Регрессия: теория и практика

С примерами на R и Stan

Гельман Э., Хилл Дж., Вехтари А.

ДМК Пресс издательство

Эндрю Гельман, Дженнифер Хилл, Аки Вехтари

Регрессия: теория и практика

С примерами на R и Stan

Regression and Other Stories

ANDREW GELMAN
Columbia University, New York

JENNIFER HILL
New York University

AKI VEHTARI
Aalto University, Finland

CAMBRIDGE
UNIVERSITY PRESS

Регрессия: теория и практика

С примерами на R и Stan

ЭНДРЮ ГЕЛЬМАН
Колумбийский университет, Нью-Йорк

ДЖЕННИФЕР ХИЛЛ
Нью-Йоркский университет

АКИ ВЕХТАРИ
Университет Аалто, Финляндия

ДМК ПРЕСС
издательство

Москва, 2022

УДК 303.724.32
ББК 78.36
Г32

Эндрю Гельман, Дженнифер Хилл, Аки Вехтари
Г32 Регрессия: теория и практика. С примерами на R и Stan / пер. с англ. В. С. Яценкова. – М.: ДМК Пресс, 2022. – 748 с.: ил.

ISBN 978-5-97060-987-3

В большинстве учебников по регрессии основное внимание уделяется теории и простейшим примерам. Однако настоящие задачи прикладной статистики сложнее и многограннее. Эта книга не о теории регрессии — а об использовании ее для решения реальных задач сравнения, оценки, предсказания и причинного вывода. Книга обеспечивает плавный переход к логистической регрессии и обобщенным линейным моделям. Вместо вывода формул основное внимание уделяется практическим вычислениям в средах R и Stan, а исходный код доступен для скачивания.

Издание предназначено широкому кругу специалистов по анализу и обработке данных, а также может служить учебником для студентов технических вузов.

УДК 303.724.32
ББК 78.36

ISBN (анг.) 978-1-107-02398-7
ISBN (рус.) 978-5-97060-987-3

Оглавление

Предисловие

Существующие учебники по регрессии обычно содержат смесь практических рецептов и математических выкладок. Мы написали эту книгу, потому что увидели новый способ поделиться знаниями, сосредоточившись на *понимании* регрессионных моделей, *применении* их к реальным проблемам и *выполнении* моделей на пробных придуманных данных, чтобы понять, насколько эти модели подходят к данным определенного типа. Прочитав эту книгу и проработав упражнения, вы сможете строить собственные регрессионные модели на компьютере, использовать их для решения прикладных задач и – что немаловажно – подвергать их строгой критической оценке.

Другой особенностью нашей книги, помимо широкого набора примеров и сосредоточенности на компьютерном моделировании, является ее широкий охват, включающий основы статистики и измерений, линейную регрессию, множественную регрессию, байесовский вывод, логистическую регрессию и обобщенные линейные модели, экстраполяцию от выборки к генеральной совокупности и причинно-следственный вывод. Для нас линейная регрессия является лишь отправной точкой, и мы не будем останавливаться на достигнутом: если вы уловили основную идею статистического прогнозирования, то лучший способ закрепить понимание – применять новые знания разными способами и в разных контекстах.

После прочтения первой части этой книги вы получите необходимые знания о базовых инструментах математики, статистики и вычислений, которые позволят вам работать с регрессионными моделями. Эти первые главы послужат мостом между методами и идеями, которые вы, как мы надеемся, усвоили во вводном курсе статистики. В первой части книги будет рассказано про отображение и исследование данных, вычисление и построение графиков линейных отношений, сущность основных распределений вероятностей и статистических выводов, а также про моделирование случайных процессов для имитации погрешностей выводов и прогнозов.

После прочтения второй части вы должны научиться создавать, настраивать, целенаправленно использовать модели регрессии и оценивать их качество. В главах этой части книги представлены соответствующие статистические и вычислительные инструменты в контексте нескольких примеров прикладных и смоделированных данных. Завер-

шив изучение третьей части, вы сможете аналогичным образом работать с логистической регрессией и другими обобщенными линейными моделями. Часть IV посвящена сбору данных и экстраполяции от выборки к совокупности. А в части V мы рассмотрим причинный вывод, начиная с основных методов, использующих регрессию для контролируемых экспериментов, а затем обратимся к более сложным методам с поправкой на дисбаланс в данных наблюдений или с использованием натурных экспериментов. В части VI представлены более сложные регрессионные модели, а в приложениях мы делимся советами и предлагаем обзор программного обеспечения для подгонки моделей.

