **Переменная контроля:** переменная, которую вы включаете в свой статистический анализ, пытаясь уменьшить погрешность в оценке причинного эффекта.
- **Смещение неучтенных переменных:** смещение, возникающее из-за вашей неспособности учесть некоторые искажающие факторы при попытке оценить причинный эффект.
- **Локальный средний эффект воздействия (LATE):** средний эффект воздействия для некоторой конкретной выборки из генеральной совокупности.

Упражнения

10.1. Загрузите файл HouseElectionsSpending2018.csv и связанный с ним файл README.txt, описывающий переменные в этом наборе данных, по адресу *http://www.press.princeton.edu/thinking-clearly*.

a) Постройте регрессию доли голосов действующего президента (зависимая переменная) как по расходам действующего политика, так и по расходам претендента.

i) Обратите внимание: если расходы претендентов положительно коррелируют с более высокой долей голосов за претендентов, они должны отрицательно коррелировать с долей голосов за действующего политика. В свете этого как нам следует интерпретировать найденные коэффициенты, связанные с вашими независимыми переменными?

ii) Отличаются ли результаты, которые вы получили сейчас, от тех, которые вы получили раньше, когда строили отдельные регрессии доли голосов за действующего политика и за претендентов в зависимости от расходов претендентов в главе 9? Объясните свой ответ.

b) Добавьте в свою регрессию элементы контроля, чтобы попытаться получить более надежные оценки эффекта расходов на кампанию. Как вы, возможно, знаете, 2018 г. был удачным для демократов на выборах в Палату представителей.

i) Является ли общая хорошая результативность демократов в 2018 г. потенциальной помехой для вашей регрессии?

ii) Создайте новую переменную, указывающую, является ли действующий президент республиканцем, – назовите ее «Республиканец». Она должна принимать значение 1, если действующий политик является республиканцем, и значение 0, если он является демократом.

iii) Повторно выполните регрессию, но включите эту переменную в качестве контроля.

iv) Интерпретируйте расчетный коэффициент, связанный с вашей новой переменной «Республиканец».

v) Изменяет ли включение этой контрольной переменной значимым образом ваши расчетные процентные коэффициенты (т. е. коэффициенты расходов действующих политиков и претендентов)? Как вы думаете почему?

c) Теперь добавьте контрольную переменную для доли голосов, которую партия действующего президента получила в этом округе на президентских выборах 2016 г.

i) Какую проблему может вызвать добавление этой переменной контроля?

ii) Интерпретируйте расчетный коэффициент, связанный с этой переменной.

iii) Изменяет ли добавление этой контрольной переменной значимым образом расчетные коэффициенты перед интересующими нас переменными (т. е. коэффициенты при переменных расходов действующих политиков и претендентов)? Объясните свой ответ.

10.2. Составьте таблицу регрессии, в которой показаны результаты каждой регрессии из упражнения 1, а также количество наблюдений и r^2.

10.3. В главе 2 мы обсуждали исследование, обнаружившее корреляцию между посещением углубленных курсов по математике в средней школе и успешным окончанием колледжа, которую исследователи представили как свидетельство причинно-следственной связи. Конечно, нас может беспокоить то, что студенты, которые посещают углубленные курсы по математике, по ряду факторов отличаются от тех, кто этого не делает, поэтому авторы исследования строят регрессии, которые учитывают пол, социально-экономический статус, расу, результаты тестов на когнитивные способности и оценки по языку и математике.

   a) Уменьшают ли эти контрольные переменные ваши опасения по поводу потенциальных факторов, искажающих результат?

   b) Даже после учета этих фоновых переменных назовите потенциально упущенный искажающий фактор, который вас беспокоит. Каково вероятное направление предвзятости, связанной с этим потенциальным искажающим фактором?

Дополнительное чтение и ссылки

Для получения более подробной информации об учете искажающих факторов, а также об экспериментах, инструментальных переменных, так называемой разнице в различиях и разрыве регрессии (темы, которые мы рассмотрим в следующих трех главах) мы рекомендуем книгу:

Joshua Angrist and Jorg-Steffen Pischke. 2014. *Mastering 'Metrics*. Princeton University Press.

Для получения дополнительной информации о политической поляризации, включая подробности об усилении поляризации в Конгрессе США за последние семь десятилетий или около того, мы рекомендуем монографию:

Nolan McCarty. 2019. *Polarization: What Everyone Needs to Know.* Oxford University Press.

Дополнительную информацию о LATE и ATE, включая защиту достоверных оценок LATE, см. в статье:

Guido W. Imbens. *Better LATE than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009).* Journal of Economic Literature 48 (2): 399–423.

Исследование связи между социальными сетями и субъективным благополучием:

Hunt Allcott, Luca Braghieri, Sarah Eichmeyer, and Matthew Gentzkow. 2020. *The Welfare Effects of Social Media.* American Economic Review 110 (3): 629–76.

Глава 11

Рандомизированные эксперименты

О чем эта глава

- Рандомизация воздействия может обеспечить условия для объективной оценки причинно-следственных связей.
- Все инструменты статистических выводов и проверки гипотез работают в экспериментальных условиях и позволяют определить разницу между реальными эффектами и шумом.
- Даже при проведении рандомизированного эксперимента могут возникнуть многочисленные осложнения, которые необходимо предвидеть и тщательно к ним готовиться.
- Если испытуемые не выполняют поставленные перед ними экспериментальные задачи, важно провести сравнения на основе рандомизированного распределения.
- Даже когда исследователи не могут реализовать идеальный эксперимент, иногда они обнаруживают случаи, когда интересующий их показатель воздействия или отбора был рандомизирован естественным образом для иных целей. Такие «естественные эксперименты» часто являются плодотворной, хотя и непреднамеренной возможностью ответить на важные причинно-следственные вопросы.

Введение

Нам нравится регрессия, поэтому мы неплохо провели время в главе 10. Но мы понимаем, вас могло разочаровать утверждение о том, что вы никогда не сможете получить объективные оценки причинно-следственных связей, просто контролируя искажающие факторы. Мы постараемся наверстать упущенное в следующих трех главах и покажем вам, что есть более эффективные способы убедиться в существовании причинно-следственной связи. Эти способы называются *планом исследования* (research design) или *планом эксперимента* (experiment design). Планирование эксперимента для изучения причинно-следственных связей часто требует немалой изобретательности и изворотливости. Это делает планирование экспериментов одной из самых увлекательных тем, которые вам необходимо освоить, чтобы критически относиться к данным.

В этой главе мы рассматриваем план, называемый *рандомизированным экспериментом* (randomized experiment). Рандомизированные эксперименты прекрасны, потому что если вы можете рандомизировать воздействие, то у вас есть шанс полностью устранить искажающие факторы. Таким образом, вы можете устранить смещение в своих оценках. Аналогия с рандомизированными экспериментами также помогает объяснить, почему подходы, которые мы рассмотрим в следующих двух главах, называются планами. Планируя рандомизированный эксперимент, вы буквально сами определяете способ назначения воздействия.

В главах 12 и 13 мы обратимся к другим планам исследований, которые менее «запланированы». В частности, эти планы представляют собой способы узнать о причинно-следственной связи на основе данных, которые мы наблюдаем в независимом от нас окружающем мире, т. е. когда мир, а не экспериментатор решает, какое воздействие будет назначено конкретному объекту. Но, прежде чем мы перейдем к этой теме, нужно обсудить рандомизированные эксперименты и разобраться, как они работают.

Грудное вскармливание

На момент написания этой книги в развитых странах мира существовала практически святая уверенность в том, что младенцев следует кормить грудью. Возьмем, например, официальное заявление Всемирной организации здравоохранения: «Взрослые, которых в детстве кормили грудью, часто имеют более низкое кровяное давление и уровень холестерина, а также более низкие показатели избыточного веса, ожирения и диабета 2-го типа. Есть доказательства того, что люди, которых кормили грудью, лучше справляются с тестами на интеллект». Аналогичным образом в 2011 г. главный санитарный врач США выступил с призывом к поддержке грудного вскармливания, которое, по его словам, является «одной из наиболее эффективных профилактических мер, применяемых ответственными матерями для защиты здоровья своего ребенка». В сопроводительном отчете утверждается, что грудное вскармливание предотвращает множество детских болезней, включая ушные инфекции, экзему, диарею, респираторные заболевания, астму, ожирение, диабет 2-го типа, лейкемию и синдром внезапной детской смерти. Действительно, в огромном количестве научных публикаций задокументирована положительная корреляция между грудным вскармливанием и хорошими показателями здоровья детей.

Но, прежде чем делать поспешные выводы о причинно-следственной пользе грудного вскармливания, учтите один факт. В развивающихся странах грудное вскармливание, похоже, коррелирует с худшими, а не лучшими последствиями для здоровья детей. В таких разных странах, как Гана, Кения, Египет, Бразилия, Перу, Боливия и Таиланд, было обнаружено, что грудное вскармливание коррелирует с недоеданием и снижением роста и веса.

В чем тут дело? Может ли быть так, что грудное вскармливание полезно для детей в промышленно развитых странах и вредит детям в развивающихся странах? Возможно, но давайте убедимся, что мы критически относимся к фактам. Вы уже знаете, что корреляция не обязательно подразумевает причинно-следственную связь. И в данном случае сравнивать мам, кормящих и не кормящих грудью, скорее всего, не корректно.

Сначала рассмотрим положение дел в развивающихся странах, где грудное вскармливание отрицательно коррелирует с физическим состоянием детей. Одно из объяснений заключается в том, что грудное вскармливание само по себе вызывает такие неблагоприятные последствия. Также возможно, что какой-то искажающий фактор, такой как бедность, является причиной как грудного вскармливания, так и недоедания. Грудное вскармливание требует больших затрат времени матери, но не требует больших денег. Искусственное детское питание, напротив, стоит больших денег, но отнимает меньше времени. Было бы логично предположить, что в экономически неблагополучных семьях матери с большей вероятностью будут кормить своих детей грудью. А дети из тех же экономически неблагополучных семей могут быть более склонны к проблемам со здоровьем по причинам, совершенно не связанным с грудным вскармливанием. Также не исключена и обратная причинно-следственная связь. Возможно, плохое состояние здоровья младенца приводит к тому, что мать с большей вероятностью будет кормить грудью.

Судя по всему, искажающие факторы и обратная причинно-следственная связь действительно присущи развивающимся странам. Исследование 1997 г., опубликованное в Международном журнале эпидемиологии, отслеживало состояние здоровья 238 малышей в деревне в Перу. Данные исследования включали информацию о росте ребенка, грудном вскармливании, приеме прикорма и диарее. Исследование выявило отрицательную корреляцию между грудным вскармливанием и ростом: дети, находившиеся на грудном вскармливании, в среднем были меньше, что позволяет предположить худшее здоровье. Эта связь была самой сильной среди тех детей, которые получали меньше всего прикорма и были наиболее болезненными. Оказывается, поскольку широко распространено мнение, что грудное вскармливание полезно для здоровья, матери, чьи дети болели или не имели доступа к прикорму, позже отлучали своих детей от груди. Следовательно, дети, которые *уже* были больны и недоедали, с большей вероятностью получали грудное вскармливание. Таким образом, делается вывод в исследовании, что не грудное вскармливание является причиной того, что дети в развивающихся странах плохо растут. Скорее, дети, которые плохо растут из-за того, что они больны и недоедают, с большей вероятностью будут находиться на грудном вскармливании.

Теперь перейдем к развитым странам, где родители завалены информацией о том, что грудное вскармливание полезно для их детей. Помните, что грудное вскармливание снижает риск сердечно-сосудистых заболеваний, астмы, ожирения, лейкемии, СВДС, ушных инфекций и множества других заболеваний. К сожалению, доказательства, лежащие в основе этой общеизвестной мудрости, снова не выдерживают тщательной проверки. Ведь можно назвать множество причин, по которым сравнение детей, находящихся на грудном вскармливании, с детьми, которых не кормят грудью, не может быть корректным. Например, когда авторитетные организации опубликуют отчеты об эффективности грудного вскармливания, мы ожидаем, что богатые, эрудированные и образованные матери с большей вероятностью услышат эту новость и последуют рекомендациям. Но их дети в большинстве случаев изначально будут иметь лучшие показатели здоровья.

Кормить грудью или нет – это очень важное решение, и на него влияет так много разных факторов, что в мире невозможно найти сравнение яблок

с яблоками. Однако, возможно, мы могли бы провести собственное корректное сравнение с помощью рандомизированного эксперимента. Именно это попыталась сделать группа исследователей из Беларуси.

Стратегия команды заключалась в проведении рандомизированного эксперимента. Очевидно, что как по этическим, так и по практическим причинам они не могли заставить матерей кормить или не кормить грудью просто ради своего исследования. Но они могли бы повысить вероятность того, что случайно выбранная группа матерей решит кормить грудью посредством случайно назначенного поощрения. Для этого в некоторых случайно выбранных больницах исследователи внедрили программу по поощрению и облегчению грудного вскармливания. В других случайно выбранных больницах эту программу не внедряли. Во всех больницах они подробно регистрировали способы кормления детей и отслеживали различные показатели здоровья детей на протяжении всего эксперимента. И действительно, матери в больницах, где применялась программа поддержки грудного вскармливания, гораздо чаще кормили грудью своих новорожденных.

Несмотря на заявления Всемирной организации здравоохранения, главного санитарного врача, Американской академии педиатрии и представителей родительского сообщества в целом, ученые нашли на удивление скудные доказательства того, что грудное вскармливание приносит большую пользу. У младенцев из больниц, принимавших участие в программе, вероятность развития экземы и желудочно-кишечных инфекций была немного ниже, но исследователи получили нулевые результаты (т. е. отсутствие статистически значимых доказательств влияния грудного вскармливания) по гораздо большему количеству исходов. На втором этапе исследования, проведенном, когда детям было от шести до семи лет, исследователи изучали, демонстрируют ли дети, находившиеся на грудном вскармливании, лучшие результаты по каким-либо наблюдаемым физическим, психологическим или когнитивным показателям. Они не обнаружили никаких доказательств того, что грудное вскармливание приносит пользу с точки зрения меньшего риска экземы, аллергии, астмы, ожирения, эмоциональных проблем, проблем с поведением, гиперактивности или проблем со сверстниками. На самом деле данные содержали некоторые ограниченные доказательства отрицательной связи между грудным вскармливанием и этими результатами. Единственное доказательство в поддержку грудного вскармливания, которое удалось обнаружить, заключалось в том, что дети из больниц, осуществлявших программу грудного вскармливания, показали немного лучшие результаты по тестам IQ. Но, размышляя об этом открытии, не забывайте об уроках завышения значимости, которые мы обсуждали в главе 7. Если вы рассмотрите большое количество различных показателей, вы, скорее всего, обнаружите хотя бы один статистически значимый результат просто из-за шума.

В целом мы считаем, что экспериментальные данные не подтверждают однозначную пользу грудного вскармливания. Хотя существуют сильные корреляции между грудным вскармливанием и последствиями для здоровья в различных условиях, а также некоторые разумные аргументы в отношении биологических механизмов, посредством которых может работать грудное вскармливание, наиболее достоверные данные свидетельствуют о том, что средний эффект грудного вскармливания невелик. Без проведения рандо-

мизированных экспериментов было бы легко переоценить или недооценить пользу грудного вскармливания.

Рандомизация и причинно-следственный вывод

Что делает рандомизированные эксперименты таким мощным инструментом для изучения причинно-следственных связей? Чтобы найти ответ, давайте вернемся к обсуждению потенциальных исходов и пресловутых наклеек Body Vibes.

Предположим, мы хотим узнать, как какое-то воздействие, скажем Body Vibes, повлияет на какой-то наблюдаемый результат, скажем на здоровье кожи. В целом оценить эффект Body Vibes сложно из-за всех проблем, упомянутых в предыдущих главах. Мы хотим знать, насколько будет отличаться кожа человека, который использует Body Vibes, по сравнению с человеком, который их не использует. К сожалению, для каждого конкретного человека мы можем наблюдать только один из этих потенциальных результатов. Например, если человек использует Body Vibes, мы можем наблюдать за состоянием его кожи в этой ситуации, но мы не знаем, какой была бы его кожа, если бы он не использовал Body Vibes.

Если мы просто сравним среднее состояние кожи людей, которые используют и не используют Body Vibes, это будет некорректное сравнение. Существует множество факторов, предполагающих, что эта разница в средних значениях не является объективной оценкой среднего эффекта воздействия. Например, возможно, те, кто использует Body Vibes, просто больше заботятся о своей коже, а также используют больше увлажняющего и солнцезащитного крема. Или, может быть, смещение направлено в другую сторону. Возможно, у людей, использующих Body Vibes, плохая кожа, они перепробовали все остальные средства и впали в отчаяние. В любом случае, из-за таких факторов мы не можем получить объективную оценку эффекта Body Vibes, просто сравнивая среднее состояние кожи людей, которые их используют и не используют.

Один из способов, возможно, даже лучший способ избавиться от этой предвзятости и быть уверенным, что мы проводим корректное сравнение яблок с яблоками, – это рандомизировать воздействие. Наше сравнение тех, кто использует и не использует Body Vibes, является смещенным, поскольку эти группы, вероятно, имеют базовые различия. То есть в среднем у них, скорее всего, будет разное состояние кожи еще до начала эксперимента с использованием Body Vibes. Однако если мы случайным образом выберем группу людей, которые будут использовать Body Vibes, то эти две группы, как ожидается, будут одинаковыми с точки зрения исходного состояния кожи и всех других исходных параметров. То есть не будет искажающих факторов. Почему это так?

Если запланированное воздействие определяется подбрасыванием монеты, генератором случайных чисел на вашем компьютере или другим случайным процессом, то единственное, что отличает людей в группах воздействия и контроля, – это чистая случайность. Нет никаких причин, по которым люди в группе, испытавшей воздействие, могут быть *систематически* выше, умнее, богаче, более мотивированы или иметь лучшую кожу, чем люди в контрольной группе.

Что касается обозначения наших потенциальных исходов, предположим, что мы случайным образом выбираем одну группу (группа воздействия T), которая получает Body Vibes, а другая группа (контрольная группа U) – нет.

Мы наблюдаем среднее состояние кожи в группе, получавшей Body Vibes, $\overline{Y}_{1т}$, а также среднее состояние кожи в группе, не получавшей Body Vibes, $\overline{Y}_{0и}$. Таким образом, если мы сравним средний уровень здоровья кожи в двух группах, то получим разницу средних значений:

$$\overline{Y}_{1т} - \overline{Y}_{0и}.$$

Но благодаря рандомизации не существует систематических различий между любой из этих групп и генеральной совокупностью в целом. Таким образом, среднее состояние кожи группы пациентов, получивших Body Vibes, представляет собой объективную оценку среднего состояния кожи всего населения в гипотетическом мире, где все носят Body Vibes:

$$\overline{Y}_{1т} = \overline{Y}_1 + \text{Шум}_1.$$

И аналогично для группы, не подвергнутой воздействию:

$$\overline{Y}_{0и} = \overline{Y}_0 + \text{Шум}_0.$$

Следовательно, наблюдаемая разница в средних значениях представляет собой несмещенную оценку среднего эффекта воздействия.

$$\overbrace{\overline{Y}_{1т} - \overline{Y}_{0и}}^{\substack{\text{Наблюдаемая разница} \\ \text{в средних значениях}}} = \overbrace{\overline{Y}_1 - \overline{Y}_0}^{\text{ATE}} + \text{шум},$$

где шум представляет собой просто разность двух шумовых компонентов в уравнениях выше.

Как и в примерах из предыдущих глав, шум может возникать из-за изменчивости выборки. Возможно, нас интересует более широкая совокупность, но в нашем эксперименте мы случайно получили необычную выборку испытуемых. Шум также может быть связан с ошибкой измерения. Когда мы проводим эксперимент, шум также возникает из-за случайного назначения экспериментального воздействия испытуемым. Даже для одной и той же выборки испытуемых разные рандомизации могли дать разные оценки, и это также вносит свой вклад в шум.

Из-за шума в любом мелкомасштабном эксперименте будут некоторые различия в средних потенциальных результатах между экспериментальной и контрольной группой – просто в силу случая. Но эти различия не будут систематическими: если бы нам довелось провести много итераций эксперимента, мы бы не ожидали, что увидим одну и ту же картину различий, повторяющуюся снова и снова. Именно это мы подразумеваем под словом «ожидание».

Вспомните наше любимое уравнение:

$$\text{Оценка} = \text{Оцениваемая величина} + \text{Смещение} + \text{Шум}.$$

Рандомизация гарантирует, что смещение равно нулю. Таким образом, шум является единственной причиной того, что оценка, которую мы получаем при сравнении среднего результата в группах, испытавших и не испытавших воз-

действие, в правильно рандомизированном эксперименте отличается от истинного причинного эффекта (т. е. оценки).

По мере увеличения количества испытуемых в каждом конкретном эксперименте две группы будут становиться все более и более изначально похожими. То есть по мере увеличения размера выборки шум становится меньше.

Рандомизация дает вам возможность сравнить группы эксперимента и контроля, генерируя несмещенные оценки. Выборки большого размера дают очень небольшое количество шума, что позволяет получить точные оценки. Таким образом, рандомизация в сочетании с большим размером выборки дает вам возможность сопоставления групп, генерируя оценки, которые с большой вероятностью будут близки к истинной оценке.

Если хорошо вдуматься, становится ясно, что рандомизация – это, по сути, единственный способ гарантировать объективную оценку причинно-следственных связей. Предположим, вы попытались провести эксперимент, но вместо того, чтобы случайным образом распределить своих подопытных по группам эксперимента и контроля, вы попытались тщательно разделить группы так, чтобы они были максимально похожи друг на друга. Поскольку вы не можете наблюдать и количественно оценить все значимые характеристики ваших испытуемых, придется сделать некоторые допущения. Может быть, у вас это получится очень хорошо, а может быть, нет. Что, если ваши собственные подсознательные предубеждения заставят вас поместить несколько разных людей в группы, подвергнутые и не подвергнутые воздействию – возможно, потому что вы подсознательно стремитесь получить максимальный эффект? У вас не будет возможности узнать, действительно ли вы хорошо поработали. Поэтому зачем рисковать? Почему бы в самом деле не подбросить монетку и не назначить воздействие случайным образом? Если сейчас этот момент кажется очевидным, то в прошлом он не был очевиден для многих умных людей. Только после работы Р. А. Фишера в 1920-х гг. ученые поняли ценность рандомизации.

Есть один хороший способ обосновать необходимость попыток сделать группы как можно более похожими по наблюдаемым характеристикам. Как мы знаем из нашего любимого уравнения, оценки могут отличаться от истинного эффекта воздействия по двум причинам: из-за смещения и шума. Рандомизация устраняет смещение. Но шума все равно может быть много, особенно если размер выборки невелик или участники эксперимента сильно отличаются друг от друга по характеристикам, которые имеют значение для результата. То есть в любой итерации эксперимента группы, подвергнутые и не подвергнутые воздействию, в действительности могут выглядеть совершенно по-разному, даже если в ожиданиях они одинаковы.

Один из способов сгладить эту проблему – начать с группировки людей на основе их наблюдаемого сходства. Затем вы можете случайным образом выбрать в этих группах людей, испытавших и не испытавших воздействие. Этот прием называется *разбиением на блоки*, или *стратификацией*. Например, вас может беспокоить то, что мужчины и женщины в среднем имеют очень разные уровни здоровья кожи. Ваш эксперимент будет очень зашумленным, если в экспериментальной группе случайно окажутся преимущественно мужчины, а в контрольной группе – женщины, или наоборот. Этого не будет в ожидании (т. е. если вы проведете эксперимент бесконечное количество раз, средняя доля

мужчин и женщин будет одинаковой в обеих группах). Но это может произойти в любой итерации вашей рандомизации. Чтобы устранить этот источник шума, вы можете начать с разделения популяции на две части по биологическому полу. Затем вы случайным образом распределяете половину мужчин в экспериментальную группу, а другую половину – в контрольную и то же самое делаете с женщинами. Воздействие останется полностью рандомизированным, поэтому вы по-прежнему будете получать объективные оценки. Но при этом вы уменьшаете шум, гарантируя, что группы схожи по критерию биологического пола не только в ожиданиях, но и в реальности.

Развивая эту логику, аналитик мог бы выделить различные страты субъектов со схожими характеристиками до воздействия и провести рандомизацию внутри этих страт. Самым радикальным вариантом этого подхода является *метод подбора пар*, при котором аналитик определяет пары людей, которые, по его мнению, наиболее похожи друг на друга, и в каждой паре случайным образом выбирает одного участника для воздействия, а второго оставляет для контроля. Это может быть отличным способом повысить точность оценок. Но необходимо следить за тем, чтобы воздействие назначалось случайным образом внутри каждой пары.

Оценка и вывод в экспериментах

В главе 6 мы обсуждали выводы о взаимоотношениях на основе статистических критериев. Все эти соображения применимы и в случае экспериментальной оценки. В самом простом сценарии мы можем проанализировать результаты эксперимента, рассчитав разницу в средних значениях, т. е. сравнивая средний эффект в экспериментальной и контрольной группе. Фактически, как мы видели в главе 5, если построить регрессию результата по бинарному показателю статуса воздействия, коэффициент регрессии, связанный с переменной воздействия, будет просто разницей средних значений. А поскольку коэффициенты регрессии и различия в средних значениях представляют собой всего лишь количественные отношения, мы можем применить все статистические инструменты главы 6 и к экспериментам.

Стандартные ошибки

Предположим, мы проводим рандомизированный эксперимент и оцениваем средний эффект воздействия, сравнивая средний результат для объектов, испытавших воздействие, со средним результатом для объектов, не испытавших такового. Эта оценка является несмещенной. Но она вполне может быть неточной (т. е. может содержать много шума).

Нам, конечно, хотелось бы знать, насколько близки наши оценки к истинному эффекту (т. е. оцениваемой величине). Мы можем найти стандартную ошибку, связанную с нашей экспериментальной оценкой, точно так же как мы находили стандартную ошибку результатов опроса и коэффициентов регрессии в главе 6. Стандартная ошибка дает нам представление о том, насколько в среднем оценка будет далека от истины вследствие шума, если мы повторим наш эксперимент бесконечное количество раз, всегда используя одну и ту же процедуру для оценки эффекта воздействия. Подобно нашему примеру с ре-

зультатами опроса, истинная стандартная ошибка зависит от величин, которые не наблюдаемы, но существуют различные приближения, которые аналитики используют для оценки стандартной ошибки.

Вам не нужно запоминать формулы расчета стандартных ошибок; вы всегда можете найти их или просто поручить компьютеру рассчитать их за вас. Тем не менее полезно знать, как различные особенности экспериментов влияют на количество шума. Предположим, мы проводим эксперимент с N объектами, из которых m подвергаются воздействию, а ($N - m$) – нет. При прочих равных условиях чем больше N, тем меньше шум и, следовательно, меньше стандартная ошибка. Это должно быть интуитивно понятно. Когда размер выборки больше, экспериментальная и контрольная группа будут более похожи друг на друга по прочим характеристикам, что снижает шум и приближает наши оценки к истинному причинному эффекту.

А что насчет m? Предположим, в нашем исследовании участвуют 500 человек. Сколько из них следует поместить в экспериментальную группу и сколько – в контрольную? Очевидно, что мы не можем поместить их всех в одну из групп, потому что тогда не сможем провести сравнение (помните, что корреляция требует вариаций). Кроме того, нам нужно, чтобы в обеих группах не было слишком мало участников. Если одна из двух групп очень мала, то наши оценки будут неточными, поскольку средний результат для любой небольшой группы будет весьма чувствителен к несистематическим отклонениям характеристик нескольких участников. Обычно наиболее точные оценки можно получить, когда размеры экспериментальной и контрольной группы примерно равны.

При этом часто встречаются случаи, когда оптимальный план эксперимента может содержать разное количество испытуемых в каждой группе. Предположим, у вас есть 100 000 потенциальных испытуемых. Ваших ресурсов достаточно для того, чтобы поместить в экспериментальную группу только 100 человек, но включить больше участников в контрольную группу не составит труда. Вы можете случайным образом распределить 100 человек в экспериментальную группу, а всех остальных отправить в контрольную. Ваши оценки не будут такими точными, как если бы в каждой группе было по 50 000 человек, но они будут гораздо точнее, чем в эксперименте со 100 участниками в каждой группе.

Последний фактор, влияющий на зашумленность экспериментальных оценок, – это то, насколько изменчивы результаты в обеих группах. Если мы изучаем результаты с небольшой дисперсией в рамках каждого условия воздействия, оценки будут более точными, чем если бы мы изучали результаты с большей дисперсией. Это связано с тем, что если результат не сильно различается в зависимости от характеристик, не связанных с воздействием, то для шума остается очень мало места – мы получим одинаковые результаты для каждой группы на всех итерациях эксперимента. Вот почему, например, врачи и государственные регулирующие органы часто имеют точные оценки влияния некоторых препаратов на кровяное давление (результат с относительно низкой дисперсией), но неточные оценки влияния тех же препаратов на сердечные приступы (результат с высокой дисперсией). Конечно, иногда мы не можем на это повлиять и вынуждены довольствоваться тем, что есть. Но в других случаях можно подобрать показатели или методы измерения этих показателей, уменьшающие шум.

Проверка гипотезы

Мы также можем применить инструменты проверки гипотез, которые изучили в главе 6, к экспериментальным результатам для оценки статистической значимости. Например, деление оценки на стандартную ошибку дает значение, называемое *t*-критерием, которое можно использовать для оценки *p*-значения. А поскольку мы часто проверяем гипотезы с помощью экспериментальных результатов, необходимо постоянно помнить о рисках завышения значимости, занижения отчетности и возврата к среднему значению. В главе 7 мы говорили, что аналитики могут снизить эти риски, заранее точно сформулировав интересующие вопросы, для решения которых предназначен эксперимент, заранее определив гипотезы, которые они планируют проверить, и регрессии, которые они планируют построить (поэтому они не могут просто заняться поиском важного открытия в статистических данных), и сообщая о результатах независимо от того, совпадают ли они с ожиданиями.

Все сказанное в предыдущих главах относится и к интерпретации результатов экспериментов. Если аналитики не раскрывают шаги, которые они предприняли, чтобы избежать завышения значимости эффектов и занижения научной отчетности, нам следует скептически относиться к их выводам. И чем удивительнее результаты, тем более скептически нам следует к ним относиться. Помните, что подтверждение экстрасенсорного восприятия появилось в результате экспериментального исследования! Или, что более серьезно, подумайте еще раз об эксперименте с грудным вскармливанием, с которого мы начали эту главу. Это исследование имело много достоинств. Но одна потенциальная проблема заключается в том, что, поскольку авторы исследования собрали информацию о большом количестве эффектов, мы не видим доказательств влияния грудного вскармливания на экзему, аллергию, астму, ожирение, эмоциональные проблемы, проблемы поведения, гиперактивность или проблемы общения со сверстниками, но зато видим влияние на IQ. У нас есть все основания полагать, что кажущееся влияние на IQ возникло случайно.

Проблемы, возникающие при экспериментах

На практике редко что-то работает так прекрасно, как в наших идеализированных примерах. Теоретически вы можете разработать рандомизированный эксперимент и оценить средний эффект воздействия, просто сравнивая средние значения. Однако на практике возникают проблемы, которые делают анализ и интерпретацию не такими простыми. Давайте обсудим некоторые проблемы и способы, с помощью которых вдумчивые аналитики могут с ними справиться. Эти способы пригодятся вам и за пределами экспериментов, поскольку аналогичные проблемы могут возникнуть практически при любой стратегии оценки причинно-следственных связей.

Несоблюдение условий и инструментальные переменные

Одной из распространенных проблем является несоблюдение испытуемыми условий эксперимента. Для краткости будем называть его дальше просто *несоблюдением* (noncompliance). Например, в медицинских исследованиях до-

вольно часто некоторые испытуемые просто прекращают прием лекарств. В эксперименте по грудному вскармливанию также наблюдались случаи несоблюдения. Напомним, поскольку неэтично принуждать мать кормить грудью или отказаться от грудного вскармливания, исследователи случайным образом распределили матерей на группы, где они получали более или менее ощутимое поощрение к грудному вскармливанию. Подобные схемы поощрения позволяют исследователям экспериментально изучать множество тем, которые в противном случае были бы недоступны по организационным или этическим причинам. Но такие исследования неизбежно влекут за собой дополнительные осложнения, возникающие из-за несоблюдения условий эксперимента, поскольку некоторые матери получали поощрение, но все равно отказывались от кормления грудью, а некоторые матери, которых не поощряли, продолжали кормить ребенка грудью.

Предположим, мы разработали рандомизированный эксперимент, чтобы оценить влияние наклеек Body Vibes на здоровье кожи. Мы случайным образом отобрали некоторых людей в экспериментальную группу: дали им Body Vibes и пытаемся убедить их носить наклейки во имя науки. Мы также случайным образом отобрали некоторых людей в контрольную группу: им не дают Body Vibes и говорят вести нормальную жизнь. В итоге, несмотря на все наши усилия, некоторые из участников экспериментальной группы забывают или просто отказываются носить свои Body Vibes. И что более удивительно, некоторые самые доверчивые участники контрольной группы узнали больше о Body Vibes, купили их на свои деньги и начали носить! Что же нам делать?

Одна из очевидных идей заключается в том, чтобы просто сравнить людей, которые носили и не носили Body Vibes, игнорируя, к какой группе изначально относился нарушитель. Но этот подход не сработает. Он возвращает нас к проблеме, которую мы пытались решить с помощью рандомизации эксперимента. Люди, самовольно решившие носить или не носить Body Vibes, вероятно, отличаются друг от друга, поэтому сравнение этих двух групп не является корректным.

Другая идея состоит в исключении из эксперимента субъектов, которые не соблюдают условия. Другими словами, мы могли бы исключить из нашего анализа людей, которые были включены в экспериментальную группу, но не носили Body Vibes, и людей, которые были отнесены к контрольной группе, но носили их. После этого мы могли бы продолжать действовать как обычно, сравнивая среднее состояние здоровья кожи у остальных членов экспериментальной и контрольной группы.

К сожалению, и этот подход не решает проблему. Чтобы понять почему, подумайте о людях, которые были в экспериментальной группе, но отказались носить Body Vibes. Они могут быть особенными во многих важных отношениях – например, у них может быть здоровая кожа или они менее доверчивы. Вероятно, в контрольной группе были такие же люди, как они. Но, поскольку мы вообще не просили этих людей носить Body Vibes, мы не можем узнать, так ли это. Поэтому не можем удалить соответствующую часть контрольной группы. Значит, если мы исключим только нарушителей из экспериментальной группы, то сравнение групп уже не будет паритетным. Люди, которые не стали бы носить наклейки Body Vibes, даже если дать их бесплатно, останутся в контрольной группе, но покинут экспериментальную.

Что же мы можем сделать в случае несоблюдения требований? Кое-какой выход есть. Мы всегда можем оценить эффект от попадания в экспериментальную группу (в отличие от эффекта самого воздействия). Иногда мы называем это эффектом *намерения воздействовать* (intent-to-treat, ITT) или эффектом *редуцированной формы* (reduced-form). Мы делаем это, сравнивая результаты для людей, отнесенных к той или иной группе, независимо от того, действительно ли они соблюдают условия эксперимента. Это сравнение не даст нам объективной оценки эффекта от ношения Body Vibes. Но это даст нам объективную оценку совокупного эффекта от ношения Body Vibes и поощрения их ношения.

Бывают ситуации, когда политика или чиновника, принимающего решения, на самом деле больше заботят эффекты намерения воздействовать, чем фактические эффекты воздействия. Предположим, благотворительная организация пытается решить, следует ли ей предлагать бесплатные Body Vibes старшеклассникам с плохой кожей. Они знают, что не все, у кого есть Body Vibes, будут их носить. Более того, все, что они могут сделать с этической точки зрения, – это предложить Body Vibes; они не могут заставить кого-либо использовать их. Они проводят эксперимент, чтобы оценить преимущества бесплатных Body Vibes. Какую величину следует принять за критерий положительного эффекта, чтобы проинформировать их о том, хорошая ли это политика? Это не средний эффект от использования Body Vibes для человека, который в самом деле их использует. Это средний эффект от *предоставления* Body Vibes, независимо от того, использует ли человек их или нет, поскольку именно это на самом деле может сделать благотворительная организация. Таким образом, релевантным критерием будет эффект от намерения воздействовать. А если серьезно, во многих случаях все, что может сделать политик или организация, – это предоставить возможность; они не могут заставить людей воспользоваться ей. В любой подобной ситуации эффект от намерения воздействовать фактически может оказаться наиболее адекватным показателем.

Однако в других ситуациях нас интересует реальный эффект воздействия, а не только эффект намерения воздействовать. Предположим, например, что мы пытаемся решить, следует ли нам самим носить Body Vibes. Или, что более серьезно, предположим, что кто-то решает, стоит ли пробовать экспериментальное медицинское лечение, новый режим обучения, новую методику преподавания или новую стратегию управления, повышающую производительность. В таких случаях мы хотим знать больше, чем просто эффект от воздействия. Мы хотим знать вероятный эффект от начала воздействия – желательно как можно раньше. Так что же еще мы можем сделать с результатами наших экспериментов, несмотря на то что они страдают от проблем несоблюдения правил?

Чтобы продвинуться в этом направлении, давайте подумаем о том, как испытуемый может реагировать на экспериментальное предложение носить или не носить Body Vibes. Наша выборка состоит из четырех разных типов людей.

1. Есть *исполнители* (complier), которые будут носить Body Vibes, если им назначено воздействие, и не будут носить, если не назначено.
2. Всегда найдутся *активисты* (always-taker), которые будут носить Body Vibes независимо от того, выбраны они для воздействия или нет.
3. Есть *отрицатели* (never-taker), которые не будут носить Body Vibes независимо от того, выбраны они для воздействия или нет. (Мы оба идейные отрицатели, если речь идет о Body Vibes.)

4. И в принципе, может существовать извращенная группа *провокаторов* (defier), которые ведут себя наоборот – не будут носить Body Vibes, если они находятся в экспериментальной группе, но будут носить Body Vibes, если попадут в контрольную группу.

Очевидно, что, планируя эксперимент, мы надеемся на большое количество исполнителей. Вся идея эксперимента состоит в том, что мы хотим случайным образом назначить воздействие, а соблюдающие его условия участники готовы позволить нам это сделать.

Каждый испытуемый в эксперименте четко вписывается в одну (и только одну) из этих категорий. Однако мы не можем просто посмотреть на наших подопытных и выяснить, кто из них исполнитель, активист, отрицатель или провокатор. Почему? Предположим, мы видим, что кто-то находится в контрольной группе и не носит Body Vibes. Мы знаем, что он или исполнитель, или отрицатель. Но у нас нет возможности узнать, кто именно, потому что мы не знаем, носил бы он Body Vibes, если бы оказался в экспериментальной группе. Проблема с участниками эксперимента Body Vibes кратко представлена в табл. 11.1.

Таблица 11.1. Кто принимает участие в эксперименте Body Vibes?

	Экспериментальная группа	**Контрольная группа**
Носят Body Vibes	Исполнители и активисты	Активисты и провокаторы
Не носят Body Vibes	Отрицатели и провокаторы	Исполнители и отрицатели

Разделение людей на эти группы помогает нам ясно понять, когда мы проводим или не проводим корректное сравнение. В частности, чтобы гарантировать отсутствие путаницы, мы стараемся, чтобы в группах, которые мы сравниваем (скажем, группы, испытавшие и не испытавшие воздействие), была одинаковая доля представителей всех четырех групп.

Чтобы понять, как это помогает нам понять проблему, давайте начнем с предположения, что каждый участник либо исполнитель, либо отрицатель. Другими словами, никто из тех людей, которые могли бы покупать и носить Body Vibes самостоятельно, в нашем эксперименте не участвовал. (Мы немного расслабимся.) В табл. 11.2 показано, как выглядит наш эксперимент в мире, где есть только те, кто всегда соблюдает условия, и те, кто никогда их не соблюдает.

Таблица 11.2. Кто принимает участие в эксперименте Body Vibes, если предположить, что есть только исполнители и отрицатели?

	Экспериментальная группа	**Контрольная группа**
Носят Body Vibes	Исполнители	–
Не носят Body Vibes	Отрицатели	Исполнители и отрицатели

Теперь давайте вернемся к различным способам обращения с подопытными, которые ведут себя не в соответствии с назначением воздействия. Легко понять, почему мы не можем просто сравнивать людей, которые носили и не

носили Body Vibes, игнорируя запланированное распределение по группам. Группа, которая носит Body Vibes, состоит только из участников, испытывающих воздействие. Группа людей, которые не носят Body Vibes, представляет собой комбинацию тех, кто соблюдает правила и отправлен в контрольную группу, и тех, кто был направлен в экспериментальную группу, но никогда не носит наклейки. Таким образом, сравнение тех, кто носит Body Vibes, с теми, кто его не носит, не является корректным.

Точно так же легко понять, почему мы не можем просто исключить людей, которые явно не подчиняются нашему эксперименту. Нам не составит труда исключить тех, кто никогда не носит наклейки, из экспериментальной группы. Но мы не сможем определить, кого исключить из контрольной группы для достижения равновесия. В результате будем сравнивать тех, кто соблюдает правила эксперимента в экспериментальной группе, с комбинацией тех, кто соблюдает правила, и тех, кто никогда не носит наклейки, в контрольной группе – опять же, это не некорректное сравнение.

Похоже, что мы застряли в ситуации, когда все, что можно сделать, – это сравнить экспериментальную группу с контрольной, оценивая эффект от назначенного воздействия. Но на самом деле мы можем добиться большего. Давайте посмотрим, как это сделать.

Ключевым шагом на пути к улучшению ситуации является оценка доли участников, соблюдающих правила, в нашей выборке. Мы не знаем точно, кто именно соблюдает правила. Но в нашем упрощенном примере, в котором представлены только те, кто соблюдает правила, и те, кто никогда не соблюдает их, мы можем оценить, какую долю в выборке составляют лица, соблюдающие требования эксперимента. Мы делаем это, анализируя экспериментальную группу. Нас интересует доля исполнителей в экспериментальной группе. Благодаря случайному распределению доля исполнителей в контрольной группе должна быть такой же. Следовательно, доля исполнителей в экспериментальной группе представляет собой несмещенную оценку доли таковых во всей выборке (т. е. в экспериментальной и контрольной группах вместе взятых).

Теперь у нас есть объективные оценки как эффекта от намерения воздействовать (путем сравнения средних результатов в экспериментальной и контрольной группе), а также доли добросовестных исполнителей в выборке. Как это нам поможет?

Мы хотим знать влияние Body Vibes на некоторые результаты, допустим, на здоровье кожи. Если мы предположим, что единственный способ, которым назначенное воздействие могло повлиять на здоровье кожи, – это фактическое использование Body Vibes, то каков будет эффект от намерения воздействовать? По нашему предположению, на тех, кто никогда не носил наклейки, назначение воздействия не повлияло, а эффект назначения воздействия для тех, кто соблюдал его, является просто эффектом Body Vibes. Таким образом, ожидаемый эффект от назначенного воздействия – это средний эффект Body Vibes для аккуратных исполнителей, умноженный на долю этих лиц в выборке. Это означает, что если мы разделим эффект намерения воздействовать на нашу оценку доли лиц, соблюдающих правила, то получим несмещенную оценку среднего эффекта воздействия для участников, соблюдающих правила.

Давайте приведем небольшой пример, демонстрирующий этот подход в действии. Представьте, что Body Vibes действительно работает. (Напомним, что большая часть этой книги посвящена вымышленным мирам.) В частности, предположим, что вы можете измерить здоровье кожи по шкале от 1 до 10, где 10 – идеальная кожа, а 1 – очень плохая кожа.

Теперь давайте представим, что мы провели эксперимент на 100 участниках, чтобы изучить влияние Body Vibes. Мы случайным образом распределили 50 человек в экспериментальную группу и 50 – в контрольную. Участники из экспериментальной группы получили Body Vibes. Участники из контрольной группы остались без наклеек. Месяц спустя мы измерили состояние кожи каждого человека по шкале от 1 до 10. Предположим, данные выглядели так, как в табл. 11.3.

Таблица 11.3. Наблюдаемые различия между двумя группами

	Экспериментальная группа	**Контрольная группа**
Средний уровень здоровья кожи	7.8	6.2

Исходя из этих данных, наша оценка эффекта от намерения воздействовать составляет 1.6, т. е. в среднем у людей, получавших Body Vibes, показатель здоровья кожи был на 1.6 балла выше, чем у людей, не получавших Body Vibes.

Затем мы копнули немного глубже и обнаружили, что в контрольной группе никто не купил Body Vibes за свои деньги и не начал носить вопреки правилам, но только 40 из 50 человек в экспериментальной группе действительно начали их применять. Исходя из этого, мы подсчитали, что доля тех, кто соблюдает требования, в выборке составляет 80 % (40/50), а доля тех, кто никогда не принимает воздействие, – 20 %. Теперь вы можете оценить истинное влияние Body Vibes на добросовестных участников.

Как это делается? Чтобы убедиться в правильном понимании происходящего, давайте вернемся к обозначению потенциальных исходов. Пусть Y_{0c} – среднее состояние кожи участника без воздействия (т. е. без Body Vibes); пусть Y_{1c} – среднее состояние кожи участника, подвергнутого воздействию (т. е. Body Vibes); и пусть Y_{0n} будет средним состоянием кожи отрицателя, который никогда не соглашается на воздействие и попал в контрольную группу. Учитывая, что у нас 80 % соблюдают условия и 20 % никогда не соглашаются на воздействие, мы имеем следующие два уравнения:

$$7.8 = 80\,\% \cdot Y_{1c} + 20\,\% \cdot Y_{0n},$$
$$6.2 = 80\,\% \cdot Y_{0c} + 20\,\% \cdot Y_{0n}.$$

Первое уравнение показывает, что среднее состояние кожи у тех, кто был отнесен к экспериментальной группе (7.8), представляет собой средневзвешенное среднее значение здоровья кожи у исполнительных участников, получающих наклейки (с весом 80 %), и отрицателей, не получавших наклейки (с весом 20 %). Аналогичным образом средний показатель здоровья кожи у тех, кто был отнесен к контрольной группе (6.2), представляет собой средневзвешенное среднее значение состояния кожи у исполнительных участников, не получающих наклейки (с весом 80 %), и отрицателей, не получавших наклейки (с весом 20 %).

Мы можем вычесть левые и правые части этих двух уравнений друг из друга, чтобы получить

$$1.6 = 80\,\% \cdot (Y_{1c} - Y_{0c}).$$

Левая часть представляет собой эффект от намерения воздействовать: разницу в средних результатах между экспериментальной и контрольной группой. В правой части 80 % представляет долю участников, соблюдающих условия. А член в скобках означает *средний эффект воздействия для лиц, соблюдающих правила* (complier average treatment effect, CATE). Таким образом, мы можем восстановить средний эффект воздействия, разделив обе части на 80 %:

$$\frac{1.6}{80\%} = \overbrace{Y_{1c} - Y_{0c}}^{\text{CATE}} = 2.$$

Важно отметить различие между средним эффектом воздействия на послушного участника и общим средним эффектом воздействия. Вполне возможно, что ношение Body Vibes одинаково влияет на здоровье кожи у всех. В этом сценарии мы бы сказали, что существуют *однородные (гомогенные) эффекты воздействия*. Но это не обязательно так: Body Vibes могут по-разному влиять на здоровье кожи разных людей, и средний эффект может сильно различаться для тех людей, которые никогда не будут их использовать (отрицатели), и тех, кто будет использовать, если их побуждать к этому (исполнители). В этом случае мы говорим, что существуют *неоднородные (гетерогенные) эффекты воздействия*. Как показывает приведенная выше алгебраическая формула, деление эффекта от намерения воздействовать на долю участников, соблюдающих условия эксперимента, позволяет оценить средний эффект воздействия на этих участников. Если мы наблюдаем однородные эффекты воздействия, средний эффект воздействия на исполнительного участника аналогичен общему среднему эффекту воздействия. Но если существуют гетерогенные эффекты воздействия, они не одинаковы, и необходимо иметь в виду, что мы можем оценить средний эффект только для этой конкретной подгруппы. Этому есть простое объяснение. На самом деле только те, кто соблюдает правила, меняют свое поведение в ответ на воздействие. Так что они – единственная часть популяции, о которой мы действительно получаем информацию.

Было относительно легко увидеть, как все это работает в упрощенном мире, где каждый участник либо соблюдал правила, либо всегда отрицал любое воздействие. Но мы можем сделать то же самое, даже если отойдем от этого упрощенного мира и допустим существование участников, которые всегда испытывают воздействие (даже если его не назначали). Пока давайте продолжим считать, что «провокаторов» в нашем эксперименте нет, потому что они мутят воду. (Существует множество ситуаций, включая этот гипотетический эксперимент Body Vibes, в котором, как мы думаем, будет мало провокаторов или вообще их не будет.)

В табл. 11.4 показано, как в нашей экспериментальной выборке представлены различные типы участников эксперимента в этом более сложном мире.

Таблица 11.4. Кто принимает участие в эксперименте Body Vibes, если предположить, что провокаторов нет?

	Экспериментальная группа	**Контрольная группа**
Носят Body Vibes	Исполнители и активисты	Активисты
Не носят Body Vibes	Отрицатели	Исполнители и отрицатели

Как мы можем оценить долю тех, кто соблюдает правила эксперимента (исполнители), когда, кроме них, есть еще отрицатели и активисты? Во-первых, члены экспериментальной группы, которые на самом деле носят Body Vibes, – это либо исполнители, либо активисты. Таким образом, размер этой группы дает нам оценку доли исполнителей и активистов. Во-вторых, члены контрольной группы, носящие Body Vibes, явно относятся к активистам. Таким образом, размер этой группы дает нам оценку доли активистов. Вычитая второе число из первого, мы получаем оценку доли исполнителей. Зная эту долю, мы можем снова поступить, как указано выше, – рассчитать эффект от намерения воздействовать и разделить его на долю исполнителей, чтобы получить CATE.

Таким образом, наша общая процедура оценки среднего эффекта воздействия заключается в следующем. Мы начинаем с оценки эффекта намерения воздействовать, т. е. влияния отнесения к экспериментальной группе на наблюдаемый эффект. Затем проверяем, какая доля членов экспериментальной группы фактически подвергается воздействию. Иногда это называют *эффектом первого этапа* (first-stage effect). Если предположить, что провокаторов нет, мы получаем объективную оценку доли тех, кто соблюдает условия эксперимента. Затем восстанавливаем оценку CATE, разделив эффект от назначенного воздействия на долю участников, соблюдающих правила. Это соотношение называется *оценщиком Вальда* (Wald Estimator) в честь статистика Абрахама Вальда, который впервые разработал его, хотя и в другом контексте.

Оценщик Вальда – это частный случай так называемого анализа *инструментальных переменных* (instrumental variable, IV). Этот вид анализа подходит, когда рассматриваемый метод воздействия не назначается случайным образом, но есть какая-то другая переменная (называемая *инструментом*), которая (1) влияет на рассматриваемый метод воздействия, (2) не влияет на наблюдаемый эффект, кроме как посредством воздействия, и (3) назначается случайным образом (или существует какой-то другой способ достоверно оценить ее влияние на воздействие и результат).

Если быть более точным, есть четыре ключевых условия, которые должны соблюдаться для того, чтобы анализ инструментальных переменных работал должным образом.

1. **Экзогенность:** инструмент должен быть назначен случайным образом, что позволяет нам получить несмещенные оценки эффектов как первого этапа, так и редуцированной формы (соответствует намерению воздействовать).
2. **Ограничение исключения:** все эффекты редуцированной формы должны возникать только в результате воздействия. Другими словами, у инструмента не должно быть другого пути влияния на эффект, кроме

как через влияние на воздействие. Если это не так, то эффект в редуцированной форме включает как влияние воздействия на результат для субъектов, соблюдающих правила эксперимента, так и другие пути. Следовательно, даже после того, как мы разделим эффект первого этапа на долю исполнителей, полученная оценка все равно будет включать эти другие пути и, таким образом, не будет отражать CATE.

3. **Наличие исполнителей:** в эксперименте обязательно должны быть участники, строго соблюдающие требования.
4. **Отсутствие провокаторов:** если есть провокаторы, то наша оценка даст нам средневзвешенное значение среднего эффекта для тех, кто выполняет условия эксперимента, и среднего эффекта для тех, кто сознательно поступает наоборот, но при этом такие провокаторы получают отрицательный вес (поскольку их поведение изменилось в неправильном направлении). Негативные последствия от присутствия провокаторов зависят от того, сколько их в группе и насколько различны эффекты воздействия для исполнителей и провокаторов. Если провокаторов очень мало, то смещение, возникающее из-за их присутствия, незначительно. Но если провокаторов много, это большая проблема для инструментального анализа.

В случае с экспериментом Body Vibes разделение на экспериментальную и контрольную группу было отличным инструментом. Оно явно удовлетворяло экзогенности, поскольку мы рандомизировали воздействие. Также кажется маловероятным, что попадание в экспериментальную группу могло каким-либо образом повлиять на здоровье кожи, кроме как через Body Vibes, поэтому оно вполне удовлетворяло ограничению исключения. Таким образом, пока существуют исполнители (т. е. люди, которые действительно использовали Body Vibes, потому что им было поручено) и не было провокаторов, наш анализ давал оценку среднего эффекта воздействия на исполнителей.

Существуют более гибкие способы реализации инструментального анализа, чем оценщик Вальда. В частности, его можно реализовать с помощью регрессии, что важно, поскольку позволяет нам при необходимости учитывать переменные контроля, а также ситуации с несколькими инструментами или методами воздействия и инструментами, которые не являются бинарными.

Некоторые аналитики рассматривают инструментальный анализ как отдельный метод или план исследования. Например, аналитик может реализовать описанную выше схему и сказать, что он оценил эффект Body Vibes, используя инструментальные переменные. Технически это верно, но вводит в заблуждение. Ключевым элементом исследования в нашем примере является рандомизированный эксперимент. Мы используем инструментальные переменные для борьбы с помехами со стороны тех, кто не соблюдает правила эксперимента, признавая дополнительные допущения (помимо рандомизации), которые для этого необходимы. В частности, ограничение исключения в нашем примере оправдано, поскольку все, что было сделано в ходе эксперимента, – это раздача нелепых наклеек одним людям, а не другим. Однако в других ситуациях ограничение исключения будет труднее обосновать или реализовать, что потребует более тщательного планирования эксперимента. Мы вернемся к этому позже, когда будем обсуждать естественные эксперименты.

Случайный дисбаланс

Рандомизация гарантирует, что группы эксперимента и контроля будут ожидаемо одинаковыми с точки зрения потенциальных исходов. Но тут присутствует коварное слово «ожидаемо». Тот факт, что две группы одинаковы в ожидании, не означает, что они одинаковы на самом деле. Как мы уже говорили, в любом конкретном эксперименте группа, испытавшая воздействие, могла во многом отличаться от контрольной группы просто по случайности, и мы могли бы назвать это случайным дисбалансом. Вот почему в нашем любимом уравнении, помимо члена смещения, присутствует член шума.

Экспериментаторы часто оценивают *баланс* между группами, сравнивая их с точки зрения измеримых характеристик до воздействия. Например, в нашем эксперименте Body Vibes мы могли сравнить средний возраст, пол, вес, диету и состояние кожи субъектов в группах, испытавших и не испытавших воздействие до начала эксперимента. Мы могли бы даже проверить статистически значимые различия. Разумеется, всегда хочется надеяться, что мы не обнаружим никаких различий. Если это не так, следует задуматься о том, что, хотя наша оценка и не смещена, она тем не менее может сильно отличаться от истинного эффекта из-за шума.

Что должен делать ответственный аналитик, если, несмотря на рандомизацию, оказывается, что группы контроля и эксперимента существенно или статистически значимо различаются? Давайте рассмотрим три возможных ответа.

1. **Отвергнуть «испорченный» эксперимент.** У вас были благие намерения, когда вы проводили эксперимент, но вам не повезло, и теперь вы не можете доверять своим результатам, поэтому следует просто забыть об эксперименте и двигаться дальше. Возможно, стоит провести еще один эксперимент в надежде на лучший баланс между группами.

   Мы считаем, что это неудачный ответ. Помните о проблеме завышения значимости. Если вы проверите баланс достаточного количества переменных до воздействия, вы практически гарантированно обнаружите статистически значимый дисбаланс по некоторым из них. Следовательно, чем больше переменных до воздействия вы сможете измерить, тем больше вероятность, что вам придется отказаться от эксперимента – извращенная логика, не так ли? Более того, даже «испорченные» эксперименты содержат информацию. Важно отметить, что они не смещены (помните, что смещение позволяет получить правильный ответ в среднем на протяжении многих итераций эксперимента). Таким образом, информацию можно объединить с другими данными (возможно, из других итераций того же эксперимента) и включить в более масштабный анализ, который в конечном итоге внесет свой вклад в знания.

   Мы предполагаем, что аналитик уверен в полностью случайном назначении воздействия. Если бы это было не так, наши рекомендации изменились бы. Допустим, вы (или ваш компьютер) не выполнили рандомизацию напрямую. Вместо этого предположим, что вы проводите крупномасштабный эксперимент, и рандомизация была реализована большой командой или партнерской организацией. В подобной ситуации, если обнаружится значительный дисбаланс, вы можете заподозрить, что запланированная

рандомизация не была реализована должным образом. В этом случае может быть целесообразным отказаться от эксперимента (вероятно, после дополнительной проверки того, обоснованы ли ваши подозрения).

2. **Действовать как обычно.** Несмещенность – это ожидаемое свойство, поэтому экспериментальная оценка по-прежнему не смещена. Вы можете сообщить о дисбалансе ради прозрачности эксперимента, продолжая при этом оценивать эффект воздействия, как вы изначально планировали. Конечно, группы, испытавшие и не испытавшие воздействие, иногда случайно различаются. Именно поэтому мы сообщаем о стандартных ошибках или других показателях шума.

   Эта стратегия может показаться неразумной. Как вы помните из главы 6 и нашего любимого уравнения, даже несмещенная оценка может быть очень далека от истины. После обнаружения дисбаланса между группами, который, по нашему мнению, тесно связан с изучаемым эффектом, можно предположить, что этот случайный дисбаланс приводит к оценке, далекой от истины, несмотря на отсутствие смещения. Тем не менее все же есть смысл действовать по плану и сообщить о своей объективной (хотя, возможно, и совершенно ошибочной) оценке. Это особенно верно, если мы говорим об эксперименте, который будет повторяться много раз, так что отсутствие баланса в любой конкретной итерации будет в конечном итоге сведено на нет за счет усреднения по многим итерациям эксперимента.

   Но мы также можем задаться вопросом, есть ли какой-то способ объяснить дисбаланс и получить оценку, которая, вероятно, будет ближе к правильному ответу прямо сейчас, – что приводит нас к третьему возможному ответу.

3. **Использовать методы, описанные в главе 10, для контроля любых несбалансированных переменных.** Как было сказано в главе 10, выявление искажающих переменных до воздействия может повысить точность за счет учета дисперсии результата, обусловленной этими переменными. В этом смысле контроль может помочь вам приблизиться к истине. Но у него есть и недостатки. Благодаря рандомизации вы можете быть уверены, что оценка эффекта воздействия без учета мешающих переменных (например, простое сравнение среднего результата в группах контроля и эксперимента) приводит к объективной (хотя потенциально очень далекой от правильной) оценке истинного эффекта. Напротив, контроль переменных постфактум может привести к смещенной (хотя, возможно, и более точной) оценке. Это означает, что, если бы вы проводили эксперимент много раз и всегда контролировали бы те переменные, которые оказались несбалансированными, среднее значение ваших оценок могло бы не сойтись с истинным эффектом. Поэтому иногда приходится искать компромисс между уменьшением шума и увеличением смещения.

   Еще одна проблема, связанная с этим подходом, заключается в том, что, контролируя переменные до воздействия, исследователь пользуется дополнительными степенями свободы, что должно вызывать опасения по поводу завышения значимости и занижения отчетности. Как мы говорили в главе 7, сообразительные потребители данных должны реагировать скептически, когда видят, что аналитик манипулирует условиями,

и, если результат эксперимента зависит от определенного набора переменных контроля, которые отсутствовали в исходном плане, нам, вероятно, не следует доверять этому результату.

