ГЛАВА 1

Знакомство с причинно-следственным анализом

В первой главе я представлю многие фундаментальные концепции причинно-следственного анализа, а также характерные проблемы и возможности его практического применения. Вы узнаете многие термины, которые будут встречаться в остальной части книги. Кроме того, я хочу, чтобы вы всегда помнили, для чего нужен причинно-следственный анализ и что можно сделать с его помощью. В этой главе речь пойдет не о программировании, а об очень важных базовых концепциях причинно-следственного анализа.

Что такое причинно-следственный анализ

Многие представляют собой причинно-следственные связи как опасную эпистемологическую область, от которой лучше держаться подальше. Возможно, ваш преподаватель статистики неоднократно повторял, что «корреляция — это не причинно-следственная связь» (или «корреляция — это не каузация»), и если вы будете путать эти два понятия, то вас подвергнут академическом остракизму или по крайней мере будут на вас косо смотреть. Но есть один нюанс: *иногда корреляция является причинно-следственной связью*.

Нам, людям, это отлично известно, потому что мы явно склонны принимать корреляцию за причинно-следственную связь. Когда вы решаете не пить четвертый бокал вина, вы совершаете правильный логический вывод, что это приведет к тяжелым последствиям на следующий день. Вы руководствуетесь прошлым опытом: теми вечерами, в которых вы выпили слишком много и наутро проснулись с головной болью; теми вечерами, когда вы выпили всего один бокал или вообще ни одного и ничего не произошло. Вы узнали, что между выпивкой и похмельем существует нечто большее, чем корреляция. Вы сделали вывод о существовании причинности.

С другой стороны, в предупреждениях вашего преподавателя статистики есть доля истины. Причинность — скользкая штука. В детстве я дважды ел кальмара в кляре, и оба раза это кончилось очень плохо, что привело меня к выводу, что у меня аллергия на кальмаров (а также мидий, осьминогов и других морских беспозвоночных). Я осмелился попробовать это блюдо снова только спустя более 20 лет. На этот раз оно было не только вкусным, но и не принесло никакого вреда. В этом случае я спутал корреляцию с причинно-следственной связью. Ошибка была безвредной, так как всего лишь на несколько лет лишила меня лакомого блюда, но неумение различать эти два понятия может иметь куда более серьезные последствия. Если вы занимаетесь инвестициями, то, вероятно, уже оказывались в ситуации, когда вы решали вложить деньги незадолго до резкого повышения цены или вывести их перед самым обвалом. И возможно, у вас появилась мысль, что вы можете предчувствовать колебания рынка. Если вам удалось побороть это чувство — мои поздравления. Но многие попадаются на удочку и думают, что их интуиция причинно связана с хаотичным изменением курсов акций. Иногда эта вера заставляет вас раз за разом повышать ставки и в конце концов ведет к краху.

В двух словах корреляция означает, что две величины или случайные переменные изменяются одинаково, тогда как причинность (обусловленность) означает, что изменение одной переменной приводит к изменению другой. Например, можно установить корреляцию между количеством нобелевских лауреатов в стране и потреблением шоколада на душу населения, но даже если эти переменные перемещаются одинаково, было бы нелепо считать, что изменение одной становится причиной для изменения другой. Легко понять, почему корреляция не всегда подразумевает причинно-следственную связь, но и уравнивать эти два понятия не следует. Причинно-следственный анализ *представляет собой научный метод выведения причинно-следственной связи из корреляции и понимания того, когда и почему они различаются*.

Для чего нужен причинно-следственный анализ

Причинно-следственный анализ можно выполнять просто для лучшего понимания реальности. Тем не менее в нем часто присутствует нормативный компонент. Для чего вы делаете вывод, что избыток алкоголя приводит к головной боли? Чтобы изменить свои привычки и избежать неприятностей. Компания, на которую вы работаете, хочет знать, приводят ли расходы на маркетинг к росту доходов, потому что эта информация поможет повысить прибыль. Вообще говоря, *необходимо знать причинно-следственные связи, чтобы влиять на причины для достижения желаемого эффекта*. Если перенести причинно-следственный анализ в практическую область, то его основной сферой применения станет теория принятия решений.