Благодарности

Мы благодарим студентов и коллег, которые помогли нам понять и реализовать эти идеи, в том числе всех, кого упоминали ранее на страницах нашей предыдущей книги «Анализ данных с использованием регрессии и многоуровневых/иерархических моделей». Кроме того, мы благодарим Пабло Арготе, Билла Бермана, Данило Бздока, Андреса Кастро, Девина Кауги, Зада Чоу, Дика Де Во, Винса Дори, Сандера Гренланда, Дафну Харель, Мерлин Хайдеманнс, Кристиану Хеннига, Дэвида Кейна, Катарину Ханну, Лидию Красильникову, Стефано Лонго, Дженни Фам, Эрика Поташа, Фила Прайса, Малгожату Роос, Майкла Собеля, Мелинду Сонг, Скотта Спенсера, Мирейю Тригуэро, Ясу Вехтари, Зейна Вольфа, Лиззи Волкович, Адама Зелизера, Шули Чжан, а также студентов и помощников преподавателей с которыми мы встречались в течение нескольких лет, пока читали лекции, за полезные комментарии и предложения. Благодарим Алана Чена за помощь с главой 20; Андреа Корнехо, Зарни Хтета и Руи Лу – за помощь в разработке симуляционных упражнений для глав о причинности; Бена Сильвера – за помощь с предметным указателем; Бета Морела и Клэр Деннисон – за редактирование исходного текста; Люка Кила – за пример из раздела 21.3; Кайзера Фунга – за пример из раздела 21.5. Спасибо Марку Броди за данные о гольфе в упражнении 22.3; Майклу Бетанкуру – за демонстрацию измерения силы тяжести в упражнении 22.4; Джерри Рейтеру – за обмен идеями по обучению студентов регрессии; Лорен Коулз – за многочисленные полезные предложения по структуре этой книги. И особая благодарность Бену Гудричу и Йохану Габри за разработку пакета `rstanarm`, который позволяет подгонять регрессионные модели в Stan с использованием знакомой нотации R.

Мы благодарим разработчиков R и Stan, а также Национальный научный фонд США, Институт педагогических наук, Управление военно-морских исследований, Агентство перспективных оборонных исследовательских проектов, Google, Facebook, YouGov и Фонд Слоуна за финансовую поддержку.

Но больше всего мы благодарны нашим семьям за их любовь и поддержку во время написания этой книги.

Краткое содержание книги

Эта книга содержит описания моделей и примеров, чтобы после каждой главы у вас появлялись новые навыки подгонки, интерпретации и визуализации моделей.

- **Часть I.** Обзор основных инструментов и понятий математики, статистики и вычислений.
  - Глава 1: Общее представление о целях и задачах регрессии.
  - Глава 2: Исследование данных и знакомство с проблемами измерения.
  - Глава 3: Совершаем рывок и знакомимся с основными математическими инструментами и распределениями вероятностей.
  - Глава 4: Знакомство со статистической оценкой и оценкой погрешности, а также проблемой проверки гипотез в прикладной статистике.
  - Глава 5: Моделирование вероятностных моделей и погрешности их выводов и прогнозов.
- **Часть II.** Построение моделей линейной регрессии, использование их в реальных задачах, оценка допущений и степени соответствия данным.
  - Глава 6: Различия между описательной и причинной интерпретациями регрессии в историческом контексте.
  - Глава 7: Простая линейная регрессия с одним прогностическим параметром.
  - Глава 8: Аппроксимация методом наименьших квадратов – определение и выполнение на компьютере.
  - Глава 9: Вероятностное прогнозирование и простое байесовское агрегирование информации, а также знакомство с априорными распределениями и байесовским выводом.
  - Глава 10: Создание, настройка и интерпретация линейных моделей с несколькими прогностическими параметрами.
  - Глава 11: О важности различных допущений регрессионных моделей и умения проверять модели и оценивать их соответствие данным.
  - Глава 12: Более эффективное применение линейной регрессии путем преобразования и комбинирования прогностических параметров.
- **Часть III.** Построение и применение моделей логистической регрессии и обобщенных линейных моделей.
  - Глава 13: Подбор, интерпретация и визуализация моделей логистической регрессии для бинарных данных.
  - Глава 14: Построение, интерпретация и оценка логистических регрессий с взаимодействиями и другими усложняющими факторами.
  - Глава 15: Подгонка, интерпретация и визуализация обобщенных линейных моделей, включая пуассоновскую и отрицательную биномиальную регрессию, упорядоченную логистическую регрессию и другие модели.