Не существует простого ответа или быстрого решения проблемы, возникающей из-за случайного дисбаланса после рандомизации. Мы считаем, что вам, вероятно, следует использовать комбинацию вариантов 2 и 3. Кроме того, когда это возможно, следует попытаться повторить эксперименты несколько раз. Несмотря ни на что, будьте честны и прозрачны в своем выборе. Конечно, нам бы очень хотелось избежать этих трудных решений, в первую очередь избегая случайного дисбаланса. И есть способы сделать это. Если вы сможете заранее определить и измерить важные характеристики, то сможете спланировать свой эксперимент так, чтобы обеспечить баланс. Мы уже кратко упомянули, как это сделать: используя в своем эксперименте стратифицированные выборки. Перед воздействием разделите выборку на группы на основе значимых характеристик, а затем выполните рандомизацию внутри этих групп. Напомним, ранее в этой главе мы предположили, что Body Vibes по-разному влияют на мужчин и женщин, поэтому следовало убедиться, что группы эксперимента и контроля сбалансированы по биологическому полу. Это достигается путем разделения выборки на группы, состоящие только из мужчин и только из женщин. Затем мы случайным образом назначаем воздействие внутри этих групп. Это гарантирует, что группы контроля и эксперимента будут хорошо сбалансированы по критерию пола (уменьшение шума), при этом воздействие по-прежнему назначается случайным образом (сохраняя беспристрастность). Мы можем избавить себя от множества проблем, заранее выполнив подобную процедуру для всех характеристик объектов, способных вызвать дисбаланс между группами.

Нехватка статистической мощности

Иногда даже отлично организованный эксперимент дает неубедительные результаты, потому что стандартная ошибка настолько велика, что за ней сложно разглядеть реально существующий эффект. В этом случае мы говорим, что эксперименту не хватило *статистической мощности* (statistical power) для обнаружения интересующего эффекта. В идеале экспериментатор должен заранее подумать об этой проблеме и предпринять шаги для повышения точности и статистической мощности эксперимента, например за счет увеличения размера выборки.

Тем не менее иногда из-за нехватки бюджета или других ограничений оказывается, что вы провели эксперимент с недостаточной мощностью. Как быть, если вы уже провели эксперимент и получили неточные оценки? Ситуация напоминает рассмотренное ранее обнаружение случайного дисбаланса. Вы можете попытаться повысить точность, контролируя некоторые переменные, но, как мы уже обсуждали, у этого подхода есть свои недостатки. Иногда будет проще признать, что у вас нет убедительного ответа на вопрос и вы мало что узнали, даже после проведения эксперимента.

Вспоминая главу 7, можете задаться вопросом, способствует ли сокрытие результатов недостаточно мощного эксперимента пресловутой «проблеме картотеки» (когда результаты неудачного эксперимента прячут от научного сообщества на дальнюю полку). Ответ: да. И это хороший повод не проводить эксперименты с недостаточной мощностью. Но если результаты эксперимента

настолько неточны, что не говорят нам практически ничего нового, то нет особого смысла их публиковать. Дело в том, что отказ от публикации из-за того, что эксперимент не принес новых данных, не так вреден для науки, как отказ от публикации из-за того, что эксперимент не дал желаемого результата.

Убыль в ходе эксперимента

Иногда люди выбывают из эксперимента после назначения воздействия. Такая *убыль* (attrition) существенно отличается от несоблюдения условий эксперимента. Несоблюдение условий касается людей, которые должны были подвергнуться воздействию, но сознательно решили не делать этого. По крайней мере, мы можем наблюдать за результатом такого поведения. Когда люди выбывают из эксперимента, мы не наблюдаем вообще ничего.

Предположим, например, что, вопреки ожиданиям, Body Vibes влияет не только на кожу, но и на психику, и заставляет некоторых людей чувствовать себя такими молодыми и беззаботными, что они забывают прийти на следующую встречу, где мы планировали измерить состояние их кожи. Это плохо. Если отток происходит случайным образом (т. е. не связан с воздействием или потенциальными эффектами), то мы все равно можем получить объективную оценку эффекта воздействия, сравнивая оставшихся членов групп эксперимента и контроля. Мы просто теряем некоторую статистическую мощность, потому что наша выборка стала меньше. Если убыль не случайна, но на нее не влияет назначение в экспериментальную группу, то мы можем, по крайней мере, оценить средний эффект воздействия для того типа людей, которые остались в эксперименте. Это реальный эффект, только отвечающий на немного другой вопрос. И конечно же, в большинстве случаев, если происходит отток участников, мы беспокоимся о том, что отток неслучаен и зависит от воздействия. Например, участники могут покидать эксперимент, потому что наклейки Body Vibes работают настолько хорошо, что люди полностью перестают беспокоиться о здоровье кожи. В этом случае если бы мы сравнили оставшихся членов экспериментальной и контрольной группы, то получили бы смещенную оценку эффекта. (Именно это может легко произойти в медицинском исследовании, если исследователи не будут осторожны.) Как и во многих других случаях, гораздо лучше, если вы сможете предвидеть и смягчить убыль на этапе планирования эксперимента, а не пытаться объяснить ее постфактум.

Если убыль неизбежна, что должен делать аналитик? Прежде всего вы можете проверить, повлияло ли экспериментальное воздействие на скорость отсева. Если да, то у вас больше нет корректного сравнения между группами. И, соответственно, вы можете увидеть, существует ли систематическое различие между оставшимися участниками эксперимента из разных групп по другим ковариатам, которые могут быть связаны с результатом.

Если у вас есть основания полагать, что воздействие действительно послужило причиной убыли, как вы можете поступить? Неужели просто придется отказаться от эксперимента? Не обязательно – есть последнее средство, которое не требует от аналитика каких-либо предположений о природе убыли. Вы можете попытаться ограничить величину смещения, возникающего из-за убыли.

Чтобы понять, как это работает, представьте себе эксперимент с бинарным результатом (1 = здоровая кожа, 0 = нездоровая кожа). Предположим, что 50 %

испытуемых как в экспериментальной, так и в контрольной группе имеют здоровую кожу, что позволяет предположить отсутствие эффекта Body Vibes, но 5 % испытуемых в каждой группе так и не пришли на измерение состояния своей кожи. Мы не знаем, послужило ли экспериментальное воздействие причиной убыли. Но можно задаться вопросом, насколько серьезным могло бы быть смещение при его наличии.

Наилучшим исходом в пользу гипотезы о том, что Body Vibes полезны для здоровья кожи, была бы ситуация, когда у всех участников экспериментальной группы, которые не пришли на осмотр, была здоровая кожа, а у всех участников контрольной группы, которые не пришли на осмотр, была больная кожа. В этом сценарии 52.5 % участников экспериментальной группы будут иметь хорошее здоровье кожи по сравнению с 47.5 % в контрольной группе, что подразумевает положительное влияние Body Vibes на здоровье кожи в 5 процентных пунктов. В качестве альтернативы при худшем для Body Vibes исходе эти числа будут перевернуты, и отрицательный эффект составит 5 процентных пунктов. Мы не можем быть уверены, что убыль не искажает наши оценки, но можем утверждать, что смещение из-за убыли не превышает 5 процентных пунктов в любую сторону.

Взаимное влияние

Взаимное влияние (*интерференция*, interference) возникает, когда воздействие на одного участника эксперимента оказывает влияние на другого участника, и наоборот. Это может исказить результаты эксперимента. Чтобы понять, что имеется в виду, познакомьтесь с историей, что мы услышали от нашего коллеги Криса Блаттмана о пилотном исследовании, которое он провел в Либерии.

Блаттман хотел разобраться, как помочь молодым людям, подвергающимся высокому риску участия в преступлениях или насилии в постконфликтных ситуациях. В частности, он пытался оценить влияние двух видов воздействия: получения молодыми людьми небольших денежных грантов для открытия собственного бизнеса и посещения ими когнитивно-поведенческой терапии.

Для проверки того, как сработают эти подходы, вы могли бы создать организацию, предлагающую гранты и поведенческую терапию. Затем вы могли бы сравнить тех, кто испытал одно из этих воздействий (или оба) с теми, кто этого не сделал, чтобы проверить, изменился ли кто-то из них в лучшую сторону.

Однако такой подход не позволяет выполнить корректное сравнение. Вполне возможно, что молодые люди, которые сами решили получить гранты или терапию, изначально отличаются от среднего молодого человека в выборке. Они могут быть более амбициозными, более здоровыми, умными или что-то в этом роде. Таким образом, было бы ошибкой объяснять всю разницу в поведении между теми, кто получил гранты или терапию, и теми, кто этого не сделал, причинным эффектом воздействия.

Чтобы решить эту проблему, Блаттман разработал рандомизированный эксперимент, в котором он случайным образом назначил различные воздействия разным группам либерийских молодых людей. Каждому испытуемому выплачивалась небольшая сумма просто за участие в эксперименте. Далее, некоторые участники не получали ничего, кроме этой оплаты (контрольная группа), в то время как среди остальных участников некоторые получали денежный

грант в размере около 200 долл., некоторые получали терапию, а некоторые получали и грант, и терапию.

План Блаттмана состоял в том, чтобы сравнить уровни преступности и бездомности среди молодых людей, отнесенных к разным группам. Идея заключалась в том, что, если бы молодые люди, испытавшие одну из форм воздействия, демонстрировали лучшие результаты, чем контрольная группа, это было бы убедительным доказательством полезного эффекта воздействия. На первый взгляд все правильно.

Проблемы начались, когда молодые люди, участвовавшие в исследовании, узнали, что около половины из них получат 200 долл., а остальные – нет. Они объяснили, что не хотят играть в эту лотерею. Они предпочли бы, чтобы каждый получил по 100 долл., что исключает риск не получить ничего. Но если каждому участнику дать по 100 долл., эксперимент теряет смысл. Ведь цель состояла в том, чтобы случайным образом дать некоторым больше, чем другим, и посмотреть, действительно ли те, кто получил больше, добились большего. Поэтому команда Блаттмана распределила денежные гранты в соответствии с протоколом эксперимента – случайным образом выдав только половине участников 200 долл.

Но эти молодые люди были на шаг впереди исследователей. Судя по всему, они договорились предоставить друг другу своего рода страховку. В результате этого страхового договора каждый из победителей лотереи отдал часть своих денег проигравшим, которые ничего не получили. Такого рода вмешательство исказило оценки, полученные в результате эксперимента, поскольку теперь тщательно составленная контрольная группа фактически получила часть воздействия, а тщательно составленная экспериментальная группа отказалась от части воздействия.

Этот пример показывает, насколько сложно провести чистый эксперимент. Иногда ваши подопытные или другая внешняя сила сведут на нет ваши усилия.

Неудавшийся пилотный проект Блаттмана является наглядным примером интерференции. При планировании эксперимента вы случайным образом назначаете нужный метод воздействия разным объектам наблюдения (например, отдельным людям, домохозяйствам, чашкам Петри). Вы исходите из предположения, что эти объекты наблюдения независимы друг от друга. Однако если от факта воздействия на один объект зависит результат воздействия на другой объект, это интерференция, и она может исказить результаты вашего эксперимента. В эксперименте Блаттмана проблема интерференции заключалась в том, что воздействие в группе, получившей денежные гранты, повлияло на результаты в контрольной группе, поскольку субъекты, испытавшие воздействие, фактически разделили часть воздействия с субъектами, которым это воздействие не было назначено.

Как осторожные аналитики справляются с интерференцией? Иногда ситуация складывается настолько интересно, что сама интерференция становится объектом исследования. Влияют ли налоги в одном штате на экономическое развитие соседнего штата? Если избирательная кампания мобилизует группу сторонников, мобилизует ли это впоследствии группу противников? Если в соответствии с программой общественного здравоохранения вакцинируют детей в одной школе, поможет ли это защитить детей в другой школе? Иногда исследователи могут разработать план эксперимента с целью оценить подобные побочные эффекты. Например, Блаттман мог случайным образом

распределить несколько групп друзей, где один человек получал денежное вознаграждение, и другие группы друзей, где никто не получал денежные, вознаграждений. Затем он мог бы проверить, вели ли себя люди, которым не давали денег, по-другому, когда получали деньги от друга.

В целом осторожным аналитикам необходимо предвидеть интерференцию и планировать свои исследования таким образом, чтобы минимизировать возможность ее возникновения. Именно поэтому исследователи проводят пилотные исследования. В случае Блаттмана, когда он расширил масштаб исследования после проблемного пилотного эксперимента, он позаботился о том, чтобы осталось в секрете, кому из испытуемых были назначены денежные гранты, а кому нет, снизив таким образом риск интерференции.

Естественные эксперименты

Нам хотелось бы узнать о причинно-следственных связях во многих важных ситуациях; однако эксперимент может быть неосуществимым, неэтичным, нереалистичным или непомерно дорогим. Но иногда мир создает для нас что-то вроде экспериментальной рандомизации даже без наших усилий по организации эксперимента. Мы уже видели один пример такого рода *естественного эксперимента* при обсуждении влияния чартерных школ на успеваемость в главе 9. Хотя ни один количественный аналитик не смог бы провести эксперимент, в котором он случайным образом отправлял бы одних детей в чартерные школы, а других – в государственные, многие чартерные школы сами распределяют прием случайным образом. Школы рандомизировались не по научным причинам, а потому, что этого требовал закон. Авторы закона, очевидно, переживали по поводу социальной справедливости и равных возможностей, а не о качестве причинно-следственных выводов. Но, независимо от мотивации, эти лотереи создают рандомизацию «в дикой природе», которая позволяет нам оценить эффект от посещения чартерных школ по сравнению с обычными государственными школами более достоверно, чем если бы мы сравнивали успеваемость учащихся в двух типах школ и пытались контролировать все многочисленные потенциальные искажающие факторы.

Естественные эксперименты почти всегда предполагают определенный уровень несоблюдения требований – например, не каждый, кто выигрывает в лотерею чартерной школы, в конечном итоге посещает эту чартерную школу, а некоторые люди, проигравшие в лотерею одной чартерной школы, выигрывают ее в другой. Таким образом, мы обычно либо оцениваем эффект намерения воздействовать (т. е. зависимость в редуцированной форме между выигрышем в лотерее при поступлении и академическими результатами), либо используем подход инструментальных переменных для оценки среднего эффекта воздействия для участников. В этом примере инструментом будет выигрыш в лотерею, воздействием – посещение чартерной школы, а результатом – некоторая мера школьной успеваемости (например, результаты тестов).

Применяя подход инструментальных переменных, нам нужно серьезно относиться к соблюдению условий, о которых мы говорили ранее. Если существует естественная рандомизация, мы можем быть уверены в экзогенности. То есть мы можем достоверно оценить влияние выигрыша в приемной лотерее

на успеваемость и посещение чартерной школы. Но нужно очень тщательно подумать об ограничении исключения. То есть существуют ли способы, которыми выигрыш в приемной лотерее может повлиять на успеваемость, кроме как через посещение чартерной школы?

Вполне возможно, что в примере с чартерными школами ограничение исключения является разумным и что мы действительно можем оценить средний эффект воздействия на тех, кто выполняет условия эксперимента. Но позвольте нам привести вам еще один пример, где ограничение исключения чревато более серьезными последствиями.

Военная служба и будущие доходы

Влияние военной службы на будущие доходы представляет значительный интерес для экономистов. Но, конечно, люди, которые служат и не служат в армии, различаются во многих отношениях, которые имеют значение для заработка. Следовательно, сравнение заработков ветеранов и остальных граждан (даже с учетом многих факторов) безнадежно запутано. Такое сравнение не дает правдоподобной объективной оценки причинного эффекта.

К счастью (для социологов), есть естественный эксперимент, который может пригодиться. Во время войны во Вьетнаме мужчинам, имеющим право на призыв, случайным образом присваивались номера призывника. На самом деле людей призывали только в том случае, если случайно присвоенный номер призывника было достаточно небольшим. Следовательно, мы имеем источник рандомизации призыва на военную службу.

Конечно, абсолютно строгого соблюдения призывной лотереи не было. Например, некоторые молодые люди добровольно пошли служить в армию, несмотря на большой номер призывника. (В нашей прежней терминологии такие люди называются *активистами*.) А другие, с небольшим номером призывника, покинули страну или иным образом избежали призыва. (В нашей прежней терминологии такие люди называются *отрицателями*.) Итак, если мы хотим получить оценку причинного влияния военной службы на доходы в будущем (а не редуцированную форму в виде влияния лотерейных номеров на доходы, что вряд ли интересно), нам необходимо использовать подход инструментальных переменных, что и было сделано во многих исследованиях. Идея состоит в том, чтобы использовать призывной номер как инструмент, военную службу как воздействие, а будущие доходы как эффект воздействия.

В этом контексте экзогенность вполне правдоподобна. Насколько мы можем судить, правительство действительно присваивало номера призывникам случайным образом. (Технически они случайным образом выбирали дни рождения, поэтому все, у кого один и тот же день рождения, были в одной лодке, но выбор дня рождения был случайным.) Таким образом, мы действительно можем оценить влияние номера призывника на военную службу и на будущие заработки.

А как насчет ограничения исключения? Для того чтобы ограничение исключения действовало, необходимо, чтобы номер призывной лотереи не влиял на будущие доходы никаким другим путем, кроме как через службу в армии. Как это может быть нарушено?

Одним из вариантов является реакция людей на получение номера с небольшим значением (т. е. приоритетного для призыва). Такие люди могли

с большей вероятностью совершить определенные действия, позволяющие им избежать призыва. Например, они могли с большей вероятностью бежать из страны. Или они могли с большей вероятностью продолжить высшее образование, чтобы получить студенческую отсрочку от призыва в армию. Если человек стал эмигрантом или поступил в университет, это наверняка повлияет на его будущие доходы. По сути, это альтернативные пути, по которым номер призыва может повлиять на будущие доходы, помимо военной службы. Из-за подобных нарушений правила исключения вполне может случиться так, что даже при случайном присвоении номеров призывников подход инструментальных переменных не позволит нам использовать призывную лотерею для достоверной оценки влияния военной службы на будущие доходы.

Подведение итогов

Есть причина, по которой мы называем эксперименты золотым стандартом причинно-следственных выводов. Назначая воздействие случайным образом, мы гарантируем, что группы контроля и эксперимента будут иметь одинаковое ожидание потенциальных результатов, а это означает, что мы можем получить объективные оценки причинно-следственной связи.

Даже при проведении рандомизированного эксперимента могут возникнуть серьезные проблемы. Поэтому планирование и анализ экспериментов требует бдительности и критического мышления. Аналогичные сложные проблемы могут возникнуть вне контекста экспериментов, поэтому нам нужно продолжать помнить о них, переходя к другим исследовательским проектам.

К несчастью для науки, идеальный эксперимент, который мы хотели бы провести, часто оказывается непрактичным, неосуществимым или неэтичным. Что нам делать в таком случае? В следующих двух главах обсуждаются особые обстоятельства, при которых мы все же можем получить достоверные оценки причинно-следственных связей даже без какой-либо рандомизации.

Ключевые термины

- **План исследования:** подходы к получению объективных оценок эффекта воздействия или других оценок.
- **Случайное назначение (закрепление):** случайно сгенерированное решение о том, каким объектам эксперимента будет назначено воздействие (например, путем подбрасывания монеты или с помощью генератора случайных чисел).
- **Стратифицированное случайное распределение:** процесс разделения объектов эксперимента на разные группы или страты (обычно это группы, которые, по вашему мнению, имеют схожие потенциальные результаты) и последующая рандомизация воздействия и контроля внутри каждой из этих групп. Это может значительно повысить точность ваших оценок. Если вероятность воздействия варьируется в зависимости от страты, придется это учитывать, чтобы получить несмещенные оценки.
- **Несоответствие:** когда субъект эксперимента выбирает статус воздействия, отличный от того, который ему был назначен.

- **Исполнители**: субъекты, которые строго соблюдают правила эксперимента.
- **Активисты**: субъекты, которые всегда подвергаются воздействию независимо от того, было ли оно им назначено.
- **Отрицатели:** субъекты, которые никогда не подвергаются воздействию независимо от того, было ли оно им назначено.
- **Провокаторы**: субъекты, которые всегда придерживаются статуса воздействия, противоположного назначенному.
- **Эффект намерения воздействовать**, или **эффект редуцированной формы**: средний эффект от попадания в группу, подвергнутую воздействию, а не в контрольную группу. Он не всегда равен среднему эффекту воздействия из-за несоблюдения правил эксперимента отдельными субъектами.
- **Эффект первого этапа:** средний эффект от включения в экспериментальную группу при начале воздействия. Он соответствует доле участников, соблюдающих правила.
- **Средний эффект воздействия для субъектов, соблюдающих правила (CATE)**: особый вид LATE.
- **Инструментальные переменные:** набор процедур для оценки CATE при наличии субъектов, не соблюдающих правила. Оценщик Вальда – это особый случай инструментальных переменных. Все варианты метода инструментальных переменных требуют, чтобы мы могли достоверно оценить влияние инструмента на воздействие и на результат (экзогенность), чтобы инструмент влиял на воздействие (исполнители), чтобы инструмент влиял на результат только через свое влияние на воздействие (ограничение исключения) и чтобы в эксперименте было минимальное количество «провокаторов», которые ведут себя строго наоборот относительно назначенного воздействия.
- **Экзогенность:** инструмент является экзогенным, если он назначен случайным образом или условно считается назначенным случайным образом, так что мы можем получить несмещенную оценку эффектов как первого этапа, так и редуцированной формы.
- **Ограничение исключения:** инструмент удовлетворяет ограничению исключения, если он влияет на результат только через влияние на воздействие, а не каким-либо другим путем.
- **Случайный дисбаланс:** ситуация, когда, несмотря на случайное распределение, группы эксперимента и контроля существенно различаются из-за шума.
- **Статистическая мощность:** технически определяется как вероятность отклонения нулевой гипотезы об отсутствии эффекта, если истинный эффект имеет определенную ненулевую величину. В разговорной речи мы говорим, что исследование имеет низкую статистическую мощность, если маловероятно, что оно даст статистически значимый результат, даже если исследуемый эффект велик.
- **Убыль:** ситуация, когда испытуемые выбывают из эксперимента, в результате чего вы не наблюдаете за результатами этих испытуемых. Убыль по своей природе отличается от несоблюдения требований.

- **Интерференция:** ситуация, когда сам факт воздействия на один объект оказывает влияние на эффект другого объекта.
- **Естественный эксперимент:** когда рандомизация выполнялась не для исследовательских целей, но осторожные аналитики тем не менее могут использовать эту рандомизацию для ответа на интересный причинно-следственный вопрос.

Упражнения

11.1. Предположим, психологическая лаборатория пытается изучить феномен поведенческого прайминга. В частности, они хотят знать, начинают ли подопытные ходить медленнее, когда им говорят слова, связанные со старением и старостью. Они набирают испытуемых в свою лабораторию и платят им за выполнение задания на словесную ассоциацию. Половина испытуемых отнесена к контрольной группе, слова которой не имеют ничего общего со старостью, а другая половина испытуемых отнесена к экспериментальной группе, для которой многие слова связаны со старением и старостью.

После того как испытуемые выполнили свою задачу, один из ассистентов без их ведома засекал, сколько времени им понадобится, чтобы пройти коридор, ведущий к выходу из здания. План исследователей состоит в том, чтобы проверить, приводит ли вербальное воздействие к замедлению ходьбы.

Ниже приведены некоторые факты об эксперименте. Подумайте, какое значение каждый факт имеет для эксперимента. Является ли он проблемой для исследователей? Если да, то в чем проблема? Что они могли бы сделать при планировании эксперимента или анализе данных, чтобы решить эту проблему?

a) Группа испытуемых представляла собой широкий срез общества, поэтому некоторые из испытуемых были старыми, некоторые молодыми, некоторые спортивными, некоторые неуклюжими, некоторые худыми, некоторые имели избыточный вес. В экспериментальной группе было больше пожилых и менее спортивных людей, чем в контрольной.
b) Некоторые испытуемые не обращали пристального внимания на словесную ассоциацию, давали бессмысленные ответы и просто проходили задание как можно быстрее.
c) Некоторым испытуемым потребовалось очень много времени, чтобы пройти коридор, потому что они остановились, чтобы поговорить с кем-то или проверить свой телефон.
d) Некоторые из испытуемых вообще не пересекали коридор, поскольку в задней части здания был другой выход.
e) Ассистенты-исследователи, которые измеряли скорость ходьбы испытуемых, знали гипотезу исследователей, и это были те же люди, которые проводили эксперимент.
f) Некоторые испытуемые обсуждали друг с другом задание на словесную ассоциацию, перед тем как выйти из здания.

11.2. Загрузите файл GOTV_Experiment.csv и связанный с ним файл README.txt, описывающий переменные в этом наборе данных, по адресу *http://www.press.princeton.edu/thinking-clearly*.

Мы будем анализировать данные рандомизированного эксперимента, чтобы оценить влияние воздействия в форме прямого призыва выходить на голосование на явку избирателей.

Несколько факторов усложняют анализ этого конкретного эксперимента.

Во-первых, вероятность случайного назначения воздействия была разной в городских и пригородных районах. Во-вторых, некоторые люди, которым было назначено воздействие, не получили его (до них не дошло письмо с призывом). И в-третьих, мы не можем наблюдать явку некоторых субъектов. Дополнительную информацию см. в файле README.txt.

a) Рассчитайте среднее значение явки для людей, которые испытали и не испытали воздействие, и интерпретируйте неявный эффект воздействия призыва к голосованию на фактическую явку. Подумайте о возможных смещениях, возникающих в результате трех перечисленных выше осложнений. Как вы думаете, эти осложнения ведут к недооценке или переоценке эффекта при анализе? Поясните свой ответ.
b) Используя знания, полученные в главе 10, попытайтесь учесть тот факт, что вероятность воздействия различается в городе и пригородной местности. Как изменилась ваша оценка? Почему?
c) Используя знания, полученные в этой главе, попробуйте учесть несоблюдение требований. Во-первых, попытайтесь оценить эффект от намерения воздействовать (редуцированная форма) и степень соблюдения условий воздействия (первый этап). Теперь разделите первое значение на второе, чтобы оценить средний эффект воздействия для участника.
d) Подумайте о проблеме убыли участников. Что вы косвенно предполагаете, если просто отбрасываете субъектов, по которым не наблюдаете явку? Посмотрите, как оценка изменяется при различных предположениях. Оцените средний эффект воздействия для исполнительных участников, предполагая, что ни один из испытуемых, покинувших эксперимент, не проголосовал бы. Какова будет, по вашему мнению, эффективность кампании по повышению явки при наихудшем сценарии? А при наилучшем сценарии?

Дополнительное чтение и ссылки

В качестве подробного руководства по проведению экспериментов, особенно полевых, мы рекомендуем книгу:

Alan S. Gerber and Donald P. Green. 2012. *Field Experiments: Design, Analysis, and Interpretation.* W. W. Norton.

Исследование, показывающее, что болезненные дети в Перу позже отлучались от грудного вскармливания:

Grace S. Marquis, Jean-Pierre Habicht, Claudio Franco, and Robert E. Black. 1997. *Association of Breastfeeding and Stunting in Peruvian Toddlers: An Example of Reverse Causality.* International Journal of Epidemiology 26 (2): 349–56.

Рандомизированный эксперимент по грудному вскармливанию в Беларуси:

Michael S. Kramer, Tong Guo, Robert W. Platt, Stanley Shapiro, Jean-Paul Collet, Beverley Chalmers, Ellen Hodnett, Zinaida Sevkovskaya, Irina Dzikovich, and Irina Vanilovich. 2002. *Breastfeeding and Infant Growth: Biology of Bias?* Pediatrics 110 (2): 343–47.

Есть много статей о лотерее во время войны во Вьетнаме. Вот две из них (одна классическая, другая относительно недавняя):

Joshua D. Angrist. 1990. *Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records.* American Economic Review 80 (3): 313–36;

Joshua D. Angrist and Stacey H. Chen. 2011. *Schooling and the Vietnam-era GI Bill: Evidence fromthe Draft Lottery.* American Economic Journal: Applied Economics 3 (2): 96–118.

Если первое задание в упражнениях к главе заставило вас задуматься, действительно ли поведенческий прайминг может повлиять на скорость ходьбы человека, мы рекомендуем следующее исследование. Оказывается, результат зависит от того, замеряет ли время машина или человек, знающий гипотезу. Другими словами, исследователям легко обмануть себя, искренне думая, что они что-то обнаруживают, хотя они знают, что им предстоит найти.

Stephane Doyen, Olivier Klein, Cora-Lise Pichon, and Axel Cleeremans. 2012. *Behavioral Priming: It's all in the Mind, butWhose Mind?* PLoS ONE 7 (1): e29081.

Глава 12

Модели разрывной регрессии

О чем эта глава

- Даже когда эксперименты невозможны, все же существуют особые ситуации, которые позволяют нам объективно оценивать причинные эффекты.
- Одной из таких ситуаций является случай, когда соотношение долей объектов, подвергнутых и не подвергнутых воздействию, меняется скачкообразно при достижении определенного порога. Здесь может пригодиться метод разрывной регрессии.
- Методы разрывной регрессии оценивают локальный средний эффект воздействия для объектов, близких к порогу изменения воздействия.

Введение

В главе 11 вы увидели несколько примеров того, как грамотно продуманные естественные эксперименты могут помочь нам узнать о причинно-следственной связи, даже если мы не можем провести настоящий эксперимент. Идея состоит в том, чтобы искать в окружающем мире ситуации, в которых мы можем выполнять корректные сравнения, не проводя эксперимента. Иногда, как в случае с чартерными школами, мир делает это буквально – посредством фактической рандомизации. В других случаях вам придется быть немного хитрее.

В этой главе мы обсудим одну особую ситуацию, которая поможет нам получить достоверные причинно-следственные оценки, – когда наблюдаемое воздействие меняется скачкообразно при известном пороге. В следующей главе рассмотрим еще одну подобную ситуацию – когда воздействие меняется со временем для одних объектов наблюдения и остается постоянным для других.

В главе 10 мы обсуждали попытки узнать о причинно-следственных связях, контролируя искажающие факторы. Обычно мы не особо верим в такие подходы, потому что очень сложно измерить и учесть все факторы. А если вы не можете что-то измерить, вы не можете это контролировать. Однако встречаются редкие ситуации, когда у нас есть объемная информация о назначении воздействия, помогающая сделать оценку эффекта более правдоподобной. Одним из примеров является рандомизированный эксперимент, рассмотренный в главе 11. Если мы знаем, что воздействие было назначено строго случайным образом, то можем быть уверены, что никаких искажающих факторов нет. В центре внимания данной главы находятся ситуации, в которых воздействие назначается в соответствии с каким-то строгим правилом. В таких ситуациях

мы могли бы узнать об эффекте воздействия, используя *метод разрывной регрессии* (regression discontinuity design).

Предположим, что каждый объект эксперимента связан с какой-то оценкой и воздействие определяется этой оценкой. Объекты, чья оценка находится по одну сторону порога, получают воздействие, а объекты, чья оценка находится по другую сторону порога, – нет. Это создает ситуацию, когда метод разрывной регрессии помогает нам оценить причинно-следственную связь. Очень близкие к этому порогу объекты с обеих сторон, вероятно, в среднем будут похожи друг на друга. Соответственно, сравнение этих двух групп (одна из которых получила воздействие, а другая нет) можно считать очень близким к корректному.

Давайте добавим конкретики. Предположим, мы хотим оценить влияние получения стипендии на обучение в колледже на будущие доходы. В целом это сложно, потому что студенты, получающие стипендии за заслуги, вероятно, отличаются во многих отношениях, которые имеют значение для будущих заработков (интеллект, способности, амбиции, трудовая этика), от тех, кто стипендию не получает. И конечно же, мы не можем измерить и контролировать все эти различия.

Но что, если присуждать стипендию в соответствии со строгими правилами подсчета баллов? Стипендиальный комитет присваивает рейтинг от 0 до 1000 для каждого заявителя на основе средней успеваемости, результатов тестов, общественных работ и внеклассной деятельности. Стипендию получают все, кто набрал рейтинг в 950 баллов и выше, а те, кто ниже, – нет. Теперь, несмотря на то что явная рандомизация отсутствует, мы могли бы узнать об эффекте получения стипендии для тех кандидатов, рейтинг которых близок к порогу в 950 баллов. Как это работает?

Предположим, что стипендиальный комитет и претенденты не могут точно манипулировать оценками. То есть студенты прилагают усилия, не зная заранее, как будут подсчитаны их баллы, а комиссия честно оценивает студентов, также не зная заранее, каким будет порог назначения стипендии. Кроме того, ожидается, что студенты с рейтингом 950 почти идентичны студентам с рейтингом 949. Явная рандомизация отсутствует, но, вероятно, существует множество специфических факторов, которые могли бы легко поднять 949 до 950 или наоборот. Если бы студенты с рейтингом 949 сдали стандартный тест в немного менее напряженный день, отработали бы еще один час общественных работ из сотен, учились бы у более лояльного преподавателя, который бы поставил чуть более высокую оценку за свой курс, они бы получили 950 баллов и выиграли бы стипендию. Точно так же, если бы у получивших рейтинг 950 случилось единственное незначительное и несистемное событие, ухудшающее оценку, они бы стали 949-ми и потеряли стипендию. Поэтому кажется разумным сказать, что в среднем 949-е, по сути, такие же, как 950-е, до принятия решения о стипендии. И поэтому мы имеем что-то вроде естественного эксперимента. Сравнение людей, находящихся прямо на пороге – некоторые из которых получили стипендию (рейтинг 950), а некоторые не получили (рейтинг 949) по практически случайным причинам, – это корректное сравнение яблок с яблоками. Сравнивая будущие доходы этих двух групп, мы можем оценить причинный эффект получения стипендии за заслуги, по крайней мере, для студентов с баллами, близкими к пороговому.

Рассмотрим более общий подход к подобной ситуации. Мы хотим оценить влияние бинарного воздействия на некоторый исход. Назначение воздействия полностью определяется некоторой третьей переменной (например, упомяну-

тым выше рейтингом студента), которую мы называем *скользящей переменной* (running variable). В частности, если скользящая переменная превышает некоторый порог для данного объекта, то этот объект подвергается воздействию ($T = 1$), а если скользящая переменная ниже этого порога, то объект не испытывает воздействия ($T = 0$). В таком случае могут быть получены данные, похожие на рис. 12.1, где черные точки соответствуют экспериментальным объектам (подвергнутым воздействию), а серые точки – контрольным объектам. На рисунке порог расположен в нулевом значении скользящей переменной.

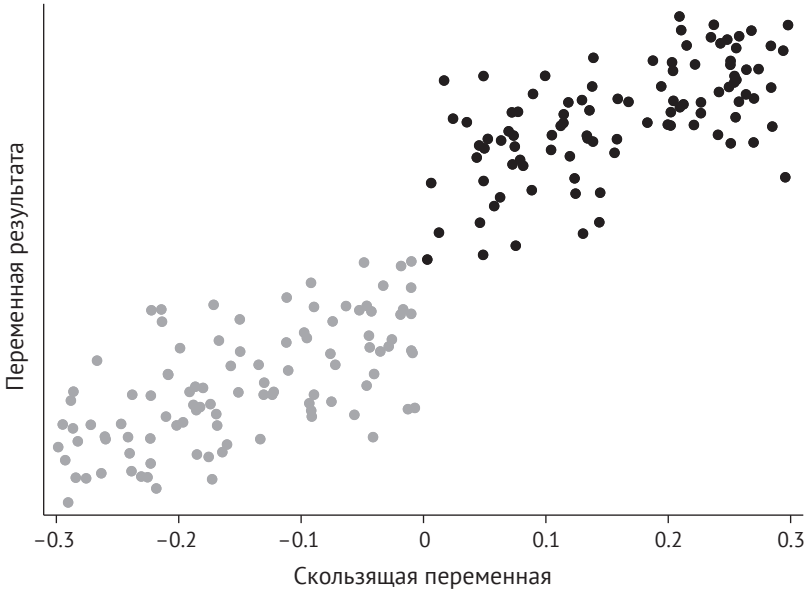


Рис. 12.1. Диаграмма рассеяния с воздействием, определяемым непрерывной переменной. Черные точки – это экспериментальные объекты. Серые точки – контрольные объекты

Как оценить эффект лечения в такой ситуации?

На первый взгляд кажется, что мы почти бессильны. Скользящая переменная сильно коррелирует с наблюдаемым результатом. В примере со стипендией это имеет смысл, поскольку комитет хочет отобрать людей с высокими способностями, и неудивительно, что критерии, которые они используют для выставления оценок, сильно коррелируют с будущими доходами независимо от того, выиграет ли студент стипендию. Комитет использует правило отсечения, поэтому каждый, кто получает стипендию, имеет более высокие значения скользящей переменной, чем тот, кто ее не получает. Очевидно, что если сравнивать тех, кто испытал и не испытал воздействие в виде стипендии, то наши входные данные будут источником искажения. И из-за правила отсечения мы не можем провести корректное сравнение, найдя студентов с одинаковым значением скользящей переменной, среди которых одни испытали воздействие (т. е. получили стипендию), а другие – нет. Все студенты с одинаковым баллом имеют одинаковый статус воздействия.

Но сдаваться рано. Давайте подумаем, что еще можно сделать. Мы можем оценить ожидаемую величину эффекта для заданного значения скользящей переменной. Для участников, чье значение скользящей переменной превышает порог,

мы получим ожидаемый эффект при воздействии с этим значением скользящей переменной. Мы можем вычислить эту величину для каждого значения скользящей переменной вплоть до порога. Аналогично для участников, чья оценка скользящей переменной ниже порогового значения, мы узнаем ожидаемый эффект без воздействия при этом значении скользящей переменной. Мы можем вычислить эту величину для каждого значения скользящей переменной вплоть до порога. Таким образом, прямо на пороге у нас есть оценки ожидаемого эффекта с воздействием и без него. Разница между этими двумя значениями вполне может быть хорошей оценкой эффекта воздействия, по крайней мере, для тех участников, у которых значение текущей переменной находится прямо на пороге.

Мы могли бы оценить эту величину, сравнивая участников по обе стороны от порога, имеющих значения скользящей переменной, очень близкие к порогу. Именно в этом заключалась идея сравнения 949- и 950-балльных студентов, чтобы узнать о влиянии стипендий за заслуги. Но на самом деле есть более совершенные методы.

Одна из стратегий состоит в том, чтобы построить две регрессии эффекта от скользящей переменной: одну для наблюдений без воздействия ниже порогового значения и одну для наблюдений с воздействием выше порогового значения. Затем мы можем использовать эти две регрессии для прогнозирования результатов с воздействием и без него прямо на пороге. На основе этих прогнозов мы можем оценить «скачок», или «разрыв», эффекта, когда скользящая переменная пересекает пороговое значение. Этот разрыв является оценкой причинного эффекта воздействия для участников, находящихся прямо на пороге. По этой причине мы называем данный подход методом *разрывной регрессии* (regression discontinuity, RD). Иллюстрация этой идеи показана на рис. 12.2.

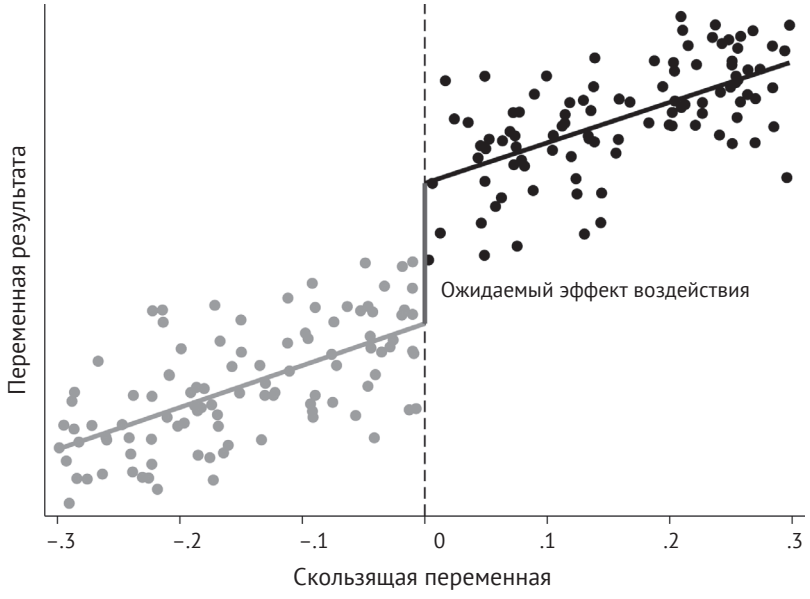


Рис. 12.2. Метод разрывной регрессии оценивает скачок ожидаемого эффекта в пороговой точке и рассматривает его как причинный эффект воздействия на объекты, находящиеся непосредственно на пороге

Стоит подчеркнуть одну вещь: *локальность* среднего эффекта воздействия, который оценивается с помощью модели разрывной регрессии. Возможно, что средний эффект воздействия различен при разных значениях скользящей переменной, как показано на рис. 12.3. На этом рисунке для каждого наблюдаемого объекта показаны оба потенциальных исхода. Для каждого объекта Y_1 показан черным, а Y_0 – серым. Фактические результаты, которые мы наблюдаем, показаны сплошными кружками, а контрфактические результаты, которые не наблюдаем, – пустыми кружками. Размер разрыва различен для каждого значения скользящей переменной.

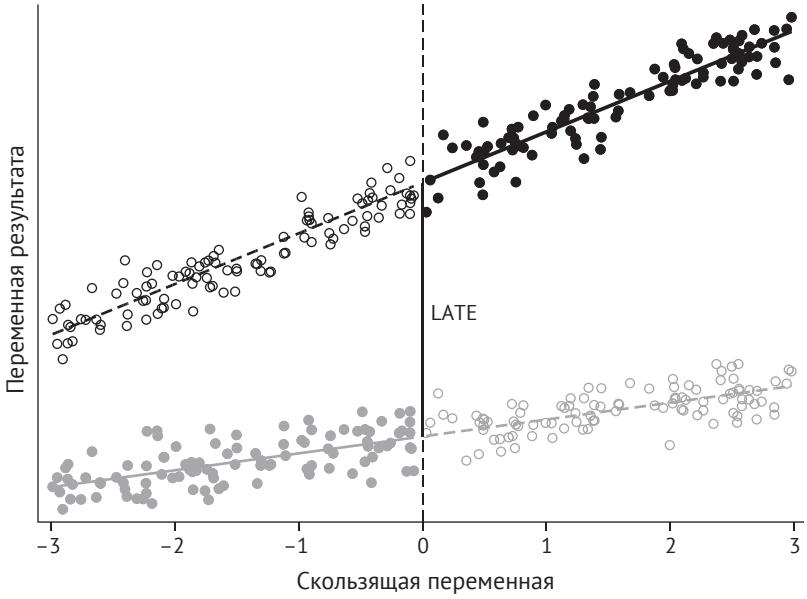


Рис. 12.3. Модель разрывной регрессии оценивает LATE в пороговой точке. Это не обязательно будет общий средний эффект воздействия, поскольку средние эффекты воздействия могут различаться для разных значений скользящей переменной

Если говорить более конкретно, то в нашем примере влияние выигрыша стипендии на будущие заработки может быть разным для студентов с низкой и высокой успеваемостью. Расчетная величина разрыва регрессии представляет собой средний эффект воздействия для объектов, у которых значения скользящей переменной находятся прямо на пороге. В текущем примере мы оцениваем влияние выигрыша стипендии на будущие заработки студентов, набравших 950 баллов, при этом влияние на студентов, набравших, скажем, 700 баллов, может быть совершенно другим. Мы называем эту оценку *локальным средним эффектом воздействия* (local average treatment effect, LATE). Как всегда, LATE может отличаться от общего среднего эффекта воздействия в генеральной совокупности. Поэтому при использовании метода разрывной регрессии важно понимать, действительно ли найденная оценка эффекта – это та величина, которая вас интересует.

Методы разрывной регрессии важны в различных ситуациях. Одним из распространенных применений является оценка последствий государственных программ. Многие государственные программы применяются скачкообраз-

но при заданных пороговых значениях. Например, получение адресного государственного пособия зависит от того, по какую сторону порога находится показатель постоянного дохода или бедности конкретного человека. Административная политика на уровне округа часто определяется пороговым значением численности населения или долей жителей определенного типа. Методы разрывной регрессии предоставляют простой способ оценить эффекты этих программ. Более того, в этих планах оценивается эффект программ для тех людей или мест, которые нас волнуют больше всего, – для граничного объекта, который едва получил право на участие в программе или, наоборот, немного не дотянул. Таким образом, если политики пытаются выяснить, следует ли им сократить или расширить конкретную государственную программу, оценка методом разрывной регрессии может быть очень информативной.

Реализация метода разрывной регрессии

Аналитикам доступны разные способы реализации моделей разрывной регрессии, и у каждого из них есть свои достоинства и недостатки.

Самый простой подход, как упоминалось выше, – просто сравнить средний эффект от воздействия для небольших диапазонов текущей переменной (иногда называемых *интервалами* или *бинами*) по обе стороны от порога. Например, мы могли бы сравнить средний заработок кандидатов, набравших от 950 до 954 баллов, со средним заработком кандидатов, набравших от 945 до 949 баллов. По причинам, которые вы вскоре увидите, мы часто называем этот подход *наивным*.

Явным преимуществом наивного подхода является его простота. Наивным его делает тот факт, что он практически гарантированно дает необъективные оценки. Почему? Скользящая переменная обычно коррелирует с потенциальными результатами. Зачем членам комитета использовать баллы для распределения стипендий, если они не верят, что баллы соответствуют способностям, усилиям, мотивации или какому-то другому фактору, который, вероятно, коррелирует с заработком в будущем?

Поскольку скользящая переменная коррелирует с потенциальными результатами, всегда будет некоторая базовая разница между группами чуть выше и чуть ниже порога. Конечно, по мере того как размер сравниваемых интервалов (иногда называемый *полосой пропускания*) уменьшается, смещение должно уменьшаться, но оно никогда не исчезнет.

Мы уже видим, что одно из важных решений, которое должен принять аналитик, – это выбор полосы пропускания. Принимая это решение, он часто сталкивается с выбором между уменьшением смещения и повышением точности. Меньшая полоса пропускания обычно дает менее смещенные, но и менее точные оценки, поскольку они используют меньше данных.

Потенциально менее смещенной альтернативой наивному подходу является локальный линейный подход. Здесь мы снова выбираем полосу пропускания и для наблюдений в пределах этой полосы строим линейную регрессию эффекта от скользящей переменной отдельно по обе стороны от порога. Мы используем эти оценки, чтобы получить прогнозируемые значения эффектов с воздействием и без него точно на пороге, а разница в этих прогнозируемых значениях является нашей оценкой эффекта воздействия для объектов, находящихся в пороговой точке.

При таком подходе мы допускаем возможность существования связи между скользящей переменной и результатом, мы допускаем, что эта связь будет разной по обе стороны от порога, и предполагаем, что эта связь приблизительно линейная (по крайней мере, для небольшого окна, в рамках которого мы анализируем данные). Именно такой подход мы использовали на рис. 12.2.

Существует способ реализовать этот локальный линейный подход с помощью одной регрессии вместо построения двух отдельных регрессий. Сначала измените масштаб скользящей переменной, чтобы пороговое значение было равно нулю (т. е. вычтите значение порогового значения из скользящей переменной). Затем сгенерируйте переменную воздействия, указывающую, находится ли наблюдение выше или ниже порога. Потом сгенерируйте интерактивную переменную, перемножив переменную воздействия и масштабированную скользящую переменную. И наконец, регрессируйте результаты воздействия, масштабированной текущей переменной и их взаимодействия для наблюдений в пределах вашей полосы пропускания. Расчетный коэффициент при переменной воздействия равен искомому разрыву.

Третий распространенный способ реализации метода разрывной регрессии – это полиномиальная регрессия. Аналитик может построить регрессию результата воздействия, скользящей переменной и полиномов более высокого порядка (т. е. скользящей переменной во второй степени, третьей степени и т. д.). Этот подход учитывает возможную нелинейную связь между текущей переменной и результатом. Недостатком является то, что точки данных, находящиеся далеко от порогового значения, могут оказать большое влияние на оценку разрыва.

При реализации выбранного метода разрывной регрессии исследователю приходится неоднократно делать важный выбор, поэтому он должен стараться избежать проблем завышения значимости и занижения отчетности. Ваши конкретные решения должны зависеть от ваших знаний предметной области и убеждений о взаимосвязи между скользящей переменной и результатом, а также от того, какое смещение вы готовы принять в обмен на повышение точности или наоборот. Лучший подход – обосновать свой выбор с помощью сочетания теории, знаний предмета исследования и анализа данных и, что, возможно, наиболее важно, продемонстрировать результаты для различных сочетаний параметров метода. Если ваши оценки надежны при различных полосах пропускания и прочих настройках, это придаст вашим результатам дополнительную достоверность. Если ваш результат справедлив только для одного очень конкретного сочетания параметров метода, вам следует отнестись к нему скептически.

Чтобы проиллюстрировать, как можно исследовать устойчивость оценки в разных полосах пропускания, на рис. 12.4 показан анализ в одной из статей Энтони, написанной в соавторстве с Харитцем Гарро и Йоргом Спенкухом. Они надеялись установить, получают ли фирмы выгоду от политических связей, проверив, растет ли цена акций фирмы, когда политический кандидат, на предвыборную кампанию которого фирма внесла пожертвования, побеждает или проигрывает с минимальным перевесом голосов. Таким образом, результатом является изменение цены акций фирмы, скользящей переменной является доля голосов за кандидата, а воздействием – выигрыш или проигрыш кандидата.

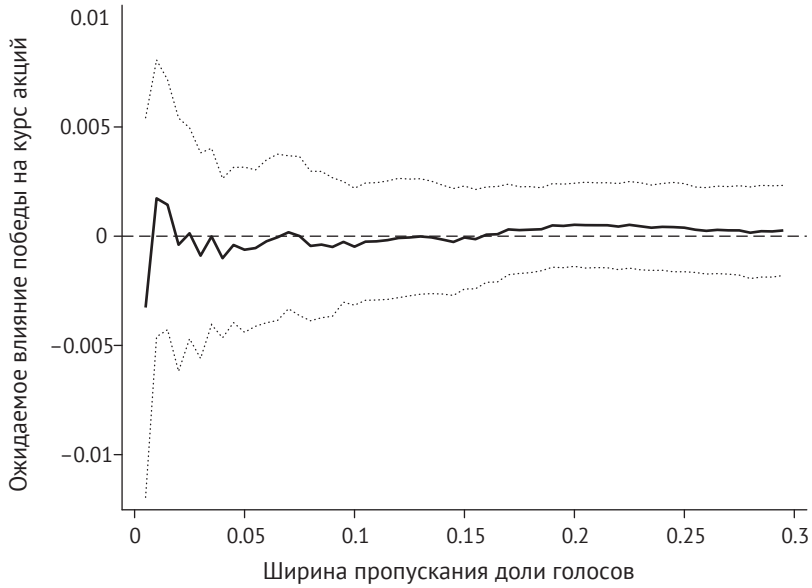


Рис. 12.4. Визуализация зависимости оценки разрывной регрессии (сплошная линия) и доверительного интервала (пунктир) от полосы пропускания

Они использовали локальный линейный подход, но хотели убедиться, что их выводы устойчивы к изменению полосы пропускания. На рис. 12.4 показаны предполагаемые эффекты, а также верхняя и нижняя границы 95-процентного доверительного интервала для 60 различных возможных полос пропускания от 0.5 до 30 процентных пунктов. Как и следовало ожидать, доверительные интервалы больше, а оценки более изменчивы для меньших полос пропускания, но оценки становятся точнее по мере увеличения полосы пропускания и включения большего количества данных. К счастью, оценки схожи почти для всех полос пропускания, что обнадеживает. Если бы оценка существенно изменилась по мере увеличения пропускной способности, это означало бы компромисс между смещением и точностью, и нам пришлось бы дальше думать о том, каким оценкам мы доверяем больше.

Дальше мы на примере обсудим реализацию и интерпретацию модели разрывной регрессии. Победители и проигравшие на выборах определяются исключительно по доле голосов, поэтому, если мы хотим оценить влияние определенного типа кандидатов на результат выборов, модель разрывной регрессии может оказаться особенно полезной.

Какие кандидаты более успешны – радикальные или умеренные?

В преддверии президентских выборов 2016 и 2020 гг. демократическая партия участвовала в жарких дебатах о возможности избрания политически радикальных и умеренных кандидатов. В частности, либеральное крыло партии было разочаровано выдвижением кандидатур Хиллари Клинтон и Джо Байдена, которых они сочли слишком умеренными. Они утверждали, что путь к победе на выборах

заключается не в том, чтобы апеллировать к избирателям-центристам. Наоборот, партии должны выдвигать идеологически радикальных кандидатов, способных создать электоральную базу. Утверждалось, что у Берни Сандерса было больше возможностей победить Дональда Трампа на всеобщих выборах, чем у любого из его более умеренных соперников. Конечно, некоторые политические заявления Сандерса могли отпугнуть какое-то количество умеренных избирателей. Но Сандерс с лихвой компенсировал бы эти потери, мобилизовав прогрессивных избирателей, которые довольно вяло поддерживали Клинтон и Байдена.

Как узнать, верен ли этот аргумент? С одной стороны, умеренные кандидаты могут убедить больше избирателей-центристов поддержать их партию. С другой стороны, радикалы могут мобилизовать новых избирателей. Итак, если вы хотите максимизировать шансы на победу вашей партии на всеобщих выборах, кого вам следует поддержать на первичных выборах? Конечно, невозможно с уверенностью сказать, что произошло бы, если бы Сандерс, наоборот, выиграл номинацию от демократической партии в 2016 или 2020 гг. (вспомните фундаментальную проблему причинно-следственных связей из главы 3). Но, возможно, мы сможем сказать больше о том, что в среднем происходит, когда партия выдвигает более радикального, а не более умеренного кандидата.

Чтобы попытаться разобраться в этом, давайте обратимся к выборам в конгресс, по которым у нас гораздо больше данных, чем по президентским выборам. На первый взгляд кажется, что сторонники радикальных кандидатов в чем-то правы. В конце концов, похоже, что в конгрессе много идейных пуристов[1]. Если умеренность является выигрышной стратегией, почему у власти так много радикалов?

Для начала нужно убедиться, что мы не забываем урок из главы 4: корреляция требует вариаций. Тот факт, что многие конгрессмены придерживаются чересчур крайних политических взглядов, не означает положительной корреляции (не говоря уже о причинно-следственной связи) между идеологическим радикализмом и успехом на выборах. Чтобы выяснить интересующую нас корреляцию, нам нужно сравнить электоральные успехи радикальных и умеренных кандидатов. Конечно, одно из возможных объяснений большого количества радикалов в конгрессе состоит в том, что радикализм действительно коррелирует с победой. Но, кроме того, в выборах изначально участвует очень мало умеренных кандидатов.

Более того, может быть неверно рассуждать о радикализме и умеренности в национальном масштабе. Скорее, с целью планирования избирательной стратегии нам нужно знать, придерживается ли кандидат крайних или умеренных взглядов по отношению к предпочтениям его конкретного электората или округа. Сандерс, безусловно, является крайним либералом по сравнению со средним избирателем в Соединенных Штатах. Но когда он баллотируется от Вермонта в Сенате, возможно, он лишь немного левее центра. Действительно, возможно, многие политики кажутся идеологически умеренными по отношению к своим избирателям, но идеологически радикальными по отношению к стране в целом. Это может произойти, если избирательные округа сами по себе идеологически экстремальные по сравнению со страной – некоторые крайне левые, другие – крайне правые. Но в данном случае нам не следует интерпретировать присутствие большого количества политиков с крайними убеждениями в конгрессе как свидетельство того,

[1] В политике пуристы – сторонники максимально строгой и чистой идеологии без компромиссов. – *Прим. перев.*

что крайние убеждения сами по себе являются эффективной избирательной стратегией, потому что победившие кандидаты в конгресс не были бы восприняты как идеологические радикалы избирателями, которые за них голосовали.

Учитывая эти опасения, на самом деле нам нужно знать не корреляцию между убеждениями и успехом на выборах, а влияние радикальности кандидата на успех среди его электората. Чтобы получить объективную оценку этого, нам нужно сравнить, какие результаты партии показывают на выборах, когда они выдвигают крайнего и умеренного кандидата при прочих равных условиях. В среднем что лучше для партии – выдвинуть крайнего или умеренного кандидата?

Конечно, наивное изучение корреляции между результатами выборов и идеологическим радикализмом кандидатов не является пустой тратой времени. По-видимому, время, место и ситуации, когда партия выдвигает умеренного кандидата, отличаются от тех, когда партия выдвигает радикала по разным причинам, которые имеют важное значение для результатов выборов. Например, наиболее вероятно, что либеральные демократы выиграют праймериз в более либеральных местах, где демократическая партия сильнее, а умеренные демократы победят в более консервативных местах, где партия слабее. Таким образом, если бы мы обнаружили, что радикалы добиваются большего успеха на всеобщих выборах, это не означало бы, что партии выигрывают, когда выставляют на выборы радикалов. Причинно-следственная интерпретация этой корреляции, очевидно, была бы запутанной. Мы могли бы попытаться контролировать различия во времени и пространстве, но нас всегда будет беспокоить тот факт, что между округами, выдвигающими крайних и умеренных кандидатов, все еще существуют ненаблюдаемые базовые различия. Мы можем добиться большего, используя модель разрывной регрессии.

Кандидаты в конгресс от основных партий избираются на первичных выборах. А результаты выборов определяются резким порогом. Предположим, мы анализируем большую выборку первичных выборов, на которых один радикальный кандидат противостоит одному умеренному кандидату. Наблюдаемое воздействие – выдвижение кандидата с крайними убеждениями. Мы хотим знать, как такое воздействие повлияет на долю голосов партии на всеобщих выборах. Чтобы построить разрывную регрессию, определите скользящую переменную как долю голосов за радикального кандидата на первичных выборах. Если эта доля голосов ниже половины, партия будет иметь умеренную позицию на всеобщих выборах; если она превышает половину, партия займет крайнюю позицию. Теперь мы можем оценить эффект от выдвижения радикального кандидата, применив модель разрывной регрессии и сравнивая результаты всеобщих выборов, когда партия с минимальным перевесом выдвинула радикала, с ситуацией, когда она с минимальным перевесом выдвинула на праймериз умеренного кандидата.

Эндрю Холл сделал именно это в исследовании 2015 г. Он обнаружил значительный отрицательный разрыв в результатах всеобщих партийных выборов в пороговой точке. То есть в среднем партия, выдвигающая кандидата с крайними убеждениями вместо умеренных, значительно снижает свои результаты на всеобщих выборах. Несмотря на прогнозы сторонников Сандерса, факты свидетельствуют о том, что выдвижение кандидатур радикалов в среднем является плохой избирательной стратегией.

Схема анализа Холла показана на рис. 12.5. Две линии представляют собой отдельные линейные регрессии по обе стороны от порога 50 %. Каждый ма-

ленький серый кружок соответствует одному наблюдению – партийным выборам. Большие черные кружки показывают среднюю долю голосов на всеобщих выборах для интервалов с перевесом в 0.02 пункта. Большой отрицательный разрыв в пороговой точке представляет собой предполагаемый эффект от выдвижения на всеобщих выборах радикального кандидата вместо умеренного в предвыборной гонке при условии, что голоса на первичных выборах были равномерно разделены между умеренными и крайними кандидатами.

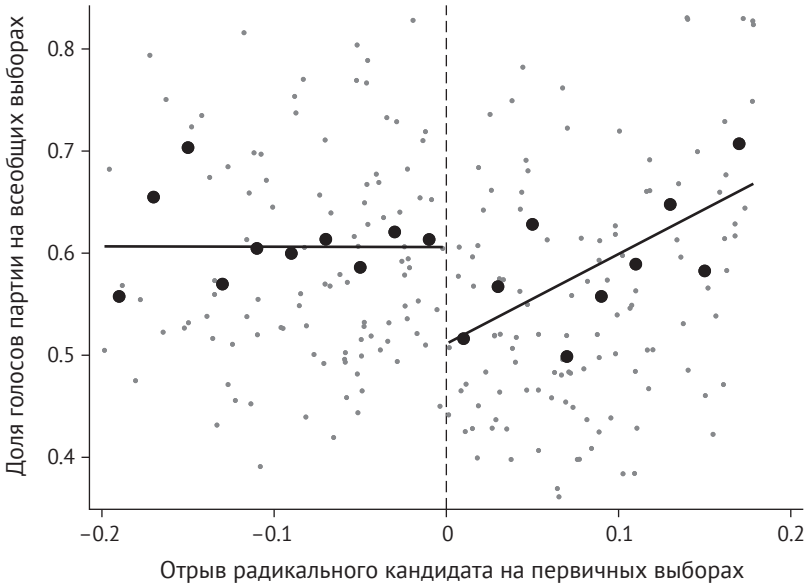


Рис. 12.5. Влияние выдвижения кандидата с крайними взглядами на перспективы голосования

Чем объясняется этот результат? В другом исследовании Холл и Дэн Томпсон исследуют ситуацию дальше. Используя аналогичную модель разрывной регрессии, они изучают влияние выдвижения кандидатуры радикала на явку избирателей. Интересно, что, вопреки прогнозам сторонников Сандерса, нет никаких доказательств того, что кандидаты-радикалы отрицательно влияют на электоральную базу партии. Точнее говоря, выдвижение радикальной кандидатуры действительно влияет, но не так, как ожидалось. Когда одна партия выдвигает кандидата-радикала, заметно возрастает количество сторонников другой партии, которые идут голосовать за своего кандидата. Таким образом, можно предположить, что, если бы Берни Сандерс выиграл праймериз демократической партии в 2016 или 2020 г., его результаты были бы немного хуже, чем у Клинтон и Байдена. Он, вероятно, потерял бы некоторых избирателей-центристов, которые предпочитали Клинтон или Байдена Трампу, но, что не менее важно, он побудил бы избирателей-республиканцев явиться в большем количестве на выборы.