Так как эта книга ориентирована на практику, в ней рассматривается часть причинно-следственного анализа, направленная на понимание последствий вмешательства. Что произойдет, если вы установите на товар новую цену вместо текущей? Что произойдет, если вы перейдете с диеты с пониженным содержанием сахара, на которой вы сейчас сидите, на диету с пониженным содержанием жиров? Что произойдет с прибыльностью банка, если он повысит кредитный лимит для клиентов? Следует ли правительству раздавать планшеты детям в школе, чтобы повысить их успеваемость, или лучше строить традиционные библиотеки? Улучшит ли брак ваше финансовое положение или женатые пары богаче только потому, что богатому проще найти партнера? Все эти вопросы имеют практическую природу. Они происходят от желания изменить что-либо в бизнесе или личной жизни, чтобы добиться лучших результатов.

Машинное обучение и причинно-следственный анализ

Внимательнее присмотревшись к вопросам, на которые вы хотите получить ответы с помощью причинно-следственного анализа, вы увидите, что они в основном относятся к типу «что если». Как ни прискорбно, машинное обучение (МО) очень плохо подходит для подобных вопросов.

МО очень хорошо отвечает на прогнозные вопросы. Как отмечают Аджай Агравал (Ajay Agrawal), Джошуа Ганс (Joshua Gans) и Ави Голдфарб (Avi Goldfarb) в своей книге «Prediction Machines» (Harvard Business Review Press)[1], «новая волна искусственного интеллекта предоставляет нам не интеллект, а важнейший его компонент — прогнозирование». С машинным обучением можно проделать много всего эффектного. Единственное условие — сформулировать задачу в виде прогноза. Нужен перевод с английского на португальский? Постройте модель МО, которая прогнозирует предложения на португальском для заданных английских фраз. Нужно распознать лица? Создайте модель МО, прогнозирующую присутствие лица в части изображения.

И все же МО не панацея. Оно может творить чудеса в жестких границах и при этом давать чудовищные сбои, если данные немного отклоняются от привычных моделей. Процитируем еще один фрагмент из «Prediction Machines»: «Во многих отраслях низкие цены коррелируют с низкими продажами. Например, в гостиничном бизнесе цены остаются низкими в несезон, но становятся высокими при высоком спросе и заполненных отелях. Если руководствоваться этими данными,

[1] Агравал А., Ганс Д., Голдфарб А. «Искусственный интеллект на службе бизнеса. Как машинное прогнозирование помогает принимать решения».

наивный прогноз может предположить, что повышение цены приводит к повышению количества бронирований».

Машинное обучение использует корреляции между переменными для прогнозирования одной переменной на основании другой. Оно замечательно работает при условии, что переменные, используемые моделью для построения прогнозов, не изменяются. Однако это полностью противоречит цели применения прогнозного МО для принятия решений, подразумевающих вмешательство.

Тот факт, что многие специалисты по работе с данными хорошо знают МО и почти ничего — о причинно-следственном анализе, привел к частым применениям моделей МО, плохо подходящих для своих задач. Одна из главных целей компаний — *повышение* продаж или эффективности использования. Да, модель МО, которая только прогнозирует продажи, нередко оказывается бесполезной — если не вредной — для этой цели. Модель даже может привести к совершенно бессмысленным выводам, как в примере, где большие объемы продаж коррелировали с высокими ценами. Но вы не представляете, как много компаний применяют прогнозные модели МО, при том что интересующая их цель не имеет никакого отношения к прогнозам.

Это не значит, что МО полностью бесполезно для причинно-следственного анализа. Это значит лишь то, что при наивном применении оно часто приносит больше вреда, чем пользы. Но если взглянуть на МО под другим углом — как на инструментарий для построения мощных моделей, а не как на чисто прогнозный механизм, — вы начнете видеть, как они связаны с целями причинно-следственного анализа. В части III я покажу, на что следует обращать внимание при сочетании МО и причинно-следственного анализа и как адаптировать популярные алгоритмы МО (такие, как деревья принятия решений и градиентный бустинг) для его проведения.