- **Часть IV.** Разработка исследований и более эффективное использование данных в прикладных задачах.
  - Глава 16: Как использовать теорию вероятностей и моделирование для принятия решений о собираемых данных и не попадать в ловушку нереалистичных уровней определенности.
  - Глава 17: Использование постстратификации для обобщения от выборки к генеральной совокупности и применение регрессионных моделей для вставки недостающих данных.
- **Часть V.** Внедрение и понимание основных статистических схем и анализов для причинно-следственного вывода.
  - Глава 18: Предположения, лежащие в основе причинно-следственного вывода, с акцентом на рандомизированные эксперименты.
  - Глава 19: Моделирование причинно-следственных связей в простых условиях с использованием регрессий для оценки эффектов воздействия и взаимодействий.
  - Глава 20: Проблемы, связанные с выводом причинно-следственных связей из данных наблюдений, и статистические инструменты для корректировки различий между экспериментальной и контрольной группами.
  - Глава 21: Допущения, лежащие в основе более сложных методов, использующих вспомогательные переменные или определенные структуры данных для выявления причинности, и умение согласовывать эти модели с данными.
- **Часть VI.** Обзор более продвинутых регрессионных моделей.
  - Глава 22: Общее представление о направлениях, в которых линейные и обобщенные линейные модели могут быть расширены для решения различных классов прикладных задач.
- **Приложения**
  - Приложение A: Первые навыки работы в статистическом пакете на языке R с акцентом на обработку данных, статистические графики, а также доводку и использование регрессионных моделей.
  - Приложение B: Идеи и советы, которые пригодятся вам при работе с регрессионными моделями.

Прочитав эту книгу, вы научитесь выбирать, создавать, интерпретировать и оценивать линейные и обобщенные линейные модели и использовать их, чтобы делать прогнозы и выводы, включая причинно-следственные связи.

Более увлекательные названия глав

В оглавлении книги вы видите сухие и строгие названия глав, которые носят описательный характер. Мы решили немного нарушить традицию и в качестве альтернативы предлагаем вам более эмоциональные названия, которые, как мы надеемся, вызовут у вас удивление и пробудят заинтересованность.

- **Часть I**
  - Глава 1: Прогнозирование как объединяющая тема в статистике и причинно-следственных связях.
  - Глава 2: Правильный сбор и визуализация данных важнее, чем вы думаете.
  - Глава 3: Математика, которую вам действительно нужно знать.
  - Глава 4: Забудьте все, что вы раньше знали о статистике.
  - Глава 5: Вы не поймете свою модель, пока не выполните имитацию.
- **Часть II**
  - Глава 6: Давайте серьезно задумаемся о регрессии.
  - Глава 7: Нельзя просто *работать* с регрессией, ее нужно *понимать*.
  - Глава 8: Наименьшие квадраты и все такое.
  - Глава 9: Откровенно о погрешности и априорных знаниях.
  - Глава 10: Вы не просто *выбираете* модели, вы их *создаете*.
  - Глава 11: Попробуйте убедить нас довериться вашей модели.
  - Глава 12: Только глупцы работают с данными без масштабирования.
- **Часть III**
  - Глава 13: Моделирование вероятностей.
  - Глава 14: Советы профессионалов по логистической регрессии.
  - Глава 15: Создание моделей – взгляд изнутри.
- **Часть IV**
  - Глава 16: Чтобы понять прошлое, вы должны узнать будущее.
  - Глава 17: Хватит рассказывать о данных. Лучше расскажите о генеральной совокупности.
- **Часть V**
  - Глава 18: Как подбрасывание монеты помогает оценить причинно-следственные связи?
  - Глава 19: Использование корреляции и предположений для вывода причинно-следственной связи.
  - Глава 20: Причинный вывод – это просто своего рода предсказание.
  - Глава 21: Больше допущений – больше проблем.
- **Часть VI**
  - Глава 22: Что нас ждет впереди?
- **Приложения**
  - Приложение A: Беглое знакомство с R.
  - Приложение B: Наши любимые советы и навыки. А что умеете вы?