Непрерывность в пороговой точке

Для того чтобы метод разрывной регрессии мог обеспечить объективную оценку причинно-следственной связи, необходимо, чтобы статус воздействия резко менялся в пороговой точке и ничто другое, что имеет значение для результатов,

не менялось. Если исходные характеристики столь же прерывисто изменяются в пороговой точке, то любые различия в средних результатах вблизи порога могут быть связаны с этими изменениями исходных характеристик, а не с воздействием. То есть сравнение экспериментальной и контрольной групп больше не будет корректным даже в пороговой точке, потому что эти две группы будут различаться не только по статусу воздействия. Но если средние исходные характеристики объектов изменяются непрерывно (а не дискретным скачком) по мере прохождения скользящей переменной через порог, то мы можем получить несмещенную оценку эффекта воздействия, потому что единственное, что будет отличать объекты, находящиеся по одну или другую сторону порога, в среднем – это статус воздействия. Требование, чтобы исходные характеристики не менялись скачком при достижении порога, мы называем *непрерывностью в пороговой точке* (или просто *непрерывностью* для краткости).

Давайте посмотрим, почему непрерывность имеет решающее значение. Рисунок 12.6 иллюстрирует, как выглядит регрессия, если условие непрерывности выполнено. Как и на рис. 12.3, заштрихованные точки – это данные, которые мы фактически наблюдаем. Сплошные линии, проведенные через них, представляют собой функции среднего потенциального результата (для соответствующего значения назначенного воздействия). Пустые кружки – это данные, которые мы не никогда наблюдаем. Пунктирные линии, проведенные через них, представляют собой функции среднего потенциального результата (опять же, для соответствующего значения назначенного воздействия). Непрерывность обеспечивается, поскольку эти функции не имеют скачков. То есть средние потенциальные результаты как при воздействии, так и при отсутствии воздействия непрерывны в пороговой точке. Все, что меняется на пороге, – это переход от объектов без воздействия к объектам под воздействием (т. е. от контрольной группы к экспериментальной).

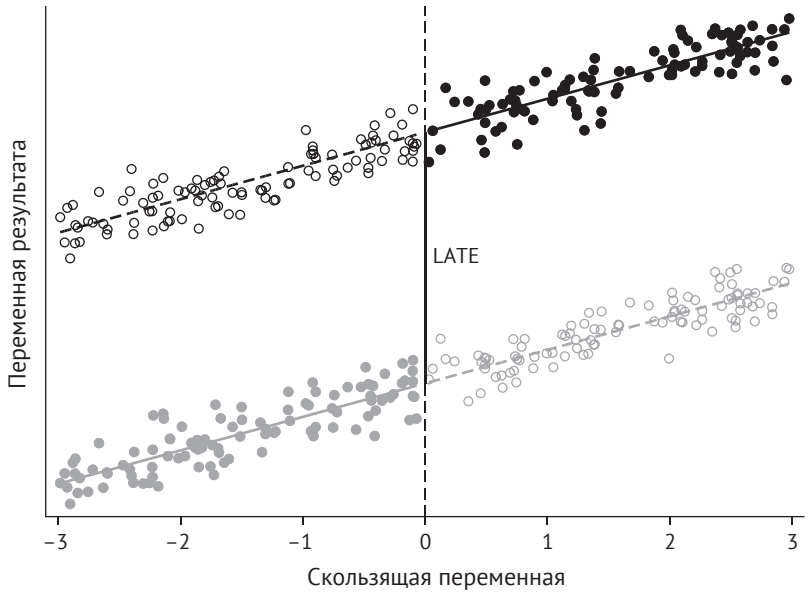


Рис. 12.6. Случай, когда функции средних потенциальных результатов являются непрерывными

Важно отметить, что если непрерывность сохраняется, то разрыв между серой и черной линией в пороговой точке фактически является значением LATE для этого порога, а это именно то, что нам нужно.

Но что, если непрерывность не сохраняется и потенциальные результаты будут выглядеть так, как показано на рис. 12.7?

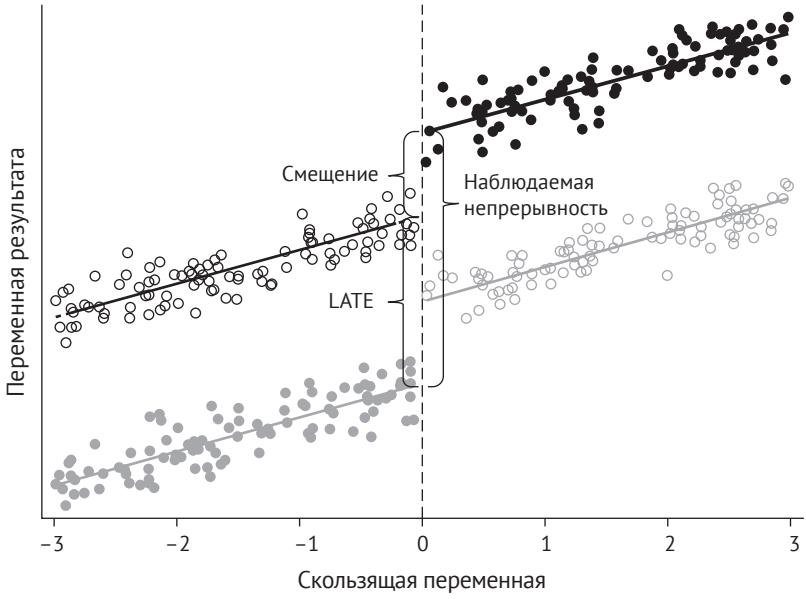


Рис. 12.7. Случай, когда функции средних потенциальных результатов не являются непрерывными в пороговой точке

Истинный средний эффект воздействия в пороговой точке представляет собой разницу между сплошными серыми и пустыми черными кружками в этой точке. (Вы также можете определить его как разницу между сплошными черными и пустыми серыми кружками.) Но на пороге потенциальные результаты резко возрастают даже при отсутствии изменения в воздействии. Неважно, по какой причине, но что-то, помимо воздействия, меняется точно в пороговой точке. Как следствие, не весь наблюдаемый разрыв – т. е. скачок между сплошными серыми и сплошными черными точками – является результатом изменения воздействия. Отчасти это результат каких-то других изменений. Таким образом, этот разрыв представляет собой смещенную оценку LATE (в данном случае это сильно завышенная оценка истинного эффекта воздействия), поскольку разрыв включает в себя как эффект изменения воздействия, так и эффект остальных изменений. Таким образом, без непрерывности в пороговой точке разрывная регрессия даст смещенную оценку локального среднего эффекта воздействия.

Когда дело доходит до реализации разрывной регрессии, аналитик может пойти разными путями. Однако, если действовать правильно, как только исследователь убедился в непрерывности потенциальных результатов в пороговой точке, его задача ясна. Используя методы, которые мы уже обсуждали (например, регрессию), ему просто нужно сгенерировать объективные оценки двух величин – среднего результата с воздействием и без воздействия в пороговой точке.

Возникает вопрос, когда схема разрывной регрессии действительно уместна. Это зависит от того, когда непрерывность в пороговой точке является правдоподобной, а когда нет. Стоит отметить, что требование непрерывности для достоверной оценки причинно-следственных связей менее строгое, чем можно было ожидать. Например, не требуется, чтобы воздействие назначалось случайным образом. В нашем случае со стипендией мы смогли использовать разрывную регрессию, хотя для каждого отдельного студента назначение воздействия было детерминированным (т. е. не было никакой случайности вообще). Непрерывность также не требует, чтобы результат не был связан со скользящей переменной. Опять же, в нашем примере со стипендией скользящая переменная отражает подлинную академическую успеваемость и, таким образом, положительно коррелирует с будущими доходами. Наконец, разрывная регрессия не требует, чтобы объекты воздействия не имели влияния на значение скользящей переменной или чтобы испытуемые не знали порогового значения. В нашем примере со стипендией студенты могли делать что угодно, чтобы повлиять на скользящую переменную (например, усерднее учиться, больше заниматься общественными работами).

Так что же может пойти не так, если непрерывность не сохранится?

Предположим, что объекты экспериментального воздействия имеют чрезвычайно точный контроль над значением скользящей переменной, так что определенные типы объектов могут группироваться чуть выше или чуть ниже порога. Потенциально это может быть проблемой. В нашем примере со стипендией вызывает беспокойство тот факт, что более привилегированные или более амбициозные студенты имеют лучшую информацию о системе оценок и могут сделать ровно столько, чтобы превысить порог (и ни шага больше). Нас также беспокоит, что у комитета могут быть особые причины предоставить стипендии студентам с определенными характеристиками (например, детям доноров, спортсменам, определенным расовым или этническим группам) и немного манипулировать баллами или порогом, чтобы получить желаемый результат. В обоих этих случаях люди, оказавшиеся чуть выше порога, не будут сравнимы с теми, кто находится чуть ниже. Вместо этого они были бы отсортированы (сами по себе или другими) вокруг порога по другим базовым характеристикам, которые имеют значение для результатов. В этом случае разрыв регрессии не дает объективной оценки причинного эффекта.

Что-то может пойти не так даже без сортировки по порогу просто потому, что на пороге меняется не только статус воздействия. Вот довольно интересный пример из реальной жизни. Во Франции (и во многих других странах) зарплата мэра зависит от численности населения города. Например, по закону зарплата мэра увеличивается, если в городе проживает более 3500 жителей.

Похоже, это возможность использовать разрывную регрессию, чтобы узнать о влиянии зарплаты мэра на самые разные показатели. Например, мы можем узнать, станет ли управление городами лучше или будут ли выборы более конкурентными, если мэрам станут платить больше. Для любого из этих исходов наблюдаемым воздействием является заработная плата мэра. Скользящей переменной является численность населения. И мы знаем, что по закону наблюдается разрывный скачок в воздействии, когда скользящая переменная пере-

секает порог в 3500 жителей. Естественно, города с населением 3400 жителей и города с населением 3600 жителей в среднем очень схожи.

Выглядит хорошо, разве нет? Но есть проблема с непрерывностью. Проблема не в том, что города обусловливают зарплату мэра численностью населения. Дело в том, что зарплата мэра не единственная характеристика городского управления, которая меняется по закону при достижении порога в 3500 жителей. К другим характеристикам, которые тоже меняются в этой точке, относится размер городского совета, количество заместителей мэра, правила выборов, процесс рассмотрения бюджета, требования гендерного паритета к городскому совету и т. д. Таким образом, любой разрыв в результатах при пороге в 3500 жителей не может служить объективной оценкой влияния зарплаты мэра, поскольку другие характеристики, которые могут иметь значение для этих результатов, также скачком изменяются при достижении порога.

Очевидно, что, прежде чем интерпретировать результаты разрывной регрессии как объективную оценку причинно-следственной связи, важно оценить правдоподобность предположения о непрерывности. Есть несколько способов сделать это. Самое главное – мыслить предметно. Лучший способ обнаружить возможные нарушения непрерывности – это знать предметную область в деталях, чтобы вовремя заметить манипуляции, подгонку, намеренную сортировку и другие изменения, возникающие в пороговой точке. В нашем примере со стипендией, если бы вы лично присутствовали на заседании комитета или обладали глубокими знаниями о тех характеристиках, которые комитет требовал от получателей стипендии, вы были бы больше готовы оценить правдоподобность предположения о непрерывности, чем при отсутствии знаний предмета. Существуют также другие виды анализа, который можно провести, чтобы подтвердить предположение о непрерывности. Например, аналитик может посмотреть на графики измеримых характеристик перед воздействием и увидеть, есть ли у них дискретные скачки в пороговой точке. Если многие измеримые характеристики выглядят непрерывными в пороговой точке, это укрепляет нашу уверенность в том, что и неизмеренные исходные характеристики объектов также являются непрерывными. Можно также посмотреть на распределение самой скользящей переменной. Если мы обнаружим выраженный дисбаланс – т. е. значительно больше объектов, у которых значение скользящей переменной слегка превышает порог или наоборот, – это явный признак какой-то манипуляции, которая нарушает непрерывность.

Насколько серьезно нарушение предположения о непрерывности, зависит от деталей проблемы. Если есть лишь небольшой дисбаланс или разрыв в базовых характеристиках, то оценка по методу разрывной регрессии будет смещена, но, возможно, лишь незначительно. А если у исследователя много данных и он может сосредоточиться только на объектах, очень близких к пороговому значению, баланс должен быть чрезвычайно точным, чтобы не искажать результаты. Например, если мы оцениваем разрывную регрессию в нашем примере про стипендию, используя данные о студентах с баллами в диапазоне 940–949 и в диапазоне 950–959, нас больше беспокоит точность соблюдения баланса между тем и другим диапазоном, чем в ситуации, когда у нас достаточно данных, чтобы рассматривать только студентов, набравших 949 или 950 баллов.

Сохраняется ли непрерывность в разрывных регрессиях для анализа выборов?

Как мы обсуждали ранее в этой главе, выборы являются отличным рабочим полигоном для проектов разрывной регрессии, поскольку они имеют четкую скользящую переменную и четкий порог для победы на выборах. Неудивительно, что метод разрывной регрессии использовали во многих исследованиях влияния выборов на результаты, начиная с пожертвований на избирательную кампанию и насилия, связанного с наркотиками, и заканчивая выдвижением радикального или умеренного кандидата. Поэтому важно критически подумать о том, действительно ли разрывная регрессия является хорошим исследовательским методом в области выборов.

Давайте вспомним, какие условия нужно выполнить, чтобы разрывная регрессия дала объективную оценку причинно-следственной связи. Нам необходимо, чтобы все остальное, что имеет значение для изучаемого результата, было непрерывным в пороговой точке. Это гарантирует, что округа, где соответствующий кандидат (например, радикал) победил с минимальным опережением, в среднем сопоставимы с округами, где соответствующий кандидат проиграл с минимальным отставанием. При любом применении метода разрывной регрессии, включая выборы, всегда важно задаться вопросом, выполнимо ли это условие. И действительно, некоторые исследования утверждают, что в некоторых сценариях выборов непрерывность может быть нарушена. Опасения связаны с манипулированием результатами на закрытых выборах. Например, в исследовании Холла о последствиях выдвижения кандидата с крайними убеждениями, возможно, партийное руководство отдает предпочтение умеренным. Если у руководства есть способы вмешаться (скажем, оказав давление на должностных лиц, ответственных за пересчет голосов), чтобы добиться желаемых результатов выборов, оно может сделать это в пользу умеренных кандидатов. В своем исследовании Холл показывает, что это не так.

Но в другой ситуации, когда речь идет о Палате представителей США после Второй мировой войны, некоторые данные свидетельствуют о проблемах с непрерывностью. В своих исследованиях ученые пытались использовать разрывную регрессию для оценки преимущества действующей партии: насколько лучше действующая партия (лидирующая по числу представителей), чем партия, которая претендует на ее место, при прочих равных условиях? Исследователь может сравнить вероятность победы демократа на выборах в ситуациях, когда демократ выиграл или проиграл предыдущие выборы с незначительным перевесом, в надежде оценить влияние предыдущего результата выборов на результаты последующих выборов. Чтобы обеспечить достоверность выводов, нужно соблюсти условие непрерывности в пороговой точке – вероятность победы демократа над республиканцем на следующих выборах не изменилась бы скачкообразно, если бы не было влияния результатов предыдущих выборов. Но есть основания думать, что это не так. В частности, на выборах в Палату представителей, на которых был получен перевес менее 0.25 % голосов, действующая партия статистически имеет больше шансов на победу, чем партия-кандидат. Если это происходит потому, что партии способны манипулировать результатами выборов, балансирующих на грани равенства голосов, то есть

вероятность, что даже очень близко к 50-процентному порогу мы не сможем провести корректное сравнение будущих результатов выборов в округах, где партия выиграла с минимальным перевесом. В чем же причина?

Девин Коги и Джас Сехон, написавшие исследовательский отчет об этом явлении, утверждают, что имеющиеся данные указывают на манипулирование выборами: действующие власти очень точно знают ожидаемую долю голосов и действуют стратегически в день выборов или чуть раньше таким образом, чтобы перевалить через половину голосов в условиях очень равной избирательной гонки. Однако, чтобы поверить в это, вы должны допустить, что действующие политики обладают невероятным чутьем и способны различать ситуации, когда они ожидают, что процент их голосов окажется между 49.75 и 50.0 или между 50.0 и 50.25. Реальные избирательные команды вряд ли достигли такого уровня точности в своих прогнозах по выборам. Поэтому целенаправленное воздействие избирательной команды вряд ли может быть объяснением. Чем еще можно объяснить дисбаланс в пользу действующей партии? Скорее всего, это тот случай, когда шум приводит к ложному положительному результату, во многом как в случае с осьминогом Паулем в главе 7. Когда Энтони и четыре соавтора повторили те же тесты, что и Коги и Сехон, но для 20 различных избирательных условий в нескольких странах, Палата представителей США послевоенного созыва была единственной, в которой присутствовал такой дисбаланс. Поэтому мы полагаем, что разрывная регрессия выборов на самом деле является хорошим исследовательским проектом для изучения причинно-следственных связей в политике.

Несоблюдение условий и нечеткая разрывная регрессия

До сих пор мы говорили об использовании модели разрывной регрессии, когда воздействие полностью определяется скользящей переменной и порогом. В этом случае мы иногда говорим, что используем *схему с резким разрывом регрессии*.

Но, как и в экспериментах, иногда возникают проблемы несоблюдения правил в ситуациях, которые в остальном подходят для разрывной регрессии. То есть воздействие может зависеть от того, по какую сторону порога находится скользящая переменная, но эта зависимость будет не полностью детерминированной. Помимо участников, которые аккуратно соблюдают правила (*исполнители*), есть те, кто ни при каких условиях не соглашаются принять воздействие (объекты со значениями скользящей переменной выше порогового значения, но отказавшиеся испытать воздействие – мы зовем их *отрицателями*), и есть те, кто всегда испытывают воздействие (объекты со значениями скользящей переменной ниже порогового значения, но все равно испытывающие воздействие – мы зовем их *активистами*).

Когда встречается такое несоблюдение условий, необходимо объединить метод разрывной регрессии с методом инструментальных переменных, который мы обсуждали в главе 11. Мы делаем это, используя факт того, по какую сторону порога находится скользящая переменная, в качестве инструмента для назначения воздействия. Этот подход иногда называют *схемой нечеткой разрывной регрессии* (fuzzy regression discontinuity design). В качестве иллюстрации применения нечеткой разрывной регрессии рассмотрим следующий пример.

Бомбардировки во Вьетнаме

Классический вопрос борьбы с повстанцами заключается в том, является ли насилие со стороны противников повстанцев, ведущее к гибели мирных жителей, а также комбатантов, продуктивным или контрпродуктивным с точки зрения подавления сопротивления. Мелисса Делл и Пабло Керубин выполнили некоторый количественный анализ, рассматривая стратегию бомбардировок США во время войны во Вьетнаме.

Во Вьетнаме Соединенные Штаты провели обширные массированные бомбардировки, пытаясь подавить партизанские силы на севере Вьетнама. Делл и Керубин хотели оценить, сработали ли эти бомбардировки.

Для ответа на этот вопрос они могли бы провести одно сравнение: были ли повстанцы менее активны в тех частях Вьетнама, которые подверглись большему количеству бомбардировок. Но если вы поразмыслите критически, то увидите, что такое сравнение не является корректным. Очевидно, что Соединенные Штаты с большей вероятностью бомбили именно те места, где повстанцы проявляли высокую активность, и в этом случае возникнет проблема обратной причинно-следственной связи.

Чтобы лучше оценить последствия бомбардировки, Делл и Керубин использовали метод разрывной регрессии. История, лежащая в основе их схемы, весьма познавательна.

Во время войны во Вьетнаме министр обороны Роберт Макнамара был одержим количественными оценками. Макнамара был пионером в использовании количественного анализа операционной деятельности во время своего пребывания на посту президента Ford Motor Company. А в Министерстве обороны он окружил себя группой «вундеркиндов» и большой командой ученых-компьютерщиков, экономистов и исследователей с целью предоставить точные, научные и количественные рекомендации специалистам по военному планированию и военным.

Одним из таких проектов стала оценивающая система «Гамлет» (Hamlet Evaluation System, HES). В этом проекте были собраны ответы на огромную массу ежемесячных и ежеквартальных опросников о безопасности, политике и экономике. Данные были собраны региональными сотрудниками США и Южного Вьетнама, которые получили информацию, посещая деревни. Ответы на вопросы вводились с помощью перфокарт в центральный компьютер, а затем сложный алгоритм преобразовывал их в непрерывную оценку от 1 до 5, которая должна была характеризовать безопасность деревни. Однако эти необработанные результаты никогда и никуда не передавались из компьютера. Ни один человек никогда их не видел. Вместо этого компьютер округлял оценки до ближайшего целого числа, так что аналитики и командование всегда видели только оценки A, B, C, D или E. Считалось, что более высокие буквенные оценки соответствуют большей безопасности деревни. Эти оценки помогали определить, какие деревни следует подвергнуть бомбардировке, – при этом бомбардировки чаще всего были нацелены на деревни, получившие худшие оценки.

Делл и Керубин смогли реконструировать алгоритм и, используя рассекреченные данные, восстановить лежащие в его основе непрерывные оценки. Это подготовило их к использованию модели разрывной регрессии.

Рассмотрим деревни с оценками в диапазоне 1.45–1.55. Некоторые из этих деревень получили оценку чуть ниже 1.5 и символ E. Другие – оценку чуть

выше 1.5 и символ D. Оценка генерировалась на основе сложной (и во многом произвольной) комбинации ответов на 169 вопросов. Поэтому разницу, скажем, между оценками 1.49 и 1.51 можно считать достаточно условной. Следовательно, мы можем ожидать, что исходный уровень активности повстанцев в этих двух типах деревень один и тот же, т. е. мы можем рассчитывать, что потенциальные результаты в пороговой точке будут непрерывными.

Но воздействие – которое в данном случае означает бомбардировку со стороны Соединенных Штатов – в пороговой точке меняется скачкообразно. Американские военные планировщики никогда не видели лежащие в основе алгоритма непрерывные оценки. Все, что они видели, – это дискретную буквенную оценку. Они считали деревни, получившие оценку D, более опасными, чем деревни, получившие оценку E (и аналогично для D против C, C против B и B против A). Таким образом, они с большей вероятностью бомбили деревни с более низкими буквенными оценками.

Рисунок 12.8 показывает, что это было именно так. Горизонтальная ось отражает скользящую переменную – расстояние до первого десятичного знака оценки деревни относительно 0.5. Оценки «Гамлета», у которых значение скользящей переменной отрицательное (поскольку первый десятичный знак в их оценке был ниже 0.5), округлялись до ближайшей буквенной оценки, а те, у которых значение скользящей переменной положительное, округлялись в большую сторону.

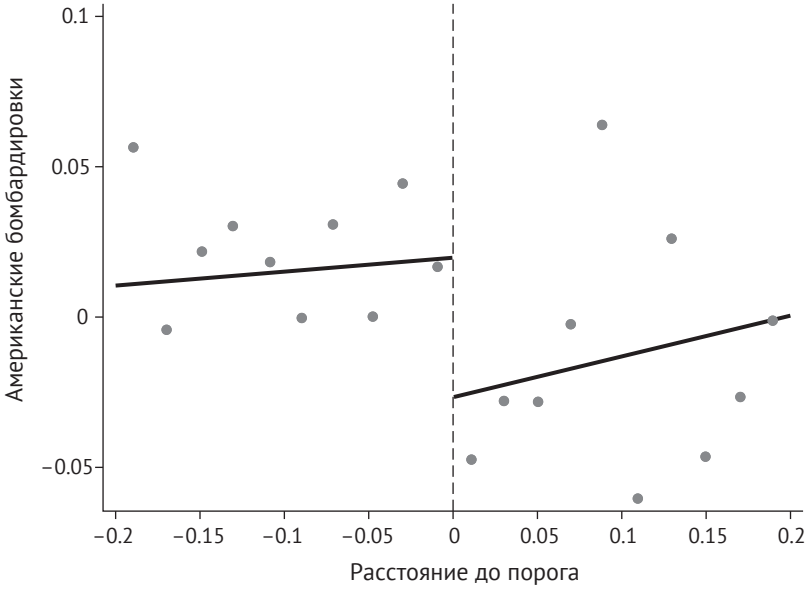


Рис. 12.8. Деревни, которые получили незначительно более высокие оценки в системе «Гамлет», подвергались бомбардировкам реже, чем деревни, получившие незначительно худшие оценки

Вертикальная ось измеряет частоту, с которой данная деревня подвергалась бомбардировкам после вычисления оценок. Серые точки соответствуют средним значениям оценок многих деревень с аналогичными значениями

скользящей переменной. Темные линии соответствуют отдельным регрессиям по обе стороны от порога. На рисунке показан прерывистый скачок частоты бомбардировок США на пороговом уровне: деревни, которые получили незначительно более высокие оценки, подвергались бомбардировкам реже, чем деревни, получившие незначительно худшие оценки.

Учитывая такое прерывистое изменение воздействия, имеет смысл использовать метод разрывной регрессии для оценки влияния бомбардировок на повстанческое движение. Рисунок 12.9 иллюстрирует эту идею. Горизонтальная ось – это та же самая скользящая переменная, что и выше. Но теперь вертикальная ось представляет собой интересующий нас результат: активность повстанцев в деревне по итогам оценивания. Как видно из диаграммы, неизбирательные бомбардировки оказались контрпродуктивными. На пороге наблюдается прерывистый спад активности повстанцев. Это означает, что деревни, которые подвергались бомбардировкам чаще (те, что слева от порога), демонстрировали большую активность повстанцев, чем аналогичные деревни, которые бомбили меньше.

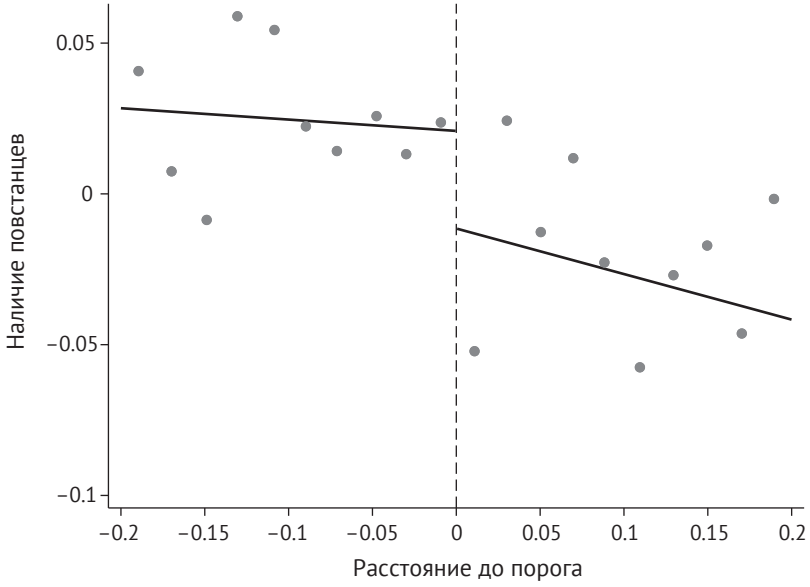


Рис. 12.9. Деревни, которые подверглись большему количеству бомбардировок, в дальнейшем демонстрировали более высокую активность повстанцев по сравнению с аналогичными деревнями, которые меньше подвергались бомбардировкам

Но обратите внимание, что здесь есть кое-что, отличающееся от нашей обычной истории о разрывной регрессии. Воздействие не является бинарным (существует континуум интенсивности бомбардировок), и переход от лучшего показателя к худшему не гарантировал увеличения количества бомбардировок. Оценка безопасности была лишь одним из факторов, влияющих на решение о бомбардировке. Следовательно, это не тот случай, когда при переходе через порог воздействие менялось от наличия к полному отсутствию. То есть, вероятно, имело место несоблюдение правил – деревни, воздействие на которые не зависело от того, по какую сторону порога оказалась их оценка.

Но мы знаем, что делать с теми, кто не соблюдает правила. Как мы обсуждали в главе 11, мы можем использовать инструментальный подход. Напомним, инструмент должен удовлетворять нескольким условиям.

1. **Экзогенность:** инструмент должен быть назначен случайным образом, что позволяет нам получить несмещенные оценки эффектов как первого этапа, так и редуцированной формы (reduced form, соответствует намерению воздействовать).
2. **Ограничение исключения:** все эффекты редуцированной формы должны возникать только в результате воздействия. Другими словами, у инструмента не должно быть другого пути влияния на эффект, кроме как через его влияние на воздействие.
3. **Наличие исполнителей:** в эксперименте обязательно должны быть участники, строго соблюдающие требования.
4. **Отсутствие провокаторов:** каким бы ни был знак эффекта первого этапа, не должно быть тех, для кого инструмент влияет на величину воздействия в противоположном направлении.

Как применить здесь подход инструментальных переменных? Идея состоит в том, чтобы использовать в качестве инструмента сторону порога, на которой находится наша скользящая переменная. Давайте убедимся, что это удовлетворяет четырем условиям, необходимым для инструмента.

Вся суть схемы разрывной регрессии – в экзогенности. Если потенциальные результаты непрерывны в пороговой точке, то разрывная регрессия позволяет нам получить несмещенную оценку как первого этапа (влияние инструмента на бомбардировку, как показано на рис. 12.8), так и редуцированной формы (влияние инструмента на деятельность повстанцев, как показано на рис. 12.9).

Ограничение исключения требует, чтобы сторона порога, на которой находится скользящая переменная, не влияла на деятельность повстанцев никаким другим путем, кроме как бомбардировкой. Здесь возникают вопросы, которые следует задать. Например, нам нужно спросить о том, использовались ли эти оценки для принятия каких-либо других военных или политических решений США. Если да, то инструмент не будет удовлетворять ограничению исключения.

Делл и Керубин предоставили два вида доказательств в поддержку наличия ограничения исключения. Во-первых, они повторили свой анализ методом разрывной регрессии для множества других видов военных операций, проводимых как американскими, так и южновьетнамскими вооруженными силами. Они не обнаружили никаких свидетельств того, что характер каких-либо других видов военных операций резко менялся в пороговой точке. Таким образом, маловероятно, что обнаруженные ими последствия являются результатом других военных действий, а не бомбардировок. Во-вторых, они изучили административную историю оценивающей системы «Гамлет». Этот обзор выявил мало свидетельств того, что баллы HES использовались для принятия каких-либо других политических решений. Единственным исключением является программа, направленная на изгнание повстанцев из наименее безопасных деревень. Но эта программа завершилась до периода выборки, охватываемого данными Делла и Керубина.

Требование о том, чтобы были аккуратные исполнители и не было провокаторов, является наиболее простым. Как из данных, так и из истории ясно, что буквенные оценки буквально и однозначно влияли на бомбардировку в со-

ответствии с воинской дисциплиной. И сложно представить, что нашлись бы деревни, которые сами потребовали бы их бомбить, вопреки высокой оценке безопасности. Однако, в отличие от наших предыдущих примеров, соблюдение условий не столь дискретно. Различные объекты (деревни) могут менять свой статус воздействия в зависимости от разного значения приборной переменной.

Учитывая все это, Делл и Керубин считают оправданным применение нечеткой модели разрывной регрессии – использование того, по какую сторону порога находится показатель безопасности деревни, в качестве инструментальной переменной для бомбардировки. При этом их оценка несколько громоздка, поскольку она отражает локальность как разрывной регрессии, так и инструментальной переменной. В частности, они оценивают локальный средний эффект бомбардировки на активность повстанцев для деревень с баллами, близкими к пороговому (LATE из разрывной регрессии), уровень бомбардировок которых реагирует на этот балл (CATE из инструментальной переменной)[1]. При этом они обнаруживают, что бомбардировка была контрпродуктивной. Для таких деревень переход от отсутствия бомбардировок к среднему уровню бомбардировок увеличивал вероятность активности повстанцев в деревне на 27 процентных пунктов.

Мотивация и успех

Давайте закончим главу забавным примером применения разрывной регрессии. Джона Бергер и Девин Поуп применили этот метод для оценки влияния психологической мотивации на результативность игроков. Они проанализировали более 18 000 профессиональных баскетбольных матчей, чтобы проверить, приводит ли мотивация отставания и необходимости догонять к более высоким результатам, чем самоуспокоенность от того, что вы впереди и просто необходимо удержать лидерство. Их скользящая переменная – это разница в очках команды хозяев в перерыве между таймами, и они проверяют, меняется ли вероятность победить в игре скачкообразно, когда команда хозяев переходит от незначительного отставания к незначительному лидерству.

На рис. 12.10 показаны результаты. Как и следовало ожидать, разница в очках в перерыве между таймами коррелирует с вероятностью окончательной победы в игре. Когда к перерыву хозяева поля опережают гостей на 10 очков, они выигрывают примерно в 85 % случаев, но когда они отстают на 10 очков, то выигрывают только в 25 % случаев. Это вполне логично, поскольку некоторые команды лучше, чем другие: хорошие команды с большей вероятностью будут впереди в середине матча и с большей вероятностью выиграют игру. Однако более интересным является сравнение, когда к перерыву между таймами разница в счете близка к нулю. Между командами, которые опережают или отстают всего

[1] Ситуация еще больше усложняется тем, что для каждой деревни не просто соблюдается или не соблюдается правило воздействия. Потенциально существует континуум соблюдения, при котором одна и та же инструментальная переменная увеличивает количество бомбардировок одних деревень значительно, других – меньше и т. д. Следовательно, вместо того чтобы думать о среднем эффекте воздействия при максимальном выполнении правила, нам на самом деле нужно думать о средневзвешенном эффекте воздействия, где каждая деревня взвешивается в соответствии со степенью, в которой бомбардировки отреагировали на рейтинг безопасности в данном случае.

на 1 очко к середине матча, в среднем существует очень небольшая разница в качестве игры. Тем не менее, оказывается, у хозяев поля больше шансов на победу, когда они отстают на 1 очко к перерыву, чем когда они опережают на 1 очко. Разрывная регрессия Бергера и Поупа показывает, что незначительное отставание увеличивает вероятность победы хозяев поля на 6 %! Возможно, знаменитые вдохновляющие речи тренера в перерыве действительно работают.

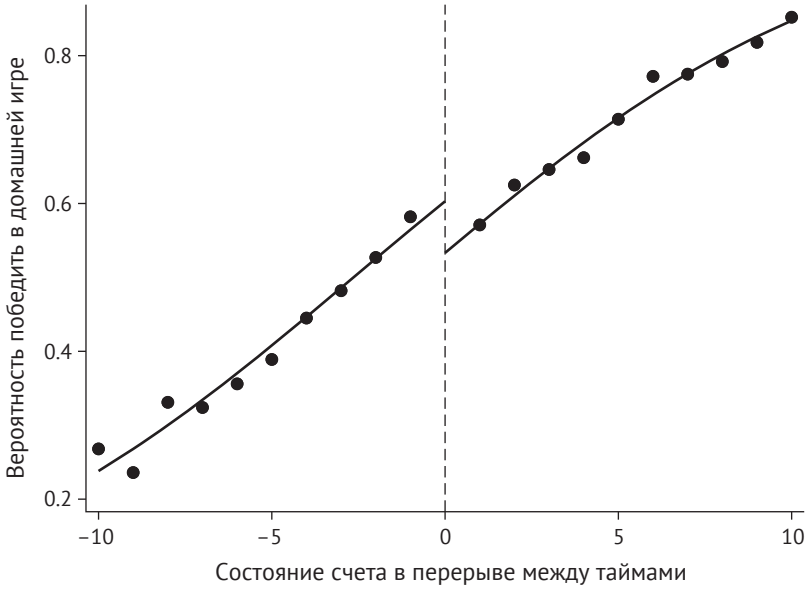


Рис. 12.10. Влияние преимущества или отставания в перерыве на победу в игре

Подведение итогов

Если нам известно, что наблюдаемое воздействие зависело (по крайней мере частично) от порогового значения некоторой переменной, то у нас есть возможность получить достоверную оценку эффекта воздействия в пороговой точке при помощи разрывной регрессии.

Подобные ситуации возникают чаще, чем вы думаете. Предположим, вы работаете в компании по производству детского питания, которая просит вас оценить эффект от ее телевизионной рекламы. Вероятно, вы не сможете убедить отдел маркетинга рандомизировать места размещения рекламы; они предпочтут размещать рекламу в местах, где она будет иметь наибольший эффект. Но, возможно, они уже решили транслировать телевизионную рекламу на всех медиарынках, где более 3 % домохозяйств имеют младенца. Это прекрасная возможность для разработки разрывной регрессии. Ничто не было случайным, отдел маркетинга все равно делал то, что хотел, но у вас есть возможность узнать об эффективности рекламы, сравнив потребление детского питания в местах чуть выше и чуть ниже этого 3-процентного порога.

Еще одна возможность получить достоверные оценки причинно-следственных связей без какой-либо рандомизации – это когда методы воздействия меняются для одних объектов и остаются неизменными для других. В таких

случаях может оказаться целесообразной схема разности различий, которая станет темой следующей главы.

Ключевые термины

- **Скользящая переменная:** переменная, которая определяет статус воздействия на объекты в зависимости от того, по какую сторону порога находится значение переменной для каждого объекта.
- **Схема разрывной регрессии:** схема исследования для оценки причинного эффекта, который оценивает прерывистый скачок результата по обе стороны от порога, определяющего назначение воздействия.
- **Непрерывность в пороговой точке:** требование, чтобы средние потенциальные исходы не менялись скачкообразно на пороге, который определяет назначение воздействия. Если непрерывность на пороге не сохраняется, то модель разрывной регрессии не обеспечивает несмещенную оценку локального среднего эффекта воздействия.
- **Четкая разрывная регрессия:** схема разрывной регрессии, в которой назначение воздействия полностью определяется тем, по какую сторону порога находится текущая переменная.
- **Нечеткая разрывная регрессия:** схема исследования, сочетающая в себе разрывную регрессию и инструментальную переменную. Нечеткая разрывная регрессия используется, когда назначение воздействия лишь частично определяется тем, на какой стороне порога находится скользящая переменная. Поэтому исследователь использует знание того, по какую сторону порога находится скользящая переменная, в качестве инструмента для назначения воздействия. В этом случае непрерывность в пороговой точке гарантирует выполнение предположения об экзогенности инструментальной переменной. Но нам все еще приходится беспокоиться об ограничении исключения и других предположениях инструментальной переменной.

Упражнения

12.1. Штат Аляска просит вас оценить влияние новой системы автоматической регистрации избирателей на их явку в избирательные пункты. Впервые система была запущена в 2017 г., но вам сообщают, что, к сожалению, изначально у правительства не было ресурсов, чтобы применить систему ко всем жителям штата. В результате сначала они применили автоматическую регистрацию только к людям, переехавшим на Аляску в течение двух лет с момента запуска системы, но еще не применили ее к людям, переехавшим на Аляску раньше. Они обеспокоены тем, что это может стать ограничением для вашего исследования, и извиняются за то, что не смогли реализовать эту политику для всех, но все равно надеются, что вы сможете помочь. Как бы вы поступили и каким образом оценили бы эффект автоматической регистрации избирателей на Аляске?

12.2. Федеральное правительство США субсидирует обучение студентов в колледжах посредством грантов Пелла. Человек имеет право на получение

гранта Пелла, если его семейный доход составляет менее 50 000 долл. США в год.

a) Как вы могли бы потенциально использовать эту информацию и внедрить модель разрывной регрессии, чтобы оценить влияние посещения колледжа на будущие доходы?
b) Какая это разновидность разрывной регрессии – четкая или нечеткая?
c) Какие данные вы хотели бы иметь в своем распоряжении?
d) Что здесь является скользящей переменной?
e) Что здесь является воздействием?
f) Что здесь является инструментом (если он есть)?
g) Что является результатом?
h) Какие предположения вам придется сделать, чтобы получить достоверные оценки?

12.3. Загрузите файл ChicagoCrimeTemperature2018.csv и связанный с ним файл README.txt, который описывает переменные в этом наборе данных, по адресу *http://www.press.princeton.edu/thinking-clearly*. Это те же данные о преступности и температуре в Чикаго в разные дни 2018 г., которые мы рассматривали в главах 2 и 5. Представьте себе, что полицейское управление Чикаго ввело в 2018 г. правило, согласно которому они прекращали патрулирование в те дни, когда средняя температура должна была быть ниже 32 °F, что соответствует 0 °C (и предположим, что у них очень хорошие прогнозы погоды, поэтому они могут очень точно предсказать в начале дня среднюю температуру на этот день). Их логика в том, что полицейским менее приятно находиться на улице в холодную погоду, да и преступности в холодные дни меньше. Используйте эту (выдуманную) информацию, чтобы оценить влияние работы полиции на уровень преступности.

a) Полезным первым шагом при реализации метода разрывной регрессии является создание вашей собственной скользящей переменной, где порог расположен в точке 0. Масштабируйте температуру так, чтобы пороговое значение было равно 0. Для этого создайте новую переменную, значение которой равно текущей температуре –32.
b) Нам также потребуется создать переменную воздействия. Создайте переменную, которая принимает значение 1, если в этот день действовало правило, и 0, если не действовало.
c) Часто бывает полезно внимательно присмотреться к данным, прежде чем проводить формальный количественный анализ. Постройте диаграмму распределения данных, на которой уровень преступности будет отложен по вертикальной оси, а температура – по горизонтальной. Сосредоточьтесь только на днях, когда температура была в пределах 10° от порогового значения правила, и проведите линию через пороговое значение. Похоже ли, что в области порога есть разрыв?
d) Существует несколько различных способов формальной реализации разрывной регрессии. Самый простой – сосредоточиться на узком окне вокруг порога и просто сравнить средний результат по обе стороны. Ориентируясь только на те дни, когда температура находилась в пределах 1° от порогового значения, вычислите среднее количество

преступлений чуть выше и чуть ниже порога и найдите разницу. Обратите внимание, что вы можете (если хотите) сделать это за один шаг с помощью регрессии.

e) Какие опасения у вас могут возникнуть в связи с приведенным выше наивным подходом? Подумайте о компромиссах, с которыми вы сталкиваетесь, выбирая полосу пропускания. Как изменится ваша оценка, если вы используете полосу пропускания 10° вместо 1°? Почему?

f) Другая стратегия заключается в использовании локального линейного подхода. Для дней, когда температура была менее чем на 5° ниже порогового значения, постройте регрессию преступности по скользящей переменной и вычислите прогнозируемое значение в пороговой точке. (Подсказка: поскольку вы изменили масштаб скользящей переменной, это будет точка пересечения с вертикальной осью координат.) Сделайте то же самое для дней, когда температура была менее чем на 5° выше порогового значения. Сравните эти два прогнозируемых значения. (Обратите внимание, что это также можно сделать с помощью одиночной регрессии, как описано ранее.)

g) Какие преимущества имеет этот локальный линейный подход по сравнению с наивным подходом?

h) Вы также можете постараться учесть нелинейность связи между скользящей переменной и результатом. Сгенерируйте новые переменные, соответствующие второй и третьей степени скользящей переменной. Постройте регрессию количества преступлений от правила, скользящей переменной, скользящей переменной в квадрате и скользящей переменной в третьей степени. Рассматривайте только наблюдения в пределах 10° от порога. Интерпретируйте найденный коэффициент регрессии при переменной правила.

i) Каковы достоинства и недостатки этого полиномиального подхода по сравнению с предыдущими подходами?

Дополнительное чтение и ссылки

Исследование зависимости корпоративных доходов от взносов в избирательные кампании:

Anthony Fowler, Haritz Garro, and Jorg L. Spenkuch. 2015. *Quid Pro Quo? Corporate Returns to Campaign Contributions.* Journal of Politics 82 (3): 844–58.

Обсуждение потенциальных нарушений непрерывности в исследованиях политических изменений, связанных с порогом численности населения:

Andrew C. Eggers, Ronny Freier, Veronica Grembi, and Tommaso Nannicini. 2018. *Regression Discontinuity Designs Based on Population Thresholds: Pitfalls and Solutions.* American Journal of Political Science 62 (1): 210–29.

Исследование последствий избрания радикального кандидата, по сравнению с умеренным, на первичных выборах:

Andrew B. Hall. 2015. *What Happens When Extremists Win Primaries?* American Political Science Review 109 (1): 18–42.

Исследования обоснованности применения разрывной регрессии в моделировании поведения избирателей:

Devin Caughey and Jasjeet S. Sekhon. 2011. *Elections and the Regression Discontinuity Design: Lessons from Close U.S. House Races, 1942–2008*. Political Analysis 19 (4): 385–408;

Andrew C. Eggers, Anthony Fowler, Andrew B. Hall, Jens Hainmueller, and James M. Snyder, Jr. 2015. *On the Validity of the Regression Discontinuity Design for Estimating Electoral Effects: New Evidence fromOver 40,000 Close Races*. American Journal of Political Science 59 (1): 259–74.

Исследование бомбардировок США во время войны во Вьетнаме:

Melissa Dell and Pablo Querubin. 2018. *Nation Building through Foreign Intervention: Evidence from Discontinuities in Military Strategies*. Quarterly Journal of Economics 133 (2): 701–64.

Исследование о влиянии отставания по очкам в середине матча на победу:

Jonah Berger and Devin Pope. 2011. *Can Losing Lead to Winning?* Management Science 57 (5): 817–27.

Глава 13

Метод разности различий

О чем эта глава

- Еще одна ситуация, когда мы можем объективно оценить причинные эффекты, – применение разного воздействия к разным объектам в разное время. В этом случае может оказаться целесообразным *метод разности различий* (*метод сравнения различий*, difference-in-differences).
- Метод разности различий эффективно учитывает все факторы, которые не меняются с течением времени, даже если их невозможно наблюдать или измерить.
- Модели разности различий часто могут быть полезны в качестве интуитивной проверки – простого способа узнать, насколько на самом деле убедительны доказательства некоторых причинно-следственных связей.

Введение

Разрывная регрессия не единственная креативная исследовательская идея, позволяющая нам найти причинно-следственную связь в отсутствие эксперимента. Когда одни объекты со временем меняют статус воздействия, а другие нет, мы можем узнать о наличии причинно-следственных связей, используя стратегию разности различий.

В основе этого метода лежит простая идея. Допустим, мы хотим узнать эффект от применения какого-то правила. Мы можем найти государства (или страны, или города, или отдельных лиц, или любую другую подходящую единицу наблюдения), для которых изменилось интересующее нас правило (т. е. воздействие), и измерить их показатели до и после изменения. Конечно, есть опасение, что результаты систематически меняются с течением времени по другим причинам. Но мы можем обнаружить и компенсировать это изменение, сравнивая изменение результатов для объектов, к которым применялось правило (воздействие), с изменением результатов для объектов, к которым оно не применялось. Если окажется, что эффект наблюдается только для объектов, испытавших изменение воздействия (экспериментальная группа), то мы можем использовать объекты контрольной группы в качестве источника сравнения, чтобы учесть долговременные тенденции. Таким образом, наша оценка причинного эффекта от изменения правила будет основываться на любых изменениях состояния объектов, подвергшихся воздействию, сверх той исходной тенденции, которую мы оценили по объектам, не подвергнутым воздействию. Этот подход называется методом разности раз-

личий, потому что мы сначала получаем *различия* (или изменения) характеристик с течением времени как для объектов, которые испытали воздействие, так и для объектов, которые его не испытывали. Затем находим *разность* этих различий. Таким образом, мы компенсируем путем вычитания изменения, возникшие с течением времени, и получаем в остатке чистый эффект воздействия

Как и в случае с моделью разрыва регрессии, сила подхода разности различий заключается в том, что он позволяет нам оценить причинные эффекты, даже если мы не можем рандомизировать воздействие или контролировать каждый возможный искажающий фактор. Но за все приходится платить. Схемы разности различий имеют свои собственные требования. Для разрывной регрессии нам нужна была непрерывность в пороговой точке. Для сравнения различий необходимо условие, которое мы только что описали: тенденция изменения состояния была бы в среднем одинаковой для всех объектов, если бы не изменение воздействия, коснувшееся некоторых объектов. Это условие часто называют *параллельностью трендов* (parallel trends).

Параллельность трендов

Имеет смысл подробнее разобраться в том, что на самом деле означает условие параллельности трендов. Как мы уже говорили, оценки разницы в различиях являются несмещенными до тех пор, пока линии тренда изменения состояний остаются параллельными при отсутствии каких-либо изменений в воздействии. Другими словами, соблюдение параллельности трендов на самом деле касается потенциальных исходов. При бинарном воздействии мы можем рассматривать состояние каждого объекта с воздействием и без него в каждом из двух периодов времени. Чтобы лучше уловить эту идею, давайте подумаем о потенциальных исходах для каждого объекта в каждый период времени. Мы будем называть эти два периода времени периодом *I* и периодом *II*. И давайте представим, что наша популяция разделена на две части: группу, которая при смене периода переходит от отсутствия воздействия к его наличию ($\mathcal{ИТ}$), и группу, которая остается без воздействия на протяжении обоих периодов ($\mathcal{ИИ}$).

Обозначим средний потенциальный исход в группе $\mathcal{G}$ в период p в соответствии со статусом воздействия следующим образом:

$$\overline{Y}^{p}_{T,\mathcal{G}}.$$

Мы наблюдаем состояние для выборки членов каждой группы в каждом периоде. Начнем с группы, для которой статус воздействия никогда не меняется ($\mathcal{ИИ}$). Если мы просто найдем среднее изменение состояния между двумя периодами, это даст нам оценку разницы состояний объектов без воздействия между двумя периодами:

$$\text{DIFF}_{\mathcal{ИИ}} = \underbrace{\overline{Y}^{II}_{0,\mathcal{ИИ}} - \overline{Y}^{I}_{0,\mathcal{ИИ}}}_{\text{средний тренд без воздействия для } \mathcal{ИИ}} + \text{Шум}_{\mathcal{ИИ}}.$$

Шум возникает из-за того, что мы рассматриваем выборку, а не всю совокупность.

Аналогично для группы, статус воздействия которой меняется ($\mathcal{UT}$), среднее изменение состояния между двумя периодами равно:

$$\text{DIFF}_{\mathcal{UT}} = \overline{Y}^{II}_{1,\mathcal{UT}} - \overline{Y}^{I}_{0,\mathcal{UT}} + \text{Шум}_{\mathcal{UT}}.$$

Разность различий – это в буквальном смысле разность этих двух изменений:

$$\text{Разность различий} = \text{DIFF}_{\mathcal{UT}} - \text{DIFF}_{\mathcal{UU}}.$$

Чтобы увидеть, где появляются параллельные тренды, перепишем $\text{DIFF}_{\mathcal{UT}}$, добавляя и вычитая из него $\overline{Y}^{II}_{0,\mathcal{UT}}$. Возможно, вы помните, что мы делали нечто подобное еще в главе 9, чтобы понять базовые различия. Хотя это преобразование выглядит странным, мы просим вас довериться нам на минуту. И, как в прошлый раз, этим преобразованием мы не причиняем никакого вреда, поскольку, прибавляя и вычитая один и тот же член, мы на самом деле просто прибавляем ноль. Сделав это, мы получим

$$\begin{aligned}\text{DIFF}_{\mathcal{UT}} &= \overline{Y}^{II}_{1,\mathcal{UT}} - \overline{Y}^{I}_{0,\mathcal{UT}} \\ &= \underbrace{\left(\overline{Y}^{II}_{1,\mathcal{UT}} - \overline{Y}^{II}_{0,\mathcal{UT}}\right)}_{\text{средний эффект воздействия для } \mathcal{UT}} + \underbrace{\left(\overline{Y}^{II}_{0,\mathcal{UT}} - \overline{Y}^{I}_{0,\mathcal{UT}}\right)}_{\text{средний тренд без воздействия для } \mathcal{UT}} + \text{Шум}_{\mathcal{UT}}.\end{aligned}$$

И снова наш алгебраический трюк оказался очень крутым. Теперь мы видим, что изменение с течением времени в группе $\mathcal{UT}$ состоит из трех факторов, которые соответствуют нашему любимому уравнению. Во-первых, это средний (период *II*) эффект воздействия для группы $\mathcal{UT}$. В главе 9 вы узнали, что он называется ATT – средний эффект воздействия на экспериментальные объекты. Мы можем считать его нашей оценкой. Во-вторых, существует тренд изменений, которые произошли бы в группе, даже если бы она не подвергалась воздействию. Мы можем рассматривать его как источник смещения, возникающего просто из-за наблюдения за тем, что происходит в группе $\mathcal{UT}$ до и после воздействия. Третьим слагаемым, как всегда, является шум.

Исходя из этого уравнения, теперь мы можем переписать уравнение разности различий с использованием ATT, средних трендов без воздействия для обеих групп и шума. Теперь становится ясно, что мы задумали – использовать долгосрочный тренд в группе $\mathcal{UU}$, чтобы попытаться устранить смещение, возникающее при прямом рассмотрении $\text{DIFF}_{\mathcal{UT}}$:

$$\begin{aligned}&\text{Разница различий} = DIFF_{\mathcal{UT}} - DIFF_{\mathcal{UU}} \\ &= \underbrace{\overline{Y}^{II}_{1,\mathcal{UT}} - \overline{Y}^{II}_{0,\mathcal{UT}}}_{ATT} + \underbrace{\overline{Y}^{II}_{0,\mathcal{UT}} - \overline{Y}^{I}_{0,\mathcal{UT}}}_{\text{средний эффект воздействия для } \mathcal{UT}} - \underbrace{\overline{Y}^{II}_{0,\mathcal{UU}} - \overline{Y}^{I}_{0,\mathcal{UU}}}_{\text{средний тренд без воздействия для } \mathcal{UU}} \\ &\qquad \underbrace{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad}_{\text{разность средних трендов}} \\ &\qquad + \underbrace{\text{Шум}_{\mathcal{UT}} - \text{Шум}_{\mathcal{UU}}}_{\text{Шум}}.\end{aligned}$$

Сейчас хорошо видно, что на самом деле означает параллельность трендов с точки зрения потенциальных результатов и нашего любимого уравнения. Разность различий равна ATT (оцениваемая величина) плюс разница между средними трендами при отсутствии воздействия для группы $\mathcal{UT}$ и группы $\mathcal{UU}$ (смещение) плюс шум. Когда разность различий дает нам несмещенную оценку ATT? Только в том случае, когда средние тренды в отсутствие воздействия одинаковы для обеих групп, так что разность средних трендов равна нулю.

Вот что значит параллельность трендов: *если обе группы останутся без воздействия, то изменение их состояния с течением времени будет одинаковым*. В этом случае, вычитая $\text{DIFF}_{\mathcal{UU}}$ из $\text{DIFF}_{\mathcal{UT}}$, мы устраняем общий тренд изменений, оставляя несмещенную оценку среднего эффекта воздействия (в периоде *II*) для объектов, статус воздействия которых изменился.

Обратите внимание: применяемые обозначения подчеркивают еще один тонкий момент. Разность различий не всегда позволяет оценить ATE. Она оценивает средний эффект воздействия для тех объектов, которые фактически меняют статус воздействия, т. е. ATT. Является ли это хорошей оценкой ATE, зависит от того, систематически ли различаются эффекты воздействия для объектов, которые меняют или не меняют статус воздействия. Но в любом случае это настоящий причинный эффект, и, по крайней мере, для некоторых приложений он действительно может представлять собой интересующую величину.

Два объекта и два периода

До сих пор мы рассуждали немного абстрактно. Теперь рассмотрим конкретный пример из классической работы Дэвида Карда и Алана Крюгера о влиянии минимальной заработной платы на занятость. Этот пример хорош, потому что он показывает, как работает разность различий в самой простой форме. В данном случае имеется только два объекта, два периода и одно изменение статуса воздействия.

Безработица и минимальная заработная плата

Кард и Крюгер хотели знать, приведет ли повышение минимальной заработной платы к увеличению безработицы. Их идея заключалась в том, чтобы воспользоваться тем фактом, что штат Нью-Джерси повысил минимальную заработную плату в начале 1992 г., а Пенсильвания, граничащая с Нью-Джерси, этого не сделала. Они собрали данные о среднем количестве сотрудников, трудоустроенных на эквивалент полной ставки (full-time equivalent employees, FTE), в заведениях быстрого питания (которые, как правило, платят минимальную заработную плату) как в Нью-Джерси, так и в Пенсильвании в январе 1992 г. (до того, как Нью-Джерси повысил минимальную заработную плату) и в ноябре 1992 г. (после того, как Нью-Джерси повысил минимальную заработную плату). Полученные данные сведены в табл. 13.1.

Таблица 13.1. Занятость в заведениях быстрого питания в Нью-Джерси и Пенсильвании в 1992 г.

	Январь 1992 г. *NJ и PA* *низкая минимальная зарплата*	**Ноябрь 1992 г.** *NJ – высокая минимальная зарплата* *PA – низкая минимальная зарплата*
Нью-Джерси (NJ)	20.44	21.03
Пенсильвания (PA)	23.33	21.17

Первое сравнение, которое можно провести, чтобы узнать о влиянии минимальной заработной платы на занятость, – это разница между уровнями занятости в Нью-Джерси и Пенсильвании в ноябре 1992 г. Ведь к ноябрю минимальная заработная плата в Нью-Джерси была выше, чем в Пенсильвании. Это сравнение показывает, что в заведениях быстрого питания Пенсильвании в среднем занято всего на 0.14 человек больше, чем в заведениях Нью-Джерси, что позволяет предположить, что более высокая минимальная заработная плата почти не повлияла на занятость.

Но это некорректное сравнение, поэтому мы не можем интерпретировать упомянутую разницу как эффект повышения минимальной заработной платы. Нью-Джерси и Пенсильвания могут отличаться во многих аспектах, которые имеют значение для трудоустройства, помимо минимальной заработной платы. Например, возможно, в этих двух штатах разный уровень экономического процветания, разные налоговые системы или заведения быстрого питания разного размера. А поскольку в этом сравнении штат и воздействие идеально коррелируют, любую подобную разницу между Нью-Джерси и Пенсильванией можно рассматривать как искажающий фактор.

Еще одно сравнение, которое мы могли бы сделать, – это посмотреть на изменение занятости в Нью-Джерси в период с января по ноябрь, поскольку минимальная заработная плата в Нью-Джерси изменилась за эти два месяца. Это сравнение показывает увеличение занятости на 0.59 человек на заведение, и можно предположить, что повышение минимальной заработной платы немного увеличило занятость. Преимущество этого подхода заключается в сравнении состояния с самим собой, поэтому нам больше не нужно беспокоиться о каких-либо базовых различиях между состояниями. Но теперь у нас появилась новая проблема. Возможно, январь и ноябрь различаются с точки зрения занятости в сфере быстрого питания по другим причинам – например, из-за сезонности или общих изменений в экономике в течение года. Любые такие временные тенденции будут мешать этому сравнению. Так что это сравнение также не является корректным.

В табл. 13.2 показаны два различия, которые мы обсуждали, и с точки зрения нашего любимого уравнения объясняется, почему ни одно из них не дает нам объективной оценки эффекта минимальной заработной платы. Разница между занятостью в ноябре и январе в Нью-Джерси представляет собой сумму эффекта более высокой минимальной заработной платы (оцениваемая величина), тренда от времени (смещение) и шума. Разница между занятостью в Нью-Джерси и Пенсильвании в ноябре представляет собой сумму эффекта более высокой минимальной заработной платы (оцениваемая величина), раз-

личий между штатами (смещение) и шума. Таким образом, оба различия являются искаженными.

Таблица 13.2. Два сравнения, которые не позволяют объективно оценить причинный эффект повышения минимальной заработной платы

	Январь 1992 г. *NJ и PA низкая минимальная зарплата*	**Ноябрь 1992 г.** *NJ – высокая минимальная зарплата PA – низкая минимальная зарплата*	**Разница** *ноябрь – январь*
NJ	20.44	21.03	0.59 *эффект от высокой минимальной зарплаты + тренд от времени + шум*
PA	23.33	21.17	
Разность *NJ – PA*		–0.14 *эффект от высокой минимальной зарплаты + различия между штатами + шум*	

Но мы можем добиться большего. Начнем со сравнения Нью-Джерси и Пенсильвании в ноябре. Проблема с этим сравнением заключается в том, что оно отражает как эффект более высокой минимальной заработной платы (оцениваемая величина), так и любые систематические различия между Нью-Джерси и Пенсильванией (смещение) плюс, как всегда, шум. Но предположим, что различия между Нью-Джерси и Пенсильванией не меняются со временем. Тогда различие в занятости в Нью-Джерси и Пенсильвании в январе, когда у них обоих более низкая минимальная заработная плата, отражает те же самые различия между штатами, но без эффекта более высокой минимальной заработной платы, которую Нью-Джерси ввел позже в этом году. Таким образом, мы можем использовать различие в занятости в январе, чтобы найти базовые различия между двумя штатами. И тогда, вычитая январское различие из ноябрьского (т. е. находя разность различий), мы получим несмещенную оценку эффекта более высокой минимальной заработной платы. (Конечно, в каждом различии присутствует свой шум, поэтому члены шума просто так не исчезают.)