Корреляция и причинно-следственная связь

На интуитивном уровне примерно понятно, почему корреляция не является причинно-следственной связью. Если кто-то скажет вам, что эффективность вашего бизнеса повысилась благодаря услугам первоклассного консалтингового агентства, скорее всего, он наткнется на ваш недоуменный взгляд. Как узнать, помог ли консалтинг повысить эффективность бизнеса или процветающее предприятие просто может позволить себе оплачивать услуги таких крутых агенств?

Чтобы сделать пример более осязаемым, представьте, что вы работаете в компании-маркетплейсе. Малый и средний бизнес рекламирует и продает на вашей платформе свои товары. Ваши клиенты полностью автономны в таких вопросах, как назначение цен и выбор времени продаж. Однако вы заинтересованы

в том, чтобы их дела шли как можно лучше. Вы решаете помочь своим клиентам и предоставить им рекомендации о том, как и когда запускать распродажу. Первое, что для этого необходимо знать — *как снижение цен влияет на количество проданных единиц товара*. Если выгода от увеличения продаж компенсирует потери от снижения цены, распродажа имеет смысл. Если вы еще не поняли, это вопрос причинности. Вам необходимо ответить на вопрос, сколько дополнительных единиц товара продаст клиент по сравнению с тем, как если бы цена на товар осталась прежней.

Не стоит и говорить, что это сложный вопрос; возможно, даже слишком сложный для самого начала книги. На вашей платформе работают компании разного профиля. Одни продают продукты питания; другие — одежду. Третьи — удобрения и сельскохозяйственную продукцию. Снижение цены может иметь разный эффект в зависимости от профиля компании. Например, компании, продающей одежду, разумно объявить о снижении цен за неделю до большого праздника. С другой стороны, для аграрного бизнеса такое снижение цен вряд ли приведет к заметным последствиям. Немного упростим задачу и ограничимся одним профилем: продажей игрушек. Также сосредоточимся на одном периоде времени: декабре перед Рождеством. Для начала просто попытаемся ответить на вопрос, как снижение цен в этот период отражается на объеме продаж. Полученная информация будет передана продавцам игрушек, чтобы они могли принять более взвешенные решения.

Чтобы решить, стоит ли проводить распродажу, можно воспользоваться информацией от нескольких продавцов игрушек. Эти данные доступны в датафрейме pandas. Несколько первых строк данных помогают представить, с чем вы имеете дело:

	store	weeks_to_xmas	avg_week_sales	is_on_sale	weekly_amount_sold
0	1	3	12.98	1	219.60
1	1	2	12.98	1	184,70
2	1	1	12.98	1	145.75
3	1	0	12.98	0	102,45
4	2	3	19.92	0	103.22
5	2	2	19.92	0	53.73

В первом столбце хранится уникальный идентификатор магазина. Вам доступны данные по неделям для каждого магазина (`store`) в декабре. Также имеется информация о размере каждого магазина, выраженная в среднем количестве товаров, продаваемых в нем за неделю, рассчитанном за текущий год (`avg_week_sales`). В логическом столбце `is_on_sale` (0 или 1) помечено наличие распродажи в указанный период. Последний столбец, `weekly_amount_sold`, содержит еженедельные продажи по магазину.

Объект анализа (единица анализа)

Объектом анализа в теории причинно-следственного анализа обычно является объект, на который требуется воздействовать. В большинстве случаев объектом анализа будут люди (например, когда нужно узнать, как запуск нового продукта отражается на удержании клиентов). Тем не менее нередко встречаются и другие виды объектов анализа. Так, в примере этой главы объектом анализа является компания. В этом же примере можно попытаться определить, когда лучше всего проводить продажи; в таком случае объектом анализа будет период времени (в данном случае неделя).

Воздействие и результат

Имея данные для анализа, можно переходить к техническим деталям. Обозначим T_i флаг воздействия для объекта i:

$$T_i = \begin{cases} 1 & \text{если объект } i \text{ получил воздействие} \\ 0 & \text{в противном случае.} \end{cases}$$

В данном случае «воздействие» — общий термин, обозначающий некоторое вмешательство, эффект которого нужно узнать. В нашем примере воздействием является снижение цены в одном из магазинов на вашей торговой платформе, представленное столбцом `is_on_sale`.