В этой книге мы рассказываем о различных методах и иллюстрируем их использование во многих прикладных сценариях. Мы также стараемся дать представление о том, где эти методы могут потерпеть неудачу, и стремимся передать волнение, которое мы испытали, когда впервые узнали об этих идеях и применили их к нашим собственным задачам.

Скачивание исходного кода примеров

Скачать файлы с дополнительной информацией для книг издательства «ДМК Пресс» можно на сайте www.dmkpress.com или www.дмк.рф на странице с описанием соответствующей книги.

Максимально эффективное использование книги

Для чтения этой книги не требуется глубокое знание математики. Например, чтобы изучить линейную регрессионную модель, вам следует знать алгебраические уравнения, описывающие точки пересечения и наклон прямой, но нет необходимости разбираться в матричной алгебре при выводе вычислений методом наименьших квадратов. Вы будете использовать показатели степени и логарифмы, особенно в главах 12–15 при изучении нелинейных преобразований и обобщенных линейных моделей.

Наличие навыков программирования не требуется. Вы будете немного программировать в статистической среде общего назначения R при подгонке и использовании моделей из этой книги, и некоторые из этих процедур будут выполнены с помощью программы байесовского вывода Stan, которая, как и R, является бесплатной и с открытым исходным кодом. Читатели, плохо знакомые с R или программированием, должны сначала изучить Приложение А, этого будет достаточно.

Мы подгоняем регрессионные модели с помощью функции `stan_glm` в пакете `rstanarm` в R, выполняя байесовский вывод. Это небольшое отклонение от обычных методов (включая нашу предыдущую книгу), в которых используются методы наименьших квадратов и максимального правдоподобия, например с использованием функций `lm` и `glm` в R. Мы обсуждаем различия между различными вариантами программных инструментов и между различными режимами вывода в разделах 1.6, 8.4 и 9.5. С точки зрения пользователя, переход на `stan_glm` не имеет большого значения, за исключением упрощения получения вероятностных прогнозов и распространения погрешностей вывода, а также в некоторых проблемах с коллинеарностью или разреженными данными (в этом случае байесовский подход в `stan_glm` дает более стабильные оценки) и когда мы хотим включить в анализ априорную информацию. Для большинства вычислений, выполненных в этой книге, при желании можно получить аналогичные результаты с использованием классического программного обеспечения.

В помощь преподавателю: возможная структура курсов

Материал этой книги можно разбить на односеместровые курсы по нескольким направлениям. Окончательное решение остается за преподавателем, но мы предлагаем несколько возможных вариантов.

- *Основы линейной регрессии:* главы 1–5 в качестве обзора, затем главы 6–9 (линейная регрессия с одним прогностическим параметром) и 10–12 (множественная регрессия, диагностика и построение модели).
- *Прикладная линейная регрессия:* главы 1–5 в качестве обзора, затем главы 6–12 (линейная регрессия), 16–17 (разработка и постстратификация) плюс избранный материал из глав 18–21 (причинный вывод) и главы 22 (дополнительная информация).
- *Прикладная регрессия и причинный вывод:* краткий обзор на основе глав 1–5, затем главы 6–12 (линейная регрессия), глава 13 (логистическая регрессия), главы 16–17 (дизайн и постстратификация) и избранные материалы из глав 18–21 (причинный вывод).
- *Причинный вывод:* главы 1, 7, 10, 11 и 13 для обзора линейной и логистической регрессии, затем главы 18–21 для более подробного изложения материала.
- *Обобщенные линейные модели:* краткий обзор на основе глав 1–12, затем главы 13–15 (логистическая регрессия и обобщенные линейные модели), за которыми следует избранный материал из глав 16–21 (разработка, постстратификация и причинный вывод) и 22 (дополнительная информация).

Типографские соглашения, принятые в книге

В этой книге используется несколько стилей выделения некоторых элементов текста.

`Фрагмент кода в тексте` – ключевые слова, операторы, имена переменных и функций непосредственно в тексте. Пример: «Большая часть приведенного выше кода предназначена для построения и вывода графической схемы, вероятностные вычисления выполняются в строке `y = stats.norm(mu, sd).pdf(x)`».