Та же процедура работает, если мы начнем со сравнения Нью-Джерси в ноябре и Нью-Джерси в январе. Проблема с этим сравнением заключается в том, что оно отражает как эффект более высокой минимальной заработной платы, так и любые другие различия между ноябрем и январем, которые имеют значение для занятости (плюс шум). Но предположим, что эта зависимость одинаковая для Нью-Джерси и Пенсильвании. Тогда разница в занятости в Пенсильвании в период с ноября по январь представляет собой оценку тренда занятости без какого-либо влияния минимальной заработной платы (поскольку Пенсильвания не меняла свою минимальную заработную плату в 1992 г.). Таким образом, если вычесть изменение уровня занятости в Пенсильвании из изменения уровня занятости в Нью-Джерси, мы снова получим объективную оценку эффекта от повышения минимальной заработной платы.

Как показано в табл. 13.3, в любом случае мы получим один и тот же ответ. Удивительно, но оценка, которую дает нам эта процедура, заключается в том, что более высокая минимальная заработная плата, по-видимому, увеличивает занятость на 2.75 полностью занятых сотрудника на заведение. Ключевым моментом является то, что данные по Пенсильвании свидетельствуют о значительном базовом падении занятости с января по ноябрь 1992 г. Таким образом, найденное ранее увеличение на 0.59 в Нью-Джерси является значительной недооценкой истинного эффекта от повышения зарплаты, скрытого за отрицательным общим трендом.

Таблица 13.3. Оценка влияния минимальной заработной платы на занятость в сфере быстрого питания по методу разности различий

	Январь 1992 г. *NJ и PA низкая минимальная зарплата*	**Ноябрь 1992 г.** *NJ – высокая минимальная зарплата PA – низкая минимальная зарплата*	**Разница** *ноябрь – январь*
NJ	20.44	21.03	0.59 *эффект от высокой минимальной зарплаты + тренд от времени + шум*
PA	23.33	21.17	–2.16 *долгосрочный тренд + шум*
Разность *NJ – PA*	–2.89 *различия между штатами + шум*	–0.14 *эффект от высокой минимальной зарплаты + различия между штатами + шум*	**Разность различий** 0.59 – (–2.16) = –0.14 – (–2.89) = 2.75 *эффект от повышения минимальной зарплаты + шум*

Важно отметить, что, рассчитав разность различий, мы смогли объяснить систематические различия между штатами и влияние долгосрочного тренда, даже не замечая, в чем заключались эти различия или тренд. В этом сила метода разности различий.

Конечно, это не было волшебством. Как мы уже говорили, для того чтобы этот подход давал достоверный результат, необходимо строго соблюдать условие параллельности трендов – чтобы на протяжении наблюдаемого интервала времени состояние (и, следовательно, искажающие факторы) при отсутствии воздействия совпадало у всех объектов. Но обычно это более реалистичное допущение, чем предположение, что мы действительно учли все возможные искажающие факторы. Например, в нашем примере мы не думаем, что Нью-Джерси и Пенсильвания одинаковы (или что мы напрямую учитываем любые различия) при отсутствии каких-либо различий в минимальной заработной плате. Мы также не надеемся, что отсутствуют изменения характеристик штатов с течением времени. Вместо этого мы предполагаем, что эти изменения протекают параллельно: какие бы тренды ни влияли на занятость, они действуют одинаково и в Нью-Джерси, и в Пенсильвании, по крайней мере, в ожидании.

Метод разности различий имеет много преимуществ, и существует множество ситуаций, когда мы считаем, что условие параллельных трендов вполне правдоподобно. Этот подход учитывает все различия между объектами, которые не меняются с течением времени и затрудняют сравнение двух объектов только за один период времени. Он также учитывает все факторы времени, которые мешают анализу любого отдельного объекта «до» и «после». Чего он не учитывает, так это изменяющихся во времени различий между объектами. Проблема возникает, если объекты меняются в зависимости от воздействия. Скажем, если бы штат Нью-Джерси увеличил минимальную заработную плату, потому что там думали, что экономика вот-вот испытает взрывной рост по сравнению с соседними штатами, то это было бы нарушением предположения о параллельных трендах.

Конечно, даже если предположение о параллельных трендах кажется концептуально разумным, простое рассмотрение двух объектов не особо проясняет картину. Ведь в любых двух местах в течение любых двух месяцев возникает множество случайных различий, поэтому компонент шума, скорее всего, будет большим. Чтобы добиться лучших результатов, нам нужно распространить идею, которую мы развили в этом простом примере, на ситуации, когда мы наблюдаем более двух объектов в течение более двух периодов.

N объектов и два периода

Чтобы развить и обобщить наш подход, предположим, что объектов много (например, у нас есть данные о занятости и минимальной заработной плате во всех 50 штатах), но наблюдались всего два периода времени. И предположим, что некоторые объекты никогда не испытывали воздействие, в то время как остальные испытали воздействие во втором, но не в первом периоде. Мы должны рассмотреть изменения для объектов, у которых изменилось воздействие, и сравнить их с изменениями для объектов, у которых статус воздействия не менялся. Нам доступны три разных варианта вычислений, все они алгебраически идентичны и, следовательно, дадут один и тот же ответ.

1. **Вручную:** так же, как мы это делали в приведенном выше примере, рассчитайте средний результат в каждом периоде отдельно для тех, кто никогда не испытывал воздействия, и для тех, кто испытал воздействие во втором периоде, и рассчитайте разность различий вручную.
2. **Первые различия:** поместите данные в электронную таблицу по одной строке на объект (это представление называется *широким форматом*, wide format). Рассчитайте изменение результата и воздействия для каждого объекта и постройте регрессию первого от второго. Изменение воздействия будет равно 0 для объектов, статус воздействия которых никогда не менялся, и 1 для объектов, у которых он изменился. Затем просто сравните среднее изменение для этих двух групп.
3. **Регрессия с фиксированными эффектами:** поместите данные в электронную таблицу с одной строкой на период объекта (это представление называется *длинным форматом*, long format). Постройте регрессию результатов воздействия, включив в нее фиктивные переменные для каждого объекта и периода времени. В этом примере у нас будет фиктивная

переменная, которая принимает значение 1, если наблюдение находится в периоде *II*, и 0, если наблюдение находится в периоде *I*. У нас также будут отдельные фиктивные переменные для каждого объекта. Таким образом, фиктивная переменная для объекта *i* будет принимать значение 1, если наблюдение включало объект *i*, и 0, если оно включало другой объект (у каждого объекта будет своя фиктивная переменная). Мы часто называем эти фиктивные переменные *фиксированными эффектами*. Например, если аналитик говорит, что он включил в регрессию *фиксированные эффекты состояния*, он просто имеет в виду, что использует отдельную фиктивную переменную для каждого состояния. Добавление этих фиксированных эффектов в уравнение регрессии гарантирует, что мы удалим все средние различия между объектами и все средние различия на протяжении периода времени, и, как только мы это сделаем, коэффициент, связанный с переменной воздействия, станет просто разностью различий.

Рассмотрим забавный пример с несколькими объектами.

Вредит ли просмотр телевизора детям?

Мэтью Генцкоу и Джесси Шапиро интересовались тем, как просмотр телевизора дошкольником влияет на его будущую успеваемость. Проблема, конечно, в том, что на продолжительность просмотра влияют самые разные факторы, которые также влияют на будущую успеваемость в школе. Так что простое сравнение тех, кто много смотрит телевизор, с теми, кто его не смотрит или смотрит мало, не является корректным. Чтобы получить более достоверное представление о причинно-следственной связи, они использовали различия между интервалами времени, когда телевидение первоначально стало доступным в разных местах Соединенных Штатов. Мы упростим схему их исследования, чтобы вам было легче понять их основную идею.

Вещательное телевидение впервые стало доступно в большинстве городов США в период с начала 1940-х по начало 1950-х гг. К счастью, в 1965 г. было проведено крупное исследование американских школ (так называемое исследование Коулмана), в ходе которого среди прочего были зафиксированы результаты стандартизированных тестов более чем 300 000 шести- и девятиклассников. Ученик 9-го класса в 1965 г. посещал детский сад примерно в 1955 г. Ученик 6-го класса в 1965 г. посещал детский сад примерно в 1958 г. Генцкоу и Шапиро сопоставили данные по развитию вещательной сети телевидения и данные Коулмана, чтобы узнать о влиянии просмотра телепередач в раннем детском возрасте на результаты школьных тестов.

Допустим, у нас есть данные Коулмана о результатах тестов шести- и девятиклассников в двух типах городов. В городах группы А телевидение впервые появилось в 1953 г. Таким образом, телевидение у них появилось, когда ученики 6-го и 9-го классов, участвовавшие в исследовании Коулмана, ходили в дошкольные учреждения. В городах группы В не было телевидения до 1956 г. Таким образом, телевидение в них появилось, когда в дошкольном возрасте были шестиклассники, а девятиклассники уже вышли из этого возраста. В табл. 13.4 показано, как выглядят наблюдаемые данные.

Таблица 13.4. Структура данных о просмотре телевизора и школьных тестах

	9-й класс в 1965 г. дошкольники в 1955	**6-й класс в 1965 г. дошкольники в 1958**
Город А Начало вещания в 1953 г.	Средние результаты тестов 9A	Средние результаты тестов 6A
Город В Начало вещания в 1956 г.	Средние результаты тестов 9B	Средние результаты тестов 6B

Если вы хотите узнать о влиянии просмотра телевизора в дошкольном возрасте на будущую успеваемость в школе, первое сравнение, которое вы могли бы провести, – это сравнить результаты тестов девятиклассников из городов категории В (которые не могли смотреть телевизор) с результатами тестов девятиклассников из городов А (которые могли смотреть телевизор в дошкольном возрасте). Вы можете сделать это, просто вычитая средний балл за тест девятиклассника из города В из среднего балла за тест девятиклассника из города А.

Но мы уже знаем множество причин, по которым мы не можем интерпретировать эту разность как объективную оценку причинного эффекта просмотра телевидения в дошкольном возрасте. Эти два типа городов могут различаться во многом, кроме даты начала телевещания, что имело значение для успеваемости. Например, может быть, у них разный средний уровень преподавания в школах, разный уровень индустриализации или что-то еще. А поскольку в этом примере тип города и воздействие идеально коррелируют, любая такая разница между городами является искажающим фактором.

Еще одно сравнение, которое мы могли бы провести, – это посмотреть на разницу в результатах тестов в городах В между девятиклассниками и шестиклассниками, поскольку шестиклассники имели доступ к телевидению в дошкольном возрасте, а девятиклассники – нет.

Преимущество этого подхода заключается в том, что тип города остается фиксированным, поэтому нам больше не нужно беспокоиться о систематических различиях между городами. Но теперь у нас появилась новая проблема. Возможно, когорты 9-го и 6-го классов демонстрируют различие в результатах тестов по другим причинам – например, потому, что девятиклассники старше или из-за различий, специфичных для каждой группы. Любые систематические различия между когортами, в том числе возникающие с течением времени, могут исказить это сравнение.

В табл. 13.5 кратко отражены две вышеупомянутые идеи и объясняется, почему ни одна из них не дает нам достоверной оценки истинного эффекта с точки зрения нашего любимого уравнения.

Но, как и в примере с минимальной заработной платой, мы можем добиться большего. Начнем со сравнения девятиклассников из двух типов городов. Проблема с этим сравнением заключается в том, что оно отражает как влияние просмотра телевизора, так и любые другие базовые различия между типами городов А и В. Но предположим, что эти базовые различия не меняются со временем. Тогда разница в средней успеваемости шестиклассников в двух типах городов отражает именно различия между городами без влияния телевидения, поскольку все шестиклассники могли в дошкольном возрасте смотреть телеви-

зор. Таким образом, мы можем использовать это различие между шестиклассниками, чтобы оценить различия между городами. И тогда, вычитая различие между шестиклассниками из различия между девятиклассниками (т. е. вычисляя разность различий), мы получим только эффект воздействия телевидения в дошкольном возрасте (плюс шум).

Таблица 13.5. Два сравнения, которые не дают объективной оценки эффекта от просмотра телевидения

	9-й класс в 1965 г. дошкольники в 1955	**6-й класс в 1965 г. дошкольники в 1958**	**Различие**
Город А Начало вещания в 1953 г.	Средние результаты тестов 9А	Средние результаты тестов 6А	
Город В Начало вещания в 1956 г.	Средние результаты тестов 9В	Средние результаты тестов 6В	6В – 9В *эффект от просмотра + различие между когортами + шум*
Различие	9А – 9В *эффект от просмотра + различие между городами + шум*		

Аналогичная процедура работает, если мы начнем со сравнения девятиклассников и шестиклассников из городов В. Проблема с этим сравнением заключается в том, что оно отражает как эффект воздействия телевидения в дошкольном возрасте, так и базовые различия между когортами 6-го и 9-го классов, которые имеют значение для успеваемости (плюс шум). Но предположим, что эти временны́е, или когортные, тренды одинаковы в городах А и В. Тогда разница в успеваемости между 6-ми и 9-ми классами в городах А представляет собой оценку динамики, или когортного тренда, без какого-либо влияния телевидения (поскольку в городах А обе группы детей имели доступ к телевидению). Таким образом, если вычесть различие в результатах тестов в городах А из различия в результатах тестов в городах В, мы снова получим несмещенную оценку эффекта воздействия телевидения в дошкольном возрасте.

Как показано в табл. 13.6, в любом случае мы получим один и тот же ответ.

Вам интересно, чем закончилось исследование? Генцков и Шапиро нашли доказательства того, что в 1950-х гг. просмотр телевидения дошкольниками действительно был полезен для последующего обучения в школе, особенно для детей из бедных семей. Конечно, это было в то время, когда дети смотрели такие шоу, как Howdy Doody. Скорее всего, вам не захочется слепо экстраполировать этот вывод на сегодняшнюю реальность.

Для наших целей более важно увидеть силу подхода разности различий. Вычислив разность различий, мы смогли объяснить систематические различия между городами и во времени (или когортами), даже не вникая, в чем заключались эти различия или тренды.

Таблица 13.6. Как разность различий может дать объективную оценку эффекта от просмотра телевидения

	9-й класс в 1965 г. дошкольники в 1955	**6-й класс в 1965 г. дошкольники в 1958**	**Различие**
Город А Начало вещания в 1953 г.	Средние результаты тестов 9А	Средние результаты тестов 6А	6А – 9А *различие между когортами + шум*
Город В Начало вещания в 1956 г.	Средние результаты тестов 9В	Средние результаты тестов 6В	6В – 9В *эффект от просмотра + различие между когортами + шум*
Различие	9А – 9В *эффект от просмотра + различие между городами + шум*	6А – 6В *различие между городами + шум*	**Разность различий** (6В – 9В) – (6А – 9А) = (9А – 9В) – (6А – 6В) *эффект от просмотра + шум*

N ОБЪЕКТОВ И *N* ПЕРИОДОВ

Предположим, у вас более двух наблюдаемых периодов, и предположим также, что воздействие на разные объекты меняется в разное время. Как поступить в таком случае?

Большая часть логических выкладок из приведенного выше обсуждения применима и в этом случае. Конечно, вариант 1 (вычисление разности различий вручную) уже не работает. Но вы все равно можете использовать вариант 2 (первые различия) или вариант 3 (фиксированные эффекты). Однако эти методы больше не являются математически идентичными и не обязательно дадут одинаковые ответы, если вы выйдете за пределы двух периодов. В чем дело? По методу первых различий вы строите регрессию изменений от периода к периоду и результатов воздействия в разные периоды. По методу фиксированных эффектов вы строите регрессию результата воздействия, учитывая при этом все фиксированные характеристики объектов и периоды времени. Оба метода делают фактически одно и то же, но несколько разными способами.

Какой подход имеет больше смысла, зависит от конкретного контекста. В целом стратегия фиксированных эффектов более гибкая. Например, она позволяет включать в регрессию дополнительные изменяющиеся во времени переменные (при необходимости), а также позволяет проводить полезную диагностику. Важно отметить, что в обоих случаях время наступления эффекта имеет значение именно для того, что вы оцениваете. В случае первых различий вы ищете эффекты, которые возникают сразу после изменения статуса воздействия. Если для проявления эффекта от воздействия требуется некоторое время или если величина эффекта со временем затухает или увеличивается, вы можете получить неверные оценки. Впрочем, сложности, связанные со сроками воздействия, также затрудняют точную интерпретацию того, что оценивается, когда вы используете подход фиксирован-

ных эффектов. Мы не будем вдаваться в эти вопросы подробно, поскольку на момент написания этой книги они фактически являлись темой самых передовых исследований. Однако, если вы намерены регулярно проводить количественный анализ, основанный на разности различий, возможно, вам захочется углубиться в эти вопросы. Мы предлагаем материалы для дополнительного чтения в конце главы.

Несмотря на техническую сложность некоторых деталей, предыдущие примеры должны были сформировать у вас интуитивное понимание метода разности различий. Это очень важное понимание. Если кто-то сообщает вам, что некое воздействие коррелирует с наблюдаемым результатом, вы должны быть настроены скептически из-за того, что узнали в главе 9. Метод разности различий позволяет вам проверить, коррелируют ли изменения воздействия с изменениями результата. Если это так, то вы нашли убедительное доказательство причинно-следственной связи. А если это не так, то первоначальная корреляция могла быть результатом влияния искажающих факторов.

Теперь рассмотрим пример исследования, в котором для реализации схемы разности различий используется подход с фиксированными эффектами, когда несколько объектов меняют статус воздействия в разное время.

Контрацепция и гендерный разрыв в оплате труда

Доступность оральных контрацептивов, начиная с 1960-х гг., предоставила женщинам беспрецедентный контроль над своим репродуктивным поведением и экономическими решениями. Понимание влияния этих изменений на жизнь женщин важно для понимания эволюции современной экономики и общества.

Конечно, если мы хотим оценить влияние оральных контрацептивов на решение женщин родить ребенка, представленность женщин на рынке труда или заработную плату, мы не можем просто сравнивать показатели для женщин, которые использовали и не использовали оральные контрацептивы. В конце концов, на использование медицинских препаратов влияют такие факторы, как богатство, образование, место проживания, раса и т. д. Поэтому подобные сравнения наверняка будут существенно искажены. И никто не проводил эксперимент, предоставляющий одним женщинам доступ к оральным контрацептивам и ограничивающий доступ другим. Но это не значит, что мы лишены возможности исследовать причинно-следственную связь.

В важной и широко известной статье Клаудия Голдин и Лоуренс Кац отмечают, что государственная политика создала своего рода естественный эксперимент. Оральные контрацептивы впервые стали доступны в США в конце 1950-х гг. Однако легальная доступность оральных контрацептивов для молодых женщин различалась в разных штатах. В некоторых штатах законы запрещали продажу противозачаточных средств незамужним женщинам, а в большинстве штатов женщинам, не достигшим совершеннолетия, требовалось согласие родителей на приобретение противозачаточных средств. Со временем суды и законодательные собрания штатов постепенно сняли эти ограничения и снизили возраст совершеннолетия. К счастью для наших целей причинно-следственного вывода, они делали это в разное время.

Это означало, что в штатах-первопроходцах, таких как Аляска и Арканзас, незамужняя и бездетная женщина в возрасте до 21 года могла приобрести оральные контрацептивы с 1960 г. В штате Миссури это было невозможно до 1976 г. Остальные штаты занимают место где-то посередине. Это важно, поскольку женщины в возрасте до 21 года принимают особенно важные решения о том, когда заводить детей, когда выходить замуж, получать ли высшее образование и т. д.

В другой влиятельной статье Марта Бэйли использует этот вариант для реализации модели разности различий, чтобы оценить влияние ранней доступности оральных контрацептивов на то, когда женщины впервые рожают детей, а также на то, входят ли они в состав трудоустроенного населения и в какой степени.

Основная идея проста. Представьте себе четыре группы женщин в двух штатах: Канзасе (который разрешил молодым женщинам приобретать оральные контрацептивы в 1970 г.) и Айове (не разрешал приобретение до 1973 г.). В обоих штатах в конце 1960-х гг. были женщины в возрасте от 18 до 20 лет; ни одна из этих групп не имела доступа к оральным контрацептивам. И в каждом штате в начале 1970-х гг. есть женщины в возрасте от 18 до 20 лет; женщины в Канзасе имели возможность приобретать оральные контрацептивы, а женщины в Айове – нет. Таким образом, мы можем использовать изменения показателей женщин в Айове в качестве основы для сравнения с изменением показателей женщин в Канзасе, чтобы попытаться оценить эффект ранней доступности оральных контрацептивов.

Бэйли может добиться большего, чем этот простой вывод, поскольку у нее есть данные по женщинам из многих возрастных групп всех 50 штатов, а разные штаты меняли политику в разное время. Поэтому она использует схему с фиксированными эффектами – строит регрессию показателей по фиктивной переменной, определяющей, имела ли данная группа женщин доступ к пероральным контрацептивам в возрасте от 18 до 20 лет, а также отслеживает эффекты принадлежности к штату и когорте. Это позволяет ей реализовать подход разности различий, при котором отслеживаются многие объекты, меняющие статус воздействия в разное время.

Поскольку выбор штатов, которые первыми разрешили раннее приобретение оральных контрацептивов, не был случайным, нам следует подумать о параллельности трендов. Разумно ли предполагать, что тренды к рождению детей и устройству на работу в среднем параллельны во всех штатах? Бэйли приводит некоторые аргументы в пользу этого предположения. Например, она показывает, что время начала продажи оральных контрацептивов для молодых женщин не коррелирует с широким спектром характеристик штата в 1960 г., которые теоретически могли бы повлиять на решение родить детей или устроиться на работу. К ним относятся географическое положение, расовый состав, средний возраст вступления в брак, образование женщин, рождаемость, бедность, религиозный состав семьи, безработица среди мужчин и женщин, заработная плата мужчин и женщин и т. д.

При помощи метода разности различий Бэйли получила результаты, позволяющие предположить, что доступ к оральным контрацептивам в возрасте, когда женщины принимают важные жизненные решения, на самом деле

имеет заметные последствия. В частности, по ее оценкам, возможность приобретения оральных контрацептивов в возрасте до 21 года снизила вероятность стать матерью в возрасте до 22 лет на 14–18 % и увеличила вероятность того, что женщина выйдет на рынок труда в возрасте старше 20 лет, на 8 %. Более того, женщины, которые имели доступ к оральным контрацептивам до 21 года, работали больше примерно на 70 часов в год, когда им исполнилось двадцать с небольшим. То есть, предоставляя возможность отсрочить и спланировать деторождение, оральные контрацептивы, похоже, дали женщинам свободу делать долгосрочную карьеру и больше работать.

Полезные проверки

Как мы уже говорили, чтобы разность различий позволила дать объективную оценку среднего эффекта воздействия, нам нужны параллельные тренды. То есть в контрфактическом мире, где воздействие не менялось, различие в средних результатах между объектами, для которых воздействие действительно изменилось, и объектами, для которых оно не менялось, осталось бы неизменным. Поскольку мы не наблюдаем этот контрфактический мир, мы не можем знать, правда ли это. Поэтому осторожный аналитик всегда старается сделать все возможное, чтобы проверить параллельность трендов.

Одно из предположений заключается в том, что при наличии параллельных трендов мы должны увидеть аналогичные тенденции в результатах в более ранние периоды, до того как для каких-либо объектов изменился статус воздействия. Мы можем напрямую проверить эти тенденции до начала воздействия (часто называемые *предварительными трендами*, pre-trend), сравнивая тренды результатов для объектов, которые впоследствии меняют и не меняют статус воздействия. Эту проверку можно также выполнить методом регрессии, включив в нее переменную *опережающего воздействия* (lead treatment variable), т. е. фиктивную переменную, указывающую статус воздействия в следующем периоде. Если тренды действительно параллельны до изменения воздействия, коэффициент при переменной опережающего воздействия должен быть равен нулю, а коэффициент при переменной воздействия не должен меняться, когда мы включаем эту переменную опережающего воздействия в регрессию.

Также можно немного ослабить требование параллельности трендов, допустив возможность того, что разные объекты будут следовать разным линейным трендам, чтобы посмотреть, повлияет ли это на наши результаты. Конкретные детали реализации этого подхода на данный момент не важны (вы можете прочитать о них в более сложной книге). Пока вам достаточно знать, что существуют различные стратегии для анализа различий, чтобы увидеть, оправдалось ли предположение о параллельности трендов.

Помните, что такого рода диагностические тесты являются дополнением, а не заменой критического мышления. Самой важной защитой такого предположения, как параллельность трендов, должен быть существенный аргумент. Почему для одних объектов воздействие изменилось, а для других нет? Связана ли эта причина с трендами изменения каких-либо свойств объектов, или

не зависит от них? Это важные вопросы, ответы на которые требуют глубокого предметного знания контекста задачи и данных. Обоснованные ответы абсолютно необходимы для оценки того, насколько убедительны оценки, полученные на основе разности различий.

Чтобы лучше понять, как формулировать вопросы о параллельности трендов, рассмотрим два примера.

Влияет ли поддержка газет на решение по голосованию?

Газеты регулярно поддерживают кандидатов на выборные должности. Влияет ли такая поддержка на результаты голосования?

В исследовании Джонатана Лэдда и Габриэля Ленца была предпринята попытка ответить на этот вопрос, применяя модель разности различий к данным из Соединенного Королевства. Их исследование представляет собой хорошую иллюстрацию того, как проверять параллельность предварительных трендов в качестве диагностики правдоподобности предположения о параллельных трендах.

Во время всеобщей избирательной кампании 1997 г. в Соединенном Королевстве несколько газет, которые исторически были склонны поддерживать консервативную партию, неожиданно поддержали лейбористскую партию. Лэдд и Ленц использовали этот редкий сдвиг, чтобы оценить влияние поддержки газет на выбор избирателей.

Используя модель разности различий, они сравнивают изменения в выборе кандидата между теми, кто регулярно читает газету, неожиданно поддержавшую лейбористскую партию, и теми, кто не читал ни одну из газет регулярно. Поскольку у них были данные, измеряющие партийные предпочтения одних и тех же британцев в 1992, 1996 и 1997 гг., они смогли изучить предварительные тренды и увидеть, являются ли они параллельными. Если бы люди, которые читали и не читали газеты, переметнувшиеся на сторону лейбористов в период с 1996 по 1997 г., изначально имели разные тренды в период с 1992 по 1996 г., это заставило бы нас думать, что предположение о параллельных трендах нарушается (возможно, мы бы заподозрили, что газеты сменили направление, потому что их читатели начали склоняться к лейбористской партии). Но если бы эти две группы имели схожие тренды в период с 1992 по 1996 г., это дало бы нам больше уверенности в том, что любые возникающие разности различий объясняются неожиданной поддержкой газет в 1997 г.

Проверка Лэдда и Ленца обнадеживает, о чем свидетельствует рис. 13.1. Люди, которые читали газеты, перешедшие к поддержке лейбористской партии, имели очень схожие тренды в уровне поддержки лейбористской партии в период с 1992 по 1996 гг. Но между 1996 и 1997 гг., когда газеты неожиданно поддержали лейбористскую партию, среди избирателей, читавших эти газеты, значительно возросло число голосов за лейбористов по сравнению с теми, кто эти газеты не читал. Таким образом, пока у нас нет оснований полагать, что другие вещи, помимо этой неожиданной поддержки, изменились по-разному для читателей разных газет в 1996 г., мы можем обоснованно интерпретировать разность различий как оценку причинного эффекта от поддержки газет.

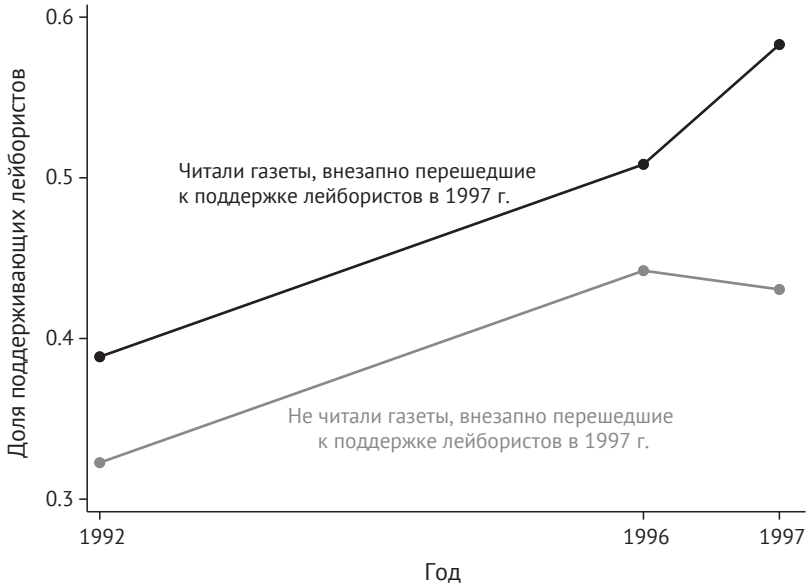


Рис. 13.1. Графическое представление предварительных трендов и использование разности различий для оценки влияния поддержки газет на выбор избирателей

Заразно ли ожирение?

Люди – социальные животные. Мы живем в сложной паутине отношений. Нам говорят, что наши связи определяют, кто мы есть. Все больше исследователей утверждают, что они могут точно измерить, как наше мышление, вкусы и поведение определяются нашими социальными связями.

Пожалуй, самое известное из этих исследований принадлежит Николасу Кристакису и Джеймсу Фаулеру. Они обнаружили, что поведение и характеристики, которые многие из нас считают глубоко личными, – курение, употребление алкоголя, счастье, ожирение, – все они, очевидно, являются социально обусловленными. Или, выражаясь более красочным языком, «ожирение заразно».

В исследовании распространения ожирения в социальных сетях, опубликованном в Медицинском журнале Новой Англии, Кристакис и Фаулер изучают взаимосвязь между изменением веса человека и изменениями веса его друзей, членов семьи или соседей. Они проводят эти сравнения с учетом личных характеристик, таких как возраст, пол и образование.

Что они обнаружили? Вероятность того, что человек приобретет ожирение, на 57 % выше, если у этого человека есть друг, который страдает ожирением, чем в отсутствие такого друга. Когда дело доходит до набора веса, дружба влияет даже больше, чем семейное родство. Если у человека есть брат или сестра, страдающие ожирением, вероятность этого человека заболеть ожирением увеличивается на 40 %. Если супруг человека страдает ожирением, вероятность того, что этот человек заболеет ожирением, увеличивается на 37 %. Наличие тучных соседей не имеет никакого эффекта. На основе этих выводов газета New York Times заявила в статье на первой полосе: «Лучший способ не толстеть – это избегать толстых друзей». Кристакису и Фаулеру такая интерпрета-

ция не понравилась. Вместо этого, как сообщает Times, Кристакис предложил: «Почему бы не дружить с худым человеком… и позволить поведению худого человека влиять на вас и вашего друга, страдающего ожирением?»

Кажется очевидным, что на ваше поведение влияют те, с кем вы взаимодействуете, на их поведение влияют те, с кем они взаимодействуют, и т. д. В этом смысле мы полностью на стороне Кристакиса и Фаулера – мы все находимся под влиянием нашего социального окружения. Но эти авторы и многие другие ученые, изучающие сетевые эффекты, выдвигают более серьезные утверждения, чем просто здравое наблюдение о том, что наши взаимодействия влияют на наше поведение. Они заявляют, что могут измерить и количественно оценить этот эффект. Как они это делают?

Подход Кристакиса и Фаулера, по сути, представляет собой модель разности различий. Они проверяют, как изменения в ожирении одного человека соответствуют изменениям в ожирении другого. Поэтому, если мы хотим критически судить о том, являются ли эти оценки «эффекта заражения» заслуживающими доверия, нам нужно решить, считаем ли мы правдоподобным предположение о параллельных трендах.

Напомним, о чем здесь говорят параллельные тренды. Они требуют, чтобы в контрфактическом мире, где не было никаких изменений в воздействии (т. е. никто из друзей не страдал относительно заметным ожирением), тенденция в результатах (личное ожирение) была бы в среднем одинаковой среди людей, которые в нашем мире испытали изменение воздействия (т. е. ожирение друзей которых изменилось), и людей, у которых воздействие не изменилось (т. е. ожирение друзей не изменилось). Если параллельность трендов сохраняется, то модель Кристакиса и Фаулера дает объективную оценку влияния друзей на ожирение конкретного человека. Но если тренды не параллельны, то их оценки смещены, поскольку некоторые различия, которые они наблюдают и приписывают сетевым эффектам, имели бы место, даже если бы ожирение друзей не изменилось.

Одной из проблем, которая часто встречается при изучении сетевых эффектов, является то, что исследователи-медики называют *гомофилией*. Люди со схожими характеристиками склонны группироваться вместе. Предположим, вы обнаружили, что люди, чьи друзья курят, с большей вероятностью сами будут курильщиками. Исследователи социальных сетей могут интерпретировать это наблюдение как свидетельство того, что наличие курящих друзей повышает вероятность курения. Но чтобы этот вывод был оправдан, нам придется сравнивать яблоки с яблоками. То есть, если не считать курения, люди в социальном окружении курящих и некурящих субъектов должны быть, по сути, одинаковыми во всем остальном. Если членами социального окружения курильщиков с большей вероятностью становятся другие курильщики, чем некурящие, то мы сравниваем яблоки с апельсинами.

Было бы логично предположить, что люди, которые входят в социальное окружение курильщиков, сами являются курильщиками просто по причине гомофилии. Курильщики вполне могут встречаться со своими друзьями в барах, где курение разрешено, на работе или в школе, где люди собираются покурить в перерыве, или в других местах, дружественных для курящих. Иными словами, наличие курящих друзей не обязательно послужило причиной того, что вы

курите. Возможно, из-за того, что вы курите, у вас появляются курящие друзья. Поскольку мы все выбираем свое социальное окружение не случайно, было бы некорректно сравнивать людей в социальных сетях курильщиков с людьми в социальных сетях некурящих.

Но одной лишь гомофилии недостаточно, чтобы создать проблему для модели разности различий Кристакиса и Фаулера. Эта схема учитывает фиксированные характеристики субъектов, такие как вероятность того, что люди с ожирением склонны дружить друг с другом, а курильщики склонны проводить время вместе. Это одна из замечательных особенностей разности различий. Их открытие более убедительно, чем просто сравнение людей, у которых много друзей с избыточным весом, с людьми, у которых меньше таких друзей. Они показывают, что, когда один человек *становится* полным, его друзья также с большей вероятностью *становятся* полными следом. Чтобы гомофилия стала проблемой, она должна заставить нас усомниться в параллельности трендов. Например, если люди, находящиеся на пути к ожирению (возможно, потому что у них схожие диеты, уровни физической нагрузки, генетическая предрасположенность, культурное давление и т. д.), с большей вероятностью будут дружить друг с другом, это будет нарушением параллельности. А если допущение о параллельности трендов нарушается, разность различий не дает объективной оценки причинного эффекта.

Мы не можем знать наверняка, приводит ли гомофилия к нарушению параллельности трендов. Но некоторые данные указывают на вероятность того, что разность различий в данном случае не является беспристрастной. Итан Коэн-Коул и Джейсон Флетчер провели исследование распространения двух индивидуальных характеристик – роста и прыщей – в социальном окружении. Используя тот же подход разности различий, который Кристакис и Фаулер используют, чтобы аргументировать социальную заразность разводов, одиночества, счастья, ожирения и многих других вещей, Коэн-Коул и Флетчер обнаружили, что и рост, и прыщи выглядят заразными для социального окружения. Довольно трудно поверить, что на рост и количество прыщей влияют социальные связи, как справедливо заметили авторы исследования. Рост и прыщи, скорее всего, не распространяются социальным путем. Это очевидное нарушение параллельности трендов, по всей вероятности, из-за гомофилии. Если у вас есть друзья с прыщами, это не значит, что у вас появятся прыщи из-за переписки в социальных сетях; просто люди с высоким риском появления прыщей, как правило, проводят время вместе. То же самое может быть справедливо и в отношении ожирения, развода, счастья и т. д.

Здесь нужно внести ясность: мы не утверждаем, что причинно-следственные сетевые эффекты не существуют. Мы уверены, что они есть. Более того, исследование Кристакиса и Фаулера, безусловно, выглядит убедительно, потому что они сравнивали изменения с изменениями, а не просто показывали, что люди, страдающие ожирением, с большей вероятностью будут дружить друг с другом. Но существует множество вариантов нарушения параллельности трендов, и некоторые из них не очевидны. Поэтому мы должны вести себя осторожно и критически оценивать обнаруженные явления, прежде чем интерпретировать модель разности различий как объективную оценку истинного причинного эффекта.

Разность различий как проверка достоверности выводов

Иногда анализ разности различий может быть полезен как способ проверить достоверность причинно-следственных связей. Представьте себе ситуацию, когда кто-то оценивает корреляцию между воздействием и эффектом, возможно, даже с учетом некоторых искажающих факторов. Вспомнив уроки главы 9, вы можете скептически отнестись к причинной интерпретации этой оценки. Возможно, вы назовете несколько других факторов, которые не поддаются наблюдению и, следовательно, не могут быть учтены. Но, даже имея такие аргументы, бывает трудно убедить людей серьезно отнестись к вашим опасениям.

Если данные включают в себя несколько наблюдений за одним и тем же объектом, разность различий может послужить интуитивной проверкой достоверности выводов[1]. Если воздействие действительно определяет наблюдаемый эффект, то мы должны ожидать наличие корреляции не только между воздействием и эффектом, но и между изменениями в статусе воздействия и изменениями эффекта. То есть мы должны ожидать, что взаимосвязь между воздействием и эффектом никуда не денется при анализе по методу разности различий.

Даже если вы обнаружите причинно-следственную связь в модели разности различий, это не служит гарантией правильности причинной интерпретации. Вам нужно будет проверить соблюдение условия параллельности трендов. Но если причинно-следственная связь исчезает при переходе к модели разности различий, то это мощный аргумент в пользу вашего скептицизма.

Подведение итогов

Мы показали, что изменения воздействия, наблюдаемые на определенном интервале времени, позволяют более достоверно оценить эффекты этого воздействия, используя метод разности различий. Чтобы этот метод сработал, требуется строгое соблюдение условия параллельности трендов: если бы не изменение статуса воздействия, средние наблюдаемые показатели в экспериментальной и контрольной группе изменялись бы одинаково. Существует несколько полезных диагностических тестов, которые помогут аналитикам оценить, правдоподобно ли это предположение, но ничто не заменит критическое мышление и знание предметной области.

Последние четыре главы были посвящены методам получения более достоверных оценок причинно-следственных связей. Оценка причинно-следственных связей – сложная и благородная задача. Но зачастую нам хочется знать больше. Мало просто знать, что эффект есть. Мы хотим знать, *почему* он возникает. В следующей главе рассматривается важная задача ответа на вопросы «почему?» с помощью количественных данных.

Ключевые термины

- **Разность различий:** схема исследования для оценки причинно-следственных связей, когда с течением времени для одних объектов меняется статус воздействия, а для других – нет.

[1] Мы полагаем, что вы уже выработали критическое отношение к выводам и будете интуитивно применять к ним все доступные способы проверки.

- **Параллельность трендов:** условие, при котором в отсутствие воздействия среднее значение наблюдаемого показателя со временем изменяется одинаково для обеих групп. Это означает, что на разницу средних значений наблюдаемого показателя (эффект) между группами должно влиять только наличие или отсутствие воздействия. Если условие параллельности трендов не соблюдается, разность различий не дает объективной оценки ATT.
- **Первые различия:** статистическая процедура для реализации метода разности различий. Она включает в себя регрессию изменения эффекта для каждого объекта от изменения воздействия для каждого объекта.
- **Широкий формат:** способ структурирования набора данных, в котором каждому объекту соответствуют несколько наблюдений и каждая строка таблицы соответствует уникальному объекту.
- **Длинный формат:** способ структурирования набора данных, в котором каждому объекту соответствуют несколько наблюдений и для каждого объекта в каждом периоде времени имеется отдельная строка.
- **Регрессия с постоянными эффектами:** статистическая процедура для реализации метода разности различий. Она предполагает регрессию эффекта воздействия, а также включение в уравнение регрессии фиктивных переменных (*постоянных эффектов*) для каждого периода времени и для каждого объекта.
- **Предварительные тренды:** тенденция изменения средних показателей до того, как для какого-либо объекта изменился статус воздействия. Если предварительные тренды изначально не были параллельны, то их параллельность при воздействии потребует отдельного доказательства.
- **Опережающая переменная воздействия:** фиктивная переменная, указывающая, что статус воздействия на объект изменится в следующий период времени.

Упражнения

13.1. В течение многих лет в штате Иллинойс проводится квалификационный тест штата (ISAT) для учащихся 3-го, 5-го и 8-го классов. Большую часть этого времени значимость теста была относительно низкой: он никак не влиял на перевод в следующий класс, вознаграждение учителей, выделение школам дополнительных ресурсов и т. д. Ставки изменились в 2002 г., когда ISAT стал тестом, который государственные школы Чикаго использовали на предмет соответствия федеральному закону «Ни один ребенок не останется без внимания».

Рассмотрим две группы учащихся: пятиклассников в 2001 г. и пятиклассников в 2002 г. Обе эти группы учащихся сдавали ISAT в 3-м классе, когда значимость теста была низкой. Учащиеся пятого класса в 2001 г. также сдавали ISAT в 5-м классе, когда значимость была низкой. Но ученики 5-го класса в 2002 г. сдавали второй тест ISAT, когда его значимость уже была высока. Составьте таблицу 2×2, показывающую, как можно узнать о среднем влиянии тестирования с более высокой ответственностью на результаты тестов учащихся, используя метод разности различий, если

имеются данные о средних результатах тестов этих двух когорт студентов, когда они были пятиклассниками и третьеклассниками.

13.2. Кроссовки Nike Vaporfly вызвали споры в мире элитных бегунов на длинные дистанции, поскольку некоторые утверждают, что эта обувь дает несправедливое преимущество тем, кто ее использует, и делает предыдущие рекорды устаревшими. Предположим, у вас есть данные о множестве разных марафонов, в которых указано время каждого бегуна, а также обувь, которую носил каждый бегун. Как вы могли бы оценить эффект Nike Vaporfly? Вы должны обязательно принять во внимание тот факт, что время прохождения марафона меняется в разные дни и на разных трассах. Вам также необходимо принять во внимание тот факт, что некоторые бегуны просто лучше и быстрее, чем другие.

a) Какой анализ вы бы провели, чтобы отделить влияние обувной технологии от других факторов, и какие предположения вам пришлось бы сделать?
b) Считаете ли вы эти предположения правдоподобными? Выскажите свои потенциальные опасения.
c) Можете ли вы что-нибудь сделать для решения потенциальных проблем?
d) Еще одна проблема заключается в том, что не все, кто начинает марафон, заканчивают его, поэтому в вашем исследовании может возникнуть естественная убыль. Что вы могли бы сделать, чтобы решить эту потенциальную проблему?
e) Можете ли вы использовать тот же подход для оценки влияния новой технологии обуви или перчаток на количество очков, набранных в профессиональном боксе? Поясните свой ответ.

13.3. Предположим, мы хотим оценить, насколько расходятся политические позиции кандидатов от демократов и республиканцев на выборах в конгресс. Другими словами, мы хотели бы знать, насколько по-разному кандидаты от демократической и республиканской партий будут представлять один и тот же набор избирателей.

a) Допустим, мы измерили, насколько консервативно каждый член конгресса голосовал по законопроектам, и провели регрессию поименного голосования по показателю принадлежности к республиканцам. Будет ли это удовлетворительным способом оценки расхождения? Какие смещения вас беспокоят?
b) Загрузите файл CongressalData.csv и связанный с ним файл README.txt, описывающий переменные в этом наборе данных, по адресу *http://www.press.princeton.edu/thinking-clearly*. Этот набор данных содержит информацию о выборах в конгресс и поименном голосовании. Используя только переменные, доступные в предоставленном наборе данных, попытайтесь оценить расхождение, исключив искажающие факторы. Подсказка: попробуйте анализировать только одну сессию конгресса за раз.
c) Используя доступные данные, теперь оцените расхождение, используя модель разрывной регрессии. Опять же, возможно, вам будет полезно сосредоточиться только на одной сессии конгресса за раз.

d) Наконец, оцените расхождение, используя метод разности различий.
e) Сравните и противопоставьте эти три разных подхода. Какой из них оценивает расхождение с помощью наиболее обоснованных предположений? Насколько ваши оценки зависят от схемы исследования?

13.4. В ходе исследования дискриминации по признаку пола при приеме на работу Клаудия Голдин и Сесилия Роуз изучают эффект «слепых» прослушиваний в симфонические оркестры, когда кандидатов помещают за ширму. Идея заключается в том, что, если люди, оценивающие кандидатов, не могут определить пол человека, проходящего прослушивание, у них не останется возможности для дискриминации.

Оказывается, как отмечают Голдин и Роуз, разные оркестры в разное время переняли практику использования такой ширмы. Подумайте, как можно использовать этот факт, чтобы узнать о причинно-следственном эффекте наличия ширмы. (Мы воспользуемся несколько иным эмпирическим подходом, чем тот, который используют Голдин и Роуз.)

a) Предположим, что для каждого оркестра и каждого года вы наблюдали долю женщин среди новых сотрудников этого оркестра и то, использовал ли этот оркестр ширму на прослушивании. Если вы просто объедините все имеющиеся данные и постройте регрессию доли женщин, принятых с использованием ширмы, можно ли придавать результатам этой регрессии причинно-следственную интерпретацию? Объясните свой ответ.
b) Предположим, что вместо этого вы решили использовать для этих данных модель разности различий. Какую регрессию вы построите?
c) Опишите предположения, которые должны быть верными, чтобы обеспечить объективную оценку причинного эффекта. (Недостаточно просто сказать «параллельные тренды»; опишите, что именно должно выполняться в данном случае, чтобы сохранилась параллельность трендов.)
d) Считаете ли вы предположение о параллельности трендов правдоподобным? Какие опасения у вас могут возникнуть?

Дополнительное чтение и ссылки

Исследование о влиянии повышения минимальной заработной платы в Нью-Джерси:

David Card and Alan B. Krueger. 1994. *MinimumWages and Employment: A Case Study of the Fast-Food Industry inNew Jersey and Pennsylvania.* American Economic Review 84 (4): 772–93.

Исследование влияния телевидения на успеваемость:

Matthew Gentzkow and Jesse M. Shapiro. 2008. *Preschool Television Viewing and Adolescent Test Scores: Historical Evidence from the Coleman Study.* Quarterly Journal of Economics 71 (3): 279–323.

Если вы хотите узнать больше о сложностях метода разности различий, когда имеется N объектов и N периодов, прочитайте следующие работы:

Andrew Goodman-Bacon. 2018. *Difference-in-Differences with Variation in Treatment Timing*. NBERWorking Paper No. 25018;

Kosuke Imai and In Song Kim. 2019. *When Should We Use Unit Fixed Effects Regression Models for Causal Inference with Longitudinal Data?* American Journal of Political Science 63 (2): 467–90.

Два исследования об оральных контрацептивах, о которых мы упомянули:

Claudia Goldin and Lawrence F. Katz. 2002. *The Power of the Pill: Oral Contraceptives and Women's Career and Marriage Decisions*. Journal of Political Economy 110 (4): 730–70;

Martha J. Bailey. 2006. *More Power to the Pill: The Impact of Contraceptive Freedom on Women's Life Cycle Labor Supply*. Quarterly Journal of Economics 121 (1): 289–320.

Исследование влияния газет в Соединенном Королевстве:

Jonathan McDonald Ladd and Gabriel S. Lenz. 2009. *Exploiting a Rare Communication Shift to Document the Persuasive Power of theNewsMedia*. American Journal of Political Science 53 (2): 394–410.

Исследования заразности ожирения и прочих сетевых эффектов:

Nicholas A. Christakis and James H. Fowler. 2007. *The Spread of Obesity in a Large Social Network over 32 Years*. New England Journal of Medicine 373: 370–79;

Ethan Cohen-Cole and Jason Feltcher. 2009. *Detecting Implausible Social Network Effects in Acne, Height, and Headaches: Longitudinal Analysis*. BritishMedical Journal 338 (7685): 28–31.

Исследование оркестровых прослушиваний, обсуждавшееся в задании 4:

Claudia Goldin and Cecilia Rouse. 2000. *Orchestrating Impartiality: The Impact of 'Blind' Auditions on Female Musicians*. American Economic Review 90 (4): 715–41.

Глава 14

Механизмы причинно-следственных связей

О чем эта глава

- Оценка среднего причинно-следственного эффекта не говорит нам, почему и как возникает этот эффект.
- Узнать о механизмах, лежащих в основе причинно-следственных связей, сложнее, чем кажется. Как правило, мы не можем просто измерить потенциальные механизмы и выяснить, какие из них наиболее важны.
- Сочетание теории, измерений и критического мышления помогает нам узнать о механизмах, лежащих в основе причинно-следственных связей.

Введение

Если говорят, что воздействие влияет на эффект (результат, исход), то имеют в виду лишь то, что изменение воздействия ведет к изменению эффекта. Но эта зависимость не обязательно должна быть прямой – влияние какого-то события на какой-то результат может быть результатом длинной цепочки отношений. Таким образом, во многих случаях, даже если мы достоверно оценили причинно-следственную связь, мы можем оставаться в неведении относительно того, почему и как воздействие влияет на результат.

Например, предположим, мы обнаружили, что посещение чартерных, а не обычных государственных школ увеличивает вероятность того, что учащиеся поступят в колледж. Это интересный и политически значимый вывод. Он говорит нам о том, что в среднем чартерные школы помогают учащимся. Но он не говорит нам, *как* они это делают, т. е. не говорит нам о механизмах, с помощью которых чартерные школы помогают учащимся. Может быть, учебная программа более инновационная, или на учащихся влияют более мотивированные сверстники, или дисциплина более строгая, или условия лучше, или есть классы дополнительного обучения для одаренных детей (advanced placement, AP), или учащиеся лучше подготовлены к стандартным тестам, или, может быть, школа мотивирует учеников работать усерднее. Обычно под механизмом воздействия мы понимаем ответы на вопросы «как?» («Как возник этот эффект?») или «почему?» («Почему это произошло?»).

Рандомизация учащихся при зачислении в чартерные школы, по сравнению с государственными, или использование какого-либо другого подходящего

метода исследования позволяет нам оценить средний эффект от посещения чартерной школы. Но он не раскроет, какие механизмы выполняют работу, хотя иногда это важно знать. Например, если мы собираемся построить больше чартерных школ, нам хотелось бы знать, какие особенности существующих чартерных школ наиболее важно воспроизвести. К чему мы должны стремиться – к более хорошим условиям, к увеличению количества классов для одаренных детей или к более строгой дисциплине? В этой главе мы рассмотрим некоторые трудности, возникающие при попытке изучить не только эффекты, но и механизмы, лежащие в основе этих эффектов.

Анализ причинной медиации

Один из подходов, который используют некоторые исследователи, пытаясь понять причинные механизмы, называется анализом *опосредованной причинной связи*, или анализом *причинной медиации* (causal mediation). Цель анализа причинной медиации состоит в извлечении информации о том, насколько важную роль тот или иной механизм играет в обеспечении определенного эффекта.

Эту идею иллюстрирует рис. 14.1, напоминающий нашу иллюстрацию механизмов на рис. 9.9. Предположим, например, что мы хотим знать, в какой степени посещение чартерной школы влияет на поступление в колледж благодаря тому, что учащиеся чартерной школы посещают больше занятий в специализированных классах. На нашем привычном языке посещение чартерной школы – это воздействие, а поступление в колледж – результат. Мы называем посещение занятий в АР *посредником* – механизмом, с помощью которого чартерные школы влияют на поступление в колледж. Идея состоит в том, что посещение чартерных школ способствует посещению занятий АР, а посещение занятий АР способствует поступлению в колледж. Таким образом, влияние посещения чартерной школы на поступление в колледж в какой-то мере зависит от механизма классов АР. Нам было бы полезно знать, какая часть эффекта обусловлена посещением АР (стрелки от чартерной школы к классам АР и далее к посещению колледжа), а какая – другими факторами (стрелка напрямую от посещения чартерной школы к посещению колледжа).

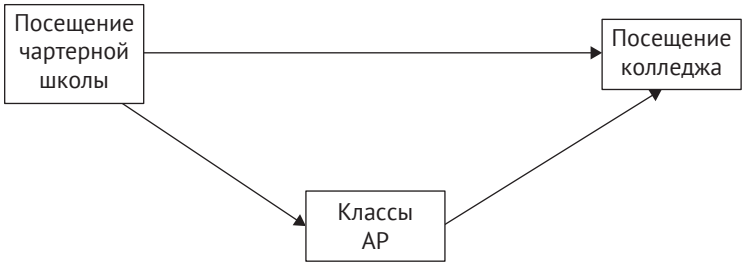


Рис. 14.1. Классы АР могут быть одним из механизмов, с помощью которых чартерные школы влияют на посещаемость колледжа

У аналитика без навыков критического мышления может возникнуть соблазн использовать методы из главы 10, чтобы попытаться ответить на этот вопрос. Он начнет с анализа вступительной лотереи, чтобы оценить влияние посещения чартерной школы на поступление в колледж. Это можно сделать, построив

регрессию поступления в колледж после выигрыша во вступительную лотерею приема в чартерную школу для группы студентов, участвовавших в лотерее. Затем аналитик подсчитает количество классов AP, которые посещал каждый ученик, и повторно запустит регрессию поступления в колледж после выигрыша в лотерее чартерной школы, но с учетом классов AP. Идея состоит в том, что если предполагаемый эффект чартерных школ уменьшается после того, как были учтены классы AP, то та часть исчезнувшего эффекта связана с механизмом классов AP, поскольку, учитывая классы AP, вы фактически сохраняете их постоянными, следовательно, они статистически исключены из вашей оценки эффекта.

На первый взгляд эта идея может показаться разумной. Мы пытаемся выяснить, в какой степени чартерные школы влияют на поступление в колледжи через классы AP. Мы хотели бы сравнить влияние чартерных школ на поступление в колледж с влиянием, очищенным от эффекта посещения классов AP. Так почему же этот подход не сработает?

Основная проблема заключается в том, что учесть посещение классов AP – это не то же самое, что устранить их влияние. Взгляните на этот вопрос шире. Чтобы отделить влияние посещений классов AP на поступление в колледж от влияния посещения чартерной школы, мы, конечно же, должны иметь способ оценить влияние занятий AP на поступление в колледж. Но мы не описали никакого плана исследования, позволяющего это сделать. Займемся этим сейчас.

Лотерея приема в чартерные школы позволяет нам оценить влияние посещения чартерных школ на поступление в колледж. Она также позволяет оценить влияние посещения чартерной школы на посещение занятий в классах AP. Но как оценить влияние посещения классов AP на поступление в колледж? Мы уверены, что можно придумать множество факторов, искажающих взаимосвязь между количеством занятий в классах AP, которые посещает ученик, и тем, поступит ли он потом в колледж. Так что простая регрессия поступления в колледж от посещения классов AP явно не поможет.

Чтобы понять, в чем проблема, давайте подумаем о крайней версии того, что может пойти не так. Предположим, что чартерные школы действительно заставляют учащихся посещать больше занятий в классах AP, но эти занятия никак не влияют на поступление в колледж. (Это, конечно, просто допущение ради аргументации.) Итак, если наша стратегия контроля помогает выявить значимость механизма, то мы не должны обнаружить никакой разницы во влиянии посещения чартерной школы на поступление в колледж, независимо от того, учитываем мы посещение классов AP или нет (поскольку оно на самом деле не является механизмом). Но предположим также, что посещение классов AP коррелирует с талантом к учебе (который мы не измеряли) и что талант влияет на поступление в колледж. Теперь, когда мы построим регрессию поступления в колледж по посещаемости чартерных школ и занятий AP, мы обнаружим, что предполагаемый эффект от посещения чартерной школы действительно ниже, если учитывать посещение классов AP. Это связано с тем, что посещение классов AP отражает (т. е. измеряет) талант к учебе. Из этого статистического результата мы ошибочно заключим, что возможность посещать классы AP является важным механизмом, с помощью которого чартерные школы способствуют поступлению в колледж. Но на самом деле мы оговорили, что классы AP не имеют такого эффекта. Посещение занятий AP коррелирует с талантом, который также коррелирует с по-

ступлением в колледж. Таким образом, наша стратегия контроля была ошибочной. Без должного критического анализа мы можем в конечном итоге принять неверные решения о том, как распределять ресурсы или организовать обучение.

В научной литературе подробно рассматривается, какие условия необходимы для проведения опосредованного причинно-следственного анализа. Если не вдаваться в технические детали, все сводится к чему-то вроде исследований, позволяющих отдельно оценить влияние воздействия на результат, влияние воздействия на механизм и влияние механизма на результат. Если вы сможете оценить все эти величины, то сможете оценить влияние воздействия на результат, транслируемое через механизм. Ключевой вывод, конечно, заключается в том, что не существует волшебного технического или статистического способа определить, какие механизмы имеют значение. Если вы хотите обнаружить реальные механизмы причинно-следственной связи, придется усердно поработать и многое критически осмыслить.

Промежуточные результаты

Единственное, что может сделать аналитик, – это проверить влияние воздействия на промежуточные результаты, которые могут дать некоторые подсказки о механизмах. Если у вас есть план исследования, позволяющий оценить эффект определенного воздействия, его, в принципе, можно применить к любому последующему результату, который вы можете измерить.

Итак, чтобы частично оценить причинные механизмы, мы можем посмотреть, на какие промежуточные результаты влияет воздействие. Возвращаясь к нашему примеру, мы можем использовать лотерею чартерных школ, чтобы оценить влияние посещения чартерной школы не только на поступление в колледж, но и на промежуточные результаты, такие как привычки в учебе, участие во внеклассных занятиях, результаты стандартизированных тестов, посещение классов AP и т. д. Конечно, это не дает нам возможности оценить влияние промежуточных результатов на конечный результат (в данном случае поступление в колледж). Для этого понадобится отдельное исследование со своим планом. Короче говоря, по причинам, которые мы только что обсудили, этот подход не скажет нам точно, какая часть эффекта объясняется тем или иным механизмом. Зато он позволяет более или менее хорошо разобраться в том, какие механизмы помогают или не помогают понять эффект. Например, если окажется, что посещение чартерной школы почти не влияет на посещение классов AP, то маловероятно, что занятия AP являются одним из механизмов, с помощью которых чартерные школы влияют на поступление в колледж.

Одним из реальных примеров использования промежуточных результатов является исследование Криса Блаттмана, Джулиана Джеймисона и Маргарет Шеридан, проведенное в Либерии.

Когнитивно-поведенческая терапия и молодежь из группы риска в Либерии

Блаттман, Джеймисон и Шеридан случайным образом назначили курс когнитивно-поведенческой терапии некоторым молодым людям в Либерии, кото-

рые, как считалось, подвергались риску участия в преступлениях или насилии, в надежде улучшить экономическое состояние и снизить уровень преступности и насилия среди молодого населения. Похоже, что терапия сработала хорошо, значительно улучшив оба показателя. Это хорошая новость для терапевтической программы, которую изучали Блаттман, Джеймисон и Шеридан. Но, помимо знания того, что терапия сработала, было бы неплохо узнать, *как* и *почему* она сработала. Экспериментальное воздействие длилось восемь недель и включало работу над различными навыками, от самоконтроля до внешнего вида. Некоторые способы воздействия даже включали денежную компенсацию. Таким образом, существует множество различных нюансов воздействия, которые могут объяснить результат. И если вы попытаетесь применить эти наблюдения еще где-то, то сразу возникнет вопрос, какие особенности этой программы важны, а какие второстепенны.

Как мы уже говорили, невозможно точно узнать, какие особенности воздействия привели к совокупному эффекту. Но авторы все же измерили ряд промежуточных результатов, которые помогают прояснить механизмы и могут пригодиться практикам, желающим применить полученные уроки в других местах. Интересно, что авторы нашли мало доказательств благотворного влияния терапии на навыки самоконтроля испытуемых, такие как импульсивность, настойчивость и добросовестность. Они пришли к выводу, что эти навыки самоконтроля вряд ли станут важным механизмом, с помощью которого терапия снижает уровень насилия или повышает благосостояние молодых людей. Напротив, они обнаружили, что их воздействие оказало большое влияние на другие промежуточные результаты, такие как формирование социальных связей и отношение к преступлениям. Это говорит о том, что данные механизмы с большей вероятностью объясняют успех воздействия, и при последующем воспроизведении программы им нужно уделить особое внимание.