Обозначение воздействия

В некоторых текстах, а также далее в книге воздействие иногда будет обозначаться буквой D вместо T. Обозначение D позволяет избежать путаницы при включении фактора времени в задачи причинности.

Кроме того, я буду называть `weekly_amount_sold` (переменную, на которую необходимо влиять) *результатом*. Результат для объекта будет обозначаться Y_i. С этими двумя новыми обозначениями цель причинно-следственного анализа можно переформулировать как процесс изучения влияния T на Y. В нашем примере это равнозначно определению влияния `is_on_sale` на `weekly_amount_sold`.

Фундаментальная проблема причинно-следственного анализа

Мы подошли к интересному вопросу. *Фундаментальная проблема причинно-следственного анализа* заключается в том, что невозможно наблюдать за одним и тем же объектом с воздействием и без него. Вы словно оказываетесь на распутье и знаете только то, куда вас приведет выбранная дорога. Чтобы в полной мере понять суть проблемы, вернемся к примеру и построим диаграмму зависимости результата от воздействия, то есть `weekly_amount_sold` от `is_on_sale`. Сразу же

становится видно, что магазины, снизившие цены, продают намного больше (рис. 1.1).

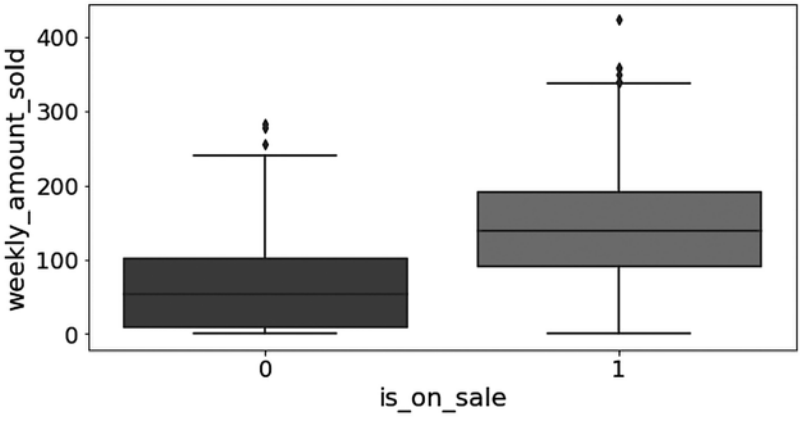

Рис. 1.1. Объем продаж за неделю во время распродажи (1) и без распродажи (0)

Этот факт совпадает с нашими интуитивными представлениями о мире: люди больше покупают по низкой цене, а распродажа (обычно) означает снижение цены. И это хорошо, так как причинно-следственный анализ идет рука об руку с экспертными знаниями. Однако не стоит увлекаться. Вероятно, это тот случай, когда скидки и распространение информации о них побуждают людей больше покупать. Но насколько больше? Из диаграммы на рис. 1.1 можно сделать вывод, что объем продаж увеличивается в среднем приблизительно на 150 единиц на время распродажи (по сравнению с обычной ценой). Оценка получается подозрительно высокой, так как без распродажи объем продаж лежит в диапазоне от 0 до 50. Немного поразмыслив, мы начинаем понимать, что здесь корреляция может ошибочно приниматься за причинно-следственную связь. Возможно, только крупные компании — на долю которых и так приходится наибольший объем продаж — могут позволить себе агрессивное снижение цены. А может, компании проводят распродажи ближе к Рождеству, когда покупатели и так тратят больше денег.

Суть в том, что можно с уверенностью определить настоящий эффект снижения цен на количество проданных единиц, только если понаблюдать за исследуемой компанией (объектом) одновременно с распродажей и без нее. Только сравнив две взаимоисключающие ситуации, можно понять последствия снижения цены. Но, как уже говорилось, фундаментальная проблема причинно-следственного анализа заключается как раз в том, что это попросту невозможно. Приходится придумывать другие способы.