Блок кода отображается в следующем формате:

```
μ = 0.
σ = 1.
X = stats.norm(μ, σ)
x = X.rvs(3)
```

Курсив – имена файлов, каталогов и прочих объектов.

Полужирный шрифт – важные (ключевые) слова, элементы пользовательского интерфейса или слова, которые выводятся на экран.

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 ре-

цензии». 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. Издательства «ДМК Пресс» и Cambrdige University Press очень серьезно относятся 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 электронной почты dmkpress@gmail.com.

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

Часть I. Основы

Глава 1

Обзор темы и знакомство с регрессией

В этой книге мы рассмотрим проблемы построения, интерпретации и использования прогнозных моделей. Оказывается, есть много тонкостей даже при подгонке простой линейной модели – построении прямой линии регрессии по точкам исходя из имеющихся данных. После обзора фундаментальных понятий из области обработки данных, измерений и статистики в первых пяти главах книги мы рассмотрим линейную регрессию с одним и несколькими предикторами, а затем логистическую регрессию и другие обобщенные линейные модели. Затем мы рассмотрим различные прикладные применения регрессии – как простые, наподобие обобщения имеющихся данных, так и более сложные, включая выборку и причинный вывод. Книга завершается знакомством с современными идеями в области моделирования и двумя приложениями, которые содержат полезные советы и краткое введение в программирование на языке R.

В этой вводной главе излагаются ключевые задачи статистического вывода в целом и регрессионного моделирования в частности. Мы представляем множество практических примеров, чтобы наглядно продемонстрировать, насколько сложной и утонченной может быть регрессия и почему нужна целая книга не только о теории регрессионного моделирования, но и о том, как применять ее на практике.

1.1. Три задачи статистики

Статистический вывод призван решить три ключевые задачи.

1. *Обобщение выборки на генеральную совокупность* – задача, связанная с наличием ограниченной выборки из потенциально более обширных данных, но фактически она возникает почти при каждом применении статистического вывода.
2. *Обобщение данных экспериментального воздействия на контрольную группу* – задача, связанная с причинным выводом, который явным или неявным образом является частью интерпретации большинства наблюдаемых нами регрессий.

Обобщение наблюдаемых измерений на интересующий нас конструкт[1], поскольку в большинстве случаев наши наблюдения не отражают в точности то, что мы в идеале хотели бы изучить.

Все три проблемы могут быть сформулированы как проблемы прогнозирования (вычисления ожидаемых показателей для новых людей или новых предметов, не вошедших в выборку, будущих откликов системы при различных потенциально возможных вариантах воздействия и скрытых конструктов, если их свойства можно достаточно точно измерить).

Мы ожидаем, что после прочтения этой книги вы получите следующие ключевые навыки:

- *понимание сути регрессионных моделей.* Вы изучите математические модели для прогнозирования выхода (переменной результата, отклика на воздействие) на основе набора предикторов, начиная с линейной аппроксимации и заканчивая различными нелинейными обобщениями;
- *умение строить регрессионные модели.* Это открытый творческий процесс, включающий множество возможных вариантов, в том числе выбор параметров, а также их преобразование и нормирование;
- *умение подгонять регрессионные модели по данным* – процесс подбора параметров модели, который мы будем выполнять с помощью программного обеспечения с открытым исходным кодом R и Stan;
- *визуализацию и интерпретацию результатов*, что требует дополнительных навыков программирования и знания математики.

Центральной темой этой книги, как и большинства книг о статистике, является *логический вывод* (infrerence) – использование математических моделей для получения общих утверждений на основе конкретных данных.