Внесем ясность: из этого исследования не следует, что навыки самоконтроля не влияют на экономическое благополучие или склонность к насилию. Они могут иметь очень большие последствия. На самом деле исследование показывает, что конкретное изучаемое воздействие мало повлияло на навыки самоконтроля, и поэтому их вряд ли можно считать механизмом воздействия на экономическое благополучие или склонность к насилию. Более того, если будущие практики соберутся повторить успех этой поведенческой терапии в Либерии, они могут извлечь из исследования урок, что по какой-то причине они не добьются большого успеха в изменении навыков самоконтроля, поэтому им, возможно, лучше сосредоточиться на других факторах, таких как социальные связи и отношение к преступлениям, где этот конкретный подход наиболее эффективен.

Независимые теоретические прогнозы

Другой способ выявить специфические механизмы воздействия – поразмыслить теоретически и провести независимые тесты, помогающие сделать выбор между различными потенциальными механизмами. В главе 7 мы обсуждали, как можно использовать независимые теоретические прогнозы, чтобы проверить, является ли некоторый предполагаемый эффект результатом шума

(т. е. ложного срабатывания). Там мы привели пример критического анализа предполагаемого влияния результатов футбольных матчей в колледжах на исход выборов. Здесь идея аналогична. Но теперь, вместо того чтобы проверять, является ли наблюдаемый эффект подлинным, мы будем искать механизмы, лежащие в основе эффекта, в подлинности которого не сомневаемся.

Исследование Сары Анциа и Криса Берри показывает, как работает такой подход.

Дискриминируют ли избиратели женщин?

В обществе существует большая обеспокоенность по поводу потенциальной дискриминации в избирательном процессе. Например, не относятся ли избиратели предвзято к кандидатам-женщинам? На этот вопрос сложно дать убедительный ответ по понятным причинам. Избиратели, склонные к дискриминации, могут не раскрывать свои предубеждения в опросах. А на реальных выборах, помимо дискриминации, могут сработать и другие факторы, объясняющие, почему женщины добиваются лучших или худших результатов, чем мужчины.

Некоторые ученые заметили, что в Соединенных Штатах, когда кандидаты-женщины баллотируются на выборные должности, их результаты в среднем аналогичны кандидатам-мужчинам. Это, по их мнению, означает отсутствие сколько-нибудь заметной дискриминации. Однако Анциа и Берри отмечают, что при наличии дискриминации только наиболее выдающиеся женщины решатся выставить свою кандидатуру, что может объяснить, почему женщины избираются примерно так же хорошо, как мужчины, даже в мире с дискриминацией. Таким образом, тот факт, что женщины избираются так же хорошо, как и мужчины, не обязательно означает, что избиратели не подвергают женщин дискриминации.

Продолжая эту линию рассуждений, Анциа и Берри пытаются сформулировать теоретические прогнозы, которые должны быть справедливы, если на выборах действительно существует дискриминация женщин. Один из прогнозов заключается в том, что при наличии дискриминации при прочих равных условиях женщины, избранные на должность, должны лучше справляться со своей работой, чем мужчины. Из-за дискриминации им придется стать лучше, чтобы победить на выборах. Конечно, у нас нет идеальных показателей эффективности работы. Но Анциа и Берри рассматривают несколько показателей эффективности работы членов конгресса, включая количество законопроектов, которые они вносят, а также объем федерального финансирования, который они приносят в свои округа. Результаты точно соответствуют теоретическим предсказаниям. Используя модель разности различий, они показывают, что в среднем (с учетом различий между округами и периодами времени, чтобы сделать сравнение как можно более корректным) женщины работают в конгрессе лучше, чем мужчины, что согласуется с предположением о необходимости преодолевать более высокие барьеры, чтобы добиться избрания из-за дискриминации избирателей.

Обнаружив интересный и захватывающий феномен, Анциа и Берри идут еще дальше. Очевидно, что в среднем среди членов конгресса женщины более продуктивны, чем мужчины, и дискриминация является потенциальным объяснением. Но существуют ли другие механизмы, которые потенциально могли бы объяснить этот эффект? Что, если, например, женщины просто лучше

справляются с некоторыми задачами, чем мужчины, независимо от какого-либо отбора или дискриминации? Или что, если к женщинам будут относиться по-другому, когда они попадут в конгресс, не из-за их способностей, а потому, что их рассматривают как символическое меньшинство?

Анциа и Берри постарались ответить и на эти вопросы. Если наблюдаемые явления действительно объясняются дискриминацией, существуют ли другие теоретические прогнозы, которые они могут проверить? Один из прогнозов заключается в том, что разрыв в эффективности между избранными женщинами и мужчинами будет тем больше, чем с большей дискриминацией сталкиваются женщины. Конечно, мы не знаем наверняка, какие избирательные округа конгресса дискриминируют женщин больше. Но одна разумная гипотеза заключается в том, что более консервативные округа будут дискриминировать кандидатов-женщин больше, чем либеральные округа. Поэтому Анциа и Берри проверили, не возрастает ли разница в эффективности между избранными в конгресс мужчинами и женщинами при переходе к рассмотрению только консервативных округов. Ответ: да.

Чтобы предоставить дополнительные доказательства существования этого механизма, Анциа и Берри отмечают, что одной большой группе женщин-представительниц конгресса – тем, кто получил должность, потому что они были вдовой недавно умершего члена конгресса, – вероятно, не пришлось преодолевать дискриминацию избирателей как кандидатам-женщинам. Таким образом, нет причин ожидать, что они будут в среднем более успешными, чем мужчины. И действительно, Анциа и Берри установили, что вдовы не работают лучше, чем мужчины – члены конгресса, и их результаты заметно хуже, чем у женщин, которые были избраны независимо от своих супругов.

Убедительность исследования Анциа и Берри базируется не на единственном достоверном исследовании или статистическом тесте, демонстрирующем наличие дискриминации. Напротив, они выявляют и объясняют интересный и правдоподобный механизм, создавая теорию дискриминации на выборах и проверяя многочисленные независимые прогнозы, следующие из этой теории.

Конечно, это еще не финал. Наблюдаемые закономерности можно объяснить и другими механизмами. Например, ряд ученых отмечают доказательства того, что женщины часто недооценивают свои способности или не склонны выдвигать себя в качестве кандидатов. Возможно также, что представительницы женского пола лучше справляются со своей работой именно при условии победы на выборах. Так что предстоит еще много работы, чтобы выяснить, какие механизмы работают на самом деле. Но, на наш взгляд, это исследование представляет собой прекрасный образец использования сочетания критического мышления и анализа данных в стремлении лучше понять причинные механизмы.

Естественные способы тестирования механизмов

В некоторых особых обстоятельствах мы можем планировать исследования таким образом, чтобы изолировать определенные механизмы. Возьмем, к примеру, продвинутое исследование Алана Гербера, Дона Грина и Кристофера Ларимера о том, как давление общества влияет на явку избирателей.

Давление общества и голосование

Гербер, Грин и Лаример разослали открытки случайно выбранной группе зарегистрированных избирателей. Открытки сообщали получателям, кто из ближайших соседей голосовал, а кто не голосовал на недавних выборах. (Возможно, вы этого не знали, но в Соединенных Штатах информация об участии в голосовании общедоступна. Секретом является только то, за кого был отдан голос.) Они также указали, что еще одна аналогичная открытка будет отправлена жителям района после завершения предстоящих выборов. Подразумевалось, что, если получатель не проголосует на предстоящих выборах, об этом узнают все его соседи. Эта необычная (и, возможно, агрессивная) открытка резко увеличила явку избирателей; люди, получившие открытку, показали явку на 8 % больше, чем контрольная группа.

У нас может возникнуть вопрос, почему открытки имели такой большой эффект, – т. е. мы хотим знать, с помощью каких механизмов открытки вызывают увеличение явки избирателей. Был ли важен социальный аспект воздействия? Неужели люди действительно не хотят, чтобы соседи знали, что они не голосуют? Люди мобилизуются только потому, что открытка напомнила им о выборах? Или люди, возможно, просто пытаются произвести впечатление на исследователей и меняют свое поведение, как только узнают, что их изучают?

Чтобы проверить важность социального механизма, исследователи усложнили свой эксперимент, включив в него еще одну случайно рассылаемую открытку. Эта открытка повторяла все особенности предыдущей, за одним исключением. Вместо информации о явке на выборы всех соседей получателя она содержала такую информацию только о членах семьи получателя, проживающих вместе с ним. Получатели подобных открыток не будут беспокоиться о том, что все их соседи узнают, если они не проголосуют. Теперь об их избирательном поведении узнают только люди, с которыми они живут. А эти люди наверняка и так знают о голосовании. Идея состоит в том, что это небольшое изменение устраняет из воздействия большую часть механизма социального давления. И действительно эта открытка тоже увеличила явку избирателей, но лишь примерно на 5 % относительно контрольной группы.

Самое интересное в этой схеме исследования заключается в том, что, включив в эксперимент несколько вариантов воздействия, авторы смогли оценить, насколько важен социальный аспект первой открытки. В частности, обнародование информации о явке избирателей по всему району, а не только внутри домохозяйства, по-видимому, составляет 3 % от общего эффекта в 8 %.

Косвенное выявление механизма

Иногда мы можем провести аналогичный анализ механизмов, даже если у нас нет возможности разработать исследование самостоятельно. Для этого, конечно, нам необходимо иметь несколько исследовательских планов, которые позволят отдельно оценить эффекты различных механизмов. Рассмотрим следующий пример.

Скачки цен на сырьевые товары и вооруженные конфликты

На протяжении десятилетий ученые изучали экономические условия, вооруженные конфликты и причинно-следственные связи между ними. Трудно при-

думать тему, где ставки были бы выше. Разумеется, человечеству хотелось бы улучшить экономические условия и сократить количество вооруженных конфликтов по всему миру, но пока не ясно, как это сделать.

Мы уже обсуждали сложность эмпирической оценки влияния экономики на конфликты в главе 9. Когда мы наблюдаем сильную корреляцию между конфликтом и плохими экономическими условиями, неясно, является ли первое причиной второго, или второе – первого, или какой-то третий фактор влияет на первые два, или же действует комбинация всех этих факторов.

Чтобы детально оценить хотя бы одну часть проблемы, многие ученые разработали исследования, направленные на изучение взаимосвязи между экономическими условиями и вооруженными конфликтами. Одна из распространенных стратегий предполагает использование шоковых скачков цен на сырьевые товары как часть расчета разности различий. Ключевая идея заключается в следующем.

Выращивание мака является ведущей отраслью промышленности в некоторых частях Афганистана. Можно предположить, что афганские фермеры и сельскохозяйственные рабочие, зарабатывающие деньги на выращивании мака, будут менее склонны воевать, потому что потеряют относительно хорошие экономические условия. Но очевидно, что мы не можем просто соотнести количество насилия в различных частях Афганистана с объемом выращивания мака, чтобы узнать о взаимосвязи между конфликтами и экономическими возможностями. Здесь много искажающих факторов. Мак является ключевым сырьем для производства героина, поэтому выращивание мака сопровождается торговлей наркотиками, что может независимо влиять на насилие. Причем мак произрастает в горных районах, а местность также может повлиять на уровень насилия.

Другая мысль заключается в том, что готовность фермеров и сельскохозяйственных рабочих включиться в вооруженный конфликт может снизиться, когда маковый бизнес особенно хорош, и повыситься, когда он плох. Но опять же такое сравнение во времени может сбить с толку. Возможно, всплеск спроса на мак также совпадает с особенностями сезона, размещением войск США или другими факторами, которые также влияют на насилие.

Но стратегия разности различий могла бы решить обе проблемы. То есть нам нужно найти различия в уровне насилия в местах, где выращивают мак, когда маковый бизнес хорош или плох, и сравнить их с такими же различиями в местах, где мак не производится. Идея состоит в том, что, принимая во внимание изменения в уровне насилия с течением времени и учитывая возможность того, что базовые уровни насилия различаются в регионах, выращивающих и не выращивающих мак, мы можем получить более достоверную оценку влияния экономического процветания на конфликты.

Чтобы реализовать подобную стратегию, конечно, нам нужна определенная оценка того, когда маковый бизнес хорош, а когда плох. Для этого исследователи используют изменения мировой цены на мак (или мировой цены на героин). Идея состоит в том, что относительно мирового сырьевого рынка большинство стран являются мелкими игроками. Таким образом, мировая цена этого товара вряд ли будет сильно зависеть от того, что происходит в конкретной стране. И поэтому, возможно, мы сможем использовать изменения мировых цен для оценки эффекта местного экономического процветания (вспомните инструментальные переменные из нашего обсуждения несоблюдения требований в главе 11).

Идея весьма привлекательная. Используя модели разности различий, мы можем оценить влияние экономических потрясений на насилие более достоверно, чем просто сравнивая уровни насилия в богатых и бедных странах даже на протяжении длительного времени.

Вместо того чтобы сделать это только для одного товара и одной страны, ученые проделали кропотливую работу по измерению того, сколько каждого из сотен товаров производит каждая страна. На основе этого они создали индекс товарного набора каждой страны. Они также собрали мировые цены на каждый товар за каждый год, чтобы можно было измерить, насколько ежегодно меняется стоимость набора товаров в каждой стране. Исходя из этого, они смогли провести гигантский анализ различий, чтобы увидеть, как уровень насилия меняется в ответ на экономические потрясения во всем мире.

Интересно, что, когда ученые сделали это, они получили множество противоречивых результатов. Иногда казалось, что положительные экономические перемены снижают уровень насилия, иногда – что они его увеличивают, а иногда они не имели заметного эффекта. Эти противоречивые и непоследовательные выводы смутили ученых. Более качественные данные и более строгие схемы исследований должны давать больше, а не меньше точных ответов.

Что происходит? Эрнесто и Педро Даль Бо предложили одно возможное теоретическое объяснение. Они отметили, что существует по крайней мере два механизма, посредством которых экономические условия могут влиять на конфликт, и они действуют в противоположных направлениях. С одной стороны, хорошие экономические условия создают больше рабочих мест лучшего качества, и рабочие могут быть менее склонны оставлять их ради борьбы. Другими словами, хорошие экономические условия увеличивают альтернативные издержки боевых действий. Если вы безработный и голодный, вы охотно присоединитесь к революции, но наличие хорошо оплачиваемой работы может удержать вас от перемен. С другой стороны, вооруженные группировки часто борются за контроль над экономическими ресурсами. Хорошие экономические условия означают, что есть за что бороться. Другими словами, хорошие экономические условия увеличивают выгоды от хищнического захвата ресурсов. Если вы живете в пустынном месте, не имеющем экономической ценности, какой смысл бороться за него? Но если на контроле над быстро развивающейся экономикой можно заработать деньги, то имеет смысл бороться за новое место под солнцем. Возможно, тот факт, что резкое изменение экономических условий может активизировать эти противоборствующие силы, объясняет неясные результаты предыдущих исследований.

Что же нам делать дальше? Осознание того, что существуют конкурирующие силы, теоретически дает много информации, но этого недостаточно для информирования политиков. К счастью, Эрнесто и Педро Даль Бо задумались об условиях, при которых одна сила может доминировать над другой. Их идея заключалась в том, что в трудоемких отраслях и экономиках должен доминировать механизм альтернативных издержек, поскольку экономические потрясения создают бóльшую потребность в рабочей силе, более высокую заработную плату и лучшие рабочие места. Но в капиталоемких отраслях и экономиках должен доминировать механизм хищничества и силового захвата, потому что улучшение экономических условий создает больше возможностей для борьбы без существенного повышения заработной платы или занятости.

В исследовании 2013 г. Оейндрила Дубе и Хуан Варгас нашли способ эмпирически проверить эти идеи, используя стратегию разности различий, которую мы уже описали. Они изучали вооруженный конфликт в Колумбии и сосредоточили большую часть своего внимания на двух основных отраслях промышленности: кофе и нефти. Кофе является относительно трудоемким процессом, требующим большого количества рабочих для выращивания и переработки. Нефть является относительно капиталоемкой: как только нефтяная скважина пробурена, производителям нефти не нужно много рабочей силы. Важно отметить, что в некоторых регионах Колумбии экономика ориентирована на кофе, а в других – на нефть. В соответствии с теоретическими предсказаниями положительные шоки мировых цен на кофе уменьшают конфликты в регионах с интенсивным производством кофе по сравнению с регионами, где кофе не выращивается. Это свидетельствует в пользу механизма альтернативных издержек. Напротив, позитивные потрясения мировых цен на нефть, похоже, усиливают конфликты в нефтедобывающих регионах по сравнению с остальными регионами. Это свидетельствует в пользу механизма хищничества.

В совокупности эти данные свидетельствуют о том, что экономические условия действительно влияют на насильственные конфликты посредством множества механизмов. Таким образом, улучшения экономики могут либо смягчить, либо усугубить конфликт, в зависимости от того, какой механизм доминирует в данном регионе. Это означает, что, если мы не видим явной взаимосвязи между экономическими потрясениями и конфликтами во всем мире, не факт, что экономические условия не имеют значения для конфликта. Скорее, эта взаимосвязь будет замаскирована множеством компенсирующих эффектов, которые мы сможем понять, только если выявим механизмы. Так что, возможно, экономический рост может снизить риск вооруженного конфликта, но только тогда, когда он сопровождается возможностями трудоустройства. Экономическая помощь, просто увеличивающая размер экономического пирога, за который могут бороться враждующие группировки, но не обеспечивающая возможностей трудоустройства, скорее всего, усугубит ситуацию. Это важное открытие, которое можно было сделать только посредством взаимодействия данных, схемы исследования, теории и критического мышления. Одних данных было недостаточно. Не обойтись и хорошей схемой исследования. Здесь были нужны все эти инструменты вместе.

Подведение итогов

В третьей части вы многому научились. Вы достигли одну из главных целей книги – по-настоящему поняли, почему корреляция не обязательно подразумевает причинно-следственную связь. Нам удалось выяснить, почему причинно-следственные выводы так сложны, а затем вы погрузились в захватывающий мир креативных исследовательских проектов, позволяющих достоверно оценить причинные эффекты и раскрыть механизмы воздействия. Это действительно важный материал, и мы надеемся, что вы остались довольны тем, что узнали.

Но даже хорошего знания причинно-следственных связей недостаточно для правильного использования количественной информации в области принятия обоснованных решений. В части IV мы обратимся к заключительным темам,

которые помогут нам задавать правильные вопросы, находить правильные свидетельства для ответа на эти вопросы и осознавать пределы возможностей количественной оценки.

Ключевые термины

- **Медиатор (посредник):** характеристика мира, на которую влияет воздействие и которая, в свою очередь, влияет на результат.
- **Анализ опосредованной причинно-следственной связи:** методы, позволяющие оценить, в какой степени наблюдаемая зависимость результата от воздействия является следствием воздействия на посредника и последующего влияния посредника на результат.

Упражнения

14.1. В 1990-х гг. Министерство жилищного строительства и городского развития США провело крупномасштабный полевой эксперимент под названием «Перемещение к возможностям». Они случайным образом выбрали несколько домохозяйств, живущих в государственном жилье в районах с высоким уровнем бедности, и предложили им жилищные ваучеры (т. е. деньги, которые можно было бы использовать для оплаты аренды), если они переедут в более богатый район. Остальным домохозяйствам ничего не дали. Целью было выяснить, принесет ли переезд в более процветающий район пользу для экономического, психического и физического состояния.

Исследователи изучили данные и обнаружили, что домохозяйства, испытавшие воздействие (получившие ваучер на переезд в более хороший район), имели лучшее физическое и психическое здоровье, а также субъективное благополучие, чем домохозяйства, не испытавшие воздействие (не получившие ваучер). Никаких существенных различий в экономическом состоянии домохозяйств не обнаружено.

a) Этот результат является убедительным доказательством того, что воздействие – получение жилищного ваучера, который вы можете использовать, только переехав в более благополучный район, – приводит к улучшению благополучия. Означает ли это, что воздействие работает через механизм физического перемещения людей в более хорошие районы? Предложите хотя бы еще один механизм, который мог бы объяснить полученный результат.

b) Как вы могли бы изменить или дополнить эксперимент, чтобы лучше выявить эффект от переезда в более благополучный район?

c) На самом деле в эксперименте рассматривалось еще одно воздействие. Другую группу людей случайным образом выбрали для получения ваучера на жилье, чтобы они могли переехать куда угодно (не обязательно в более богатый район). В свете этой информации какое сравнение вы бы провели, чтобы отделить эффект переезда в район с низким уровнем бедности от потенциальных последствий переезда как такового или получения финансовой выгоды в виде ваучера?

d) Бонусное задание: неудивительно, что в этом эксперименте встречались нарушения правил – некоторые люди, получившие ваучер, предпочли не переезжать. И, как вы можете себе представить, степень соблюдения правил воздействия была разной: 63 % домохозяйств использовали ваучер, когда не было ограничений на то, куда они могли переехать, но только 48 % использовали его, когда требовалось переехать в более благополучный район. Ситуация еще больше усложнялась тем, что некоторые домохозяйства, получившие свободу выбора, переехали в более бедные районы. Как можно оценить эффект от переезда в свете этих проблем с несоблюдением требований? (На этот вопрос нет простого ответа, но мы надеемся, что он поможет вам разобраться во всех затруднениях и осознать, насколько сложно изучать причинно-следственные механизмы.)

14.2. Существуют убедительные доказательства того, что образование увеличивает вовлеченность в политическую жизнь. Давайте подумаем, почему это может быть.

Некоторые люди предполагают, что это связано, по крайней мере частично, с механизмом дохода или богатства. Возможно, образование увеличивает экономическое процветание, богатых людей больше волнуют налоги или экономическая политика, и поэтому они с большей вероятностью будут голосовать. Проанализируйте приведенные ниже факты, которые могут быть использованы для обоснования этой гипотезы о механизмах. Насколько убедительным вы считаете каждое из них и почему?

a) Допустим, если вы построите регрессию явки избирателей от длительности обучения, то получите большой коэффициент, а если построите еще одну регрессию явки избирателей по длительности и доходу, коэффициент, связанный с обучением, будет заметно меньше.
b) Из-за требований об обязательном школьном образовании люди, родившиеся ближе к концу календарного года, как правило, получают больше образования (поскольку они молоды для своего класса и поэтому должны оставаться в школе на год дольше, прежде чем смогут прервать учебу, – такова особенность американской системы образования). Используя этот естественный эксперимент и инструментальные переменные, экономисты подсчитали, что обучение значительно увеличивает доходы. Используя другой естественный эксперимент, исследователи обнаружили, что выигрыш в лотерею увеличивает явку избирателей.
c) Предположим, вы узнали, что среди людей, получивших высшее инженерное образование, те, кто имеет более высокооплачиваемые специальности (например, в аэрокосмической, химической и нефтяной промышленности), участвуют в политической жизни больше, чем работники относительно низкооплачиваемых специальностей (например, социальные или экологические службы).

Дополнительное чтение и ссылки

Если вы хотите узнать больше про опосредованный причинно-следственный анализ, можете начать со следующих публикаций:

John G. Bullock, Donald P. Green, and Shang E. Ha. 2010. *Yes, ButWhat's theMechanism? (Don't Expect an Easy Answer).* Journal of Personality and Social Psychology 98 (4): 550–58;

Kosuke Imai, Luke Keele, Dustin Tingley, and Teppei Yamamoto. 2011. *Unpacking the Black Box of Causality: Learning about Causal Mechanisms from Experimental and Observational Studies.* American Political Science Review 105 (4): 765–89.

Исследование по поведенческой терапии в Либерии:

Christopher Blattman, Julian C. Jamison, and Margaret Sheridan. 2017. *Reducing Crime and Violence: Experimental Evidence from Cognitive Behavioral Therapy in Liberia.* American Economic Review 107 (4): 1165–1206.

Исследование о дискриминации женщин на выборах:

Sarah F. Anzia and Christopher R. Berry. 2011. *The Jackie (and Jill) Robinson Effect: Why Do Congresswomen Outperform Congressmen?* American Journal of Political Science 55 (3): 478–93.

Эксперимент по влиянию общества на явку избирателей:

Alan S. Gerber, Donald P. Green, and Christopher W. Larimer. 2008. *Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment.* American Political Science Review 102 (1): 33–48.

Исследования взаимосвязи экономического процветания и вооруженных конфликтов:

Ernesto Dal Bo and Pedro Dal Bo. 2011. *Workers, Warriors, and Criminals: Social Conflict in General Equilibrium.* Journal of the European Economic Association 9 (4): 646–77;

Oeindrila Dube and Juan F. Vargas. 2013. *Commodity Price Shocks and Civil Conflict: Evidence from Colombia.* Review of Economic Studies 80: 1384–1421.

Существует масса работ о последствиях программы переселения «Движение к возможностям». Вот одна из классических статей:

Lawrence F. Katz, Jeffrey R. Kling, and Jeffrey B. Liebman. 2001. *Moving to Opportunity in Boston: Early Results of a Randomized Mobility Experiment.* Quarterly Journal of Economics 116 (2): 607–54.

ЧАСТЬ IV

От информации к решению

Глава 15

Как наделить статистику смыслом

О чем эта глава

- Статистические данные часто представляют или визуализируют таким образом, что они вводят в заблуждение или бесполезны для принятия решений.
- Вдумчивый анализ рассматриваемого вопроса помогает преобразовать статистические показатели в более полезную прикладную информацию.
- Количественные данные сами по себе не могут сказать вам, во что верить. Ваши текущие убеждения зависят от комбинации новых фактов и предыдущих убеждений. Правило Байеса говорит нам, как обновлять убеждения в ответ на новую информацию.
- Количественные данные сами по себе не могут сказать вам, что делать. Принимая решение, нужно опираться на сочетание ваших текущих убеждений и ценностей.

Введение

Количественный анализ должен предоставить информацию, которая помогает принимать более обоснованные решения. Идеи, которые мы рассматривали до сих пор, – как установить, существует ли связь, возврат к среднему значению, разница между причинно-следственной связью и корреляцией, инструменты для оценки причинно-следственных связей и т. д., – являются важными компонентами на входе этого процесса. Но они не являются конечной точкой.

Предположим, вы подсчитали, что какое-то воздействие оказывает положительное влияние на какой-то результат. Означает ли это, что непременно следует провести воздействие? Вы не можете знать этого только на основе количественного анализа. Решение зависит от ваших убеждений и ценностей, а также от различных компромиссов, которые вам, возможно, придется рассмотреть. Чтобы перейти от фактов к действию, необходимо преобразовать статистическую информацию в предметный ответ на ваш вопрос.

В этот момент часто возникает путаница. Легко утратить критическое мышление, получив точные и авторитетные количественные данные, и прийти к неверным выводам даже на основе правильной информации. В этой главе мы рассмотрим, как избежать таких ошибок. Главное – превратить статистику

в осязаемое содержание, т. е. «овеществить» ее, чтобы быть уверенным, что вы задаете вопрос, который вас действительно волнует, и отвечаете именно на него.

Каков правильный масштаб?

Существуют разные способы точного и достоверного представления количественной информации. Но не все они одинаково полезны. То, как представлена информация, может оказать важное влияние на смысловое восприятие. Например, изменение масштабов может кардинально изменить то, кажутся ли отношения большими или маленькими, значимыми или неважными, веской причиной для принятия мер или нет. Поэтому, получив информацию, очень важно критически подумать о том, соответствует ли способ ее представления вопросу, на который вы пытаетесь ответить, или же данные нужно переосмыслить как-то иначе. Чтобы лучше понять, что мы имеем в виду, рассмотрим пару примеров.

Мили на галлон или галлоны на милю?

Допустим, вы работаете в агентстве по охране окружающей среды, контролирующем выбросы автомобилей. Ваша команда предлагает вам оценить два законопроекта. Один законопроект приведет к увеличению эффективности потребления топлива небольшими седанами из расчета 2 дополнительные мили пробега на галлон. Другой приведет к увеличению эффективности потребления топлива большими внедорожниками, тоже из расчета 2 дополнительных мили пробега на галлон. Предположим, что на дорогах одинаковое количество автомобилей этих двух видов, и в среднем каждый из них проезжает 10 000 миль в год. Седаны проезжают 30 миль на 1 галлоне (закон улучшит это значение до 32 миль на галлон), а внедорожники – 10 миль на 1 галлоне (закон улучшит это значение до 12 миль на галлон). Реализация новых правил для внедорожников будет стоить немного дороже. А поскольку каждый из этих двух законопроектов предусматривает увеличение пробега в 2 мили на галлон для одного и того же количества транспортных средств, проезжающих одинаковое количество миль в год, ваша команда рекомендует принять новые правила для седанов. Правильное ли это решение?

Давайте начнем с основного вопроса, на который вы хотите ответить. Ваша задача – сократить выбросы автомобилей за счет снижения потребления топлива. Приводит ли повышение пробега в 2 мили на галлон на седанах и внедорожниках к одинаковому сокращению расхода бензина? Сделаем статистику более осязаемой.

Внедорожники расходуют 10 миль на галлон, а это означает, что, поскольку средний водитель проезжает 10 000 миль в год, в среднем внедорожники расходуют 1000 галлонов бензина в год (10 000/10). Если вы введете в действие закон, который увеличит пробег до 12 миль на галлон, то в среднем внедорожники будут расходовать около 833 галлонов в год (10 000/12). Увеличение пробега в 2 мили на галлон экономит 167 галлонов бензина на внедорожник в год.

А что насчет седанов? Седаны проезжают 30 миль на одном галлоне, а это означает, что, поскольку средний водитель проезжает 10 000 миль в год, в среднем седаны расходуют около 333 галлонов бензина в год (10 000/30). Повы-

шение пробега до 32 миль на галлон приведет к тому, что седаны будут расходовать около 313 галлонов бензина в год (10 000/32). Увеличение пробега на 2 мили на галлон экономит всего около 20 галлонов бензина на седан в год.

Это не было очевидно, пока мы не овеществили статистические показатели, но теперь видно, что рекомендация вашей команды является неверной. Визуально одинаковое увеличение пробега на 2 мили на галлон оказывает гораздо большее влияние на расход топлива применительно к прожорливому внедорожнику, чем к относительно экономичному седану. Поэтому вам следует применить новые правила к внедорожникам, если только это не обойдется намного дороже.

На рис. 15.1 показано, сколько топлива расходует автомобиль, проезжающий 10 000 миль в год, в зависимости от пробега в милях на галлон. Как и ожидалось, расход топлива снижается по мере увеличения пробега в милях на галлон. Менее интуитивен, но важен для понимания результатов, которые мы только что показали, тот факт, что наклон этой кривой действительно крутой для низких значений миль на галлон (менее экономичные автомобили) и гораздо менее крутой для высоких значений миль на галлон (более экономичные автомобили). Вы получаете гораздо больше отдачи от вложенных средств, увеличивая пробег на галлон для неэффективных автомобилей, чем для автомобилей, которые и без того более эффективны. Это имеет интересные последствия, выходящие за рамки нашего примера. В частности, переход людей с обычных транспортных средств на несколько более экономичные транспортные средства дает гораздо больше пользы для сокращения выбросов, чем переход людей, уже пользующихся относительно экономичными транспортными средствами, на очень экономичные транспортные средства, такие как гибриды.

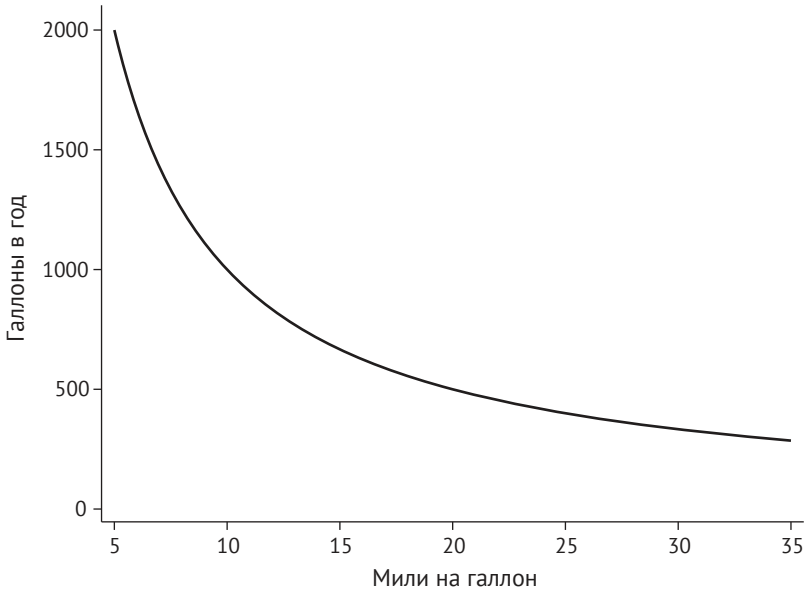


Рис. 15.1. Расход топлива при пробеге 10 000 миль в зависимости от количества миль на галлон

Вернемся к нашему примеру. В количественной информации, которую ваша команда использовала для формирования своих рекомендаций, не было ничего плохого или неправильного. Однако итоговая рекомендация была неверной. Почему? Проблема возникла из-за конкретной метрики, используемой для представления количественной информации. Мили на галлон – наиболее распространенный показатель топливной эффективности в Соединенных Штатах. Но это не особенно полезная метрика для принятия существенных решений.

Вопрос, который нас волнует, заключается в том, сколько бензина сжигает автомобиль с учетом расстояния, которое он проехал. Но количество миль на галлон показывает, какое расстояние проедет автомобиль, с учетом количества топлива, которое он сжег. Это информация, вывернутая наизнанку относительно вопроса. Нам пришлось выполнить определенные расчеты, чтобы сопоставить информацию с вопросом, на который нам действительно нужен ответ. Но большинство людей этого не сделают. На самом деле большинство людей даже не заметят разницы. И поэтому как потребители, так и регулирующие органы могут быть сбиты с толку (или обмануты), что заставит их принимать неправильные решения.

Если бы вы хотели предоставить более полезную информацию, вы бы использовали более наглядную меру топливной эффективности, например количество галлонов на сто миль, а не количество миль на галлон. Как мы только что видели, одно и то же улучшение расхода топлива на 2 мили на галлон приводит к совершенно разным улучшениям количества галлонов на милю, в зависимости от исходной топливной эффективности автомобиля. У вас не было бы проблем с принятием правильного решения, если бы ваша команда пришла к вам с анализом, где сказано, что в одном случае новый закон позволяет сэкономить около 8.3 галлона на сто миль (законопроект для внедорожников), а в другом – около 3.1 галлона на сто миль (законопроект для седанов).

В исследовании 2008 г. Ричард Ларрик и Джек Солл показали, что способ представления статистических показателей может иметь ключевое значение для принятия важных решений. Они цитируют автомобильного эксперта, который полагает, что нецелесообразно пытаться внести незначительные улучшения повышения пробега больших внедорожников, хотя на самом деле именно здесь инженеры и политики, вероятно, получат наибольшую отдачу от вложения средств в сокращение выбросов. Более того, они показывают, что потребителей часто вводит в заблуждение статистика, которая им представлена, и это может иметь серьезные последствия для решений о покупке.

В частности, Ларрик и Солл попросили студентов колледжей подумать, сколько они готовы заплатить за новую машину. Респондентам было предложено представить, что они проезжают 10 000 миль в год. Им показали машину, которая проезжает 15 миль на одном галлоне, и попросили представить, что они оценивают эту машину в 20 000 долл. Затем им показали альтернативные версии этого автомобиля, которые, как сообщается, идентичны базовой версии во всех отношениях, за исключением того, что они проезжают 19, 25, 33, 43 или 55 миль на одном галлоне. Сколько респонденты были бы готовы заплатить за эти более эффективные автомобили?

На рис. 15.2 показаны результаты. Черные точки – это средняя готовность платить (в тысячах долларов), о которой сообщили респонденты опроса. Заяв-

ленная готовность платить увеличивается примерно линейно с увеличением количества миль на галлон. Но это неправильно! Серые точки на рисунке показывают, как примерно респонденты должны были оценить эти автомобили (тоже в тысячах долларов), предполагая, что они будут использовать машину в течение десяти лет и у них будет 3-процентная ставка дисконтирования (т. е. доллар завтра будет стоить 97 сегодняшних центов). Как мы видели на рис. 15.1, увеличение топливной эффективности на 1 милю на галлон гораздо более ценно, если вы начинаете с низкого уровня эффективности, чем если вы начинаете с высокого уровня. Таким образом, респонденты в этом исследовании допускают большую ошибку в оценке этих гипотетических автомобилей.

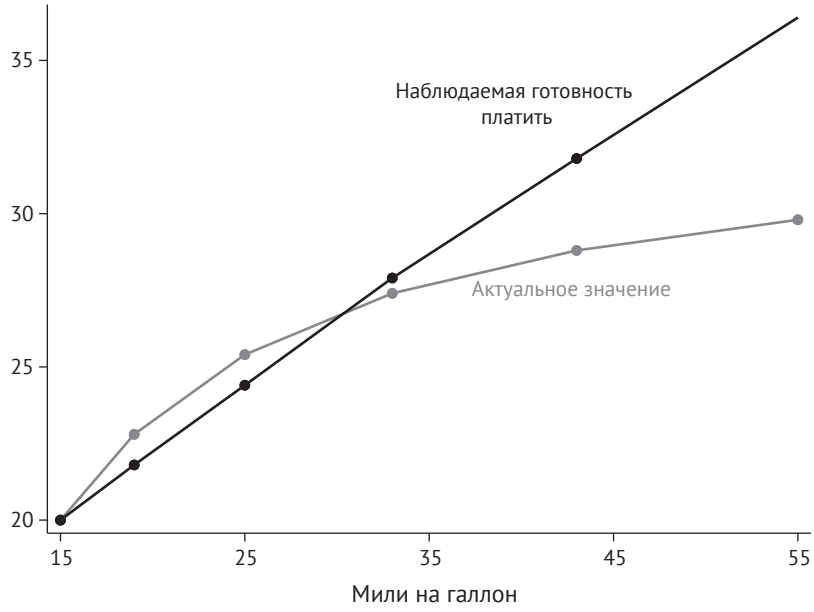


Рис. 15.2. Готовность платить по сравнению с фактической стоимостью эффективности использования топлива (в тысячах долларов)

Ларрик и Солл далее показывают, что эту ошибку легко исправить, если представлять эффективность использования топлива в галлонах на сто миль, а не в милях на галлон. Другими словами, разные способы передачи одной и той же информации могут иметь огромное значение для принятия решений, поэтому нам нужно подумать о наилучшем способе представления количественной информации, чтобы лица, принимающие решения, могли наилучшим образом воплотить свои предпочтения в действия.

Процент или процентный пункт?

Для оценки существенной важности того или иного эффекта нам нужно знать, насколько велик этот эффект. Существует как минимум два способа указать размер эффекта: изменение результата, который он вызывает, в процентах или в процентных пунктах. Это разные вещи! *Изменение в процентных пунктах* – это простая числовая разница между двумя процентными значения-

ми. *Изменение в процентах* – это отношение изменения на процентный пункт к первоначальному значению. Так, например, переход от 20 % к 22 % означает увеличение на *2 процентных пункта* (22 % – 20 %), но в то же время увеличение на *10 процентов* (2/20), что может привести к очень разным представлениям о величине эффекта. Поэтому важно проверить свое понимание, задумавшись о том, что имеет значение для вашего вопроса. Вот пример.

The Wall Street Journal сообщила о медицинском эксперименте, который показал, что новый препарат снижает «риск смерти от сердечных заболеваний, сердечных приступов и других серьезных проблем с сердцем на 44 %». Сокращение на 44 % звучит солидно. В сочетании с заголовком «Препарат от холестерина снижает риск сердечного приступа у здоровых пациентов» это указывает на необходимость сделать препарат максимально доступным для всех.

Но давайте остановимся и критически подумаем над существенным вопросом, который нас интересует при оценке такого количественного результата. Чтобы определить, стоит ли назначать конкретное лечение большой группе населения, мы хотели бы знать, сколько стоит препарат и сколько людей он спасет. Знание того, что лекарство снижает вероятность сердечного приступа среди здоровых людей на 44 %, на самом деле не говорит вам, сколько людей оно спасает. Чтобы определить это, вам также необходимо знать, насколько часто встречаются сердечные приступы среди этой группы населения.

Далее из статьи мы узнаем, что у 250 из 9000 человек, случайно попавших в контрольную группу и получавших таблетку плацебо, в ходе исследования случился сердечный приступ. Это предполагает, что базовый риск сердечного приступа составляет около 2.8 % (250/9000). Снижение частоты сердечных приступов на 44 % означает переход от примерно 2.8 % людей, перенесших сердечный приступ, к примерно 1.6 %. Поскольку сердечные приступы среди этой группы населения и без того случаются очень редко, сокращение числа сердечных приступов на 44 % означает снижение примерно на один процентный пункт – не такая уж огромная разница. И если препарат дорогой, вы вполне можете прийти к выводу, что его массовое применение нецелесообразно.

Здесь мы снова видим ценность овеществления статистики. В статье используются статистические данные, отвечающие на один вопрос (вызывает ли препарат заметное снижение количества сердечно-сосудистых заболеваний?), на который ответ положительный. Но заголовок создает впечатление, будто данные отвечают на гораздо более важный вопрос (спасет ли препарат много жизней?), на который в статье нет ответа. Преобразовав статистический показатель (процентное снижение) в суть (количество предотвращенных сердечных приступов на 100 пролеченных человек), мы можем легко обнаружить разницу и ответить на вопросы, наиболее важные для принятия решений.

Визуальное представление данных

Одним из наиболее распространенных способов представления и использования количественной информации является использование какого-либо графика, рисунка или визуального отображения. Мы тоже на протяжении всей книги использовали визуальное представление.

Точное и информативное отображение данных – это частично искусство,

частично наука. Поэтому стоит ненадолго остановиться и поразмышлять над некоторыми передовыми приемами. Есть отличные книги, почти полностью посвященные этой теме (см. раздел «Дополнительное чтение и ссылки» в конце этой главы), поэтому не будем углубляться в обсуждение. Но мы хотим затронуть некоторые основные моменты.

Самым существенным является следующее соображение: какими бы красивыми ни был графики и диаграммы, визуализация данных не заменяет критического мышления. Легко обмануться эстетически приятной, но вводящей в заблуждение графикой. Как потребитель количественных данных, вы должны сосредоточиться на критическом мышлении по существу вопроса. Какие данные и инструменты анализа привели к представленным числам и выводам? Верны ли лежащие в их основе предположения? Существуют ли другие статистические данные или представления данных, которые были бы более информативными? Отвечают ли представленные результаты на поставленный вопрос? Уместен ли масштаб, в котором представлены данные, или он был выбран для того, чтобы скрыть или преувеличить реальную величину взаимосвязи? Есть ли в графическом представлении ненужные, отвлекающие детали, которые могут ввести вас в заблуждение?

Выбор масштаба представления данных является одним из наиболее важных решений при создании визуализации. Безобидное на первый взгляд изменение масштаба может превратить график, на котором отношения или результаты выглядят огромными, в график, где они кажутся несущественными, и наоборот.

Чтобы понять, что мы имеем в виду, взгляните на рис. 15.3. На этом рисунке показаны четыре разные гистограммы, каждая из которых представляет собой просто сравнение числа 89 с числом 90. Но, изменяя масштаб – в данном случае, изменяя диапазон вертикальной оси, – мы изменяем степень увеличения или масштабирования. В результате мы можем сделать так, чтобы числа 89 и 90 сильно отличались друг от друга или были почти идентичными. И также можем заставить оба числа выглядеть очень большими или очень маленькими. Поэтому один из самых простых и важных способов убедиться, что вы правильно интерпретируете график, – это внимательно рассмотреть оси и подумать о том, что они означают на самом деле.

Важно отметить, что не существует правильного масштаба отдельно от рассматриваемого вопроса. Вы должны сами решить, что представляет собой существенно значимое различие в вашем конкретном контексте. Есть некоторые обстоятельства, при которых разница между 89 и 90 существенно велика. Например, если вы сопровождали 90 школьников на экскурсии, а домой вернулись только 89, то эта разница имеет большое значение. С другой стороны, вряд ли имеет значение, опоздает ли автобус, везущий детей домой, на 89 или 90 секунд. Правильный масштаб вашего графика зависит от того, в какой ситуации вы находитесь.

Если масштаб графика настолько велик, что вы не можете увидеть существенно значимые различия, это означает, что важная информация скрыта. Если разница в 1 балл очень важна, то диаграмма со шкалой от 88.9 до 90.1 (верхняя правая часть рис. 15.3) соответствующим образом отражает важное различие между 89 и 90, тогда как график по шкале от 0 до 1000 (нижняя левая часть) скрывает это различие.

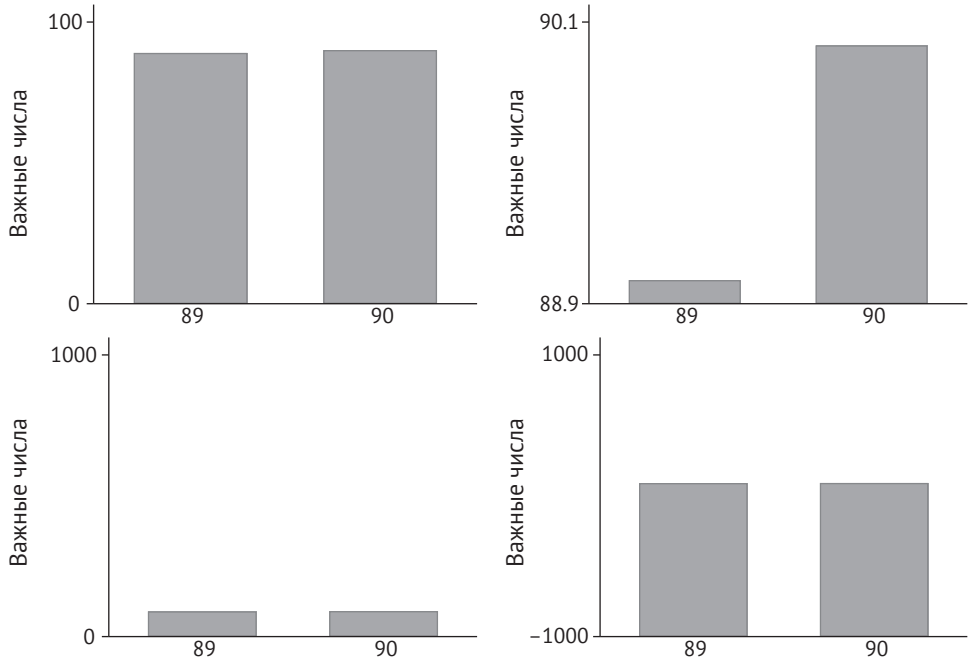


Рис. 15.3. Четыре разных способа показать разницу между 89 и 90 с помощью гистограммы

А если масштаб графика настолько мал, что различия, которые вас не интересуют, кажутся большими, вам следует задуматься о том, что результаты преувеличены. Например, если разница в 1 балл, по существу, незначительна, то график по шкале от 88.9 до 90.1 неуместно делает ее похожей на важную разницу, в то время как график по шкале от 0 до 100 точно отражает, что эти два числа, по сути, являются одинаковыми.

Опасения по поводу масштаба графиков применимы гораздо шире, чем просто эти несколько глупые гистограммы (в конце концов, вы можете просто назвать числа 89 и 90 в одном предложении и не строить графики). Изменяя масштаб осей, аналитики могут сделать корреляции сильными или слабыми, они могут сделать наклоны линий регрессии большими или маленькими и даже могут сделать линейную зависимость нелинейной или наоборот (например, с помощью выбора, показывать ли доход или логарифм дохода). Как мы уже говорили, существует множество хороших и плохих причин для преобразования переменной или тщательного выбора масштаба представления чего-либо. Аналитик всегда должен думать о том, как представить свои данные наиболее информативным образом, а потребитель должен превратить то, что ему показывают, в то, что его больше всего волнует.

Политические предпочтения и перестройка Юга

Рассмотрим пример. В книге 2016 г. Кристофер Эйкен и Ларри Бартельс утверждают, что политические взгляды избирателей мало связаны с голосованием на выборах. Они утверждают, что голосование определяется неполитическими соображениями. В качестве одного из подтверждающих доказательств Эйкен

и Бартельс утверждают, что политические взгляды не объясняют, почему белые избиратели на Юге США перешли от поддержки демократической партии к республиканской во время так называемой перегруппировки Юга, которая произошла во второй половине XX в. Доказательством этого утверждения является визуальное представление данных, которое мы попытались воспроизвести как можно точнее на рис. 15.4.

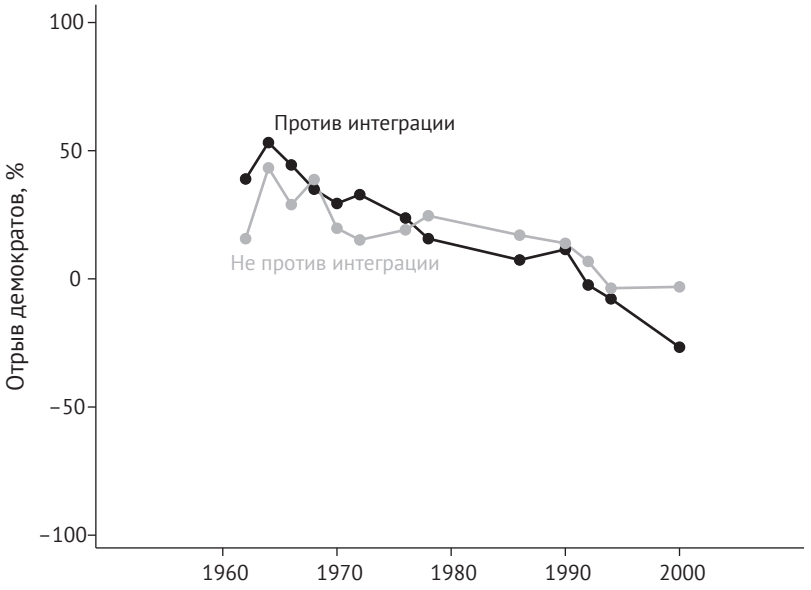


Рис. 15.4. Тенденции в партийной поддержке белых южан, которые выступали против интеграции и не выступали против нее

На рисунке отдельно показана тенденция партийной принадлежности белых южан, которые выступали против интеграции и не выступали против нее. Горизонтальная ось – годы. Вертикальная ось показывает преобладание демократов, измеряемое как процент людей, идентифицирующих себя как демократы, минус процент людей, идентифицирующих себя как республиканцы. Таким образом, чем выше точка данных по вертикальной оси, тем больше демократов по сравнению с республиканцами.

На рисунке ясно показана политическая перегруппировка Юга. В 1960 г. белые южане в подавляющем большинстве были демократами. Но со временем ситуация изменилась, и к концу XX в. они в подавляющем большинстве были республиканцами.

Ахен и Бартельс утверждают, что эти данные также показывают, что политические позиции белых избирателей по вопросам интеграции не влияют на изменение партийной принадлежности. То есть они утверждают, что эти две тенденции более или менее одинаковы. И это, по их мнению, говорит о том, что позиции избирателей даже по весьма важным политическим вопросам не влияют на партийную принадлежность.

Что вы заметили на рис. 15.4? Очевидно ли, что тенденции более или менее одинаковы? Во-первых, мы могли бы взглянуть на масштаб вертикальной оси.

Показатель партийной принадлежности – процент людей, идентифицирующих себя как демократов, минус процент людей, идентифицирующих себя как республиканцев, – теоретически может варьироваться от –100 до 100. Именно в этом масштабе и представлены данные. Однако на практике многие люди не идентифицируют себя ни как демократы, ни как республиканцы, поэтому практически в любой большой популяции мы, вероятно, не увидим отрыва демократов даже близко к теоретическому минимуму или максимуму. Поскольку диапазон оси настолько велик, не скрывается ли здесь существенная значимая разница, которую трудно увидеть, как в правой нижней части рис. 15.3?

Также обратите внимание на горизонтальную ось. Рисунок включает данные только с 1962 по 2000 г., но график достаточно широк, чтобы включить данные с 1950 по 2010 г., оставляя кучу пустого и ненужного места. Нет веской причины оставлять это место пустым. Но из-за него сжимается изображение.

Как изменились бы наши основные выводы, если бы мы удалили часть этого бесполезного пространства и перерисовали ту же количественную информацию в масштабе, который более точно отражает наблюдаемый диапазон данных? Вы можете видеть это на рис. 15.5. Мы также добавили линии линейной регрессии, которые, по нашему мнению, облегчают визуализацию средних тенденций для двух разных групп избирателей.

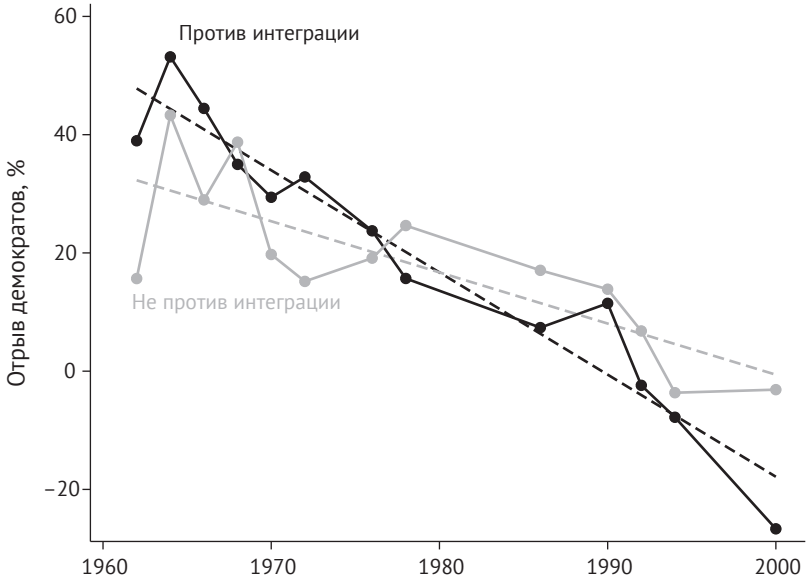


Рис. 15.5. Тенденции изменения партийной принадлежности белых южан, которые выступали и не выступали против интеграции в более подходящем масштабе и с линиями регрессии

Визуальное представление данных на рис. 15.5 предполагает существенно отличающуюся интерпретацию относительно рис. 15.4. В частности, рис. 15.5 показывает, что тенденции изменения партийной принадлежности на самом деле были совершенно разными для людей, которые выступали против интеграции, по сравнению с теми, кто не выступал против нее. Среди тех, кто выступал против интеграции, в 1960-е гг. было больше демократов. Но к концу

XX в. среди противников интеграции стало больше республиканцев. Таким образом, их линия тренда была существенно более крутой: люди, выступавшие против интеграции, меняли партийную принадлежность быстрее, чем люди, которые не выступали против интеграции. Возможно, политические взгляды действительно помогают объяснить сдвиг в партийной идентификации во время перегруппировки Юга.

Чтобы не попасться на уловку, показанную в верхней правой части рис. 15.3, нам нужно интерпретировать числа по существу и убедиться, что видимая на графике большая разница имеет существенное значение. Мало кто из нас регулярно задумывается о процентном разрыве в партийной самоидентификации, поэтому нужно попробовать найти более эффективные способы передать эту информацию. Давайте немного поразмыслим.

Мы видим, что с 1962 по 2000 гг. белые южане, выступавшие против интеграции, прошли путь от разрыва примерно 48 пунктов в пользу демократической партии до 18 пунктов в пользу республиканской партии. Среди тех, кто не выступал против интеграции, также стало больше республиканцев, но изменение было скромнее – с 32-балльного перевеса демократов до 1-балльного перевеса республиканцев. Таким образом, изменение среди противников интеграции было на 33 процентных пункта больше, или в два раза больше, чем изменение сторонников интеграции.

Но большая это или маленькая разница? В качестве ориентира, если мы посмотрим на данные за 2020 г., 33 процентных пункта – это примерно разница в количестве демократов между Массачусетсом и Айдахо. Поэтому мы считаем, что можно с уверенностью сказать, что две тенденции, которые различаются на 33 пункта по этой шкале, на самом деле имеют существенное различие, а визуализация на рис. 15.4 скрывает эту важную информацию.

Мы проиллюстрировали один набор вопросов, которые можно задать по поводу рис. 15.4, и показали, насколько они важны. Но мы лишь коснулись вопросов, которые стоит задать, когда вы пытаетесь овеществить статистические показатели. Например, почему правильным критерием для оценки политического поведения является партийная самоидентификация, а не что-то более политически значимое, например поведение на выборах? Зачем начинать этот анализ с 1962 г., когда многие считают, что перегруппировка Юга началась раньше? Является ли единственный вопрос о взглядах на интеграцию лучшим способом измерения политических предпочтений в этом контексте?

Некоторые ключевые правила визуализации данных

При интерпретации визуализированных данных есть над чем подумать. Как мы уже говорили, мы не имеем возможности дать здесь исчерпывающий обзор. Но мы перечислим некоторые ключевые принципы, которые, по нашему мнению, важно учитывать при создании или использовании графических изображений количественной информации.

- Будьте проще. Если вам не нужно несколько цветов, не используйте цвета. Если вам не нужна необычная графика, не добавляйте ее. Если третье измерение не добавляет чего-то важного, используйте двухмерный график. Если у вас есть сложные легенды и метки, попробуйте разбить их на разные графики.

- В центре внимания должно быть содержание. Вы пытаетесь передать информацию в ясной и легкоусвояемой форме. Убедитесь, что выбранный вами дизайн способствует достижению цели – дать ответ на поставленный вопрос.
- Если вы просто показываете какие-то простые числа (например, 89 и 90 или коэффициент регрессии), возможно, имеет смысл отказаться от графиков и представить числа в таблице. Оставьте рисунки для ситуаций, когда они могут передать больше информации, чем таблица.
- Показывайте данные. Одна из замечательных особенностей рисунков заключается в том, что вы можете показать гораздо более сложные взаимосвязи и гораздо больше деталей, чем это можно было бы сделать с помощью таблицы. Если цель вашего рисунка – просто показать точку пересечения и наклон линии регрессии, вы можете обойтись таблицей. Но график может многое добавить к пониманию чисел, если вы построите и линию регрессии, и данные, лежащие в основе регрессии, чтобы можно было увидеть, является ли связь приблизительно линейной или нет. Вспомните о рис. 2.5 или 5.8. Мы извлекаем намного больше из визуализаций, по сравнению с простым сообщением о коэффициенте корреляции или регрессии, именно потому, что вместе с ними отображаются основные данные.
- Если возможно, передайте неопределенность. Показ ваших данных – хороший способ сделать это. Вместо того чтобы просто показывать средние значения, попробуйте отобразить распределения. Если вы строите оценки, рассмотрите также возможность построения стандартных ошибок или доверительных интервалов, как мы это сделали на рис. 12.4.

От статистики к убеждениям

Данные никогда не говорят сами за себя. Факты всегда интерпретируются в свете наших существующих представлений о том, как устроен мир, других фактов, которые встречались ранее, и т. д. Чтобы использовать количественную информацию, важно вдумчиво и осторожно интегрировать эту новую информацию в наш существующий запас знаний, чтобы мы могли преобразовать статистику в убеждения. Ключевой инструмент, который у нас есть для этого, называется *правилом Байеса*. Чтобы показать важность правила Байеса и продемонстрировать, как оно работает, начнем с примера.

В 1964 г. в Лос-Анджелесе пожилая женщина по имени Хуанита Брукс шла по переулку, волоча за собой тележку с продуктами, а ее сумочка лежала сверху, когда женщину толкнули на землю сзади, а сумочку украли. Она не успела как следует разглядеть преступника. Примерно в это же время случайный свидетель видел, как из того же переулка выбежала женщина и села в желтую машину. Свидетель тоже не очень хорошо ее рассмотрел. Но он заметил, что бежавшая женщина была белой, со светлыми волосами, собранными в хвост, а водителем машины был чернокожий мужчина с бородой и усами. На основании показаний очевидца полиция позже арестовала Малкольма и Джанет Коллинз и обвинила их в ограблении. Малькольм был чернокожим мужчиной с бородой и усами. Джанет была белой женщиной со светлыми волосами, собранными в хвост. И они ездили на желтой машине.

Как рассказывает в статье Джонатан Келер, прокуратура привлекла математика для дачи показаний относительно вероятности того, что имеющиеся скудные свидетельства подтверждают вину Малькольма и Джанет в ограблении. Математик пришел к выводу, что вероятность невиновности пары составляет всего лишь 1 из 12 млн. Вот в чем заключалась его логика.

Если бы мы совершенно случайно арестовали невиновную пару, очень маловероятно, что муж был бы черным с бородой и усами, жена была бы белой со светлым хвостом и что они бы ездили на желтой машине. Почему это?

Доказательство начинается с некоторых количественных фактов. Если мы просто случайным образом выберем человека из населения, есть 10-процентная вероятность того, что он будет черным, потому что около 10 % населения США – чернокожие. Предположим, что 10 % всех мужчин носят бороды, поэтому вероятность того, что у него будет борода, также составляет 10 %. Возможно, есть 20-процентная вероятность того, что у него будут усы. И вероятность того, что он будет водить желтую машину, составляет всего 0.5 %, учитывая количество желтых машин на дороге.