Причинные модели

Все подобные задачи можно решать на интуитивном уровне, но если вы хотите выйти за пределы простой интуиции, вам понадобится формальная запись — своего рода повседневный язык, на котором можно рассуждать о причинности. Считайте, что это универсальное средство общения тех, кто занимается причинно-следственным анализом.

Причинная модель (causal model) представляет собой серию механизмов присваивания, обозначаемых стрелкой ←. В таких механизмах я буду использовать символ u для обозначения переменных за пределами модели, имея в виду, что я не делаю никаких утверждений о том, как они генерируются. Все остальные переменные представляют для нас интерес, поэтому включаются в модель. Наконец, имеются функции f, отображающие одну переменную на другую. Для примера возьмем следующую причинную модель:

$$T \leftarrow f_t(u_t)$$
$$Y \leftarrow f_y(T, u_y).$$

В первой формуле я утверждаю, что u_t — набор переменных, которые я явно не синтезирую (такие переменные еще называются *экзогенными*), — оказывает воздействие T через функцию f_t. Затем T вместе с другим набором переменных u_y (которые я тоже решил не синтезировать) совместно приводят к результату Y через функцию f_y. Присутствие u_y в последней формуле говорит о том, что результат определяется не только воздействием. В нем также отражаются другие переменные, даже если я решил их не синтезировать. В примере с продажами это будет означать, что результат `weekly_amount_sold` обусловлен воздействием `is_on_sale` и другими факторами, не указанными в этой модели и представленными u. Присутствие u учитывает все вариации в переменных, которые еще не учтены переменными, включенными в модель, — такие переменные также называются *эндогенными*. В нашем примере можно сказать, что падение продаж вызвано факторами, не входящими в модель, — это может быть размер компании или что-то еще:

$$IsOnSales \leftarrow f_t(u_t)$$
$$AmountSold \leftarrow f_y(IsOnSales, u_y).$$

Я использую ← вместо =, чтобы явно указать на необратимость причинности. Со знаком равенства $Y = T + X$ эквивалентно $T = Y - X$, но я вовсе не утверждаю, что «T является причиной Y» эквивалентно «Y является причиной T». Как бы то ни было, я довольно часто избегаю такой записи, потому что она немного громоздка. Просто необходимо помнить, что из-за необратимости причин

и следствий при работе с причинными моделями, в отличие от традиционной алгебры, нельзя просто расставить переменные вокруг знака равенства, как вам заблагорассудится.

Если вы хотите явно синтезировать больше переменных, вынесите их из u и встройте в модель. Помните, я говорил, что большое различие в продажах со скидкой и без скидки может быть связано с тем, что крупные компании могут позволить себе более агрессивные продажи? В предыдущей модели размер компании (*BusinessSize*) не включался явно. Его влияние вынесено вовне, тогда как все остальное осталось в u. Однако его можно синтезировать и явно:

$$BusinessSize \leftarrow f_s(u_s)$$

$$IsOnSales \leftarrow f_t(BusinessSize, u_t)$$

$$AmountSold \leftarrow f_y(IsOnSales, BusinessSize, u_y).$$

Чтобы включить дополнительную эндогенную переменную, я сначала добавляю еще одну формулу, которая показывает, откуда берется эта переменная. Затем выношу *BusinessSize* из u_t. Это означает, что размер компании уже не рассматривается как переменная, находящаяся за пределами модели. Я явно утверждаю, что *BusinessSize* является причиной *IsOnSales* (вместе с некоторыми другими внешними факторами, которые я сознательно решил не синтезировать). Это просто формальный способ кодирования предположения о том, что крупные компании чаще снижают цены. Наконец, я также добавил *BusinessSize* в последнюю формулу. Тем самым я кодирую наше предположение о том, что более крупные компании также больше продают. Другими словами, *BusinessSize* является общей причиной как для воздействия *IsOnSales*, так и для результата *AmountSold*.