Как нам взять эти числа и превратить их в общую вероятность того, что случайно выбранная невинная пара будет иметь такое сочетание характеристик? Давайте рассмотрим аналогию с колодой карт. Какова вероятность того, что случайно выпавшая карта окажется четверкой червей? Вероятность того, что случайно вытянутая карта окажется четверкой, равна 1 к 13. И вероятность того, что случайно вытянутая карта окажется червовой масти, равна 1 из 4. Поскольку вероятности четверки и червовой масти независимы (т. е. знание того, что карта является четверкой, ничего не говорит вам о том, какой она может быть масти), вероятность того, что эти две характеристики возникнут вместе, является просто произведением вероятности возникновения каждой из них по отдельности. Таким образом, если мы вытащим случайную карту из колоды, вероятность того, что это четверка червей, равна $\frac{1}{13} \times \frac{1}{4} = \frac{1}{52}$. Вполне логично. В колоде 52 карты. Только одна из них – четверка червей.

Прокурор применил ту же логику к чете Коллинз. Он утверждал, что шансы на то, что случайно выбранный человек будет черным, будет иметь усы и бороду и будет водить желтую машину, являются произведением вероятностей индивидуальных характеристик: $\frac{1}{10} \times \frac{1}{10} \times \frac{1}{5} \times \frac{1}{200} = \frac{1}{100\,000}$. Он продолжал добавлять характеристики (наличие жены, принадлежность к межрасовой паре, женщина со светлыми волосами и хвостом и т. д.), в конечном итоге придя к вероятности 1 к 12 млн. Действительно, как указал прокурор, даже это была заниженная оценка, поскольку у пары было много других характеристик, которые не были учтены, так что вероятность невиновности, вероятно, была скорее 1 к 1 млрд! Присяжные признали супругов Коллинз виновными, а газеты похвалили прокуроров за столь строгое количественное обоснование.

Как вы думаете: отражает ли этот пример критическое мышление? Мы надеемся, что вы сказали «нет», потому что на самом деле здесь так много неправильного, что трудно решить, с чего начать. Но начать нужно.

Итак, во-первых, вышеупомянутые характеристики (в отличие от червы и четверки в колоде игральных карт) не являются независимыми друг от друга. Таким образом, вы не можете просто перемножить вероятности каждой отдельной ха-

рактеристики, чтобы получить вероятность сочетания характеристик. Например, наличие бороды положительно коррелирует с наличием усов. Таким образом, вероятность иметь бороду и усы намного выше, чем вероятность иметь бороду, умноженную на вероятность иметь усы. То есть если 1 из 10 мужчин имеет бороду, а 1 из 5 – усы, то многие из них – одни и те же мужчины. Следовательно, намного больше, чем 1 из 50 мужчин, имеют бороду и усы одновременно. На самом деле вероятность гораздо ближе к 1 из 10, поскольку почти у каждого, у кого есть борода, есть и усы. Итак, если бы мы приняли во внимание все соответствующие корреляции, возможно, мы бы не пришли к выводу, что вероятность того, что случайно выбранная пара соответствует описанию очевидца, составляет 1 к 12 млн. Но мы все равно получим довольно низкую вероятность (возможно, один на миллион). Это все еще кажется надежным доказательством для обвинения, не так ли?

Нет. На самом деле это не так. Мы еще не говорили о главной ошибке в анализе, а именно о том, что он отвечает на полностью неправильный вопрос. Если мы критически подумаем над правильным вопросом, то придем совсем к другому выводу.

Присяжным предстоит решить, признать ли чету Коллинз виновной в преступлении. Они не должны этого делать, если существует достаточная вероятность того, что пара Коллинз невиновна, но должны это сделать, если есть достаточная вероятность обратного. Итак, правильный вопрос для присяжных: насколько вероятно, что чета Коллинз *невиновна*, учитывая свидетельства очевидцев? Доказательством является тот факт, что пара Коллинз соответствует описанию очевидца. Таким образом, подходящим статистическим показателем для ответа на правильный вопрос присяжных является вероятность того, что супруги Коллинз невиновны, при условии, что их внешний вид совпадает с описанием свидетелей. Запишем это высказывание как Pr(невиновны | совпадение). Это так называемая *условная вероятность*, поскольку это вероятность того, что одно событие обусловлено другим. Выражение условной вероятности можно прочитать двумя способами. Люди говорят либо «вероятность их невиновности при условии совпадения доказательств», либо «вероятность их невиновности при наличии данных доказательств». То и другое одинаково верно. Вероятность их виновности при условии совпадения с показаниями Pr(виновны) = 1 – Pr(невиновны).

Математический анализ, который мы обсуждали до сих пор, не предоставил нам эту вероятность и, следовательно, не ответил на правильный вопрос. Анализ, проведенный на данный момент, показывает нам, насколько вероятно, что случайно выбранная пара будет соответствовать описанию очевидца. То есть он предоставил нам Pr(совпадение | невиновны), что читается как «вероятность того, что для пары совпадут свидетельства, при условии, что они невиновны». Хотя эта статистика может быть полезна для ответа на вопрос присяжных, она сама по себе не является ответом. Присяжным нужно знать Pr(невиновны | совпадение). Прокурор предоставил им Pr(совпадение | невиновны). Но присяжные (и пресса) не заметили разницы, потому что они не привыкли мыслить критически.

Давайте посмотрим, почему это важно. Предположим, мы договорились, что вероятность Pr(совпадение | невиновны) примерно равна 1 к 1 000 000. Нам нужно выяснить Pr(невиновны | совпадение). Можем ли мы это сделать?

Нам поможет табл. 15.1. Она классифицирует супружеские пары в округе Лос-Анджелес по двум характеристикам: соответствуют ли они описанию очевидца или нет и виновны они или нет. Мы знаем, что виновна ровно одна пара, и эта пара соответствует описанию очевидца. Таким образом, столбец «Виновны» заполнить легко. Столбец «Невиновны» немного сложнее. Мы согласились с тем, что анализ прокурора, а также некоторые предположения показывают, что вероятность того, что невиновная пара соответствует описанию очевидца, составляет примерно 1 к 1 000 000. Если мы приблизительно примем, что в 1964 г. в округе Лос-Анджелес насчитывалось около 2 млн невиновных супружеских пар, мы придем к выводу, что существовали примерно две невиновные пары, которые также соответствуют описанию. Остальные 1 999 998 пар в округе Лос-Анджелес попадают в последнюю ячейку: невиновные и несовпадающие.

Таблица 15.1. Распределение пар из Лос-Анджелеса по виновности и невиновности, а также по тому, соответствуют ли они свидетельствам

	Невиновны	**Виновны**
Не соответствуют	1 999 998	0
Соответствуют	2	1

Итак, насколько велика вероятность того, что пара окажется невиновной, если они соответствуют описанию, т. е. что представляет собой Pr(невиновны | совпадение)? Итак, в Лос-Анджелесе есть три пары, соответствующие описанию. Виновна ровно одна из них. Таким образом, истинная вероятность того, что пара *невиновна*, учитывая, что они соответствуют описанию очевидца, не равна 1 на 1 000 000. Это 2 из 3. Это означает, что вероятность *виновности* пары, учитывая, что они соответствуют описанию очевидца, составляет всего 1 из 3. Судя по показаниям очевидцев, пара Коллинз скорее была невиновна, чем виновна!

Расхождение возникает не потому, что математик, прокурор и пресса использовали неправильную количественную информацию, а потому, что они использовали количественную информацию для ответа на неправильный вопрос. Как заявили математик и прокурор, очень маловероятно, чтобы случайно выбранная невиновная пара совпала с описанием преступников. Но это не значит, что маловероятно, что пара, подходящая под описание преступников, невиновна. Только одна невинная пара из миллиона соответствует этому описанию. Но две пары из трех, подходящие под описание, невиновны. Если бы присяжные могли более критически отнестись к предоставленной информации, то, скорее всего, чета Коллинз не была бы осуждена. Лишь немногие присяжные согласятся отправить людей в тюрьму на том основании, что вероятность того, что они совершили преступление, составляет 1 из 3[1].

[1] Интересный факт: Малкольм Коллинз обжаловал обвинительный приговор на том основании, что прокурор использовал ошибочный математический аргумент, чтобы осудить его. Верховный суд Калифорнии отменил приговор, аргументируя важность критического мышления. В нем говорилось: «Математика, настоящий волшебник в нашем компьютеризированном обществе, помогая исследователю фактов в поисках истины, не должна околдовать его».

Правило Байеса

Анализ, который мы только что провели, является примером общего подхода к выяснению того, во что нам следует верить при наличии некоторых свидетельств. Математический инструмент, называемый *правилом Байеса* (или иногда теоремой Байеса или законом Байеса), дает нам формулу для расчета этого значения. Он назван в честь Томаса Байеса, философа и статистика XVIII в.

Правило Байеса подсказывает нам правильную формулу вычисления того, насколько вероятно, что утверждение будет верным при условии имеющихся свидетельств. Это происходит следующим образом. Предположим, мы хотим узнать вероятность того, что утверждение C истинно при наличии свидетельства E. То есть мы хотим знать $\Pr(C \mid E)$. В нашем примере утверждалось, что чета Коллинз невиновна, а свидетельства говорили о том, что внешний вид пары совпадает с описанием очевидца. Правило Байеса гласит:

$$\Pr(C \mid E) = \frac{\Pr(E \mid C)\,\Pr(C)}{\Pr(E)}.$$

Давайте вернемся к делу Коллинзов, чтобы лучше разобраться в этом уравнении. Мы хотим знать вероятность невиновности четы Коллинз при условии, что они совпадают с описанием очевидца. В этом случае правило Байеса гласит:

$$\Pr(\text{Невиновны} \mid \text{Совпадение}) = \frac{\Pr(\text{Совпадение} \mid \text{Невиновны})\,\Pr(\text{Невиновны})}{\Pr(\text{Совпадение})}.$$

Мы можем использовать табл. 15.1, чтобы найти значения для подстановки в уравнение и увидеть, как это работает.

Что такое Pr(совпадение | невиновны)? Это вероятность совпадения пары с описанием свидетеля при условии, что они невиновны. Есть 2 млн невиновных пар. Две из них совпадают с описанием. Следовательно, Pr(совпадение | невиновны) = 2/2 000 000.

Что такое Pr(невиновны)? Это общая вероятность того, что случайная пара невиновна. В округе Лос-Анджелес насчитывается 2 000 001 супружеская пара. Из них 2 000 000 невиновны. Следовательно, Pr(невиновны) = 2 000 000/2 000 001.

Наконец, что такое Pr(совпадение)? Опять же, есть 2 000 001 пара, из них 3 соответствуют описанию очевидца. Тогда Pr(совпадение) = 3/2 000 001.

Подставив эти значения в уравнение, получаем:

$$\Pr(\text{Невиновны} \mid \text{Совпадение}) = \frac{\Pr(\text{Совпадение} \mid \text{Невиновны})\,\Pr(\text{Невиновны})}{\Pr(\text{Совпадение})}$$

$$= \frac{\frac{2}{2{,}000{,}000} \cdot \frac{2{,}000{,}000}{2{,}000{,}001}}{\frac{3}{2{,}000{,}001}} = \frac{2}{3}.$$

Обратите внимание: ранее мы смогли это выяснить, не зная правила Байеса, просто взглянув на таблицу. Поэтому запоминать формулу не нужно. Но важно знать, как вычислить убеждения на основе фактов, и убедиться, что вы правильно понимаете, какой вопрос хотите задать и как на него ответить. Потому что действительно легко убедить себя, что Pr(совпадение | невиновны) – это

то же самое, что Pr(невиновны | совпадение). Но, как мы теперь увидели, они могут быть очень разными.

Информация, априорные и апостериорные убеждения

Правило Байеса срабатывает каждый раз, когда мы получаем новую информацию и хотим обновить наши убеждения о том, насколько вероятно, что какое-то утверждение верно. Прежде чем мы получим новую информацию, у нас есть то, что мы называем *априорным убеждением* (prior belief) относительно этого утверждения, т. е. мы обладаем определенным убеждением в вероятности того, что утверждение истинно, пока не получены новые свидетельства. В формуле это априорное убеждение представлено членом Pr(C) – вероятностью, что утверждение верно, без ссылки на свидетельства. После того как получена новая информация, правило Байеса дает нам то, что мы называем *апостериорным убеждением* (posterior belief): Pr($C \mid E$).

В деле «Народ против Коллинз», предварительным убеждением является базовая вероятность невиновности четы Коллинз до того, как были заслушаны показания очевидца. На тот момент не было причин подозревать их больше, чем любую другую пару, живущую в Лос-Анджелесе, поэтому априорное убеждение было очень близко к 1 – что-то около 2 000 000/2 000 001, поскольку все пары, кроме одной, были невиновны.

Мы узнали, что чета Коллинз соответствует описанию преступников. Фактически вероятность того, что случайная невиновная пара соответствует этому описанию, составляет всего 1 к 1 000 000, и это может заставить нас думать, что они почти наверняка виновны. Но правило Байеса предписывает нам подумать, прежде чем делать поспешные выводы. С одной стороны, доказательства кажутся весьма убедительными. Крайне маловероятно, что невиновная пара случайно совпадет с описанием по нескольким критериям. С другой стороны, априорное убеждение толкает нас в другом направлении. Крайне маловероятно, что какая-либо случайно выбранная пара виновна в преступлении. Чтобы выяснить, насколько вероятно, что чета Коллинз виновна, учитывая оба этих факта, мы должны задаться вопросом об относительном правдоподобии каждого из них. Если мы проигнорируем либо наше априорное убеждение, либо новые свидетельства, мы придем к неправильному выводу. Объединив и то, и другое, мы видим, что, хотя у пары Коллинз гораздо больше шансов быть виновными, чем у случайно выбранной пары, все же существует большая вероятность того, что они невиновны.

Один из способов понять, в чем подвох аргументации прокурора, заключается в том, что он говорил только о новых свидетельствах, игнорируя предыдущие. Это распространенная ошибка, которую совершают люди, не обученные правильно оперировать количественными данными.

Возвращаясь к целиакии Эйба

Еще в главе 1 мы рассказывали вам историю о сыне Итана, Эйбе, которому ошибочно поставили диагноз целиакия. Напомним основные моменты этой истории.

В детстве Эйб был чересчур мал для своего возраста, что является признаком целиакии. Его педиатры сделали два анализа крови. Один из них оказался положительным (свидетельство того, что у ребенка заболевание), другой – отри-

цательным (свидетельство того, что заболевания нет). Врачи пришли к выводу, что Эйб, вероятно, болен целиакией, поскольку у теста, давшего положительный результат, «точность более 80 %».

Тест на целиакию, оказавшийся отрицательным (назовем его Тестом 1), имел довольно низкий уровень ложноотрицательных и ложноположительных результатов, около 5 % каждый. Мы можем записать это в наших новых обозначениях. Доля ложноотрицательных результатов – это вероятность того, что вы получите отрицательный результат теста при условии, что у вас есть заболевание, т. е. Pr(отрицательный результат Теста 1 | есть целиакия) = 0.05. Доля ложноположительных результатов – это вероятность того, что вы получите положительный результат теста при условии, что у вас нет заболевания, т. е. Pr(положительный результат Теста 1 | нет целиакии) = 0.05.

Тест на целиакию, который оказался положительным (назовем его Тестом 2), имел долю ложноотрицательных результатов около 20 %, т. е. Pr(отрицательный результат Теста 2 | есть целиакия) = 0.2. Мы подозреваем, что именно отсюда и появилось утверждение о «точности 80 %». Доля ложноположительных результатов этого теста составляет 50 %, т. е. Pr(положительный результат Теста 2 | нет целиакии) = 0.5.

До анализа крови разумное предположение о вероятности заболевания Эйба целиакией, учитывая его небольшой рост, составляло примерно 1 к 100. Это априорное предположение Итана: Pr(целиакия) = 0.01.

Давайте на секунду проигнорируем Тест 1 и просто применим правило Байеса к Тесту 2. Представьте себе группу из 10 000 детей, все из которых были одинаково маленького роста. Наше априорное убеждение говорит нам, что из этих 10 000 детей около 100 (1%) будут страдать целиакией. Доля ложноотрицательных результатов Теста 2 говорит нам о том, что из 100 детей с целиакией около 20 (20 %) тем не менее будут иметь отрицательный результат, а остальные 80 – положительный. Доля ложноположительных результатов Теста 2 говорит нам о том, что из 9900 детей без целиакии около 4950 (50 %) тем не менее будут иметь положительный результат теста, и 4950 – отрицательный. В табл. 15.2 представлена сводная информация.

Таблица 15.2. Результаты теста на целиакию у 10 000 детей

	Есть целиакия	**Нет целиакии**
Отрицательный Тест 2	20	4950
Положительный Тест 2	80	4950

Так какова вероятность того, что у Эйба целиакия, учитывая, что он был маленького роста и его Тест 2 дал положительный результат? Итак, в общей сложности 4950 + 80 = 5030 детей дали положительный результат. Из них 80 страдают целиакией. Таким образом, вероятность того, что у одного из этих детей целиакия при положительном результате Теста 2, равна 80/5030, или примерно 1.6 %.

Теперь, зная правило Байеса, мы могли бы получить тот же результат, не создавая таблицу:

$$\Pr(\text{Целиакия} \mid \text{Положительный Тест 2}) = \frac{\Pr(\text{Положительный Тест 2} \mid \text{Целиакия})\ \Pr(\text{Целиакия})}{\Pr(\text{Положительный Тест 2})}.$$

Мы знаем достаточно, чтобы вычислить каждую из этих величин. Pr(положительный результат Теста 2 | целиакия) равен 1 минус доля ложноотрицательных результатов, которая равна 0.8. Pr(целиакия) – это наше априорное убеждение, которое равно 0.01.

Вычислить Pr(положительный результат Теста 2) немного сложнее. Вот как это сделать. Есть два типа людей с положительным результатом теста: дети с целиакией, которые получают правильный результат теста, и дети без целиакии, которые получают ложноположительный результат. Один процент детей страдает целиакией, и из них 80 % получают положительный результат теста. Девяносто девять процентов детей не страдают целиакией, и из них 50 % получают положительный результат теста. Отсюда

$$\begin{aligned}\Pr(\text{Положительный Тест 2}) &= \Pr(\text{Целиакия})\ \Pr(\text{Положительный Тест 2} \mid \text{Целиакии}) \\ &\quad + \Pr(\text{Нет целиакии})\ \Pr(\text{Положительный Тест 2} \mid \text{Нет целиакии}) \\ &= .01 \times .8 + .99 \times .5 \\ &= .503.\end{aligned}$$

Теперь мы можем напрямую вычислить апостериорные убеждения Итана:

$$\begin{aligned}\Pr(\text{Целиакия} \mid \text{Положительный Тест 2}) &= \frac{\Pr(\text{Положительный Тест 2} \mid \text{Целиакия})\ \Pr(\text{Целиакия})}{\Pr(\text{Положительный Тест 2})} \\ &= \frac{.8 \times .01}{.503} \\ &\approx .016.\end{aligned}$$

Конечно, на самом деле Эйб сдавал два теста. Что произойдет, если мы добавим тот факт, что тест Эйба оказался отрицательным в более точном Тесте 1? Если мы предположим, что ложноположительные и ложноотрицательные результаты в этих двух тестах независимы, то мы можем просто перемножить их вероятности, чтобы получить соответствующие значения:

$$\begin{aligned}&\Pr(\text{Целиакия} \mid \text{Отр. Тест 1 \& Полож. Тест 2}) \\ &= \frac{\Pr(\text{Отр. Тест 1 \& Полож. Тест 2} \mid \text{Целиакия})\ \Pr(\text{Целиакия})}{\Pr(\text{Отр. Тест 1 \& Полож. Тест 2})}.\end{aligned}$$

Какова вероятность того, что ребенок с целиакией получит отрицательный результат по тесту 1 и положительный по тесту 2? Тест 1 дает отрицательный результат для ребенка с целиакией (т. е. ложноотрицательный результат) только в 5 % случаев. Тест 2 дает положительный результат для ребенка с целиакией в 80 % случаев. Итак, если ложноотрицательные и ложноположительные результаты в двух тестах независимы, то

$$\Pr(\text{отр. Тест 1 \& полож. Тест 2} \mid \text{целиакия}) = 0.8 \times 0.05 = 0.04.$$

Априорное убеждение Pr(целиакия) остается прежним и составляет 1 %. И опять же, есть два типа детей, которые могут получить отрицательный результат Теста 1 и положительный – Теста 2. Во-первых, у ребенка может быть целиакия (это верно для 1 % детей). Тогда ребенок должен получить ложноотрицательный результат в Тесте 1, но правильный результат в Тесте 2. Как мы только что видели, вероятность этого равна 0.8 × 0.05 = 0.04. Во-вторых, у ребенка может не быть целиакии (это верно для 99 % детей). Тогда ребенок должен получить правильный результат в Тесте 1 и ложноположительный результат в Тесте 2. Это происходит с вероятностью 0.99 × 0.5 = 0.495. Теперь мы можем вычислить общую вероятность этих двух результатов теста:

$$
\begin{aligned}
&\Pr(\text{отр. Тест 1 \& полож. Тест 2}) \\
&= \Pr(\text{целиакия})\Pr(\text{отр. Тест 1 \& полож. Тест 2} \mid \text{целиакия}) \\
&\quad + \Pr(\text{нет целиакии})\Pr(\text{отр. Тест 1 \& полож. Тест 2} \mid \text{нет целиакии}) \\
&= 0.01 \times 0.04 + 0.99 \times 0.495 \\
&= 0.49045.
\end{aligned}
$$

Подставляя все это в правило Байеса, получаем

$$
\begin{aligned}
&\Pr(\text{Целиакия} \mid \text{Отр. Тест 1 \& Полож. Тест 2}) \\
&= \frac{\Pr(\text{Отр. Тест 1 \& Полож. Тест 2} \mid \text{Целиакия})\ \Pr(\text{Целиакия})}{\Pr(\text{Отр. Тест 1 \& Полож. Тест 2})} \\
&= \frac{.05 \times .01}{.49045} \\
&\approx .001.
\end{aligned}
$$

Вероятность того, что у Эйба была целиакия, учитывая результаты двух тестов, составляла примерно 1 к 1000[1].

Теперь, когда вы знаете правило Байеса, вы можете видеть, что врачи не совсем правильно понимали, что на самом деле означают свидетельства тестов.

Поиск террористов в аэропорту

В годы, последовавшие за террористическими атаками 11 сентября 2001 г., правительство США вложило большие средства в обеспечение безопасности аэропортов. Одна из крупных новых программ называлась «Проверка пассажиров

[1] Мы получили бы тот же ответ, если бы применили правило Байеса итеративно. Мы могли бы начать с априорного убеждения, что у Эйба есть целиакия, еще до того, как увидели какие-либо свидетельства, изменить наши убеждения в соответствии с данными Теста 1, рассматривать это апостериорное убеждение как новое априорное убеждение, а затем снова изменить наши убеждения в соответствии с данными Теста 2. И порядок, в котором мы это делаем, не имеет значения. В конечном итоге мы пришли бы к тем же убеждениям, если бы начали с Теста 2, а затем перешли к Тесту 1. В качестве бонусного упражнения вы можете попробовать перепроверить это самостоятельно, чтобы убедиться, что вы понимаете, как применять правило Байеса.

с помощью технологий наблюдения» (screening of passengers by observation techniques, SPOT).

Идея SPOT заключалась в том, чтобы использовать поведенческие сигналы для поимки потенциальных террористов до того, как они сядут в самолет. Офицеры по контролю поведения наблюдали за людьми в очереди на досмотре в аэропортах, выискивая признаки того, что человек нервничает или вызывает какие-либо подозрения. Разным видам подозрительного поведения было присвоено разное количество баллов. Если совокупная оценка подозрений превышала определенный порог, этого человека подвергали дополнительному допросу, обыску и проверке.

К 2010 г. около 5 % годового бюджета Управления транспортной безопасности (TSA), т. е. сотни миллионов долларов в год, шло на финансирование программы SPOT. Давайте воспользуемся правилом Байеса, чтобы понять, почему это было не очень хорошее использование денег.

Программа должна отвечать на такие вопросы, как «Учитывая набор моделей поведения и характеристик, насколько вероятно, что рассматриваемое лицо является террористом?» Другими словами, TSA пытается сформировать апостериорное убеждение о вероятности того, что путешественник является террористом, учитывая некоторые свидетельства, полученные в результате наблюдения за поведением путешественника. Чтобы правильно сформировать такие апостериорные убеждения на основе наблюдений, TSA необходимо знать как минимум три части информации.

1. Насколько вероятно, что случайный путешественник окажется террористом?
2. Насколько вероятно, что террорист покажется подозрительным офицеру по контролю поведения?
3. Насколько вероятно, что обычный путешественник (не террорист) покажется подозрительным офицеру по расследованию поведения?

К сожалению, по данным Главного контрольного управления (GAO) – независимого внепартийного агентства, работающего на конгресс и отвечающего за расследование того, как федеральное правительство тратит деньги налогоплательщиков, – TSA не знает ответов ни на один из этих вопросов. Никакие существующие научные исследования не подтверждают, тем более количественно, полезность наблюдения за поведением для выявления террористов. Зато мы точно знаем, что, даже согласно собственному отчету TSA, призванному продемонстрировать эффективность программы SPOT, ни один террорист никогда не был пойман с ее помощью. Со своей стороны GAO сообщает, что самой распространенной причиной задержания людей, которых офицеры по контролю поведения выбрали для дополнительной проверки, была нелегальная иммиграция.

Таким образом, у правительства нет данных, которые нам нужны для расчета апостериорных убеждений на основе доказательств, собранных программой SPOT. Но мы и без этих данных видим, что программа SPOT никогда не будет работать так, как ожидалось. Давайте спросим, насколько хорошо программа будет работать в лучшем случае. То есть мы подготовим демонстрационные данные, будучи чрезвычайно щедрыми по отношению к программе SPOT во всех наших предположениях, и посмотрим, будет ли программа хорошо рабо-

тать при этих допущениях. Если окажется, что ответ отрицательный даже при таких щедрых предположениях, то мы можем быть уверены, что ответ будет отрицательным и при более реалистичных сценариях.

Во-первых, по данным GAO, ежегодно через аэропорты США проходят около 2 млрд пассажиров. Для удобства предположим, что их 2 млрд плюс 100. Предположительно подавляющее большинство этих людей – невинные путешественники. Очень немногие путешественники пытаются угнать самолеты или участвовать в других формах терроризма. Давайте проявим щедрость по отношению к правительству и предположим, что каждый год 100 потенциальных террористов оказываются в аэропортах США и пытаются угнать самолеты. Итак, вот наше априорное убеждение: Pr(террорист) = 100/2 000 000 100.

Во-вторых, нам необходимо знать, насколько вероятно, что эти террористы будут демонстрировать подозрительное поведение, которое ищут сотрудники службы безопасности. Конечно, мы понятия не имеем. Но все научные данные свидетельствуют о том, что такого рода поведенческие сигналы весьма ненадежны. Опять же, давайте проявим щедрость и подведем итоги в пользу SPOT. Предположим, что 99 % всех террористов демонстрируют поведение, которое ищет TSA, т. е. Pr(подозрительный | террорист) = 0.99. На самом деле это число, конечно, намного меньше.

Наконец, нам нужно знать, насколько вероятно, что невинные путешественники проявят подозрительное поведение. Как мы уже говорили, такое поведение является ненадежным индикатором, поэтому обычные путешественники будут его демонстрировать достаточно часто. Но мы хотим проявить щедрость к SPOT. Итак, предположим, что только 1 % невиновных людей демонстрирует подозрительное поведение, например Pr(подозрительный | не террорист) = 0.01. Опять же, на самом деле эта доля наверняка намного выше. В этом упражнении мы предполагаем, что SPOT – это невероятно точная программа поведенческого скрининга.

Насколько вероятно, что человек, который ведет себя подозрительно, является террористом? Даже при таких чрезвычайно щедрых предположениях вероятность невелика. В табл. 15.3 показаны данные, которые вы получите на основе наших предположений.

Таблица 15.3. Сколько террористов и обычных путешественников кажутся подозрительными

	Не террорист	**Террорист**
Неподозрительный	1 980 000 000	1
Подозрительный	20 000 000	99

Из 2 000 000 100 пассажиров только 100 относятся к террористам. Девяносто девять из них проявят подозрительное поведение. Остальные 2 млрд поездок совершаются невинными путешественниками. Из них 1 % проявит подозрительное поведение. А ведь этот 1 % составляет 20 млн человек! Целых 20 000 099 человек ведут себя подозрительно. Из них только 99 – террористы. Таким образом, вероятность того, что кто-то является террористом, учиты-

вая, что он действовал подозрительно, равна 99/20 000 099. То есть примерно 0.000005 – около 1 на 200 000.

Мы могли бы аналогичным образом вычислить эту вероятность непосредственно на основе правила Байеса:

$$\Pr(\text{Террорист} \mid \text{Подозрительный}) = \frac{\Pr(\text{Подозрительный} \mid \text{Террорист})\Pr(\text{Террорист})}{\Pr(\text{Подозрительный})}$$

$$= \frac{\frac{99}{100} \cdot \frac{100}{2{,}000{,}000{,}100}}{\frac{20{,}000{,}099}{2{,}000{,}000{,}100}}$$

$$= \frac{99}{20,000,099}.$$

Напомним, что приведенные выше цифры основаны на предположениях, которые чрезвычайно щедры по отношению к правительству. Невозможно, чтобы террористы на самом деле демонстрировали поведение, которое программа SPOT выявляет в 99 % случаев. И не может быть, чтобы обычные люди проявляли подозрительное поведение, на которое настроена программа SPOT, только в 1 % случаев. Таким образом, вероятность того, что подозрительный человек является террористом, на самом деле намного ниже, чем 1 на 200 000. Действительно, если террористы демонстрируют подозрительное поведение в 75 % случаев, а обычные люди – в 10 % случаев, вероятность обнаружить настоящего террориста при подозрительном поведении составит примерно 1 на 37 млн:

$$\Pr(\text{Террорист} \mid \text{Подозрительный}) = \frac{\Pr(\text{Подозрительный} \mid \text{Террорист})\Pr(\text{Террорист})}{\Pr(\text{Подозрительный})}$$

$$= \frac{\frac{75}{100} \cdot \frac{100}{2{,}000{,}000{,}100}}{\frac{200{,}000{,}075}{2{,}000{,}000{,}100}}$$

$$= \frac{75}{200,000,075}$$

$$\approx \frac{1}{37,000,000}.$$

Примечательно, что даже с этой вероятностью мы все еще слишком щедры. Согласно исследованию Национальной академии наук, специалисты по скринингу, которые ищут только одну характеристику лица (а не множество вещей, которые ищут специалисты по скринингу SPOT), в идеальных условиях дают правильную оценку только примерно в 60 % случаев. В более реалистичных условиях они дают правильную оценку лишь примерно в 30 % случаев. Учитывая такой уровень точности и небольшую долю людей, являющихся террористами, мы можем с уверенностью сказать, что более 1 млрд долл., выделенных на программу SPOT, были выброшены на ветер. В этом легко убедиться, если мы задаем правильные вопросы.

Давайте закончим эту неприятную историю еще одним неудобным моментом, который возвращает нас к ключевому уроку из главы 4: корреляция требует изменений. Счетная палата правительства – это наблюдательная организация, которая должна следить за тем, чтобы государственные учреждения тратили деньги надлежащим образом. После изучения программы она также может предоставить соответствующему государственному органу рекомендации о том, как ее улучшить. Именно так и сделали в GAO после оценки программы SPOT.

Одной из проблем, вызвавших обеспокоенность GAO, было отсутствие научной основы для поведенческих характеристик, которые TSA поручила проверять специалистам SPOT. По данным GAO, в TSA понятия не имели, действительно ли террористы с большей вероятностью будут демонстрировать то поведение, которое им нужно, или нет. (Как мы только что увидели, даже если это так, программа SPOT – пустая трата денег.) Итак, вот что GAO рекомендует TSA для повышения точности:

> «Изучение видеозаписей поведения людей, стоящих в очереди на досмотр и проходящих через контрольно-пропускные пункты аэропорта, которым впоследствии были предъявлены обвинения или признали себя виновными в преступлениях, связанных с терроризмом, может дать представление о поведении, которое может быть распространенным среди террористов».

Предположим, вы посмотрели эти видео и обнаружили, что, например, все люди, оказавшиеся террористами, были в темных очках и выглядели взволнованными, стоя в очереди на досмотр. Вы собираетесь начать арестовывать всех, кто соответствует этому описанию? Мы надеемся, что нет. Как мы знаем из главы 4, корреляция требует вариаций. Если вы хотите лучше справиться с (безнадежной) задачей выявления характеристик, позволяющих предсказать, является ли человек террористом, вам, по крайней мере, необходимо сравнить характеристики террористов и обычных людей. Нельзя просто изучать террористов.

Правило Байеса и количественный анализ

Одним из наиболее интересных применений правила Байеса является анализ того, насколько мы должны быть уверены в истинности некоторой научной гипотезы в свете доказательств, представленных в научном исследовании. Конечно, мы уже обсуждали один из подходов к этому вопросу в главе 6. Там вы узнали, что *p*-значение говорит нам, насколько вероятно, что данный результат получен случайно. Но если вдуматься, то станет ясно, что это не ответ на правильный вопрос. Фактически, когда аналитик обнаруживает низкое *p*-значение и приходит к выводу, что результат, скорее всего, достоверен, он совершает ту же ошибку, что и математик с прокурором в деле «Народ против Коллинзов». Аналитик рассчитал вероятность того, что он нашел бы связь в своих данных, даже если бы в мире не было реальной связи, т. е. Pr(результат | связь отсутствует). Но на самом деле ему нужно знать, насколько вероятно отсутствие реальной связи при имеющемся результате, то т. е. Pr(связь отсутствует | результат). Вероятность существования реальной связи при условии данного результата равна 1 – Pr(связь отсутствует | результат).

Давайте воспользуемся правилом Байеса, чтобы внести ясность в понимание этого нюанса. Предположим, мы собираем некоторые данные, проверяем наличие связи и получаем статистически значимый результат на уровне 0.05 (т. е. $p < 0.05$). Какова вероятность того, что предполагаемая взаимосвязь отражает реальное явление, а не появляется в данных из-за шума? Об этом нам говорит правило Байеса:

$$\Pr(\text{связь существует} \mid \text{результат}) = \frac{\Pr(\text{результат} \mid \text{связь существует})\ \Pr(\text{связь существует})}{\Pr(\text{результат})}.$$

И, как и раньше, мы можем разбить Pr(результат) на два компонента. Один из способов получить результат состоит в том, что взаимосвязь реальна и тест правильно ее определил. Вероятность этого равна Pr(связь существует) × Pr(результат | связь существует). Другой способ, которым мы могли бы получить результат, заключается в том, что взаимосвязи не существует, но тест ошибочно идентифицирует ее как реальную из-за шума. Вероятность этого равна Pr(связь отсутствует) × Pr(результат | связь отсутствует). Следовательно, мы можем записать правило Байеса следующим образом:

$$\Pr(\text{связь существует} \mid \text{результат}) = \frac{\Pr(\text{результат} \mid \text{связь существует})\ \Pr(\text{связь существует})}{\Pr(\text{связь существует})\ \Pr(\text{результат} \mid \text{связь существует}) + \Pr(\text{связь отсутствует})\ \Pr(\text{результат} \mid \text{связь отсутсвует})}.$$

Мы знаем Pr(результат | связь отсутствует). Это всего лишь уровень значимости, используемый в нашей проверке гипотезы. Если мы объявляем результат статистически значимым при условии $p < 0.05$, то Pr(результат | связь отсутствует) < 0.05.

Остальные числа найти сложнее. Величина Pr(связь существует) – это наше априорное убеждение в том, что настоящая связь существует, до того, как мы увидели какие-либо новые свидетельства. Величина Pr(результат | связь существует) – т. е. вероятность того, что вы найдете взаимосвязь в ваших данных при условии, что она действительно существует – называется *статистической мощностью* теста. Статистическая мощность представляет собой ответ на следующий вопрос: какова вероятность того, что мы обнаружим статистически значимый результат в данных, при условии, что взаимосвязь реальна? Существуют способы оценки статистической мощности при наличии более подробной информации о данных и тестах. Например, можно провести компьютерное моделирование, чтобы определить, насколько вероятно статистически обнаружить эффект определенной величины.

Теперь мы можем еще раз переписать формулу правила Байеса с учетом более предметной интерпретации величин:

$$\Pr(\text{связь существует} \mid \text{результат}) = \frac{\Pr(\text{результат} \mid \text{связь существует})\ \Pr(\text{связь существует})}{\Pr(\text{связь существует})\ \Pr(\text{результат} \mid \text{связь существует}) + \Pr(\text{связь отсутствует})\ \Pr(\text{результат} \mid \text{связь отсутсвует})}$$

$$= \frac{\text{Мощность} \times \text{Априорное убеждение}}{\text{Мощность} \times \text{Априорное убеждение} + \text{Значимость} \times (1 - \text{Априорное убеждение})}.$$

Давайте применим эту формулу на практике и посмотрим, что она означает в отношении наших апостериорных убеждений в свете новых, статистически значимых научных данных. Предположим, у нас есть догадка о некой причинно-следственной связи. Пока это лишь слабое предположение. Мы считаем, что вероятность существования этого эффекта составляет 5 % (наше априорное убеждение составляет 0.05). Затем проводим рандомизированный эксперимент. Мы хотим быть уверены в ответе, поэтому обеспечиваем большой размер выборки, такой, что статистическая мощность нашего теста будет равна 0.8 (у нас будет 80-процентная вероятность обнаружить эффект, если он действительно существует). Следуя общепринятому соглашению, мы используем порог статистической значимости 0.05. Теперь можно задать вопрос: каким должно быть наше апостериорное убеждение относительно вероятности того, что эффект реален, при условии, что наблюдается статистически значимый результат?

Подставив эти числовые значения в приведенное выше уравнение, мы получим

$$\Pr(\text{эффект реален} \mid \text{результат}) = \frac{.8 \times .05}{.8 \times .05 + .05 \times .95}$$

$$\approx .46.$$

Что случилось? Даже при условии получения результата, статистически значимого на уровне 95 %, вероятность того, что эффект, который мы наблюдаем, существует на самом деле, составляет всего лишь 46 %! Логика та же, что лежит в основе вывода о возможной невиновности четы Коллинз, хотя вероятность того, что случайная пара совпала с описанием, составляла всего один на миллион. Значение *p*, как и одна миллионная, – это всего лишь одно из чисел, которые нам нужны для формирования наших апостериорных убеждений. Если мощность мала или наши априорные убеждения низки, наши апостериорные убеждения, вероятно, также будут низкими.

Эти рассуждения помогают нам лучше понять кризис репликации во многих научных дисциплинах, который мы описали в главах 7 и 8. Помните исследование экстрасенсорного восприятия? Каково было ваше априорное убеждение относительно людей, обладающих экстрасенсорным восприятием, до того как вы увидели результаты этого исследования? Вероятно, довольно низкое, не так ли? Как следствие, ваше правильное апостериорное убеждение в том, что эффект реален, даже при наличии статистически значимых доказательств, не так уж велико. Рисунок 15.6 дает наглядное представление об этом. Вертикальная ось представляет собой апостериорную вероятность того, что наблюдаемая связь реальна. Горизонтальная ось – это априорная вероятность того, что она реальна. Кривая отображает правильное апостериорное убеждение как функцию вашего априорного убеждения при условии, что исследование со статистической мощностью 0.8 и порогом значимости 0.05 дало статистически значимые доказательства существования взаимосвязи.

Наши априорные убеждения чрезвычайно важны для формирования апостериорных убеждений. В самом деле, если у вас чрезвычайно низкое априорное убеждение относительно существования экстрасенсорного восприятия

(как и у нас), то, возможно, даже не имеет смысла изучать экстрасенсорное восприятие, потому что результаты исследования практически не окажут влияния на ваши убеждения.

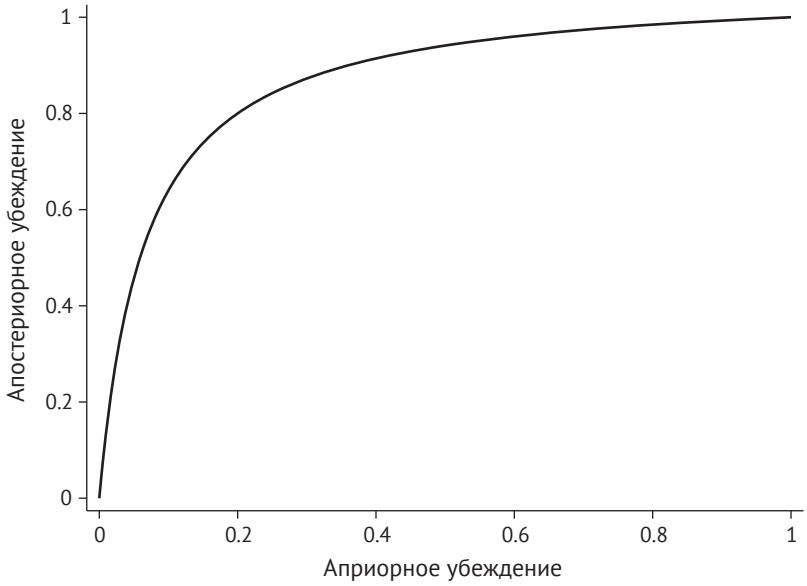


Рис. 15.6. Апостериорное убеждение в том, что эффект реален при наличии статистически значимых доказательств, как функция априорного убеждения

На рис. 15.7 показано, как изменение убеждений в ответ на новые данные соотносится с априорным убеждением. Здесь изображено ваше апостериорное убеждение в том, что реальная взаимосвязь существует, минус ваше априорное убеждение в том, что реальная взаимосвязь существует, для различных значений априорного убеждения при условии, что вы наблюдали статистически значимые доказательства в пользу этой взаимосвязи. Как видите, если ваше априорное убеждение уже очень близко к 0 или 1, изменить его очень сложно. Эффект новых данных является наибольшим для умеренно неожиданных результатов (т. е. результатов, для которых ваше априорное убеждение было около 0.2).

Рисунок 15.7 также показывает, что два человека могут (и должны) совершенно по-разному реагировать на одну и ту же информацию, если у них разные априорные убеждения. Некоторые люди могут увидеть какое-то свидетельство об экстрасенсорном восприятии, последствиях глобального потепления или вмешательстве России в американские выборы и резко изменить свои убеждения, в то время как другие могут увидеть те же доказательства и вообще почти не изменить свои убеждения. Когда мы сталкиваемся с этим в повседневной жизни, то часто приходим к выводу, что люди, которые реагировали иначе, чем мы, неразумны или иррациональны. Но правило Байеса говорит нам, что совершенно естественно, когда разные люди по-разному реагируют на одну и ту же информацию, если вначале у них были разные априорные убеждения.

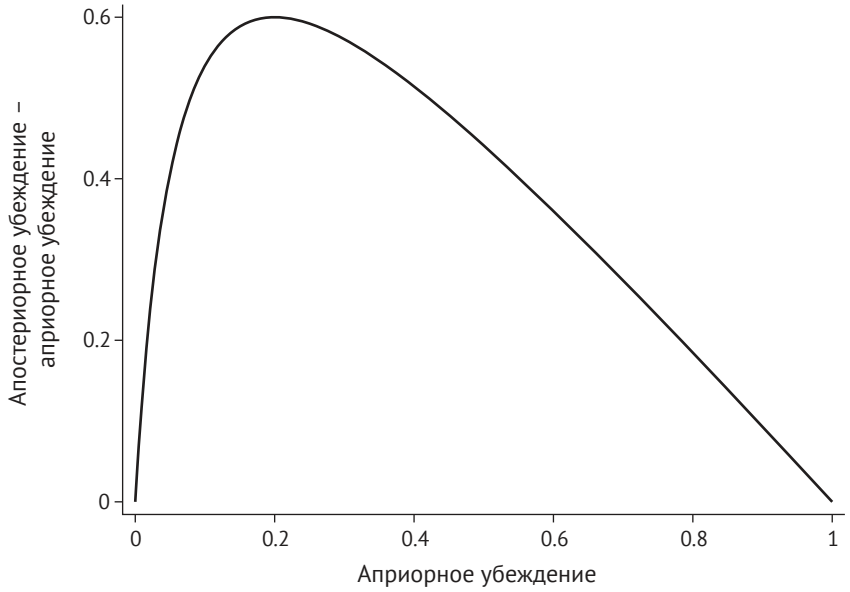


Рис. 15.7. Изменение апостериорных убеждений в ответ на новые доказательства в зависимости от априорных убеждений

Выводы из наших рассуждений могут вызвать у вас дискомфорт. Разве мы, аналитики данных, не должны позволить данным говорить непредвзято, не навязывая собственных предубеждений? И откуда берутся эти априорные убеждения, если не из данных? Это трудные вопросы. Но обойти их невозможно. Если вы хотите что-то сказать о вероятности существования подлинной причинно-следственной связи при наличии каких-либо доказательств, вам необходимо иметь априорные убеждения о вероятности этой связи. Вы не можете просто игнорировать свои априорные убеждения. Потому что, как мы видели, вероятности Pr(результат | связь отсутствует) и Pr(связь отсутствует | результат) могут сильно различаться.

Вот еще один важный момент. В большинстве случаев нас не интересует вероятность того, что какое-то явление существует или не существует (хотя в случае экстрасенсорного восприятия вопрос именно в этом). Обычно мы хотим знать, насколько существенно важен или велик причинный эффект, а не просто знать, что он существует. То есть вместо того, чтобы просто узнать, оказывает ли стратегия кампании реальное влияние, скажем, на долю голосов, мы хотим знать *величину* влияния стратегии кампании на долю голосов. Сколько сторонников может получить партия, проведя поквартирный обход избирателей округа? Повысится ли явка на 0.1, 1 или 10 процентных пунктов? В таких ситуациях мы также можем использовать байесовский вывод, но это сложно. Когда вы думаете о величине причинно-следственной связи, ваше предварительное убеждение – это не просто одно число, как было раньше, когда вы думали о вероятности существования отношений. Теперь это убеждение относительно того, насколько вероятна каждая возможная величина взаимосвязи. И когда вы обновляете свои убеждения, необходимо обновить свои представления о каж-

дой из этих вероятностей. Некоторые аналитики делают это формально, определяя априорное распределение убеждений обо всех возможных значениях, а затем выполняя сложные вычисления для оценки апостериорных значений (это называется *байесовской статистикой*.) Альтернативный подход состоит в том, чтобы продолжать использовать традиционную статистику, подобную той, которая описана в главе 6 (так называемая *частотная статистика*), при этом стараясь соблюдать осторожность при интерпретации результатов.

Ожидаемые затраты и выгоды

Ваши убеждения относительно причинных эффектов – это лишь один из факторов, влияющих на решение. Даже если вы убедились, что все данные имеют правильный масштаб, что вы отвечаете на правильный вопрос и сформировали правильные апостериорные убеждения, основанные на фактах и ваших априорных убеждениях, количественная информация все равно не говорит сама за себя. Чтобы использовать информацию и свидетельства в процессе принятия решений, вам необходимо объединить свои убеждения, основанные на фактических данных, со своими ценностями и целями.

В каком-то смысле это очевидно. Предположим, что качественно спланированная серия исследований убеждает вас в том, что определенная политика в отношении учебных заведений увеличивает посещаемость колледжей на 30 процентных пунктов. Это большой эффект. Но сам по себе он ничего не говорит о том, насколько хороша или разумна политика. Чтобы ответить на этот вопрос, вам как минимум нужно оценить затраты на реализацию политики и сопоставить их с ожидаемой отдачей от увеличения посещаемости.

В процессе формирования убеждений, основанных на сложном анализе данных, легко забыть о затратах, выгодах, ценностях и целях. Большой эффект сам по себе может показаться убедительным. Но важно не попасть в эту ловушку, потому что очевидное на первый взгляд доказательство может оказаться не таким уж очевидным после детального анализа. Рассмотрим пример такой ситуации.

Скрининг: часто и точно

Пока мы пишем этот раздел, пандемия коронавируса распространяется по всему миру. Одна из центральных задач в борьбе с пандемией касается тестирования, в частности достаточно быстрого выявления инфицированных людей, чтобы их можно было изолировать, прежде чем они распространят инфекцию среди слишком большого числа других людей.

Как мы неоднократно подчеркивали в этой книге, при рассмотрении эффективности теста для диагностики заболевания важны как ложноположительные, так и ложноотрицательные показатели. Чем ниже *каждый* показатель, тем точнее диагноз. Неудивительно, что регулирующие органы, такие как Управление по санитарному надзору за качеством пищевых продуктов и медикаментов (FDA), требуют предоставлять тесты с низким уровнем ложноположительных и ложноотрицательных результатов, не допуская их на рынок, если они слишком неточны по любому из параметров.

В большинстве случаев это вполне разумно. Мы ведь не хотим, чтобы больные люди думали, что они здоровы (ложноотрицательные результаты), или здоровые люди заключали, что они больны (ложноположительные результаты). И мы не хотим скомпрометировать идею тестирования, заставив людей прийти к выводу, что они не могут доверять тестам в целом.

В первые месяцы распространения коронавируса ученые-медики испробовали различные подходы к тестированию. Некоторые из них имели низкий уровень ложноположительных результатов. Но полимеразная цепная реакция (ПЦР) с использованием мазков из носа имела дополнительное преимущество: низкий процент ложноотрицательных результатов. Дело в том, что тест ПЦР способен обнаружить присутствие вируса при очень низкой концентрации характерных веществ. Поскольку ПЦР-тесты удовлетворяли требованиям FDA по низкому уровню ложноположительных и ложноотрицательных результатов, они были быстро одобрены и стали фактическим стандартом тестирования.

Конкурирующей технологии – тестам, в которых слюна наносилась на бумажную полоску, – было сложнее получить одобрение. Причиной был более высокий уровень ложноотрицательных результатов. Тесты с бумажными полосками могли обнаружить вирус только при более высоких уровнях концентрации. Таким образом, они с большей вероятностью пропускали кого-то, кто был инфицирован, особенно в первые дни заражения, когда вирусная нагрузка человека была еще относительно низкой.

В отношении многих заболеваний позиция FDA имеет под собой серьезные основания. Если мы проводим тестирование на целиакию или рак, имеет смысл утверждать только самые точные тесты. Но случай с коронавирусом, возможно, отличается во многих отношениях, над которыми стоит задуматься.

При сравнении достоинств двух диагностических тестов важны частота ложноположительных и ложноотрицательных результатов. Но это не единственные важные критерии. Следует также учитывать относительную стоимость двух тестов. И, особенно в случае такого высокоинфекционного заболевания, как коронавирус, следует также учитывать скорость проведения теста. Одно дело – подождать неделю или две результатов теста на целиакию. Другое дело – подождать неделю результатов теста на коронавирус. За это время человек, о котором идет речь, может заразить множество других людей.

Как оказалось, хотя тесты с бумажными полосками имеют более высокий уровень ложноотрицательных результатов, чем тесты ПЦР, они намного дешевле, их можно проводить дома и давать результаты менее чем за час, по сравнению с 5–10 днями, которые люди проводили в ожидании результатов ПЦР[1]. Если мы объединим эти дополнительные фрагменты информации с разницей в показателях ложноотрицательных результатов, мы можем прийти к совершенно иному выводу о том, правильно ли поступило FDA, отложив одобрение тестов на бумажных полосках.

Чтобы вникнуть в суть проблемы, подумайте о разнице в цене. По некоторым данным, тесты на бумажных полосках стоят 1–5 долл., а ПЦР-тесты –

[1] Быстрые тесты ПЦР появились в широком обиходе уже после написания этой книги, их цена быстро снизилась до приемлемого уровня, и теперь этот пример далек от реальности. – *Прим. перев.*

50–100 долл. Поэтому сравнение одного ПЦР-теста с одним тестом на бумажной полоске вряд ли справедливо. Мы могли бы сделать по крайней мере десять тестов на бумажных полосках вместо одного теста ПЦР.

Основная причина ложноотрицательного результата такого теста – это если ваша вирусная нагрузка слишком низка, чтобы ее можно было обнаружить с помощью теста. ПЦР-тест имеет более низкий уровень ложноотрицательных результатов, поскольку он более чувствительный и может обнаружить вирус в гораздо меньшей концентрации. Но коронавирус очень быстро развивается в организме человека. Ученые полагают, что требуется всего день или около того, чтобы перейти от вирусной нагрузки, которую можно обнаружить с помощью ПЦР-теста, к той, которую можно надежно обнаружить с помощью теста на бумажных полосках.

Если это так, то один из способов понять разницу между двумя тестами заключается в следующем. Предположим, вы можете позволить себе N тестов на бумажных полосках относительно одного ПЦР-теста. Чтобы сохранить затраты равными, давайте представим, что мы проводим тест с бумажными полосками каждый день или ПЦР-тест раз в N дней. В качестве аргумента давайте примем $N = 10$ и проигнорируем задержки в получении результатов тестов. Вам придется выбирать между сдачей ПЦР-теста в 1-й, 11-й, 21-й день и т. д. и ежедневным тестом на бумажной полоске. Сосредоточьтесь на днях с 1 по 10. При схеме ПЦР, если у вас низкая вирусная нагрузка в первый день, вы обнаруживаете вирус с помощью ПЦР-теста в этот же день, но тест на бумажной полоске будет давать отрицательный результат еще день-два. Если ваша вирусная нагрузка низкая на второй день, вы не обнаружите свое заболевание с помощью ПЦР-теста раньше 11-го дня, но вы обнаружите заболевание с помощью теста на бумажной полоске на третий день. То же самое верно для дней с 3-го по 9-й. Если ваша вирусная нагрузка низкая на 10-й день, вы узнаете, что больны, с помощью любого теста на 11-й день. Таким образом, в целом вероятность того, что вы обнаружите заболевание быстрее с помощью ПЦР-теста, составляет 1 из 10. Вероятность обнаружить заболевание быстрее с помощью теста на бумажных полосках, составляет 8 из 10. Вероятность того, что вы обнаружите заболевание одновременно двумя тестами, также составляет 1 из 10.

Конечно, помимо низкой вирусной нагрузки, могут быть и другие причины, по которым люди получают ложноотрицательные результаты. Вот еще один способ провести сравнение тестов. Предположим, опять же ради наглядности, что оба теста имеют нулевой уровень ложноположительных результатов. Поэтому нас беспокоит только уровень ложноотрицательных результатов. Пусть p – доля ложноотрицательных результатов теста ПЦР, а q – доля ложноотрицательных результатов теста с бумажными полосками. Вероятность того, что ПЦР пропустит инфицированного человека, равна p. Насколько вероятно, что десять тестов с бумажными полосками пропустят этот случай? Это зависит от того, насколько коррелируют ложноотрицательные результаты тестов одного и того же человека. Если они идеально коррелируют (что, конечно же, неверно, поскольку вирусная нагрузка человека со временем увеличивается), то, один раз получив ложноотрицательный результат, вы всегда будете получать ложноотрицательный результат. В этом случае вероятность того, что десять тестов с полосками бумаги пропустят заболевание, равна вероятности q для одного те-

ста с бумажными полосками. Если, напротив, ложноотрицательные результаты совершенно не коррелируют между случаями (что также, безусловно, неверно, поскольку у некоторых людей вирусная нагрузка ниже, чем у других, и поэтому их случаи труднее обнаружить), тогда вероятность того, что десять тестов с бумажными полосками пропустят инфицированного человека, равна q^{10}. Так, например, если бы тест ПЦР имел долю ложноотрицательных результатов в одну десятую процента, а тест с бумажными полосками имел долю ложноотрицательных результатов в 20 %, десять тестов с бумажными полосками имели бы гораздо больше шансов выявить инфицированного человека, чем один ПЦР-тест ($0.001 > 0.2^{10} \approx 0.0000001$). Истина, конечно, лежит где-то посередине.

При оценке этих двух подходов к тестированию необходимо учитывать еще больше факторов. Во-первых, как мы уже указывали, ложноотрицательные результаты более вероятны на ранних стадиях заражения человека, когда вирусная нагрузка низкая. Но ведь при этом и люди менее заразны. Таким образом, значимость разницы между ПЦР и тестом на бумажных полосках уменьшается.

Во-вторых, скорость теста – невероятно важная часть расчета затрат и выгод. Основное преимущество тестирования заключается в том, чтобы люди не заражали других после начала заболевания. Коронавирус быстро развивается в организме инфицированного человека. Таким образом, возможность проводить тесты дома и получать результаты менее чем за час дает огромные преимущества.

По всем этим причинам исследования, моделирующие распространение заболевания при различных схемах тестирования, показывают, что различия в частоте и скорости тестирования могут быть намного более важны, чем различия в доле ложноотрицательных результатов. Следовательно, разумное в целом правило одобрения диагностических тестов FDA, возможно, не было самым правильным в данном случае.

У нас может вызвать сомнения мысль о том, что, в отличие от тестов ПЦР, тесты на бумажных полосках имеют долю ложноположительных результатов, отличную от нуля. Если тест будет давать много ложноположительных результатов, ежедневное тестирование может привести к дорогостоящему и ненужному самокарантину. Ложноположительные показатели трудно изучить, но, по крайней мере, некоторые данные свидетельствуют о том, что они были низкими, даже для тестов с бумажными полосками. Но даже если уровень ложноположительных результатов высок, можно найти разумное решение в виде комбинации двух технологий. Для достижения максимальной эффективности тесты с бумажными полосками не следует рассматривать как окончательный ответ. Если все люди начнут проходить тест с бумажными полосками каждый день, у некоторых из них окажутся ложноположительные результаты. В таком случае их можно отправить на карантин и сразу же провести более точный ПЦР-тест. Поскольку нагрузка на лаборатории снижается за счет сокращения ПЦР-тестирования, сроки обработки теста могут даже ускориться. И, как следствие, ложные срабатывания можно исправить относительно быстро, с минимальными неудобствами.

Цель этого обсуждения не в том, чтобы предложить окончательное решение в сложной области, в которой мы не являемся экспертами. Мы постарались проиллюстрировать тот факт, что нам приходится учитывать множество

различных доводов за и против при принятии решений, и окончательное решение во многом зависит от шкалы ценностей человека или общества. Легко зациклиться на одном конкретном количественном показателе, например на доле ложноотрицательных результатов, но обычно это ведет к принятию ошибочных решений. Мы вернемся к этим темам в последних двух главах.

Подведение итогов

Овеществление формальных статистических показателей помогает нам критически подумать о том, на какие вопросы отвечают факты и совпадают ли они с вопросами, которые мы хотели задать. Не научившись постоянно держать эти вопросы в центре внимания, вы не сможете правильно использовать количественную информацию для принятия решений. На самом деле об этом нельзя забывать не только при интерпретации результатов анализа, но и при выборе способа измерения, выборе образцов, которые мы изучаем, и принятии решения о том, в каких случаях применимы наши выводы. Эти вопросы являются темой главы 16.

Ключевые термины

- **Изменение в процентных пунктах:** простая арифметическая разность между двумя процентными значениями.
- **Изменение в процентах:** способ представления степени изменения. Это разница между первоначальным значением и новым значением, деленная на исходное значение (и умноженная на 100). В отличие от изменения в процентных пунктах изменение в процентах очень чувствительно к исходному значению.
- **Условная вероятность:** вероятность события, обусловленная некоторой другой информацией или событием. Мы записываем вероятность C, обусловленную E, как $\Pr(C \mid E)$.
- **Априорное убеждение:** ваше убеждение в чем-либо до получения новых свидетельств.
- **Апостериорное убеждение:** ваше убеждение в чем-то после получения новых свидетельств.
- **Правило Байеса:** формула для расчета апостериорного убеждения на основе новых свидетельств и вашего априорного убеждения. В частности, $\Pr(C \mid E) = \frac{\Pr(E \mid C)\Pr(C)}{\Pr(E)}$. Иногда ее называют теоремой Байеса или законом Байеса.
- **Статистическая мощность:** вероятность обнаружения статистически значимого результата в данных при условии, что причинно-следственная связь действительно существует.

Упражнения

15.1. Газета сообщает: «Экономический рост в стране А в прошлом году был на 20 % выше, чем в стране В».

Типичный способ, которым экономисты измеряют экономический рост, – это процентное изменение ВВП от года к году. То есть мы бы сказали, что экономический рост составил 3 % в конкретной стране и в конкретном году, если ВВП в конце года был на 3 % выше, чем в начале.

a) Предположим, рост ВВП в стране B составил 10 %. Каков был рост ВВП в стране A?
b) Предположим, рост ВВП в стране B составил 0.1 %. Каков был рост ВВП в стране A?
c) Как вы посоветуете написать заголовок, чтобы он раскрывал разницу между сценариями, описанными в пунктах (a) и (b)?

15.2. Теперь рассмотрим две другие страны, C и D. Предположим, что рост в стране C составляет 1 %, а рост в стране D – 0.1 %.

a) Какова процентная разница в росте? Какова разница в процентных пунктах?
b) Напишите два заголовка, каждый из которых будет включать правдивые статистические факты об этих двух странах. Исходите из того, что разница в их экономическом росте действительно имеет большое значение. Остальное не важно.
c) Теперь предположим, что при статистическом обзоре выяснилось, что рост в стране D, оказывается, составил всего лишь 0.001 % вместо 0.1 %. Какова сейчас процентная разница в темпах роста между двумя странами? Какова разница в процентных пунктах? Какой из этих двух статистических показателей лучше передает существенное значение сдвига с 0.1 % к 0.001 %? Почему?

15.3. Во время пандемии коронавируса правительства и частные организации по всему миру бросились создавать диагностические тесты. Эти тесты различались по своей точности. Возьмем один из этих тестов, который, как сообщалось, имел 1 % ложноположительных и 10 % ложноотрицательных результатов.

Мы не знаем скрытый уровень заболеваемости коронавирусом среди тех, у кого нет симптомов. Допустим, вероятность того, что у человека без симптомов есть коронавирус, равна некоторому числу q, т. е. априорное убеждение, что любой данный человек болен, составляет $\Pr(\text{болен}) = q$.

a) Используя приведенную выше информацию о частоте ложноотрицательных результатов, найдите вероятность того, что человек получит положительный результат, при условии, что у него действительно есть коронавирус (обозначим как $\Pr(+ \mid \text{болен})$)? Подсказка: для ответа на этот вопрос не требуется правило Байеса.
b) Есть два варианта получения положительного результата теста. Человек с коронавирусом может получить правильный результат теста. А ложноположительный результат может получить человек, у которого нет коронавируса. Рассчитайте общую вероятность того, что у человека без симптомов результат теста окажется положительным:

$$\Pr(+) = \Pr(\text{болен}) \cdot \Pr(+ \mid \text{болен}) + \Pr(\text{здоров}) \cdot \Pr(+ \mid \text{здоров}).$$

(В вашем ответе будет буква q, поскольку он будет зависеть от априорного убеждения, что человек, у которого нет симптомов, болен.)

c) Теперь используйте правило Байеса для расчета Pr(болен) – вероятности того, что у человека без симптомов результат теста окажется положительным. (Ваш ответ опять же будет содержать q.)

d) На самом деле мы не знаем q. Рассмотрите разные варианты.

i) Рассчитайте Pr(болен | +), если $q = 0.005$ (т. е. если 0.5 % бессимптомного населения имеет коронавирус).

II) Рассчитайте Pr(болен | +), если $q = 0.01$ (т. е. если у 1 % бессимптомного населения есть коронавирус).

III) Рассчитайте Pr(болен | +), если $q = 0.05$ (т. е. если у 5 % бессимптомного населения есть коронавирус).

IV) Нарисуйте график с q на горизонтальной оси (от 0 до 1), который отображает Pr(болен | +).

15.4. Дискриминация определенных групп людей на рынке труда является серьезной социальной и политической проблемой. Многие исследования направлены на получение количественных данных, подтверждающих масштабы такой дискриминации.

Давайте подумаем над очень простым примером. Представьте себе общество с двумя одинаково большими и одинаково квалифицированными группами: привилегированными и непривилегированными.

Используя обозначение условной вероятности, которое мы разработали ранее, запишем вероятность того, что человек получит работу при условии его членства в группе, как Pr(нанят | группа). Аналогично запишем вероятность того, что человек является членом определенной группы, при условии, что он нанят на работу, как Pr(группа | нанят).

a) Предположим, вы хотите знать, будет ли член привилегированной группы принят на работу с большей вероятностью, чем член непривилегированной группы, если они претендуют на одно и то же место. То есть вы хотите знать, верно ли для тех, кто претендует на работу, следующее неравенство:

$$\Pr(\text{нанят} \mid \text{есть привилегии } \& \text{ претендует}) > \Pr(\text{нанят} \mid \text{нет привилегий } \& \text{ претендует}).$$

I) Используйте правило Байеса, чтобы переписать Pr(нанят | есть привилегии & претендует) как функцию трех членов: Pr(есть привилегии | нанят & претендует), Pr(нанят | претендует) и Pr(есть привилегии | претендует).

II) Используйте правило Байеса, чтобы переписать Pr(нанят | нет привилегий & претендует) как функцию трех членов: Pr(нет привилегий | нанят & претендует), Pr(нанят | претендует) и Pr(нет привилегий | претендует).

b) Предположим, исследование показывает, что люди, занимающиеся определенной работой, с одинаковой вероятностью могут быть привилегированными и непривилегированными. Запишите это, исполь-

зуя наши обозначения. Исходя из ответа на пункт (a), какую дополнительную информацию вам нужно знать?

c) Достаточно ли информации в пункте (b), чтобы определить, будет ли член привилегированной группы принят на работу с большей вероятностью, чем член непривилегированной группы, если они претендуют на одно и то же место? Исходя из ответа на пункт (a), какую дополнительную информацию вам нужно знать?

d) Предположим, вы узнали, что на эту работу претендовало одинаковое количество членов обеих групп. А теперь вы можете дать ответ?

Дополнительное чтение и ссылки

Исследование по статистике топливной эффективности:

Richard P. Larrick and Jack B. Soll. 2008. *The MPG Illusion*. Science 320: 1593–94.

Статья в Wall Street Journal о препарате от холестерина:

Ron Winslow. *Cholesterol Drug Cuts Heart Risk in Healthy Patients.* Wall Street Journal, Nov. 10, 2008. *https://www.wsj.com/articles/SB122623863454811545*.

Чтобы узнать больше о том, как создавать информативные визуализации данных и как не дать себя обмануть плохими графиками, мы рекомендуем следующие книги:

Carl T. Bergstrom and Jevin D.West. 2020. *Calling Bullshit: The Art of Skepticism in a Data-DrivenWorld.* RandomHouse;

KieranHealy. 2019. *Data Visualization: A Practical Introduction.* PrincetonUniversity Press;

Edward R. Tufte. 2001. *The Visual Display of Quantitative Information, 2nd Edition.* Graphics Press.

Данные о партийных тенденциях на Юге США взяты из работы:

Christopher H. Achen and Larry M. Bartels. 2016. *Democracy for Realists: Why Elections Do Not Produce Responsive Government.* Princeton University Press.

История статистических ошибок, допущенных в суде над четой Коллинзов, изложена в статье:

Jonathan J. Koehler. 1995. *One inMillions, Billions, and Trillions: Lessons fromPeople v. Collins (1968) for People v. Simpson (1995).* Journal of Legal Education 47 (2): 214–23.

Для получения дополнительной информации о программе SPOT ознакомьтесь с двумя отчетами Главной бухгалтерской службы:

the 2010 report: *https://www.gao.gov/assets/310/304510.pdf*;

the 2013 report: *https://www.gao.gov/assets/660/658923.pdf*.

Анализ схем тестирования на коронавирус можно найти в статье:

Daniel B. Larremore, Bryan Wilder, Evan Lester, Soraya Shehata, James M. Burke, James A. Hay, Milind Tambe, Michael J. Mina, and Roy Parke. 2020. *Test Sensitivity Is Secondary to Frequency and Turnaround Time for COVID-19 Surveillance. https://www.medrxiv.org/content/10.1101/2020.06.22.20136309v3*.

Ранняя запись в блоге об этой идее находится здесь:

Alex Tabarrok. *Frequent, Fast, and Cheap Is Better than Sensitive.* Marginal Revolution. July 24, 2020. *https://marginalrevolution.com/marginalrevolution/2020/07/frequent-fast-and-cheap-is-better-than-sensitive.html*.

Глава 16

Измерение показателей вашей миссии

О чем эта глава

- Важно, чтобы вы оценивали результаты и методы воздействия, которые соответствуют вашей миссии.
- Если вы оцениваете результат не полностью, видимые улучшения могут ввести в заблуждение.
- Данные всегда поступают в определенном контексте. Применяя уроки, извлеченные из данных, к новому контексту, важно подумать о том, достаточно ли схожи контексты, чтобы уроки не потеряли смысл.
- Иногда в мире встречаются взаимосвязи, которые могли бы помочь вам в достижении цели. Но как только вы начнете совершать действия, эти связи исчезнут сами по себе и станут бесполезными.

Введение

Когда вы используете доказательства для обоснования своих решений, вы преследуете какую-то цель. Эта цель – и есть ваша стратегическая задача, которую мы для краткости будем называть *миссией*. Как ее можно измерить и почему это важно?

Предположим, у вас есть доказательства причинно-следственной связи; какое-то действие влияет на какой-то результат предсказуемым образом. Если изменение этого результата означает, что вы достигли своей цели, – т. е. если при измерении результата вы измерили свою миссию, – тогда знание этой причинно-следственной связи просто полезно. Но что, если изменение измеренного вами результата не обязательно означает достижение вашей цели или если оно соответствует только одной части вашей цели? В таком случае не всегда ясно, какие действия будут способствовать дальнейшему развитию вашей миссии.

То же самое касается корреляций. Предположим, ваша миссия заключается в попытке предсказать какой-то результат, но вы измерили другой, хотя и связанный результат. Уверены ли вы, что корреляции измеренного результата помогут предсказать интересующий вас результат?

В этой главе мы рассмотрим несколько причин, по которым что-то может пойти не так, если у нас есть убедительные свидетельства того, что может оказаться неправильным. Каждый из этих примеров иллюстрирует причины, по

которым важно как можно лучше оценить свою миссию и попытаться использовать фактические данные для принятия более эффективных решений.

Оценка неправильного результата или воздействия

Самый простой способ ошибочно оценить свою миссию – это измерить результат или воздействие, которое не совсем соответствует тому, что вас действительно интересует. Здесь мы рассмотрим три причины, по которым это обычно происходит.

Частичные измерения

Часто наша миссия состоит в том, чтобы изменить какой-то показатель (скажем, достижения в области образования, национальной безопасности или здравоохранения), который трудно измерить в целом. Например, у нас может не быть комплексного показателя общего уровня образования, но, возможно, мы сможем измерить, улучшаются ли результаты стандартизированных тестов. Бесспорно, такие частичные измерения могут быть полезны. Но мы должны быть осторожны с интерпретацией, потому что улучшение результатов тестов не является нашей миссией. Наша миссия – улучшение образования.

Во многих случаях есть веские основания полагать, что улучшение одного показателя может сопровождаться ухудшением других показателей. То есть, сделав лучше в одной части проблемы, мы можем сделать хуже в других частях. Наиболее очевидная причина этого – ограниченность ресурсов. Предположим, ваша главная цель – сделать местный парк красивее. У вас есть бюджет для осуществления этой миссии. Если вы потратите больше ресурсов на вывоз мусора, то останется меньше денег на благоустройство территории. Таким образом, улучшение в одном измерении означает ухудшение в другом. А если вы располагаете лишь частичной оценкой вашей миссии (скажем, количество мусора на земле), то, по мере того как вы тратите больше денег на вывоз мусора, может возникнуть ощущение, что вы лучше справляетесь с выполнением своей миссии. Но это заблуждение, поскольку в сфере ландшафтного дизайна ситуация ухудшается в результате выделения большего количества ресурсов на вывоз мусора.

Помимо ограниченности ресурсов, существуют и другие причины отрицательной корреляции между аспектами проблемы. Возможно, наиболее интересной является *стратегическая адаптация*: необходимость прилагать усилия по улучшению результатов в каком-то аспекте заставляет людей корректировать свое поведение, чтобы обойти эти усилия. Это также может затруднить частичные измерения. Давайте посмотрим, как это происходит, на примере.

Металлодетекторы в аэропортах

Начиная с середины 1960-х гг. угоны самолетов стали серьезной проблемой в гражданской авиации США. Только в 1969 г. угонщики захватили более 80 самолетов. Среди угонщиков были американцы, хорваты, кубинцы, японцы, северокорейцы, палестинцы и многие другие. Их мотивы варьировались от простого выкупа до националистических, левых и других глобальных политических целей. В начале 1970-х гг. в ответ на растущую угрозу безопасности полетов Соединенные Штаты усилили меры безопасности в аэропортах. Са-

мое главное, что в начале 1973 г. металлодетекторы были установлены во всех крупных аэропортах США.

Представьте, что вы правительственный чиновник, которому поручено оценить эффективность этих повышенных мер безопасности. Вы можете задать естественный вопрос: привели ли они к значительному снижению количества угонов самолетов. Рисунок 16.1, показывающий количество угонов самолетов за квартал с 1968 по 1978 гг., предполагает, что ответ – да. До 1973 г. (обозначенного пунктирной вертикальной линией) в квартал совершалось в среднем почти 20 угонов самолетов. Но после 1973 г. это число упало до менее десяти в квартал.

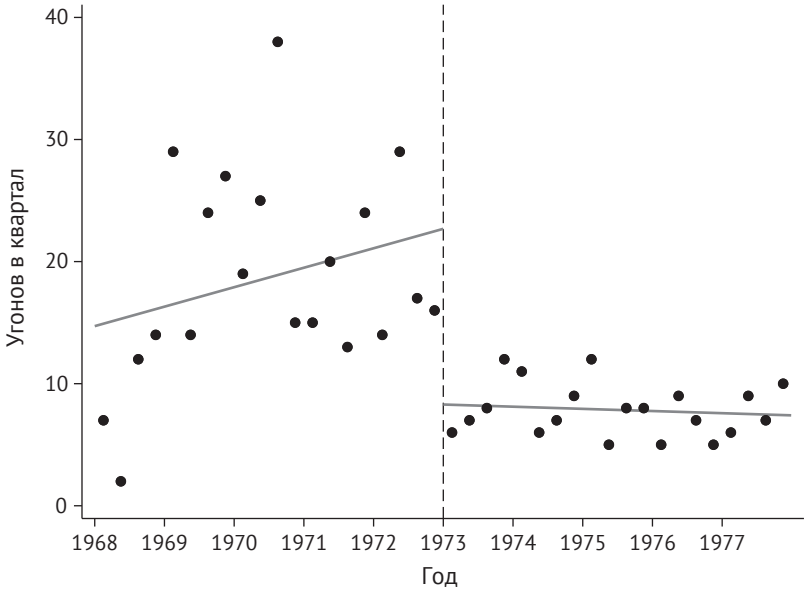


Рис. 16.1. Ежеквартальное количество угонов самолетов в 1968–1978 гг., а также отдельные линии регрессии для кварталов до и после первого квартала 1973 г. Вертикальная пунктирная линия указывает, когда в аэропортах США были установлены металлодетекторы

Давайте подумаем, удалось ли нам измерить свою миссию. Мы можем заявить, что наша миссия заключается в уменьшении количества угонов самолетов. В этом случае количество угонов самолетов является правильным показателем для измерения, и достигнутый результат выглядит как успех. Но если мыслить шире, то наша истинная миссия состоит в том, чтобы снизить частоту всех захватов заложников террористами, а не только путем угона самолетов. В этом случае угоны самолетов являются лишь частичной мерой миссии, поскольку существует множество других способов захвата заложников.

Более того, это как раз тот случай, когда улучшение одного аспекта проблемы (здесь – угоны самолетов) сопровождается ухудшением других аспектов проблемы (здесь – другие виды террористических атак). Причина в стратегической адаптации. Поскольку безопасность в аэропортах улучшается, есть все основания беспокоиться о том, что террористы заменят захват самолетов другими видами захвата заложников. Если это так, то очевидное сокращение количества угонов самолетов может ввести в заблуждение в качестве показателя

того, насколько успешным было усиление безопасности в аэропортах с точки зрения общей контртеррористической миссии.

Судя по всему, так и получилось. На рис. 16.2 показан вывод, основанный на работе Уолтера Эндерса и Тодда Сэндлера: после того как в аэропортах США были установлены металлодетекторы, захваты заложников стали более частыми! Итак, если у нас есть более всеобъемлющая, а не частичная оценка нашей миссии, мы приходим к несколько иным выводам.

Конечно, это не означает, что политика установки металлодетекторов провалилась. Замена угонов самолетов другими захватами заложников не выглядит равнозначной. Угоны самолетов в среднем более опасны, чем другие виды захвата заложников. Так что установка металлодетекторов привела к частичному успеху в борьбе с терроризмом. Но это далеко не такая драматичная победа, как можно было бы подумать, глядя только на количество угонов самолетов, а не на более полную оценку миссии.

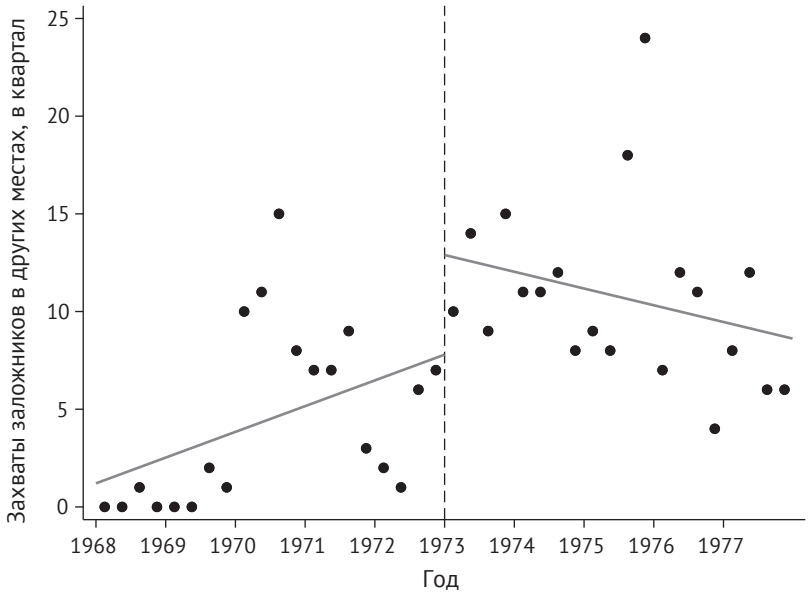


Рис. 16.2. Ежеквартальное количество захватов заложников не в самолетах, 1968–1978 гг., а также отдельные линии регрессии для кварталов до и после первого квартала 1973 г. Вертикальная пунктирная линия указывает, когда в аэропортах США были установлены металлодетекторы

Промежуточные результаты

Часто измерение результатов, связанных с вашей миссией, затруднено, занимает много времени или просто не дает достаточного количества данных. Распространенным решением является измерение промежуточных результатов – шагов на пути миссии, которые, как мы надеемся, будут указывать на долгосрочную цель.

Предположим, вы проводите политическую кампанию и пытаетесь максимизировать вероятность победы вашего кандидата. Вы хотите протестировать несколько разных объявлений, чтобы увидеть, какие из них наиболее

эффективны. Можете запустить рекламу на разных медиарынках и посмотреть, где вы добьетесь лучших результатов в день выборов. Но это не принесет особой пользы. Вам нужно знать, какую рекламу показывать до того, как состоятся выборы. Поэтому нужно измерить какой-то промежуточный результат, который даст вам некоторое представление о том, какая реклама лучше всего работает во время избирательной кампании.

Один из естественных вариантов, который поможет вам определиться со стратегией, – опросы общественного мнения. Вместо того чтобы случайным образом запускать рекламу и смотреть, что происходит с общим количеством голосов при голосовании, вы можете запускать рекламу случайным образом, проводить опросы общественного мнения и смотреть, какая реклама лучше отражается на результатах опросов. В этом нет ничего плохого. Это хорошая идея. Но вы должны иметь в виду, что опросы сами по себе вас не волнуют. Вас интересует количество голосов. Таким образом, ровно в той степени, в которой результат опросов отражает будущее распределение голосов, изучение этого промежуточного результата информативно для вашей миссии. Но может случиться так, что ваша реклама повлияет на базовые показатели опросов как таковых – какие люди готовы отвечать на опрос или будут ли они говорить правду социологам, не меняя при этом фактический выбор при голосовании. В этом случае влияние на промежуточный результат может вообще не совпадать с влиянием на конечный результат, который вас волнует. Следовательно, всякий раз, когда вы используете промежуточный результат вместо меры вашей конечной миссии, следует подумать о том, насколько вы уверены в том, что он действительно является шагом на пути к вашим реальным целям.

Давайте подумаем о примере из медицины, где невозможность изучения фактического целевого результата часто бывает особенно острой.

Артериальное давление и сердечные приступы

Предположим, цель нового лекарства – уменьшить количество сердечных приступов. К сожалению для исследовательских целей (но к счастью по другим причинам), сердечные приступы случаются редко. Лишь относительно небольшое количество людей в любой выборке столкнется с сердечными приступами во время испытаний препарата. Таким образом, очень сложно напрямую узнать, уменьшает ли лекарство вероятность сердечного приступа, даже в хорошо спланированном эксперименте, в котором рандомизировано, кто получает препарат, а кто нет.

Так что же делают медицинские исследователи? Альтернативой двадцатилетнему наблюдению за тем, снижается ли вероятность возникновения сердечных приступов у пациентов, которым назначили препарат, является изучение промежуточного, или суррогатного, показателя, такого как артериальное давление. Поскольку повышенное кровяное давление коррелирует с частотой сердечных приступов, считается, что, если лекарство снижает кровяное давление, оно, вероятно, уменьшит частоту приступов.

Но мы должны быть осторожны. Во второй части книги вы узнали, что корреляция не обязательно подразумевает причинно-следственную связь. Игра в баскетбол коррелирует с ростом, но экспериментальное увеличение количества игр в баскетбол не приводит к увеличению роста. Точно так же корреляция

между частотой сердечных приступов и кровяным давление не означает, что препарат, снижающий кровяное давление, уменьшит частоту сердечных приступов. Для этого вам потребуются убедительные доказательства того, что артериальное давление оказывает причинное влияние на сердечные приступы.

В настоящее время действительно есть веские основания полагать, что артериальное давление оказывает причинное влияние на сердечные приступы. Но для многих других промежуточных показателей, используемых в медицинских исследованиях, причинно-следственная связь может быть менее ясной.

В обзоре свидетельств 1994 г. Томас Флеминг иллюстрирует точку зрения на обсуждение исследований рака. Часто, изучая методы лечения рака, ученые не могут ждать достаточно долго, чтобы оценить влияние лечения на смертность. Поэтому они изучают влияние на промежуточный показатель. Одним из таких популярных промежуточных показателей является размер опухоли.

Например, Флеминг описывает медицинское испытание препарата, предназначенного для лечения рака простаты. Исследователи определили, что, если в качестве показателя они будут рассматривать смертность, им понадобится выборка из 40 000–100 000 человек, чтобы обнаружить значимый эффект, поскольку смерть от рака простаты встречается редко, а заболевание протекает медленно. Поскольку для своего исследования они смогли привлечь только 18 000 мужчин, вместо этого они решили использовать для оценки эффективности препарата размер опухоли, измеренный с помощью биопсии простаты.

Одна из проблем, как обсуждает Флеминг, заключается в том, что размер опухоли простаты является лишь очень слабым показателем реальной миссии, которая заключается в предотвращении смерти от рака. Тридцать процентов мужчин старше 50 лет дают положительный результат теста на опухоль простаты. Но только 3 % на самом деле умирают от рака простаты. Многие опухоли простаты растут очень медленно. Поэтому другие заболевания, такие как сердечные приступы, волнуют людей намного больше. Эксперимент показал, что тестируемый препарат значительно уменьшил размер опухоли. Но вполне возможно, что большая часть зарегистрированных результатов относится к тем опухолям, которые не причинили бы фактического вреда субъектам. Поэтому на самом деле мы не знаем, способствовал ли прогресс в достижении этого промежуточного результата прогрессу в выполнении миссии по предотвращению смерти от рака простаты.

Конечно, мы не хотим сказать, что изучение промежуточных результатов – плохая идея. Действительно, зачастую это лучшее, что можно сделать, учитывая другие ограничения. Но при интерпретации связи между каким-либо действием и промежуточным результатом следует критически оценивать дальнейшую связь между промежуточным результатом и реальной миссией.

Плохо определенные миссии

Часто вашу миссию бывает сложно определить. В частности, иногда существует несколько разумных способов измерения того, что может выглядеть одной и той же миссией. Но ваш выбор может иметь большое значение. Поэтому важно тщательно подумать о том, какие результаты и методы воздействия действительно определяют вашу миссию.

Предположим, вы студент колледжа и обдумываете свой выбор образования и карьеры с целью максимизировать будущий заработок. Первое, что вам

может прийти в голову, – это изучить список самых богатых людей мира, по версии Forbes, и попытаться пойти по их стопам. Вы можете сделать вывод, что хороший способ максимизировать свои доходы – это бросить университет и основать технологическую компанию. Именно эту стратегию выбрали Билл Гейтс, Марк Цукерберг и Ларри Эллисон, трое из восьми богатейших людей мира на момент написания этой книги. Но вы не совершите этой ошибки, поскольку в главе 4 вы узнали, что корреляция требует вариаций. Чтобы узнать, коррелирует ли отказ от учебы в колледже и создание технологической компании с успехом, вы не можете ограничиться изучением самых успешных людей.

Предположим, вы пошли дальше и попытались понять, сколько людей из генеральной совокупности бросили колледж и основали собственную технологическую компанию. Вы наверняка обнаружите, что менее 0.01 % всех людей бросили колледж и основали собственную технологическую компанию, и тем не менее 37.5 % из восьми самых богатых людей мира сделали это. Таким образом, похоже, существует сильная корреляция. Люди, которые бросают учебу и создают собственную технологическую компанию, с гораздо большей вероятностью войдут в число восьми богатейших людей мира, чем люди, которые остаются в колледже или никогда не создают технологическую компанию.

Даже при наличии такой явной корреляции все же есть причины, по которым не стоит принимать опрометчивое решение и бросать университет. Во-первых, мы могли просто случайно заняться чем-то вроде *p*-хакинга. Мы изучили небольшую группу чрезвычайно богатых людей, искали общие черты и в конце концов нашли то, что есть общего у некоторых из них. Но это может быть просто совпадением. Возможно, корреляция, которую мы наблюдаем сегодня, не сохранится в будущем, и в этом случае бросать учебу и создавать технологическую компанию может оказаться плохой идеей.

Еще одна причина, по которой мы бы не рекомендовали бросать учебу и создавать компанию, заключается в том, что мы прибегли к некорректному сравнению. Люди, которые бросают университет и создают компанию, наверняка существенно отличаются от тех, кто этого не делает, и у нас мало шансов узнать, добились бы они такого же успеха, если бы не бросили учебу. То есть, в соответствии с уроками главы 9, эта корреляция не является объективной оценкой причинно-следственной связи.

Но даже если оставить в стороне все эти причины, в самом образе мышления есть фундаментальная проблема, связанная с правильной оценкой вашей миссии. Какой результат вас действительно волнует? Это ваш ожидаемый доход или ваша вероятность стать мультимиллиардером? Поскольку отказ от учебы в вузе и создание собственной технологической компании повышает вероятность того, что вы станете одним из самых богатых людей в мире, это, вероятно, также увеличивает вероятность того, что вы окажетесь в серьезных долгах. И, насколько нам известно, отказ от обучения в вузе может значительно снизить ваши ожидаемые доходы, даже если увеличит ваши шансы стать очень богатым. Готовы ли вы пойти на такую рискованную игру?

Мы здесь не для того, чтобы говорить вам, каковы должны быть ваши конкретные цели. У некоторых людей может быть глубокое желание стать миллиардером, что заставляет их идти на значительный риск. Но мы подозреваем, что большинство людей менее склонны к риску и предпочитают максимизировать

свои ожидаемые доходы или, возможно, даже минимизировать свои шансы оказаться в бедности. Ваша конкретная цель должна определять проводимый вами анализ. Если ваша цель – максимизировать ожидаемый доход, было бы огромной ошибкой изучать корреляты попадания в список самых богатых людей Forbes. Вместо этого вам нужно собрать данные о доходах, чтобы увидеть, как различные варианты образования и карьеры в среднем соответствуют доходам. Мы подозреваем, что окончание университета и, возможно, даже поступление в профессионально-техническое училище является лучшим показателем будущего заработка, чем отказ от учебы и открытие собственной компании.

Эту ошибку, связанную с изучением неправильного результата, можно совершить и в другом направлении. Если вы управляете предвыборной кампанией или тренируете спортивную команду, вас не особо заботит ожидаемый разрыв по очкам или точная доля голосов. Что вас волнует, так это победа, поэтому следует выбирать стратегии, которые максимизируют эту цель. Например, если у кандидата, на которого вы работаете, плохие результаты опросов и до конца предвыборной кампании осталась всего неделя, возможно, вы захотите сделать ставку на опрометчивую в противном случае стратегию, чтобы дать ему шанс на победу. Возможно, вы решите озвучить очень агрессивные предвыборные обещания, которые могут не понравиться избирателям. В среднем такая стратегия наверняка уменьшит общее количество ваших голосов. Но есть небольшой шанс, что избирателям понравится ваша дикая идея и вы победите. Если вас на самом деле не волнует доля голосов (проигрыш на пять или десять пунктов все равно будет проигрышем), а заботитесь вы только о победе, оптимальной может оказаться стратегия, которая повредит вашей ожидаемой доле голосов.

И конечно же, эта проблема измерения касается не только результатов. Все сказанное также относится к измерению воздействия. Это, пожалуй, наиболее очевидно, когда измеряемые нами показатели отражают абстрактные понятия. Мы должны ясно понимать, что именно мы измеряем, когда оцениваем некоторые страны как более или менее демократичные, чем другие, или некоторые школьные классы, более или менее способные, чем другие. Но эта проблема может возникнуть и при измерении более конкретных показателей. Вот пример.

Изменение климата и экономическая продуктивность

Многих людей интересуют долгосрочные последствия изменения климата для экономического роста. Изменение климата, конечно, происходит в течение длительного периода времени, и поэтому его трудно измерить и изучить. Но связанные с ними явления, такие как погода и температура, часто меняются. Поэтому ученые иногда используют изменения погоды, чтобы попытаться узнать о последствиях изменения климата.

Например, Маршалл Берк, Соломон Сян и Эдвард Мигель оценивают влияние неожиданных колебаний температуры на рост ВВП, используя метод разности различий. То есть они сравнивают ВВП внутри страны в те годы, когда она подвергается воздействию более высоких и более низких температур, чем в среднем, из-за естественных колебаний климата. Они обнаружили, что экономическая производительность максимальна при среднегодовой температуре 13 °C и что она резко снижается с повышением температуры. Они приходят к выводу, что «если будущая адаптация будет повторять наблюдаемую адапта-

цию, ожидается, что всеобщее потепление изменит глобальную экономику за счет сокращения средних мировых доходов примерно на 23 % к 2100 г.».

Это важное исследование и важный вывод. Но авторы делают интересную оговорку («если будущая адаптация будет повторять наблюдаемую адаптацию»), указывая на критическую проблему измерения.

Авторов интересуют последствия изменения климата. Но воздействие, которое они измеряют, – это колебания температуры. Изменение климата происходит медленно, давая людям и обществу время адаптироваться. Локальные колебания температуры происходят быстро, что затрудняет адаптацию. Более того, в отличие от колебаний температуры, изменение климата связано с изменениями в картине погоды, переносчиками болезней, распространенностью стихийных бедствий и т. д. Таким образом, во многом колебания температуры не являются показателем интересующего воздействия. И в частности, они не измеряют правильное воздействие способами, которые имеют отношение к вопросу экономической эффективности. В свете этих проблем с измерением нам, вероятно, не следует слишком доверять предполагаемому эффекту в 23 %.

Чтобы оценить это различие, рассмотрим разницу между влиянием на экономическую производительность жаркого дня и жаркого столетия. Мы живем в Чикаго, где в основном довольно прохладно. Если бы нас приятно удивил особенно теплый день, у Энтони мог бы возникнуть соблазн уйти с работы пораньше, чтобы поиграть в гольф. Но если бы изменение климата означало, что большинство дней станут одинаково теплыми, он бы не бросил работу чтобы каждый день играть в гольф. А если это означает, что дни станут теплыми, но с грозами и ливнями, кто знает, что случится с его игрой в гольф. Тот факт, что неожиданные жаркие дни снижают продуктивность, не обязательно говорит нам о долгосрочных последствиях изменения климата, потому что мы не измерили и не изучили правильные вещи.

Есть ли у вас подходящая выборка?

Изучение правильного результата и правильного воздействия – это еще не все, что нужно для измерения нашей миссии. Также необходимо убедиться, что у нас есть подходящая выборка.

Применяя доказательства для принятия решений, нам почти всегда приходится брать знания, полученные в каком-то месте и времени, и пытаться применить эти знания, чтобы понять, что произойдет в другом месте и времени. По сути, мы проводим аналогию между контекстами, в которых были получены свидетельства, и контекстами, в которых теперь хотим применить уроки, извлеченные из этих свидетельств. Поэтому нам всегда приходится задаваться вопросом, достаточно ли схожи эти контексты, чтобы такая аналогия была верной. В противном случае мы можем предпринимать действия, соответствующие достижению миссии, но только в совершенно другом контексте, чем тот, в котором мы действуем.

Внешняя валидность

Основная проблема здесь в том, что взаимосвязи могут различаться от контекста к контексту. До сих пор в этой книге мы уделяли много времени тому, что иногда называют *внутренней валидностью*. Внутренняя валидность отражает

достоверность оценки оцениваемой величины (например, является ли оценка несмещенной?). Но даже если вы сделали все правильно в отношении внутренней валидности, вам все равно нужно критически подумать о том, будет ли эта связь существовать в контексте, в котором вы надеетесь ее применить. В широком смысле это проблема *внешней валидности*. Внешняя валидность заключается в том, есть ли веские основания полагать, что взаимосвязь, оцененная на основе данных из одного контекста, будет сохраняться и в каком-то другом контексте. Следующий пример иллюстрирует эту мысль.

Недоедание в Индии и Бангладеш

В 1980-х гг. Всемирный банк реализовал комплексный проект обеспечения питанием штата Тамил Наду (TINP) в регионе на юге Индии, где недоедание было эндемичным. Хотя проект включал в себя некоторые ресурсы для поставок дополнительного питания, основное внимание уделялось помощи матерям – основным лицам, принимающим решения в домохозяйстве относительно покупки и приготовления продуктов питания, – более эффективно использовать ресурсы, уже имеющиеся в их распоряжении. Всемирный банк рассматривает TINP как крупный успех. И, несмотря на некоторые споры, многие считают, что именно он внес значительный вклад в сокращение недоедания и неполноценного питания в штате Тамил Наду.

Этот очевидный успех вдохновил запуск Бангладешского комплексного проекта обеспечения питанием (BINP) в 1990-х гг. К тому времени Бангладеш, граничащая с Индией на востоке, была одной из самых голодающих стран на земле. Имеющиеся данные свидетельствуют о том, что в начале 1990-х гг. почти две трети бангладешских детей в возрасте до пяти лет имели задержку роста из-за недоедания.

Поскольку TINP был тщательно проанализирован и выяснилось, что он внес значительный вклад в проблему недоедания, BINP был выстроен непосредственно на основе TINP. Поэтому как ученые, так и практики были удивлены, когда эффект BINP не оправдал ожиданий. Несмотря на то что проект BINP был разработан с целью воспроизвести, возможно, самое успешное в истории мероприятие по борьбе с неполноценным питанием, тщательная оценка практически не выявила влияния BINP на недоедание. Что пошло не так?

Конечно, существует множество возможных ответов. И практически невозможно точно узнать, почему программа потерпела неудачу. Но одним важным фактором, по-видимому, являются культурные различия между Тамил Наду и Бангладеш. Как мы уже упоминали, в штате Тамил Наду матери, как правило, являются главными лицами, принимающими решения относительно покупки и приготовления продуктов питания. Поэтому основные усилия TINP по просвещению в области питания были направлены на матерей.

Этот акцент на матерях был напрямую перенесен из TINP в BINP. Но во многих семьях в Бангладеш решение о покупке или приготовлении еды принимают отец ребенка или свекровь (т. е. мать отца), а не мать. Поскольку в Тамил Наду было иначе, этот важный момент не учитывался в BINP. Таким образом, проект в Бангладеш потерпел неудачу, по крайней мере частично, потому что решение о таргетировании, которое имело смысл в одной ситуации, оказалось неподходящим в другой.

Этот пример особенно интересен для нас, поскольку он указывает на возможность взаимодополняемости между количественными данными и качественными знаниями. Оценка воздействия TINP потребовала количественного подхода. Но попытка применить эти знания к контексту Бангладеш оказалась неудачной из-за отсутствия знаний о ключевых культурных и институциональных различиях между Тамил Наду и Бангладеш. Команда, объединившая людей с опытом количественной оценки, правильно оценивающих эффект TINP, и людей с глубокими качественными знаниями культурных контекстов, могла бы привести к лучшему результату, чем по отдельности.

Ограниченная выборка

Наиболее распространенная причина, по которой люди в конечном итоге оценивают свою миссию в неправильном контексте, – это использование ограниченных выборок. *Ограниченная выборка*, или *отобранные образцы* (selected sample), – это выборка наблюдений, которая не получена случайным образом из интересующей совокупности, а целенаправленно отобрана для рассмотрения, поскольку она обладает некоторым набором характеристик. Проблема, конечно, в том, что ограниченные выборки могут не быть репрезентативными для совокупности в целом. И связи, которые наблюдаются внутри этих выборок, могут отсутствовать в более широкой совокупности. Если ваша миссия состоит в том, чтобы предсказать, понять или повлиять на поведение более широких слоев населения, все может пойти не так, если вы будете полагаться на данные ограниченных выборок.

Поступление в университет

Вот пример, который близок и дорог нашему сердцу. Результаты стандартизированных тестов, хорошо это или плохо, были важной частью процесса поступления в высшее учебное заведение на протяжении десятилетий. Однако в 2018 г. наш университет объявил, что больше не будет требовать от абитуриентов предоставления таких оценок. (Несколько других колледжей и университетов сделали то же самое.) Одно из нескольких обоснований необязательности тестов было основано на фактических данных. Руководители университетов изучили студентов, посещавших университет, и обнаружили слабую корреляцию между результатами тестов и успеваемостью. Итак, рассуждали они, возможно, результаты тестов не являются хорошим предиктором успеваемости в колледже.

Миссия приемной комиссии вуза многогранна. Частью этой миссии является выявление наиболее талантливых студентов из числа претендентов. Чтобы выполнить эту задачу, приемная комиссия хотела бы знать, коррелируют ли некоторые характеристики *абитуриентов* (в данном случае, их результаты тестов) с академической успеваемостью в вузе. Но это не тот вопрос, который рассматривается в данной ситуации. Скорее, эти рассуждения отвечают на вопрос, коррелируют ли некоторые характеристики *зачисленных студентов* (а именно их результаты тестов) с академической успеваемостью в колледже. Но ответ на эти два вопроса не обязательно должен быть одинаковым.

Зачисленные студенты представляют собой ограниченную выборку из совокупности абитуриентов. Студентов принимали в колледж на основании результатов тестов и других факторов, таких как умение писать эссе, рекомендации

учителей, оценки, участие в общественной деятельности и целеустремленность. Тот факт, что результаты тестов использовались при поступлении, может привести к принципиально разной корреляции между результатами тестов и академической успеваемостью у ограниченной выборки зачисленных студентов по сравнению с более широким кругом абитуриентов.

Чтобы понять, почему так получается, подумайте о студентах с низкими результатами тестов, которые тем не менее были приняты в университет. У этих студентов, должно быть, были какие-то другие характеристики, из-за которых приемная комиссия не обращала внимания на их низкие баллы. Возможно, они написали блестящие эссе, заслужили особенно сильные рекомендации от учителей или получили отличные оценки в старшей школе. Аналогичным образом студенты с особенно высокими результатами тестов, скорее всего, были приняты даже с несколько более низкими показателями по другим параметрам. По этой причине в совокупности принятых и зачисленных студентов можно ожидать отрицательную корреляцию между результатами тестов и другими показателями академической успеваемости.

Вполне вероятно, что результаты тестов являются хорошим предиктором успеваемости *абитуриентов в целом*, но умение писать эссе, рекомендации учителей и оценки в средней школе также являются хорошими предикторами. Поэтому, взглянув на ограниченную выборку *зачисленных студентов*, мы обнаружим слабую корреляцию между результатами тестов и успеваемостью. Дело в том, что в нее попали в числе прочих люди с низкими баллами, которые сильны по другим критериям. Таким образом, слабая (или отсутствующая) корреляция в ограниченной выборке зачисленных студентов не означает, что результаты тестов являются плохим предиктором академической успеваемости абитуриентов.

Проблема изучения ограниченных выборок распространена и за пределами процесса поступления в вузы. Рассмотрим еще один пример – бейсбол.

Почему питчеры высшей лиги плохо бьют?

Поклонники Высшей лиги бейсбола знают, что питчеры, как правило, являются худшими нападающими в своих командах. В Национальной лиге, где питчеры обязаны бить, менеджеры обычно заставляют своих питчеров бить последними в составе, чтобы свести к минимуму их пробежки к базе. А если питчер выбывает из игры, менеджер всегда заменяет его нападающим. В сезоне Высшей бейсбольной лиги 2017 г. средний результат среднего питчера составлял 0.125. Средний показатель обычного игрока, не являющегося питчером, составлял 0.259. Это огромная разница. В Американской лиге есть специальное правило для нападающих, согласно которому питчерам не нужно бить.

Так почему же питчеры Высшей лиги так плохо бьют? Если вы спросите эксперта по бейсболу, он, вероятно, скажет вам, что питчеры тратят так много времени на отработку подачи, что у них нет времени на отработку ударов. И возможно, в великих питчерах даже есть что-то, что делает их более слабыми нападающими. Возможно, та сила, гибкость или телосложение, которые хороши для подачи, плохи для ударов.

Эти объяснения звучат довольно убедительно, и, вероятно, в некоторой степени они верны. Но вряд ли они раскрывают все причины. Достаточно заметить, что эта закономерность не применима к школьному бейсболу.

Мы собрали данные о четырех школьных бейсбольных командах Чикаго в сезоне 2018 г. и рассчитали среднее количество ударов для не-питчеров и питчеров (определяемых как игроки, которые отработали более десяти подач за сезон). В отличие от профессионалов, среди этих старшеклассников питчеры имеют немного более высокий средний показатель результативности, чем не-питчеры: 0.322 против 0.317.

Как такое возможно? Почему корреляция между способностью принимать подачу и бить слегка положительна для школьных бейсболистов, но отрицательна для опытных профессионалов? Это наблюдение не исключает аргумент, что у профессиональных питчеров более отработанная специализация. С другой стороны, аргументы относительно особенностей телосложения одинаково применимы как в старшей школе, так и в высшей лиге. Так почему же корреляция меняется так резко и даже меняет знак с возрастом игроков?

Даже на профессиональном уровне мы видим, что не всегда существовала отрицательная корреляция между подачей и умением отбивать мяч. На рис. 16.3 изображены средние показатели количества ударов среднего питчера и среднего игрока, не являющегося питчером, в Высшей лиге бейсбола с 1871 по 2017 гг. В XIX в. питчеры и игроки на других позициях имели сопоставимые средние показатели ударов. Но, начиная с XX в., питчеры, похоже, стали хуже бить по сравнению с другими игроками, и со временем разрыв постепенно увеличивался. А в современную эпоху, как мы уже упоминали, игроки, не являющиеся питчерами, в два раза чаще попадают в базу, чем питчеры.

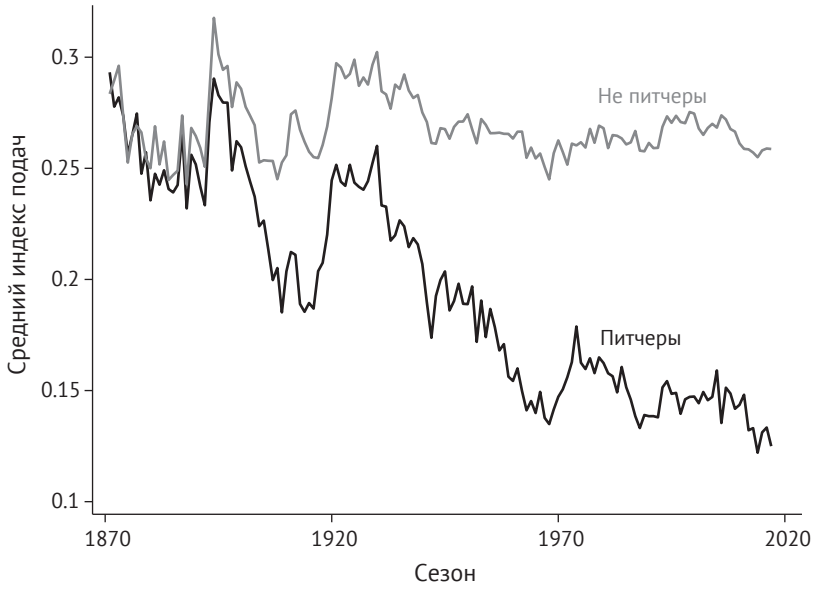


Рис. 16.3. Отрицательная корреляция между подачей и умением бить в высших лигах со временем стала сильнее

Мы подозреваем, что изменение корреляции с течением длительного времени и разница в корреляции между старшеклассниками и профессионалами объясняются одной и той же причиной. И она связана с ограниченной выборкой.

Начните с размышлений о взаимосвязи между способностью ловить подачу и отбивать мяч среди всей совокупности игроков. Предположим, мы просто случайным образом выбрали людей (скажем, подростков и старше) со всего мира и попросили их поиграть в бейсбол, чтобы мы могли измерить их способности ловить подачу и отбивать мяч. Как вы думаете, что мы найдем? Мы подозреваем, что обнаружим довольно сильную положительную корреляцию. Некоторые люди спортивны и имеют опыт игры в бейсбол. Они, вероятно, будут хороши в том и другом. Некоторые люди, напротив, не скоординированы и неопытны. Они, вероятно, будут плохо делать то и другое. Таким образом, среди населения в целом вы, скорее всего, обнаружите прямо противоположную корреляцию по сравнению с той, что наблюдается среди профессионалов.

Чтобы понять, почему это так, подумайте о том, как человек становится игроком Высшей бейсбольной лиги. Он почти наверняка играл в школе. Тренер средней школы пытается собрать лучшую команду. Сюда входит выбор игроков из числа школьников, которые обладают лучшим сочетанием навыков отбивания и подачи. Чтобы попасть в школьную команду, вы должны хорошо владеть сочетанием навыков. Но не обязательно быть выдающимся – вы можете быть хорошим нападающим (даже если вы плохой питчер), хорошим питчером (даже если вы плохой нападающий) или, по крайней мере, сносным питчером и нападающим.

Для большинства игроков путь в Высшую лигу лежит через низшие лиги. Тренеры низших лиг также стараются собрать как можно лучшую команду. И поэтому им тоже нужно лучшее сочетание ударов и подач. Итак, чтобы попасть в команду низшей лиги, вы должны быть очень, очень хороши в сочетании навыков. Это означает, что вы должны быть отличным нападающим (даже если вы плохой питчер), отличным питчером (даже если вы плохой нападающий) или, по крайней мере, неплохим питчером и надежным нападающим.

Наконец, чтобы попасть в Высшую лигу (по крайней мере, в Национальную, куда попадают питчеры), тест становится еще более строгим. Вы должны быть по-настоящему потрясающим нападающим (даже если вы плохой питчер), по-настоящему выдающимся питчером (даже если вы плохой нападающий) или довольно хорошим питчером, который также может бить.

Ниже приведена простая демонстрация (с гипотетическими данными) того, как эти все более строгие критерии отбора влияют на корреляцию между способностями к удару и подаче в различных выборках. Предположим (для простоты), что мы можем дать каждому потенциальному бейсболисту оценку, которая отдельно суммирует его способности к подаче и отбиванию мяча, и команды хотят нанять игроков с максимально возможной суммой способностей как к подаче, так и к отбиванию мяча. Насколько большой должна быть эта сумма, зависит от вашего продвижения по карьерной лестнице в бейсболе.

В верхней левой части рис. 16.4 мы нарисовали диаграмму распределения некоторых данных с сильной положительной корреляцией между подачей (горизонтальная ось) и способностью отбивать мяч (вертикальная ось). Здесь представлена вся популяция. Если бы мы взяли в команду всех подряд (как это могла бы сделать команда начального уровня для детей), мы бы увидели довольно сильную положительную корреляцию между способностью подавать и отбивать мяч. Нам это кажется правильным. Наши воспоминания о юноше-

ском спорте таковы, что дети, которые были лучшими в одном аспекте игры, часто были лучшими во всех аспектах игры.

Верхняя правая часть рис. 16.4 символизирует среднюю школу или высшую лигу XIX в. В старших классах тренер готов принимать в команду только игроков с уровнем выше среднего в популяции, что означает, что сумма навыков отбивания и подачи выше 0. Аналогично в XIX в. бейсбол не был так популярен, поэтому профессиональные тренеры не были особо привередливы. Вы можете попасть в такую команду, будучи хорошим нападающим (скажем, 3) и слабым питчером (скажем, 2); хорошим питчером (скажем, 3) и слабым нападающим (скажем, 2) или питчером и нападающим чуть выше среднего (скажем, 0.5 по обоим критериям). Но этот уровень избирательности исключает людей, которые плохи в обоих случаях. Итак, среди ограниченной выборки бейсболистов старших классов или бейсболистов XIX в. корреляция между способностями подавать и отбивать мяч незначительна.

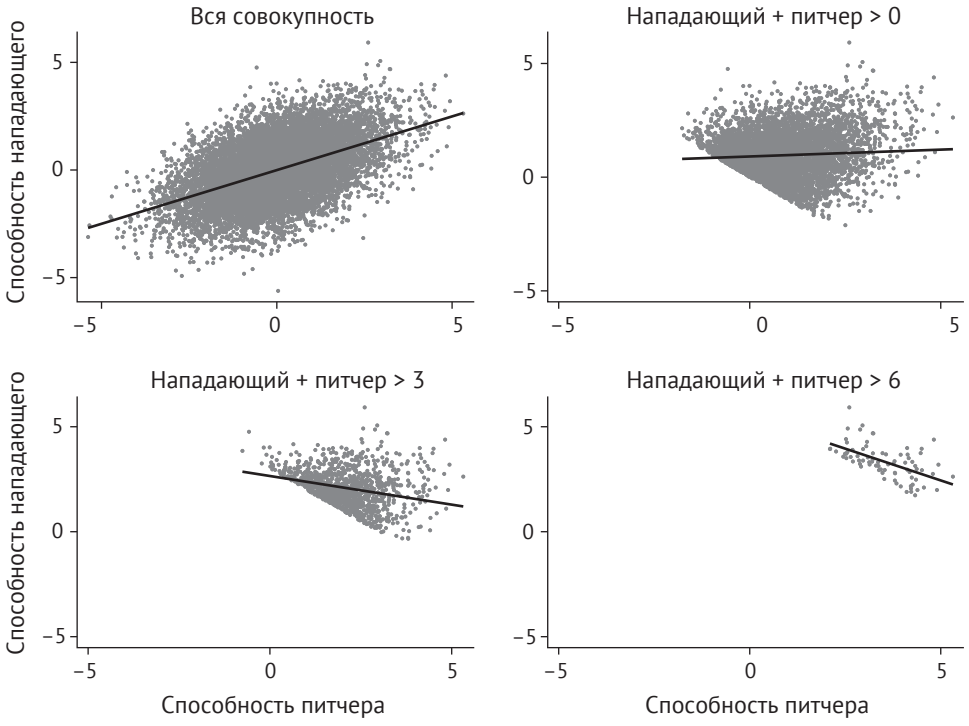


Рис. 16.4. Повышение критериев отбора может превратить положительную корреляцию в отрицательную

Левая нижняя часть рисунка представляет низшие лиги или начало XX в. Селективность возросла. Чтобы попасть в команду, игрокам нужно набрать минимум 3 балла. Таким образом, плохие питчеры должны быть по-настоящему хорошими нападающими, плохие нападающие должны быть действительно хорошими питчерами, а некоторые игроки могут попасть в команду, если будут достаточно сильны в обоих случаях. Исключая еще больше игроков-универсалов, которые хороши только в обоих случаях, этот уровень избирательности меняет

соотношение навыков внутри ограниченной выборки, вызывая небольшую отрицательную корреляцию между способностями отбивать мяч и подавать.

Наконец, правая нижняя часть представляет современную Национальную лигу. Здесь избирательность очень высока, поскольку попасть в Высшую лигу могут лишь единицы. Таким образом, плохие питчеры должны быть по-настоящему великими нападающими, плохие нападающие должны быть по-настоящему великими питчерами, а чтобы игроки попали в лигу, обладая комбинацией навыков, они должны быть потрясающими в обоих случаях. Таким образом, на самых высоких уровнях мы ожидаем увидеть сильную отрицательную корреляцию между способностью подавать и отбивать мяч, даже вне зависимости от времени, затрачиваемого на тренировки, или типа телосложения.

Рисунок 16.4 иллюстрирует довольно распространенное явление. Корреляции в ограниченных выборках часто сильно отличаются от корреляций в более широкой популяции. Это важно, поскольку зачастую у нас есть данные только по ограниченным выборкам. Но от нас могут ожидать прогнозов и выводов о более широких совокупностях.

Если вы бейсбольный скаут, у вас есть обширные данные об игроках высшей лиги. Можете попытаться использовать эти данные, чтобы предсказать, кто станет звездным игроком, и найти кандидатов среди спортсменов средней школы и колледжа. Но если вы посмотрите на корреляцию выдающихся результатов среди игроков высшей лиги, то сделаете ошибочные выводы. Например, вы можете обнаружить, что те, кто медленно бегают по полю, особенно хороши в мощных ударах. Так стоит ли искать самых медленных бегунов и приглашать их играть в профессиональный бейсбол? Конечно, нет! Причина, по которой медленные бегуны являются хорошими мощными нападающими в ограниченной выборке, та же, что и причина, по которой хорошие питчеры являются плохими нападающими. Единственный способ, которым медленный бегун может попасть в высшую лигу, – это быть выдающимся нападающим.

Стратегическая адаптация и изменение отношений

Есть еще один ключевой вопрос, на который стоит обратить внимание при измерении вашей миссии. Иногда в мире существуют отношения между событиями или объектами, которые вы могли бы использовать для достижения своей цели. Но как только вы попытаетесь это сделать, стратегическая адаптация заставит эти отношения исчезнуть или измениться, и они перестанут быть такими полезными. Чтобы понять, как это происходит, рассмотрим исторические примеры и явления.

Налоги на свет и окна

В 1696 г. английскому королю Вильгельму III понадобились деньги. Королям, конечно, всегда нужны деньги. Но эта потребность была особенно острой. Вплоть до 1660-х гг. Британия производила монеты из чеканного серебра. У этих монет была серьезная проблема: люди соскабливали или срезали ценное серебро с краев монет. В результате стоимость монет в серебре оказалась меньше их номинала. Эта широко распространенная практика обрезки монет угрожала подорвать доверие к английской валюте.

Чтобы решить эту проблему, Корона затеяла большую перечеканку монет в 1696 г., предложив выкупить обрезанные монеты в обмен на новые, обра-

ботанные на станке монеты, которые нельзя было обрезать[1]. Но выкуп обрезанных монет за настоящие монеты обходился дорого. По сути, Корона должна была компенсировать разницу между номинальной стоимостью монеты и стоимостью серебра. Итак, Короне необходимо было увеличить доходы. Но как это сделать?

Корона хотела облагать богатых более высоким налогом, чем бедных. Одним из естественных способов добиться этого является введение подоходного налога. Но англичане были против подоходного налога, поскольку оценка дохода подразумевала вторжение в личную жизнь. Поэтому Короне нужно было найти способ налогообложения богатства, более приемлемый с политической точки зрения. Решением стал налог на свет и окна, более известный как налог на окна.

Налог на окна имел то преимущество, что его можно было взимать по результатам внешнего осмотра дома, тем самым исключая любое вторжение в частную жизнь. В самой ранней версии налога Корона установила базовую плату в размере двух шиллингов за все дома. Кроме того, дома, в которых было от десяти до двадцати окон, платили дополнительно от четырех до шести шиллингов, а дома, в которых было более двадцати окон, платили дополнительно от восьми до десяти шиллингов. Окна в рабочих помещениях не учитывались. Точные пороги и сборы со временем менялись (такие налоги действовали более века), но основную идею вы поняли.

Аргументом в пользу налога на окна была очевидная корреляция между окнами и богатством (конечно, казначеи не использовали такие термины). В среднем люди, в чьих домах было больше окон, были богаче. Облагая налогом окна, Корона могла увеличить доходы таким образом, чтобы большая часть бремени ложилась на богатых и меньшая на бедных, что и было ее миссией.

Но на этом история не заканчивается. Англичане, как и многие другие народы, не любят платить налоги. И поэтому они стратегически адаптировались. Очень быстро многие окна были заколочены или заложены кирпичом, чтобы уменьшить причитающиеся налоги. Со временем архитектура изменилась. В больших домах стало меньше окон и больше комнат, которые можно было представить как рабочие помещения. Таким образом, с течением времени как доходы Короны, так и прогрессивность налога на окна снизились.

В данном случае миссией Короны было постепенное увеличение доходов. Для этого необходимо было выявить и обложить налогом более богатых людей, не вторгаясь в их частную жизнь. Авторы идеи заметили корреляцию между окнами и богатством, которая, казалось, была именно тем, что нужно для достижения цели. Но использование этой корреляции заставило домовладельцев стратегически адаптировать свое поведение так, чтобы корреляция больше не сохранялась (или, по крайней мере, сохранялась гораздо менее сильно), что подрывало миссию. Следовательно, обсуждая изменения в поведении или политике в ответ на какие-то наблюдения, всегда нужно задаваться вопросом, сохранится ли взаимосвязь, выявленная этими наблюдениями, после того как вы измените свое поведение или политику.

---

[1] Интересный факт: новые монеты нельзя было обрезать, потому что у них были рифленые края, и эта особенность сохраняется и сегодня, хотя наши монеты не сделаны из драгоценных металлов. Рифленые края придумал Исаак Ньютон, будучи в должности смотрителя Королевского монетного двора во времена великой перечеканки монет.

Сдвиг в бейсболе

Мы знаем, что уже потратили немало времени на бейсбол в этой главе. Но если вы позволите, мы хотели бы привести еще один пример. Он хорошо иллюстрирует идею стратегической адаптации, меняющей полезность статистических взаимосвязей.

Было время, когда защитники в бейсболе стояли на своем месте и ждали, прилетит ли мяч в их сторону. Конечно, полевые игроки немного корректировали свою позицию в зависимости от того, был ли отбивающий левша или правша. Но по большей части оборонительная стратегия не была слишком сложной.

Это время подошло к концу с появлением больших данных в профессиональном спорте. Теперь у команд есть подробные графики для каждого отбивающего. Эти диаграммы предоставляют данные о том, как часто каждый отбивающий попадает в различные части поля, отбивают ли они мяч по земле или в воздухе, под каким углом они соприкасаются с мячом и т. д. Используя такого рода информацию, команды могут делать точные прогнозы о том, где именно тот или иной отбивающий может ударить по мячу. И, вооружившись такими прогнозами, команды начали агрессивно корректировать свои оборонительные схемы, игра за игрой.

Самая известная версия этого изменения в оборонительной стратегии называется *сдвигом*[1]. Изучая диаграммы распределения, команды обнаружили, что, когда отбивающие (особенно мощные нападающие) отбивают мяч на земле, это почти никогда не происходит в так называемое противоположное поле (для правшей это справа, а для левшей это слева от них). Скорее, если они собираются ударить мяч на земле, то отбивающие правши бьют по нему слева, а отбивающие левши – справа. Сдвиг является очевидным ответом на эту корреляцию: в случае отбивающего-правши сдвиньте ожидаемую траекторию влево от отбивающего, а в случае отбивающего-левши – вправо. Преимущество такой стратегии заключается в том, что она значительно снижает вероятность проникновения наземного мяча через открытый коридор в инфилде. Издержки этой стратегии заключаются в том, что она оставляет большую дыру в инфилде противоположного поля. Но поскольку отбивающим очень трудно отбить наземные мячи в противоположное поле, эти затраты невелики.

Некоторые команды начали агрессивно меняться в конце 2000-х. В 2010 г. на команду Tampa Bay Rays под руководством менеджера Джо Мэддона, одного из первых сторонников продуманной защиты, приходилось 10 % всех сдвигов, хотя они были лишь одной из 30 команд. Мэддон сверился с таблицами распределения и стратегически разместил своих инфилдеров в местах, которые были оптимальны для конкретного рисунка траекторий наземных мячей, связанного с каждым отбивающим. Tampa Bay Rays и другие ранние последователи добились большого успеха. То есть существовала отрицательная корреляция между использованием сдвига и разрешенными пробегами.

Наблюдая эту корреляцию, все команды начали использовать сдвиг. В 2011 г. во всех играх Высшей лиги бейсбола было использовано всего около 2000 сдвигов. К 2014 г. это число выросло до 13 000. А в 2016 г. оно превысило 28 000.