Скорее всего, такой способ синтезирования для вас в новинку, поэтому полезно связать его с чем-то знакомым. Читателю из мира экономики или статистики может быть привычнее другой способ синтезирования той же задачи:

$$AmountSold_i = \alpha + \beta_1 IsOnSales_i + \beta_2 BusinessSize_i + e_i.$$

На первый взгляд такая запись сильно отличается, но при более внимательном рассмотрении оказывается, что она очень похожа на приведенную выше. Для начала заметим, что она всего лишь замещает последнее уравнение в предыдущей модели и раскрывает функцию f_y, явно утверждая, что эндогенные переменные *IsOnSales* и *BusinessSize* объединяются линейно и аддитивно для формирования результата *AmountSold*. В этом смысле линейная модель включает больше предположений, чем приведенная выше. Можно сказать, что она вводит функциональную форму связей между переменными. Во-вторых, она ничего не говорит о том, откуда взялись независимые (эндогенные) переменные *IsOnSales* и *BusinessSize*.

Наконец, модель использует знак равенства вместо оператора присваивания, но мы уже договорились, что это не так важно.

Вмешательство

Почему я не пожалел времени на разбор причинных моделей? Потому что если у вас есть такая модель, вы можете начать экспериментировать с ней с целью установить причинно-следственную связь. Для таких экспериментов существует формальный термин *вмешательство* (intervention). Например, можно взять очень простую причинную модель и применить ко всем объектам воздействие t_0. При этом естественные причины T устраняются и заменяются одной константой:

$$T \leftarrow t_0$$
$$Y \leftarrow f_y(T, u_y).$$

Такое вмешательство проводится как мысленный эксперимент для ответа на вопрос: «Что произойдет с результатом Y, если было бы задано воздействие t_0?» При этом не обязательно действительно задавать это воздействие (хотя это возможно и даже неизбежно, но об этом позже). В литературе по причинно-следственному анализу такие вмешательства обозначаются оператором $do(.)$. Если вы захотите проанализировать, что произойдет при вмешательстве по T, это можно записать в виде $do\ (T = t_0)$.

Ожидания

В дальнейшем я буду часто использовать ожидания и условные ожидания. Ожидания можно рассматривать как значение выборки, которое пытается оценить среднее. $E[X]$ обозначает (предельные) ожидаемые значения случайной переменной X. Оно может быть аппроксимировано выборочным средним X. $E[Y \mid X = x]$ обозначает ожидаемое значение Y при $X = x$. Значение может быть аппроксимировано средним Y при $X = x$.

Оператор $do\ (.)$ также помогает представить, почему корреляция отличается от причинно-следственной связи. Я уже приводил аргумент о том, как высокий объем продаж в компании, проводящей распродажу, $E[AmountSold \mid IsOnSales = 1]$, может привести к переоценке среднего объема продаж компании, если она примет решение о снижении цены, $E[AmountSold \mid do(IsOnSales = 1)]$. В первом случае вы рассматриваете компании, которые решили снизить цены, — скорее всего, это крупные компании. С другой стороны, вторая величина, $E[AmountSold \mid do(IsOnSales = = 1)]$, относится к тому, что произойдет, если в распродаже будут участвовать все компании, а не только крупные. Важно, что в общем случае:

$$E[AmountSold \mid IsOnSales = 1] \neq E[AmountSold \mid do\ (IsOnSales = 1)].$$

Различия между этими двумя случаями можно рассматривать как различия между выбором и вмешательством. В первом случае вы измеряете объем продаж в ограниченном подмножестве компаний, которые реально снизили цены. Во втором случае с вмешательством *do*(*IsOnSales*) вы делаете так, что все компании снижают цены, а затем измеряете объем продаж для всей выборки (рис. 1.2).

do(.) используется для определения параметров причинности, которые не всегда удается получить из наблюдаемых данных. В предыдущем примере вы не сможете наблюдать *do*(*IsOnSales* = 1) для каждой компании, так как вы не обязываете их проводить распродажи. *do*(.) в основном используется как теоретическая концепция для явного выражения интересующего параметра причинности. Так как его нельзя наблюдать напрямую, причинно-следственный анализ в значительной мере направлен на его исключение из теоретического выражения — этот процесс называется *идентификацией*.