Но произошло и кое-что еще. Поначалу корреляция, вызвавшая этот всплеск, сохранялась. Команды, которые использовали сдвиг, допустили меньше пробе-

[1] *https://wiki5.ru/wiki/Infield_shift*.

жек. Но отбивающие заметили, что новая тактика причиняет им вред. И они стратегически адаптировались, чтобы не забивать так много мячей по земле. Вместо этого они начали отбивать больше мячей на противоположную сторону поля и преимущественно в воздухе.

При нынешнем положении дел команды высшей лиги по-прежнему активно используют тактику сдвига. Но поскольку нападающие адаптировались, корреляция между сдвигами и пробегами уже не позволяет командам использовать тактику сдвига так эффективно, как раньше. Выработка защитной стратегии в ответ на корреляцию привела к изменениям в поведении, разрушившим эту корреляцию. Возможно, стоит отметить, что Джо Мэддон – один из первых новаторов, который, будучи менеджером Tampa Bay Rays, помог сделать этот сдвиг столь популярным – по-прежнему верит в защиту, основанную на фактических данных. Позже он выиграл Мировую серию в качестве тренера Chicago Cubs, где использовал сдвиг реже, чем любой другой менеджер Высшей бейсбольной лиги.

Война с наркотиками

Прежде чем оставить тему о том, изменится ли ситуация, если вы начнете действовать, стоит остановиться и подумать о совпадении этого вопроса с нашим предыдущим обсуждением проблемы частичных измерений. Это совпадение обусловлено тем фактом, что стратегическая адаптация может вызвать оба явления.

Вспомним, что нас беспокоит в случае частичных измерений. Предположим, у вас есть лишь ограниченный критерий вашей миссии (например, угоны самолетов). Вы предпринимаете действие, и ситуация по этому показателю улучшается. Но могла иметь место стратегическая адаптация, при которой улучшение в одном аспекте вашей миссии означало ухудшение в каком-то другом аспекте. Следовательно, улучшение частичного показателя не всегда совпадает с успехом вашей миссии.

Стратегическая адаптация снова оказывается в центре внимания, когда мы задумываемся о том, изменится ли ситуация, если мы будем действовать на основе каких-то доказательств. Окружающий мир состоит из сложных перепутанных отношений. Вы действуете в соответствии с этими отношениями. Люди приспосабливаются к вашим действиям, тем самым изменяя связь или заставляя ее исчезнуть.

Многие примеры могут относиться к обеим категориям в зависимости от того, с какой точки зрения на них посмотреть. Позвольте нам привести последний пример, чтобы проиллюстрировать эту мысль, на этот раз о так называемой войне Америки с наркотиками.

Как известно, большая часть нелегальных наркотиков в США попадает в страну через Мексику, страну, опустошенную десятилетней войной с наркотиками. Но так было не всегда. В 1970-х и начале 1980-х гг. очень немногие наркотики попадали в Соединенные Штаты через Мексику. Традиционный маршрут перевалки проходил через Карибское море во Флориду.

В 1980 г. правительство США начало крупное наступление на колумбийские наркокартели. Управление по борьбе с наркотиками, береговая охрана и другие ведомства задействовали тысячи сотрудников, а также значительные во-

енно-морские и воздушные силы, чтобы перекрыть карибский перевалочный маршрут. К середине 1980-х гг. поток наркотиков во Флориду резко сократился.

Но это еще не вся история. Сокращение потока наркотиков через Карибский бассейн и во Флориду в 1980-х гг. не отражает сокращение потока наркотиков в Соединенные Штаты в этот период. Действительно, наркотики продолжали поступать в Соединенные Штаты возрастающими темпами, о чем свидетельствует тот факт, что цена на кокаин упала в четыре раза на протяжении 1980-х гг., несмотря на резкий рост спроса.

Что же случилось? Колумбийские картели отказались от Карибского бассейна и Флориды в пользу Мексики. В 1989 г. треть всего кокаина в США поступала через Мексику. Всего три года спустя это число увеличилось до половины. Сегодня 90 % кокаина, продаваемого в США, поступает контрабандой из Мексики.

Эта адаптация со стороны наркокартелей имела разрушительные последствия для Мексики. На протяжении 1990-х гг. мексиканские группировки, занимающиеся незаконным оборотом наркотиков, стали крупнее и могущественнее. Они перешли от роли посредников для колумбийцев к созданию собственных цепочек поставки и сетей сбыта. Торговля наркотиками стала наиболее значимой отраслью: к середине 1990-х гг. объем торговли наркотиками в Мексике составлял примерно 20 млрд долл., затмевая крупнейший легальный товар Мексики, нефть, стоимостью около 7.5 млрд долл. По мере роста мексиканские наркогруппировки становились все более фрагментированными и жестокими. В 2010 г. мексиканский наркобизнес уносил более тысячи жизней в месяц. Правительство Мексики изо всех сил пыталось установить элементарный контроль над частями страны.

Этот пример можно одинаково эффективно рассматривать как с точки зрения *частичных измерений*, так и с точки зрения *меняющихся отношений*.

Вот как это выглядит с точки зрения частичных измерений. Перед правительством США стояла задача остановить поток наркотиков. Было отмечено, что почти все наркотики доставляются через Карибский бассейн. Поэтому правительство собирало данные о наркотиках, проходящих через Карибский бассейн, и это было лишь частичным показателем общей миссии по борьбе с наркотиками. Затем были предприняты действия, которые улучшили ситуацию в соответствии с этим частичным показателем. Но делать из этого вывод, что борьба с наркотиками удалась, было бы ошибкой. Благодаря стратегической адаптации улучшение этого частичного показателя (наркотики, поступающие через Карибский бассейн) сопровождается ухудшением других аспектов проблемы (наркотики, поступающие через Мексику). Данная история иллюстрирует, как важно не переоценивать улучшение отдельных показателей вашей миссии.

С точки зрения меняющихся отношений мы рассказываем историю немного по-другому. В мире существовала реальная корреляция: вероятность попадания наркотиков в Соединенные Штаты через Карибский бассейн была гораздо выше, чем откуда-либо еще. Правительство решило действовать на основе этой корреляции, нацелив усилия по пресечению наркотрафика на Карибы и Флориду. Наркоторговцы стратегически адаптировали свое поведение в ответ на эту акцию, перенеся трафик в Мексику. В итоге корреляция, составлявшая основу действий правительства, перестала существовать вследствие действий самого правительства.

Обе эти точки зрения верны. Какая из них более полезна, зависит от конкретного вопроса, на который вы пытаетесь ответить, и контекста, в котором вы пытаетесь на него ответить.

Подведение итогов

Измерение показателей вашей миссии, как и все другие уроки, которые мы обсуждали, является важной частью критического мышления о том, как использовать количественную информацию для принятия более эффективных решений. Но независимо от того, насколько проницательно вы мыслите, существуют пределы того, что могут вам сказать данные и доказательства. В главе 17 мы завершаем книгу, исследуя некоторые из этих ограничений.

Ключевые термины

- **Внутренняя достоверность:** оценка является внутренне достоверной, если она является достоверной оценкой оцениваемой величины (например, оценщик является несмещенным).
- **Внешняя достоверность:** оценка является внешне достоверной, если есть веские основания полагать, что взаимосвязь будет сохраняться в контексте, отличном от того, из которого взяты данные.
- **Стратегическая адаптация:** изменения в поведении, возникающие в результате попытки адаптироваться к изменению внешних правил или условий.
- **Ограниченная выборка:** выборка данных, которая была получена не случайным образом из интересующей совокупности, а отобрана целенаправленно, поскольку обладает некоторым конкретным набором характеристик.

Упражнения

16.1. Люди, которые уже заразились COVID-19 и выздоровели от него, с меньшей вероятностью заразятся снова из-за выработанного у них иммунитета. Тем не менее, исследование 2020 г., опубликованное в The Lancet, предполагает, что те редкие люди, которые заразились болезнью дважды, во второй раз испытывают худшие симптомы. Используя принципы рассуждений из этой главы и тот факт, что из-за ограниченного тестирования не все случаи COVID-19 выявляются, предложите объяснение, почему это явление может возникнуть даже если не существует биологического механизма, который делает второй случай COVID-19 хуже, чем первый. Должно ли это заставить вас скептически относиться к утверждению, что люди, как правило, испытывают худшие симптомы, когда заражаются болезнью во второй раз?

16.2. За последние несколько десятилетий вознаграждаемое тестирование стало все более важной частью американской политики в области образования. Идея вознаграждаемого тестирования состоит в том, чтобы создать определенные последствия для учащихся, учителей или школ, связанные

с результатами стандартизированных тестов. Есть надежда, что это улучшит успеваемость за счет создания стимулов для более высокой успеваемости. Результаты стандартизованных тестов в лучшем случае являются частичным показателем достижений в учебе. Приведите пример того, почему некоторые мероприятия, ведущие к улучшению результатов тестов, тем не менее не приводят к общему улучшению качества обучения.

16.3. В Академии BBC США студенты сдают одни и те же обязательные экзамены по математике, но их случайным образом распределяют к разным преподавателям. Скотт Каррелл и Джеймс Уэст показывают, что студенты, назначенные к преподавателю с более высокими оценками качества преподавания, лучше сдают экзамены по курсу. Но они также показывают, что назначение популярного преподавателя снижает успеваемость учащихся на последующих курсах математики. Чем можно объяснить эту загадочную закономерность? Как это связано с проблемой неспособности измерить свою миссию?

16.4. Важнейшее тестирование в системе начального и среднего образования обычно основано на пороговых значениях. Студент проходит тест, если он набирает балл выше минимального порога. Школа считается соответствующей стандартам, если количество учащихся, прошедших тест, превышает какой-либо другой минимальный порог.

a) Разделите учащихся на три категории: тех, кто пройдет тест, несмотря ни на что; тех, кто пройдет тест тогда и только тогда, когда они привлекут персональное внимание учителя; и тех, кто не пройдет тест, несмотря ни на что. На каких учениках учителям следует сосредоточить внимание?

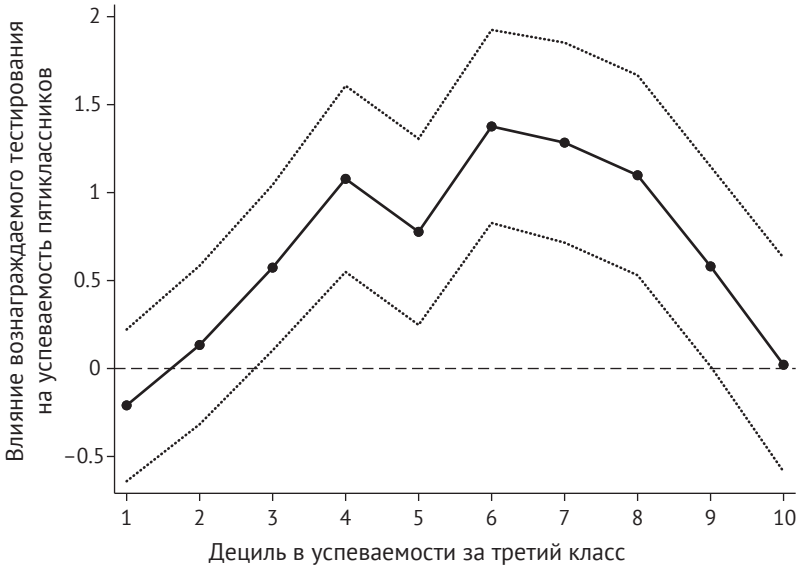
b) Дерек Нил и Дайан Уитмор Шанценбах изучали реализацию вознаграждаемого тестирования в государственных школах Чикаго, которое мы обсуждали в упражнении 1 главы 13.

Но, в отличие от этого вопроса, средний эффект от вознаграждаемого тестирования – это не совсем то, о чем они хотели знать. Они хотели знать, по-разному ли тестирование с высокими ставками влияет на разных детей.

Чтобы добиться этого, Нил и Шанценбах использовали модель «разности разностей различий». Они начали с использования тестов для третьего класса, чтобы разделить учащихся на десять групп (децили). Затем они проводят анализ разности в различиях отдельно для каждого из этих децилей. Это позволяет им найти разницу в оценке причинно-следственного эффекта вознаграждаемого тестирования среди детей в разных децилях. Их выводы отражены на рисунке ниже.

Соответствуют ли эти свидетельства вашему ответу на часть (a) этого упражнения? Дайте пояснения.

c) В свете этого является ли простой метод разности различий, в котором используется процент учащихся, прошедших стандартизированные тесты, хорошим способом оценить, достигает ли вознаграждаемое тестирование своей миссии? Объясните свой ответ.



Дополнительное чтение и ссылки

Подробнее об угонах самолетов и металлодетекторах читайте в статье:

Walter Enders and Todd Sandler. 1993. *The Effectiveness of Anti-Terrorism Policies: A Vector-Autoregression-Intervention Analaysis.* American Political Science Review 87 (4): 829–44.

Вы можете узнать больше о промежуточных результатах медицинских исследований в статье:

Thomas Fleming. 1994. *Surrogate Markets in AIDS and Cancer Trials.* Statistics in Medicine 13: 1423–35;

Thomas R. Fleming and David L. DeMets. 1996. *Surrogate End Points in Clinical Trials: Are We Being Misled?* Annals of Internal Medicine 125: 605–13.

Изучение взаимосвязи колебаний температуры и экономического роста:

Marshall Burke, SolomonM. Hsiang, and EdwardMiguel. 2015. *Global Non-Linear Effect of Temperature on Economic Production.* Nature 527 (7577): 235–39.

Сравнение проектов комплексного питания в Тамил Наду и Бангладеш:

HowardWhite and EdoardoMasset. 2007. *Assessing Interventions to Improve Child Nutrition: A Theory-Based Impact Evaluation of the Bangladesh IntegratedNutrition Project.* Journal of International Development 19 (5): 627–52.

Историческая статистика Высшей лиги бейсбола взята из банка бейсбольных данных на seanlahman.com. Статистические данные по бейсболу в средней школе взяты с сайта GameChanger на gc.com.

Если вас интересует история адаптации к налогу на окна в Англии, прочтите следующие работы:

Andrew E. Glantz. 2008. *A Tax on Light and Air: Impact of the Window Duty on Tax Administration and Architecture, 1696–1851.* Penn History Review 15 (2);

Wallace E. Oates and Robert M. Schwab. 2015. *The Window Tax: A Case Study in Excess Burden.* Journal of Economic Perspectives 29 (1): 163–80.

Об истории сдвига в бейсболе вы можете прочитать в блоге:

Travis Sawchik. 2017. *We've Reached Peak Shift.* FanGraphs. November 9. *http://blogs.fangraphs.com/weve-reached-peak-shift/*.

Статистические данные о потоках наркотиков в США взяты из двух отчетов:

United Nations Office on Drugs and Crime. 2010. *The Globalization of Crime: A Transnational Organized Crime Threat Assessment.* Chapter 4. *https://www.unodc.org/documents/lpo-brazil//noticias/2010/06/TOCTA_Report_2010_low_res.pdf*;

Office ofNational DrugControl Policy.October 2001. *The Price of Illicit Drugs: 1981 through the Second Quarter of 2000. https://obamawhitehouse.archives.gov/sites/default/files/ondcp/policy-and-research/bullet_5.pdf*.

Изучение связи между экзаменами по математике и уровнем преподавателей в военно-воздушной академии, обсуждаемое в упражнении 3:

Scott E. Carrell and James E. West. 2010. *Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors.* Journal of Political Economy 118 (3): 409–32.

Обсуждение важных вопросов тестирования и обучения учащихся в упражнении 4 самым непосредственным образом основано на статье:

Derek Neal and DianeWhitmore Schanzenbach. 2010. *Left Behind by Design: Proficiency Counts and Test-Based Accountability.* Review of Economics and Statistics 92 (2): 263–83.

Мы также опирались на исследование:

Bengt Holmstrom and Paul Milgrom. 1991. *Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design.* Journal of Law, Economics, and Organization. 7: 24–52.

Глава 17

О пределах возможностей количественной оценки

О чем эта глава

- Данные и количественные показатели не говорят нам всего, что нужно знать для принятия решений.
- Иногда необходимые количественные показатели неубедительны или отсутствуют, но это не обязательно означает, что правильным решением будет ничего не делать или придерживаться статус-кво.
- Если мы не будем осторожны, количественная оценка может иметь непредвиденные последствия для этики и справедливости.
- Данные не говорят нам, каковы наши цели. Принимая решения, следует помнить как о последствиях наших действий, так и о наших ценностях.

Введение

Возможности количественного анализа вселяют надежды на улучшение нашей жизни и окружающего мира. Но у всего есть пределы. Мы видели много примеров того, как отсутствие критического мышления мешает правильно использовать имеющиеся данные. Если вы ошибочно принимаете корреляцию за причинно-следственную связь, игнорируете возврат к среднему значению или проблему завышения статистической значимости, пытаетесь установить корреляцию без изменений или делаете вид, что данные говорят сами за себя, а не рассматриваете мышление и данные как взаимно дополняющие друг друга, количественная оценка может сбиться с пути и привести вас к принятию худших, а не лучших решений. Именно для того, чтобы избежать этих ловушек и научиться критически воспринимать количественные данные, мы так усердно работали вместе на протяжении всей книги.

В заключение мы хотим немного порассуждать о пределах возможностей количественной оценки и принятия решений на основе фактических данных. Эти ограничения возникают не из-за отсутствия ясного понимания какой-то конкретной части количественного анализа. Речь о том, что, какими бы важными ни были количественные данные, не существует такой вещи, как решение, основанное исключительно на данных. Это верно как минимум по двум причинам.

Во-первых, для многих важных решений достоверные данные ограничены или даже отсутствуют. Но решения все равно приходится принимать. На самом деле даже решение ничего не делать – это тоже решение. Поэтому важно четко понимать, что нам делать, когда мы сталкиваемся с отсутствием данных. Во-вторых, правильное решение никогда не может быть определено только на основе количественного анализа. Анализ призван стать инструментом, используемым для достижения наших целей и ценностей. Но иногда кажется, что хвост виляет собакой – мы подгоняем свои ценности под требования количественной оценки.

Это опасная ошибка, от которой следует спасаться посредством бдительности и критического мышления.

Принятие решений при ограниченных данных

Есть старая притча, которая звучит примерно так. Пьяный мужчина ползает по тротуару в поисках ключей под фонарным столбом. Прохожий спрашивает, что он делает, и мужчина отвечает: «Ищу ключи». Прохожий спрашивает: «Где вы их видели в последний раз?», на что мужчина отвечает: «Думаю, я уронил их в парке через дорогу». Прохожий резонно спрашивает мужчину, почему он ищет под фонарным столбом, если уронил ключи в парке напротив, и мужчина отвечает: «Там темно, я не могу их найти в темноте!»

Как бы банально ни звучало, эта притча иллюстрирует важный момент, связанный с количественной оценкой. Мы ищем там, где свет. Увы, не все можно легко измерить или дать количественную оценку. Ограниченное «пятно света» в мире, управляемом данными, сужает нашу точку зрения, заставляя сосредоточиться только на тех вещах, где доступны количественные свидетельства.

Но такое сужение таит в себе реальные риски. Во-первых, мы можем просто игнорировать критически важные проблемы, потому что не понимаем, как принять решение, основанное на фактических данных. Тот факт, что количественные данные не помогают ответить на вопрос, не означает, что этот вопрос неважен или его можно безопасно игнорировать. Во-вторых, требование доказательств может создать своего рода предвзятость статус-кво, т. е. стремление ничего не менять, если количественных данных недостаточно. Когда кто-то говорит «Нет никаких свидетельств о последствиях этого действия», он может иметь в виду две разные вещи. Это может означать, что множество мощных, хорошо спланированных исследований не выявили никаких эффектов. Но это может также означать, что последствия этих действий никогда раньше не изучались (или даже не рассматривались), поэтому фактически нет никаких доказательств за или против. В первом случае может быть разумно не предпринимать никаких действий. Но во втором случае просто нет доступных свидетельств, на которые может опираться ваше решение. Если есть другие веские причины полагать, что действия имеют смысл, было бы ошибкой придерживаться статус-кво лишь по причине отсутствия данных. Рассмотрим несколько примеров.

Анализ затрат и выгод и экологическое регулирование

Для Управления информации и регулирования правительства США (OIRA) в составе Управления бюджетного контроля (OMB) количественные данные

являются священным аргументом. Количественный анализ затрат и выгод необходим для многих новых нормативов, принимаемых исполнительными органами, и OIRA может, по сути, наложить вето на такие нормы, если они не подкреплены доказательствами.

Как мы уже говорили, не все можно легко измерить количественно. Но без количественных доказательств обращение в OIRA, как правило, обречено на провал. В результате, подобно пьяному мужчине, ищущему ключи под фонарным столбом, регулирующие органы вынуждены сосредоточиться на тех областях, где возможна количественная оценка, независимо от того, являются ли эти области местами, наиболее нуждающимися в их внимании.

Лиза Хайнцерлинг, бывший руководитель отдела политики Агентства по охране окружающей среды (EPA), описывает мрачные выражения, в которых это выразил бывший сотрудник EPA: «Вместо того чтобы спрашивать себя “Будет ли это полезно для защиты окружающей среды?”, мы спрашиваем “Как сделать это приемлемым для OMB?”».

Разумеется, в некотором смысле требования количественной оценки для того и нужны, чтобы сдерживать регулирующие органы от необдуманных инициатив. Предварительный количественный анализ не позволяет Агентству по охране окружающей среды (и другим агентствам) выдвигать неудачные законопроекты, когда, как говорится, лекарство хуже, чем болезнь. Но у этого подхода есть и обратный эффект. Количественные оценки искажают стимулы, сужая наше поле зрения. Критерии количественной аргументации препятствуют работе над нормативными актами, в пользу которых имеются веские аргументы, но для которых невозможно или слишком дорого количественно оценивать затраты и выгоды. Например, в типичном отчете EPA регулирующие органы могут перечислять заболевания, на развитие которых, по их мнению, влияет определенный загрязнитель. Однако эти заболевания будут включены в анализ затрат и выгод только в том случае, если у нас будет способ оценить влияние загрязнителя на риск заболевания и у нас есть количественные оценки денежных затрат, связанных с заболеванием. И если они не будут включены в анализ затрат и выгод, они не будут иметь большого влияния на принятие решений OIRA.

Хорошо известным примером являются разногласия по поводу решения EPA ужесточить нормы содержания мышьяка в воде в начале 2000-х гг. В отчете EPA, обосновывающем необходимость регулирования, перечислен широкий спектр заболеваний, риск возникновения которых, как полагают, увеличивается из-за мышьяка. К ним относятся рак мочевого пузыря, почек, легких, печени и простаты, а также ряд других заболеваний с сердечно-сосудистыми, легочными, иммунологическими, неврологическими и эндокринными нарушениями. Однако EPA отмечает, что из-за отсутствия данных «количественные доказательства», включенные в их анализ, касаются только воздействия мышьяка на «рак мочевого пузыря и легких». Остальные выгоды для здоровья от снижения воздействия мышьяка не поддаются количественной оценке. Надо сказать, что агентство мудро использовало *качественную* информацию для оценки этих более обширных последствий. Но из-за требования количественной оценки их выводы были легко отвергнуты в последовавшей полемике.

Использование зубной нити и ношение маски

Два более близких каждому из нас примера иллюстрируют, почему часто существуют веские причины действовать даже при отсутствии количественных доказательств.

Использование зубной нити

В течение многих лет Энтони каждый день тщательно чистил зубы нитью, потому что его супруга-дантист велела ему это делать и потому что он верил ей, когда она говорила, что использование нити полезно для здоровья зубов. Но затем в 2016 г., действуя во имя принятия решений, основанных на фактических данных, New York Times опубликовала статью под названием «Чувствуете себя виноватым из-за того, что не пользуетесь зубной нитью? Возможно, в этом нет необходимости». Статья предполагала, что прилежные специалисты по чистке зубной нитью, такие как Энтони, могут отказаться от лишних хлопот перед сном.

В статье, о которой идет речь, цитируется метаанализ 12 рандомизированных экспериментов, в которых исследователи сравнивали эффекты чистки зубов щеткой и зубной нитью с простой чисткой зубов щеткой. В статье сообщалось, что исследование «обнаружило только “очень ненадежные” доказательства того, что использование зубной нити может уменьшить зубной налет». Итак, вот оно – отсутствие доказательств выгоды от использования зубной нити[1].

Так почему же Энтони не перестал пользоваться зубной нитью? Одна из причин, как обсуждалось в главе 6, заключается в том, что неспособность отвергнуть нулевую гипотезу не является ее доказательством, т. е. отсутствие доказательств полезности не является убедительным доказательством ее отсутствия. Даже если у нас нет статистически значимых доказательств того, что использование зубной нити уменьшает заболевания зубов, это не означает, что использование зубной нити не дает никакого эффекта. Что, если исследования имеют низкую статистическую мощность из-за небольшого размера выборки или большого количества участников, не выполняющих условия эксперимента? Возможно, они бы не заметили эффекта, даже если использование зубной нити действительно уменьшает зубной налет.

Другая причина заключается в том, что исследователи не изучили все интересующие эффекты. Например, авторы метаанализа отмечают, что ни один из экспериментов не оценивает долгосрочные последствия и не изучает ряд важных последствий для зубов, таких как кариес, зубной камень или воспаление десен.

Но даже эти ограничения количественных исследований – это еще не все. Как и во многих решениях, хотя у нас нет достаточных количественных данных, чтобы ответить на все наши вопросы о влиянии зубной нити, существуют неколичественные аргументы, которые важно учитывать. Стоматологи предоставляют убедительные биологические и механические объяснения того, почему использование зубной нити приносит пользу. Итак, несмотря на обнаруженное журналистами отсутствие экспериментальных данных, мы вполне удовлетворены решением использовать зубную нить, вопреки тому, что коли-

[1] Метаанализ на самом деле содержал статистически значимые доказательства того, что использование зубной нити уменьшает воспаление десен, поэтому экспериментальные доказательства в пользу использования зубной нити, возможно, сильнее, чем предполагается в статье.

чественные доказательства не являются убедительными. Есть веские причины полагать, что использование зубной нити полезно, даже при отсутствии серьезных эмпирических исследований.

Ношение маски

На момент написания этой книги в нашем обществе ведутся аналогичные дебаты о последствиях ношения масок в разгар глобальной пандемии COVID-19. Как и в случае с зубной нитью, существует несколько убедительных и мощных экспериментов, демонстрирующих, что тканевые и фронтальные маски снижают передачу вируса. Есть некоторые обсервационные исследования, в которых просматриваются недостатки, упомянутые в главе 9. Некоторые из этих исследований сосредоточены только на ограниченных выборках людей, которые приходят в медицинские учреждения с симптомами, что, как обсуждалось в главе 16, также создает проблемы. Как и в случае с зубной нитью, когда исследователи пытаются провести рандомизированный эксперимент, многие люди, которым назначено воздействие, не соблюдают условия, что затрудняет оценку эффективности ношения масок. Более того, нам, вероятно, понадобится очень большой размер выборки, чтобы получить достаточно точную оценку эффекта масок или зубной нити.

Учитывая отсутствие точных данных о масках, многие люди, включая Дональда Трампа и Майка Пенса, тогдашнего президента и вице-президента США, решили отказаться от хлопот. Такие скептики иногда приводят аргументы вроде «Нет никаких доказательств того, что ношение маски приносит пользу». Но, как и в случае с зубной нитью, существуют веские теоретические и биологические основания полагать, что маски эффективны. Мы знаем, что коронавирус и многие другие вирусы передаются через дыхательные аэрозольные частицы, и у нас есть убедительные вещественные доказательства того, что маски уменьшают поток некоторых из этих частиц. Исследования также показывают, что люди, носящие маски, реже прикасаются к своим глазам, носу и рту, что является второй причиной, по которой маски, вероятно, уменьшают передачу инфекции.

Конечно, мы не уверены, что знаем правильный ответ, и надеемся, что дальнейшие исследования улучшат наше понимание последствий ношения масок. Но отсутствие количественных доказательств в пользу одного решения не является веской причиной для принятия противоположного решения, особенно если в пользу этого альтернативного решения тоже нет доказательств.

Оказавшись перед необходимостью принять решение, мудрые люди используют количественные данные, но они признают, что количественные данные говорят им не так много. Они не игнорируют определенные соображения только потому, что у нас нет хороших количественных оценок этих факторов. Они используют наилучшую доступную теорию и данные для формирования своих убеждений и принимают наилучшие решения, какие только могут, учитывая свои цели, ценности и потенциально несовершенные и неопределенные убеждения.

Последнее предложение указывает на еще одну важную мысль о принятии решений на основе фактических данных. Независимо от того, насколько хорош анализ данных, одни только факты не могут подсказать вам, как действовать. Для этого также нужно подумать о своих целях и ценностях. Мы завершаем книгу размышлениями о том, как взаимодействуют количественные оценки и эти ценности.

Количественные данные и ценности

Количественные данные должны помогать нам принимать более правильные решения, которые способствуют достижению наших целей и ценностей. Но если мы не будем осторожны, ситуация может измениться на противоположную: наши цели и ценности будут определяться доступностью количественных данных.

Мы собираемся рассказать о двух способах, с помощью которых это может произойти. Во-первых, количественные инструменты иногда могут незаметно для нас внести в процесс принятия решений ценности, с которыми мы не согласны. Во-вторых, стремление соответствовать количественной оценке может подтолкнуть нас к принятию ценностей, которые в противном случае мы могли бы отвергнуть.

Как количественные инструменты крадут наши ценности

Один из рисков количественной оценки, особенно в эпоху, когда машинное обучение и алгоритмическое принятие решений становятся все более распространенными, заключается в том, что нежелательные ценности могут проникнуть в решения незаметно для нас. Например, алгоритм может проявлять расовую или гендерную предвзятость, даже если для создания алгоритма не использовались данные о расе или поле в явном виде. Это поднимает важные вопросы о равенстве, честности и справедливости, которые заслуживают нашего внимания. В современном мире алгоритмы прогнозирующего машинного обучения используются для решения самых разных задач. Веб-сайты по трудоустройству используют такие алгоритмы для подбора соискателей работы и работодателей. Банки используют их для оценки кредитоспособности. Платформы социальных сетей используют их, чтобы решить, какой контент и рекламу предлагать пользователям. А судьи ими пользуются для обоснования приговоров по уголовным делам.

Как эти алгоритмы могут привести к этически сомнительным результатам? Алгоритмы машинного обучения – это в большей или меньшей степени просто причудливые способы использования корреляций для прогнозирования. Алгоритм, который не учитывает расовую или гендерную принадлежность, т. е. не имеет доступа к данным о расе или поле, тем не менее может в конечном итоге делать прогнозы, которые по-разному относятся к людям с разной расовой или гендерной идентичностью. Это может произойти, например, если алгоритм имеет доступ к данным о переменных, которые коррелируют с расой, или если некоторые входные данные алгоритма сами по себе подвержены предвзятости. Мы уже видели пример такого рода проблемы в главе 2, когда обсуждали, как использование корреляции между отзывами Yelp и целевыми проверками нарушений санитарных правил может привести к расовой предвзятости. Но давайте рассмотрим еще один пример.

Алгоритмы и расовые предрассудки в здравоохранении

В Соединенных Штатах крупные поставщики медицинских услуг имеют специальные программы, предназначенные для координации ухода за людьми со сложными медицинскими потребностями. Такие программы стоят дорого.

Поэтому поставщики услуг хотят отбирать только тех людей, которые имеют наибольшие потребности в уходе. Чтобы попытаться предсказать, кто эти пациенты, они используют алгоритмы машинного обучения.

Существует сильная положительная корреляция между медицинскими расходами и потребностями в медицинской помощи, поскольку более больные пациенты, как правило, получают более дорогостоящее лечение. А затраты на здравоохранение легче поддаются количественному измерению, чем потребности здравоохранения. В исследовании, о котором пойдет речь, алгоритму было предложено спрогнозировать затраты на здравоохранение. Для этого, помимо данных о расходах на здравоохранение, ему были предоставлены данные о прошлых страховых выплатах пациентов, медицинских диагнозах и лекарствах. Важно подчеркнуть, что алгоритм специально не получил никакой информации о расовой принадлежности.

Простой способ предположить, как это может работать, – провести аналогию с регрессией. Предположим, у нас есть данные о большом количестве расходов на медицинское обслуживание пациентов в году t, а также об их страховых исках, диагнозах, процедурах и лекарствах в году $t - 1$. Мы могли бы построить следующую регрессию:

$$\text{Стоимость}_t = \beta_0 + \beta_1 \cdot \text{Страховые выплаты}_{-1} + \beta_2 \cdot \text{Диагнозы и процедуры}_{t-1} + \beta_3 \cdot \text{Лекарства}_{t-1}.$$

Это даст нам расчетные коэффициенты OLS от $\hat{\beta}_0$ до $\hat{\beta}_3$.

Когда появляется новый пациент i, мы можем с помощью этого алгоритма предсказать будущие затраты на медицинское обслуживание этого пациента. Мы вводим конкретные значения страховых исков, диагнозов, процедур и лекарств этого нового пациента в наше уравнение регрессии, чтобы получить

$$\text{Прогнозируемые затраты}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{Страховые выплаты}_i + \hat{\beta}_2 \cdot \text{Диагнозы и процедуры}_i + \hat{\beta}_3 \cdot \text{Лекарства}_i.$$

Примерно это и делает алгоритм прогнозирующего машинного обучения, но цель алгоритма обычно несколько отличается от минимизации среднеквадратической ошибки, и он учитывает более сложные функции переменных, чем линейная регрессия. В статье, опубликованной в журнале Science в 2019 г., описывается поставщик медицинских услуг, использующий такие прогнозируемые значения для отбора пациентов. Пациенты, набравшие балл выше некоторого верхнего порога, сразу же включались в специальную программу. Пациенты с оценкой выше некоторого нижнего порога направлялись к врачу для дополнительного обследования.

Несмотря на то что алгоритм прогнозирования не учитывал расовую принадлежность, он систематически недооценивал потребность в медицинской помощи чернокожих пациентов по сравнению с белыми пациентами. Это показано на рис. 17.1. На горизонтальной оси расположены потребности в медицинской помощи в соответствии с прогнозами алгоритма. Показатель активных хронических заболеваний, называемый *оценкой коморбидности*, находится на вертикальной оси. Он должен служить мерой реальных потребностей в меди-

цинской помощи. Как видите, при любом заданном уровне прогнозируемых потребностей в медицинской помощи чернокожие пациенты в среднем считаются менее больными, чем белые. Таким образом, вероятность включения в специальную программу чернокожих пациентов была систематически ниже, чем у белых пациентов с аналогичным состоянием здоровья.

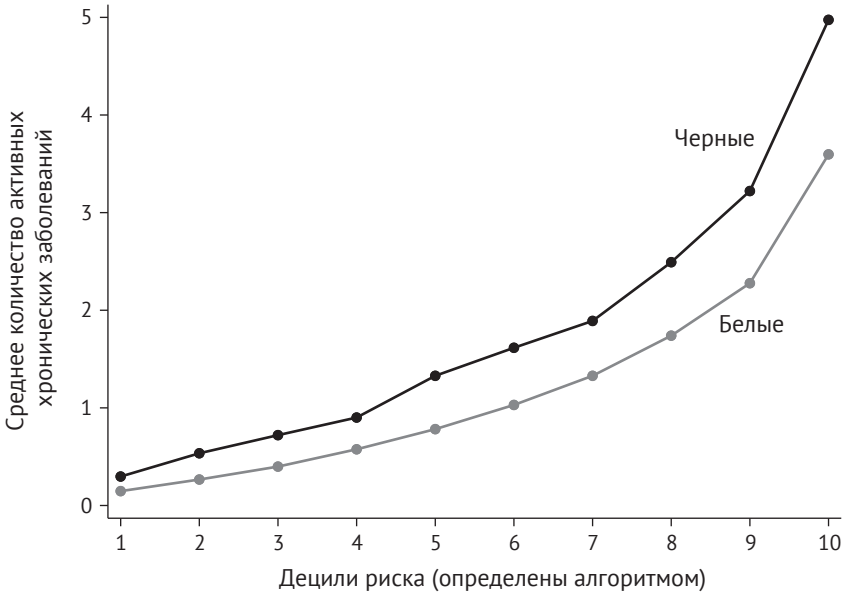


Рис. 17.1. Взаимосвязь между алгоритмическим прогнозом и фактическим здоровьем у чернокожих и белых пациентов различна

По какой причине этот не учитывающий расовую принадлежность алгоритм по-прежнему дает расово предвзятые прогнозы? Одной из возможностей является пропущенная переменная. То есть, возможно, даже при условии наличия одинаковых прошлых страховых выплат, диагнозов, процедур и лекарств чернокожие пациенты, как правило, болеют тяжелее, чем белые, по причинам, не отраженным в данных. Это может привести к тому, что алгоритм будет систематически недооценивать медицинские потребности чернокожих пациентов и переоценивать потребности белых пациентов.

Однако в данном случае, похоже, происходит нечто иное. Поставщик медицинских услуг использовал алгоритм для прогнозирования затрат на здравоохранение, поскольку затраты легко измерить и они тесно связаны с потребностями здравоохранения. Но это решение оказалось проблематичным. Систематическим фактом в системе здравоохранения США является то, что на чернокожих пациентов в среднем тратится меньше денег, чем на одинаково больных белых пациентов[1].

Это означает, что использование затрат в качестве показателя потребностей здравоохранения вносит расовую предвзятость в изначально не учитывающий

[1] В разделе «Дополнительное чтение и ссылки» вы найдете обзорную статью, документирующую множество проявлений предвзятости и дискриминации в отношении чернокожих пациентов в системе здравоохранения США.

расовую принадлежность алгоритм. Алгоритм правильно предсказывает, что затраты на здравоохранение для чернокожего пациента будут ниже, чем для белого пациента с аналогичными характеристиками (выплаты, диагнозы и процедуры, лекарства). И это создает впечатление, что чернокожие пациенты здоровее, чем такие же белые пациенты. Используя те же исходные данные, но переформулировав алгоритм для прогнозирования показателя фактического здоровья, а не затрат на здравоохранение, авторы статьи в журнале Science смогли устранить расовую предвзятость. Этот пример демонстрирует, как использование количественных инструментов может внести в процесс принятия решений критерии, которые противоречат нашей системе ценностей. Поскольку мир становится все более количественным, требуется критическое мышление и постоянная бдительность, чтобы убедиться, что наши решения не просто основаны на данных, но что ценности, лежащие в основе этих решений, совпадают с нашими ценностями.

Это подводит нас к следующей теме: каким образом количественная оценка может навязывать ценности, которые мы склонны принимать за собственные.

Как количественная оценка навязывает нам ценности

Философия морали описывает широкий спектр этических проблем, которые следует учитывать при оценке правильности или неправильности решения. Например, существуют хорошие и убедительные аргументы в пользу различных прав и обязанностей, таких как право распоряжаться своим телом или обязанность не принуждать своих собратьев силой. В современном обществе считается этичным уважать или даже поощрять такие права и обязанности, даже если отказ от них мог бы привести к более высокому общему материальному благополучию. Такую позицию, например, часто занимают принципиальные противники смертной казни, пыток или исследований стволовых клеток.

Существуют также хорошие и убедительные аргументы в пользу заботы не только об общем благополучии, но и о его распределении. Разумный человек, например, может быть готов принять более низкое общее благосостояние в обществе в обмен на большее равенство.

Но большая часть количественного анализа политических решений коренится в *велфаризме* (welfarism) – системе взглядов, согласно которой политику следует оценивать на основе ее последствий для благосостояния человека. Более того, один стандарт велфаризма преобладает над всеми остальными: *утилитаризм* – точка зрения, согласно которой политику следует оценивать на основе ее последствий для общего благосостояния человека, независимо от его распределения. И не просто утилитаризм, а то, что мы могли бы назвать вульгарным утилитаризмом, который определяет благополучие почти полностью с точки зрения материальных затрат и выгод, таких как экономическое процветание, здоровье и другие факторы, которые (относительно) легко измерить количественно, назначив денежную стоимость.

Этическая позиция, согласующаяся с количественной оценкой последствий политики, в принципе, весьма гибка; она не должна быть грубо утилитарной. Мы можем оценить различные нематериальные факторы, такие как права, обязанности, ответственность, достоинство и т. д. Более того, как только вы узнаете количественное влияние политики на благосостояние людей, вы сможете учитывать всевозможные соображения справедливости в оценке политики.

Мы могли бы, например, после количественной оценки всех эффектов определить лучшую политику как ту, которая максимизирует общее благосостояние, при условии, что уровень благосостояния каждого человека превышает некоторый минимальный порог.

Что отличает вульгарный утилитаризм от всех других нормативных рамок – даже от других форм благосостояния, – так это то, что он легко поддается количественному анализу. Трудно понять, как количественно измерить ценность прав и обязанностей или как взвесить соображения справедливости. Гораздо проще – как концептуально, так и практически – дать количественную оценку материальных затрат и выгод, а затем просто складывать и вычитать, чтобы выяснить, насколько хороша та или иная политика.

Действительно, с вульгарным утилитаризмом настолько удобно работать, что он стал частью стандартных предположений, лежащих в основе многих количественных анализов, особенно в дискуссиях о государственной политике. Стремление максимизировать чистое благосостояние – игнорирование вопросов прав, обязанностей, ответственности, справедливости, достоинства и т. д. – настолько укоренилось в нашем поведении и мышлении, что мы даже не замечаем этого. Мы просто считаем само собой разумеющимся, что хорошая политика – это та, которая максимизирует выгоды за вычетом затрат. Задумайтесь о том, что это значит. Мы преследуем не те цели и ценности, которые хотелось бы, а те, которые поддаются количественной оценке. Мы утилитаристы не потому, что таковы наши убеждения. Нас делает утилитаристами стремление работать только с измеримыми величинами.

В чем проблема позволить материалистическому утилитаризму определять наши цели? В качестве частичного ответа мы хотели бы рассказать вам одну историю.

Однажды Итан посетил научную презентацию о последствиях изъятия детей из семей, где они подвергаются жестокому обращению, и помещения их в приемные семьи. Автор презентации обнаружил, что детям из семей, где их подвергали жестокому обращению, в среднем лучше жить в приемных семьях. Более того, выгоды для детей, судя по всему, превышают затраты на содержание в приемных семьях. Поэтому исследователь пришел к выводу, что нам следует забрать детей из этих жестоких семей.

Кажется, это отличный пример данных, ведущих к принятию более эффективных политических решений. Мы можем количественно оценить выгоды для детей и выбрать политику, которая улучшит их благосостояние. Мы даже можем показать, что польза для детей превышает долларовые затраты общества. Так что это выглядит как чистая победа. Невероятно.

Один из присутствующих, занимавший несколько высоких постов в правительстве, высказал возражения. Основная критика заключалась в том, что исследователь не оценил все затраты и выгоды, чтобы дать политические рекомендации. В частности, что, если жестокие родители получают выгоду от того, что оставляют своих детей в семье (и, предположительно, продолжают жестоко обращаться с ними)? Если для них ценность сохранения детей достаточно велика, то не может ли это изменить выводы анализа затрат и выгод?

Вы можете подумать, что разумным ответом на этот вопрос будет что-то вроде «Ну, если бы кто-то был убежденным утилитаристом, это было бы

правильно. Но есть и другие ценности, и лично я считаю, что нас не должно волновать, хотят ли жестокие родители оставить детей при себе. Мы должны сосредоточиться на том, что лучше для детей и остального общества». Но исследователь ответил иначе. Вместо этого он согласился с точкой зрения критика, признав, что он действительно не может сказать, является ли изъятие детей из некомфортной семьи хорошим решением, если не известно, как это повлияет на родителей детей.

Или рассмотрим другой пример. В начале 1990-х гг. главный экономист Всемирного банка Ларри Саммерс – бывший президент Гарварда, главный экономический советник президента Обамы и министр финансов президента Клинтона – распространил меморандум, написанный его сотрудниками. В нем содержалась следующая мысль:

> «Разве Всемирный банк не должен *поощрять* дальнейший перенос грязных отраслей промышленности в малоразвитые страны? … Издержки от вредного для здоровья загрязнения определяются в том числе упущенными доходами из-за роста заболеваемости и смертности. С этой точки зрения определенное количество вредных для здоровья загрязнений должно производиться в стране с наименьшими затратами, то есть в стране с самой низкой заработной платой. Я думаю, что экономическая логика переноса токсичных производств в страну с самой низкой заработной платой безупречна, и мы должны с этим смириться».

Утверждение о том, что отправка токсичных отходов в страны с низкой заработной платой имеет «безупречную экономическую логику», заслуживает отдельного внимания. Вот три утверждения, каждое из которых по отдельности выглядит правильным:

1) очевидно, средняя готовность платить за то, чтобы избавиться от токсичных отходов, в богатых странах выше, чем в бедных;
2) следовательно, перемещение некоторых токсичных отходов из богатых стран в бедные увеличит чистое материальное благосостояние в мире;
3) если это единственные издержки и выгоды (например, если мы не считаем избегание богатыми странами ответственности за свои действия прямыми издержками) и мы утилитаристы, то это хорошая политика.

Но вряд ли можно называть эту цепочку аргументов «экономической логикой», поскольку, по крайней мере, последний шаг не имеет ничего общего с экономикой; он связан с ценностями. А первое заявление о том, что имеет смысл оценивать политику по готовности платить, также порождает тревожные размышления относительно моральных приоритетов. Мы подозреваем, что ценность условного доллара ниже для более богатых людей. Богатые люди имеют более высокую готовность платить, чем бедные, за одно и то же изменение благосостояния просто потому, что они по-разному ценят деньги. Это означает, что, если мы оцениваем затраты и выгоды на основе готовности людей платить, мы неявно предполагаем, что благополучие богатых важнее, чем благополучие бедных.

Несмотря на повсеместное распространение подобных проблем, временами количественные аналитики, кажется, упускают из виду тот факт, что принятие решений путем сравнения показателей материальных затрат и выгод не осво-

бождает нас от необходимости учитывать иные ценности. Майкл Гринстоун, известный экономист в области энергетики и окружающей среды, особенно ясно формулирует этот вопрос, приводя аргументы в пользу использования анализа затрат и выгод для принятия политических решений:

> «Я думаю, что, как только мы откажемся от анализа затрат и выгод, все начнет часто, – не всегда, но часто – перетекать в моральные решения. И глубокая проблема, с моей точки зрения, в принятии морально обоснованных решений по многим из этих вопросов заключается в том, что ваша мораль не является моей моралью, а мораль третьего человека отличается от нашей морали. И тогда у нас не останется рамок для принятия решений… У меня нет уверенности, что результат будет хорошим для всего общества».

Гринстоун делает важный вывод. Количественно оценивая затраты и выгоды, мы ограничиваем процесс принятия решений формальными рамками. Но он заходит слишком далеко, когда предполагает, что анализ затрат и выгод исключает из уравнения субъективные моральные мнения или что существует какой-то объективный научный способ судить о том, что хорошо для общества, без предварительного формирования набора оценочных суждений, которые не могут быть определены на основе данных.

Разумеется, мы можем назвать множество причин, по которым перемещение токсичных предприятий из богатых стран в бедные, даже если оно пройдет тест на рентабельность, не выглядит хорошей политикой. Возможно, мы ценим справедливость и экономическую мобильность, поэтому не думаем, что отправлять токсичные отходы в беднейшие страны – это хорошая идея. Возможно, мы думаем, что богатые страны должны взять на себя ответственность за свои действия. Возможно, мы не хотим жить в таком мире, где богатые люди могут просто платить за право не подвергаться воздействию загрязнения, являющегося побочным продуктом экономической деятельности, от которой они получили выгоду. Возможно, мы уверены, что меньшая готовность бедных людей платить за избавление от токсичных отходов означает не то, что их жизнь менее ценна, чем жизнь богатых людей, а то, что деньги – это неправильный способ измерения ценности. Точка зрения Гринстоуна отчасти верна: разумные люди могут не соглашаться со всеми этими моральными суждениями. Но разумные люди могут также не согласиться с тем, оправдывает ли максимизация выгод перенос токсичных производств из богатых стран в бедные. Хотя затраты и выгоды, измеряемые с точки зрения готовности платить, относительно легко поддаются количественному измерению, в то время как некоторые другие ценности плохо поддаются количественной оценке, нельзя думать, что первое является объективной наукой, а второе относится к субъективным суждениям. Здесь и то и другое связано с оценочными суждениями.

Чтобы внести ясность: мы не хотим сказать, что в пользу взглядов, выраженных в меморандуме Саммерса, нет никаких аргументов. Предположим, Саммерс прав в том, что перенос токсичных предприятий из богатых стран в бедные увеличит чистое материальное благосостояние. Тогда богатые смогут с избытком компенсировать бедным вред от токсичных отходов, в результате чего обе стороны будут в выигрыше от сделки. Следовательно, если у нас есть технологические возможности и политическая воля, чтобы заставить бедных

принимать токсичные отходы, а богатых платить им за это, мы могли бы создать беспроигрышную ситуацию.

Но, конечно, есть и множество других этических аргументов, которые заслуживают внимания. На наш взгляд, должно иметь значение то, как и кем будет принято это решение. Бедные страны, соглашающиеся принять токсичные отходы в обмен на компенсацию, могут относиться к этому совсем не так, как экономист в богатой стране, принимающий решения и указывающий бедной стране, что им лучше. Но в меморандуме Саммерса нет ни тени беспокойства по поводу того, получат ли бедные страны компенсацию или согласятся ли они на такую сделку. Судя по всему, ему достаточно грубого утилитарного аргумента. К его чести, в более поздних обсуждениях меморандума о токсичных отходах Саммерс выразил другую точку зрения. Например, в интервью журналу New Yorker в 1998 г. он сказал: «Нельзя считать удачной идеей прямолинейное заявление о том, что отправлять токсичные отходы в бедные страны – это хорошо. Существуют ли реальные проблемы, связанные с компромиссом между экономическим ростом и окружающей средой? Конечно. Но то, в какой форме была выражена эта мысль, совершенно неприемлемо».

Случаи жестокого обращения с детьми и переноса токсичных производств интересны по нескольким причинам. Количественная оценка часто подталкивает нас к вульгарному утилитаризму, который может привести к безжалостным и абсурдным выводам. Но дисциплина количественной оценки действительно нас чему-то учит. Для многих людей идея отправки токсичных отходов из богатых стран в бедные не заслуживает даже обсуждения с точки зрения моральных ценностей. Но количественная оценка и сравнение затрат и выгод заставляют нас увидеть серьезные аргументы в пользу этой политики (по крайней мере, той ее версии, которая предполагает согласие и компенсацию), даже если в конечном итоге некоторые из нас придут к противоположному мнению.

В обоих случаях мы считаем, что можем (и должны) использовать некоторые преимущества количественной оценки – точность, взвешивание компромиссов, состязательность. Но оба случая также иллюстрируют ключевую проблему. Права, человеческое достоинство и справедливость трудно измерить количественно. Материальные затраты и выгоды измеряются проще. На практике стремление к количественным аргументам подталкивает нас к тому, чтобы сосредоточить внимание на весьма предосудительном, грубом, материалистическом утилитаризме, который характеризует эти истории. Если мы хотим использовать количественный анализ во благо, то должны стремиться к тому, чтобы данные и количественные инструменты помогали нам оценивать важные критерии, не искажая цели и ценности, в соответствии с которыми мы делаем окончательный выбор.

Научитесь мыслить критически и помогите научиться другим

В заключение мы хотели бы призвать вас использовать инструменты и навыки, которые вы изучили, только во благо. Большая часть этой книги была посвящена тому, как критическое мышление помогает обнаружить, когда кто-то намеренно или случайно вводит вас в заблуждение с помощью данных.

Но циничный читатель может перевернуть эту благородную миссию с ног на голову, используя полученные знания как рецепт для обмана тех, кто не научился критически относиться к данным и аргументам. Если эта книга не станет мировым бестселлером, обязательным для прочтения, то большинство людей, с которыми вы общаетесь, не заметят подмену причинно-следственной связи корреляцией, о которой вы знаете, что она ошибочна, или если вы будете подгонять условия эксперимента, пока не получите желаемый результат, а затем сообщите только о нем. Пожалуйста, не делайте этого! Подумайте о ценности поиска истины и серьезно отнеситесь к своей вновь обретенной ответственности опытного количественного аналитика. Будьте откровенны в отношении сильных и слабых сторон свидетельств, которые вы приводите, будь то выводы, которые вы получили в результате собственного анализа или нашли в чужих исследованиях.

Но самое главное – найдите минутку, чтобы оценить по достоинству, как усердно вы работали над этой книгой и как далеко мы продвинулись вместе. Теперь вы являетесь членом небольшой, но растущей группы людей, которые могут ясно и непредвзято думать о проблеме выбора зависимой переменной, о разнице между статистической и существенной значимостью, возврате к среднему значению, предвзятости публикации, источниках «космического привыкания», взаимосвязи между корреляцией и причинностью, планах исследования и многом другом. Это фундаментальные навыки, которые будут служить вам всегда, даже если вы больше никогда не построите ни одну регрессию. Ведь мы живем в такое время, когда критическое мышление в отношении данных абсолютно необходимо всем, кто хочет понять мир и сделать его лучше.

Упражнения

17.1. Ваш друг Энди заметил, что каждый раз, когда он ест блины на завтрак, он хорошо сдает экзамены. Поэтому он решил, что с этого момента его рацион будет полностью состоять из блинов. На основе уроков, извлеченных из всей книги, назовите как минимум четыре ошибки в рассуждениях вашего друга.

17.2. Вы являетесь мэром крупного города, и ваши сотрудники представляют вам план по обеспечению большей комфортности проживания в районах с низкими доходами. Они говорят вам, что реализация их плана будет стоить 100 млн долл., но, по их оценкам, принесет 200 млн долл. экономической выгоды, так что экономическое обоснование безупречное. Какие вопросы вы хотели бы задать своим сотрудникам, прежде чем принять решение о реализации плана?

17.3. Подумайте о решении в своей жизни, которое вы приняли в основном без помощи количественных доказательств (например, Энтони решил продолжить пользоваться зубной нитью, несмотря на то что убедительных количественных исследований по этой теме было мало). Какие факторы побудили вас принять такое решение? Можете ли вы предложить количественное исследование, которое предоставит более убедительные доказательства? Как должны выглядеть доказательства, чтобы вы могли изменить свое решение?

Дополнительное чтение и ссылки

Вы можете прочитать отчет Агентства по охране окружающей среды об ограничении содержания мышьяка, включая длинный список не поддающихся количественной оценке последствий для здоровья, в федеральном реестре по адресу *https://www.govinfo.gov/content/pkg/FR-2001-01-22/pdf/01-1668.pdf*.

Метаанализ эффекта от использования зубной нити:

Dario Sambunjak, JasonW. Nickerson, Tina Poklepovic, Trevor M. Johnson, Pauline Imai, Peter Tugwell, and Helen V.Worthington. 2011. *Flossing for the Management of Periodontal Diseases and Dental Caries in Adults.* Cochrane Database of Systemic Reviews, Issue 12. doi.org/10.1002/14651858.CD008829.pub2.

Эта исследовательская группа обновила метаанализ, добавив еще три эксперимента по использованию зубной нити в 2019 г.:

HelenV.Worthington, LauraMacDonald, Tina Poklepovic Pericic,Dario Sambunjak, Trevor M. Johnson, Pauline Imai, and Janet E.Clarkson. 2019. *HomeUse of Interdental Cleaning Devices, in Addition to Toothbrushing, for Preventing and Controlling Periodontal Diseases and Dental Caries.* Cochrane Database of Systemic Reviews, Issue 4. doi.org/10.1002/14651858.CD012018.pub2.

Некоторые эмпирические данные о ношении масок и распространении аэрозольных частиц:

Sima Asadi, Christopher D. Cappa, Santiago Barreda, Anthony S. Wexler, Nicole M.

Bouvier, and William D. Ristenparth. 2020. *Efficacy ofMasks and Face Coverings in Controlling Outward Aerosol Particle Emission from Expiratory Activities.* Scientific Reports 10, Article 15665. doi.org/10.1038/s41598-020-72798-7.

Доказательства связи между ношением маски и частотой прикосновений к лицу:

Yong-Jian Chen,GangQin, Jie Chen, Jian-Liang Xu,Ding-Yun Feng, Xiang-YuanWu, and Xing Li. 2020. *Comparison of Face-Touching Behaviors Before and During the Coronavirus Disease 2019 Pandemic.* JAMA Network Open 3 (7);

Tiffany L. Lucas, Rachel Mustain, and Robert E. Goldsby. *Frequency of Face Touching with and without aMask in Pediatric Hematology/Oncology Health Care Professionals.* Pediatric Blood & Cancer 67 (9).

Анализ того, как раса связана с различными показателями здоровья в Соединенных Штатах:

DavidCutler,Adriana Lleras-Muney, and TomVogl. 2011. *Socioeconomic Status and Health: Dimensions and Mechanisms.* The Oxford Handbook of Health Economics, Sherry Glied and Peter C. Smith, eds. Oxford University Press.

Мы ссылались на это исследование при обсуждении того, как количественные инструменты могут проникать в значения:

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. *Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations.* Science 366 (6464): 447–53.

Вы можете найти цитату Саммерса и обсуждение памятки о токсичных отходах здесь:

John Cassid. 1998. *The Triumphalist.* The New Yorker. July 6.

Цитата Майкла Гринстоуна взята из этого подкаста:

The Value of a Life (episode 1). Pandemic Economics. Becker Friedman Institute. April 23, 2020. bfi.uchicago.edu/podcast/pandemic-economics-01.

Анализ того, как количественная оценка может формировать наши моральные ценности, во многом опирается на работу:

Ethan Bueno de Mesquita. 2019. *The Perils of Quantification*. The Boston Review. March 11. *https://bostonreview.net/forum/economics-after-neoliberalism/ethan-bueno-de-mesquita-perils-quantification*.

Предметный указатель

Книги издательства «ДМК Пресс» можно заказать
в торгово-издательском холдинге «КТК Галактика» наложенным платежом,
выслав открытку или письмо по почтовому адресу:
115487, г. Москва, пр. Андропова д. 38 оф. 10.
При оформлении заказа следует указать адрес (полностью),
по которому должны быть высланы книги;
фамилию, имя и отчество получателя.
Желательно также указать свой телефон и электронный адрес.
Эти книги вы можете заказать и в интернет-магазине: **www.galaktika-dmk.com**.
Оптовые закупки: тел. **(499) 782-38-89**.
Электронный адрес: **books@alians-kniga.ru**.

Итан Буэно де Мескита
Энтони Фаулер

Статистика без подвоха

Главный редактор *Мовчан Д. А.*
dmkpress@gmail.com
Зам. главного редактора *Сенченкова Е. А.*
Перевод *Яценков В. С.*
Корректор *Абросимова Л. А.*
Верстка *Луценко С. В.*
Дизайн обложки *Мовчан А. Г.*

Формат 70×100 1/16.
Гарнитура «PT Serif». Печать цифровая.
Усл. печ. л. 36,89. Тираж 100 экз.

Веб-сайт издательства: *www.dmkpress.com*

Увлекательное введение в науку о данных, в котором упор делается на критическое мышление, а не на статистические методы

Введение в науку о данных или статистику не должно начинаться с доказательства сложных теорем или запоминания терминов и формул, но именно так устроены многие учебники по количественному анализу. В отличие от них эта книга посвящена критическому мышлению и концептуальному пониманию; она учит читателей быть вдумчивыми потребителями и аналитиками тех видов информации и аргументов, с которыми они будут сталкиваться на протяжении всей своей жизни.

Помимо прочего, книга учит определять:

- отражают ли наблюдаемые отношения в данных подлинные отношения в мире, и если да, то являются ли они причинно-следственными;
- какие вопросы задавать тем, кто приводит аргументы, используя количественные данные;
- какая статистика особенно информативна или вводит в заблуждение;
- как количественные данные должны и не должны влиять на принятие решений.

Книга, наполненная реальными примерами, показывает, как инструменты критического анализа применяются к проблемам в самых разных областях, включая выборы, гражданские конфликты, преступность, терроризм, финансовые кризисы, здравоохранение, спорт, музыку и космические путешествия.

Прочитав эту книгу, вы узнаете, почему, несмотря на обширное применение данных в современном мире, они никогда не смогут заменить критическое мышление.

- Идеальное руководство для вводных курсов по количественным методам в науке о данных, статистике, политологии, экономике, психологии, социологии и других областях, основанных на данных.
- Знакомит с базовым набором инструментов для анализа данных, включая выборку, проверку гипотез, байесовский вывод, регрессию, эксперименты, инструментальные переменные и разрывную регрессию.
- Использует реальные примеры и данные из разных предметных областей.
- Содержит практические примеры и упражнения с данными.

Интернет-магазин:
www.dmkpress.com

Оптовая продажа:
КТК "Галактика"
books@alians-kniga.ru

издательство

www.дмк.рф

ISBN 978-5-93700-240-2