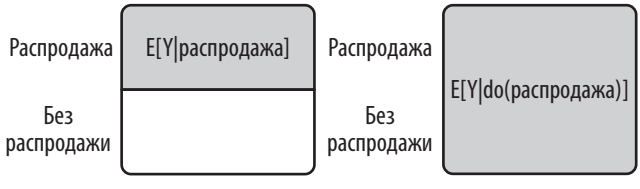

Рис. 1.2. Выбор подразумевает фильтрацию выборки на основании воздействия; вмешательство подразумевает принудительное применение воздействия ко всей выборке

Эффект индивидуального воздействия

Оператор *do*(.) также позволяет выразить *эффект индивидуального воздействия* (individual treatment effect), или влияние воздействия на результат для отдельного объекта *i*. Его можно записать в виде разности между двумя вмешательствами:

$$\tau_i = Y_i \mid do(T = t_1) - Y_i \mid do(T = t_0).$$

Эту формулу можно прочитать как «эффект τ_i перехода от воздействия t_0 к t_1 для объекта i является разностью результата этого объекта под воздействием t_1 по сравнению с t_0». Этот факт можно использовать для анализа нашей задачи вычисления эффекта от переключения *IsOnSales* из состояния 0 в состояние 1 для *AmountSold*:

$$\tau_i = AmountSold_i|do(IsOnSales = 1) - AmountSold_i|do(IsOnSales = 0).$$

Из-за фундаментальной проблемы причинно-следственного анализа можно наблюдать только один компонент приведенной формулы. Таким образом, возможность теоретического выражения этой величины не обязательно означает, что ее можно вывести из имеющихся данных.

Потенциальные результаты

Другая концепция, которую можно определить при помощи оператора `do(.)` (и возможно, лучшая и самая популярная в причинно-следственном анализе), — контрфактические, или *потенциальные*, результаты:

$$Y_{ti} = Y_i \mid do(T_i = t).$$

Это выражение следует читать как «результатом объекта i будет Y, если применить к нему воздействие t». Иногда я буду использовать функциональную запись для определения потенциальных результатов, так как нижние индексы быстро начинают нагромождаться друг на друга:

$$Y_{ti} = Y(t)_i.$$

Когда речь идет о бинарном воздействии (либо воздействие есть, либо его нет), я буду обозначать Y_{0i} потенциальный результат для объекта i без воздействия и Y_{1i} — потенциальный результат для *того же* объекта i с воздействием. Также я буду называть один потенциальный результат *фактическим* (подразумевая, что его можно наблюдать), а другой — контрфактическим (то есть ненаблюдаемым). Например, если объект i подвергается воздействию, я увижу, что с ним происходит под воздействием; то есть я увижу Y_{1i} — фактический потенциальный результат. С другой стороны, я не смогу увидеть, что произойдет, если объект i не получил воздействия. Иначе говоря, я не увижу результат Y_{0i}, так как он является контрфактическим:

$$Y_i = \begin{cases} Y_{1i} & \text{если объект } i \text{ получил воздействие} \\ Y_{0i} & \text{в противном случае.} \end{cases}$$

Иногда для тех же целей используется запись следующего вида:

$$Y_i = T_i Y_{1i} + (1 - T_i) Y_{0i} = Y_{0i} + (Y_{1i} + Y_{0i}) T_i.$$

Возвращаясь к нашему примеру, можно использовать запись $AmountSold_{0i}$ для обозначения объема продаж объекта i без распродажи, или $AmountSold_{1i}$ — при наличии распродажи. Эффект также можно определить в контексте этих потенциальных результатов:

$$\tau_i = Y_{1i} - Y_{0i}.$$

Предположения

В этой книге вы увидите, что причинно-следственный анализ всегда сопровождается предположениями. Предположениями (assumptions) называются утверждения, выражающие гипотезы о генерировании данных. Проблема в том, что обычно их нельзя проверить на основании данных; поэтому приходится принимать их на веру. Выявить предположения иногда бывает непросто, поэтому я постараюсь сделать их по возможности прозрачными.