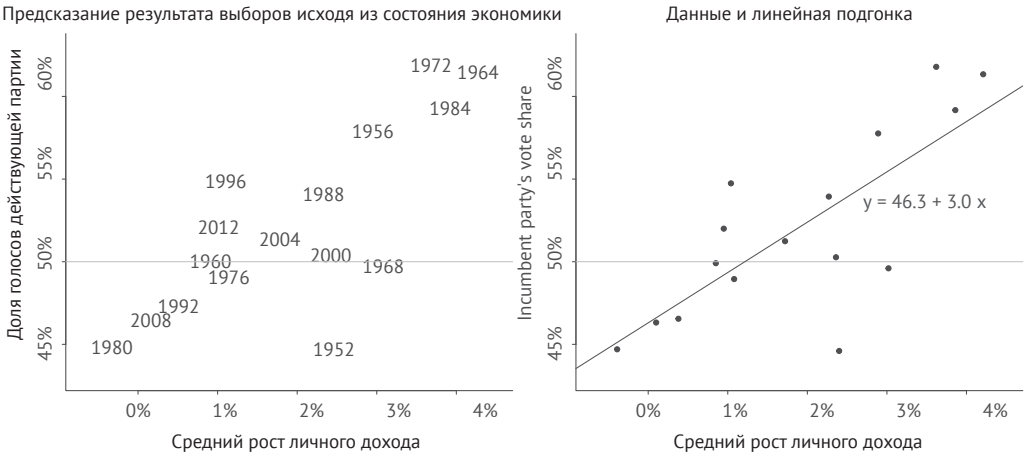


Рис. 1.1. Прогнозирование результата выборов на основе экономических данных: (а) данные, (б) аппроксимация прямой y = 46,3 + 3,0x

[1] *Конструктом* в философии восприятия называют идеальный объект, недоступный для прямого наблюдения, но гипотетически выводимый через его внешние проявления и подтверждаемый или как минимум не опровергаемый экспериментальными наблюдениями. – *Прим. перев.*

1.2. Зачем изучать регрессию?

Пример: выборы и экономика

Регрессия – это метод, который позволяет исследователям определить, как прогнозы или средние значения *выхода* (outcome) модели различаются для разных объектов, определенных набором входных данных – *предикторов* (predictor). Например, на рис. 1.1а показана доля голосов за кандидата от действующей партии в последовательном ряде президентских выборов в США в зависимости от показателя экономического роста в период, предшествующий каждому году выборов. На рис. 1.1б показана линейная регрессия, соответствующая этим данным. Эта модель позволяет нам прогнозировать итог голосования – с некоторой погрешностью – с учетом экономических показателей и в предположении, что будущие выборы в чем-то похожи на предыдущие.

Все вычисления в этой книге выполняются на языке R. Это полностью бесплатное и простое в использовании программное обеспечение. В приложении А рассказано, как установить и использовать R на вашем компьютере. Начнем с загрузки данных[1]:

```
hibbs <- read.table("hibbs.dat", header=TRUE)
```

Затем строим *диаграмму рассеяния* (scatterplot):

```
plot(hibbs$growth, hibbs$vote, xlab="Average recent growth in
personal income", ylab="Incumbent party's vote share")
```

Вычисляем регрессию $y = a + bx + error$[2]:

```
M1 <- stan_glm(vote ~ growth, data=hibbs)
```

Теперь добавляем на наш график результат подгонки линии регрессии по данным:

```
abline(coef(M1), col="gray")
```

Наш результат должен быть похож на рис. 1.1б.

Чтобы отобразить настроенную модель, наберем команду `print(M1)` и получим следующий результат:

```
            Median MAD_SD
(Intercept) 46.3    1.7
growth       3.0    0.7

Auxiliary parameter(s):
      Median MAD_SD
sigma 3.9    0.7
```

В первом столбце показаны результаты подгонки: 46,3 и 3,0 – найденные коэффициенты уравнения линии $y = 46{,}3 + 3{,}0x$ (рис. 1.1б). Во втором столбце представлены погрешности результатов с исполь-

[1] Данные и код для этого примера находятся в папке ElectionsEconomy.

[2] В разделе 1.6 представлен код R для метода наименьших квадратов и байесовской регрессии.

зованием медианных абсолютных отклонений (раздел 5.3). Последняя строка выходных данных показывает значение и погрешность σ, меры неоднородности данных, не объясненных регрессионной моделью (разброс точек выше и ниже линии регрессии). На рис. 1.1б линейная модель прогнозирует долю голосов с точностью примерно до 3,9 процентных пункта. Подробные объяснения приведенного выше кода и вывода будут представлены позже, начиная с главы 6.

При желании мы также можем выполнить подгонку модели другими способами, такими как графическое представление остатков регрессии (различий между данными и подогнанной моделью) и вычисление R^2 – доли дисперсии, объясняемой моделью, о чем пойдет речь в главе 11.

Некоторые из наиболее важных применений регрессии:

- *прогнозирование:* моделирование существующих наблюдений или предсказание новых данных. К примерам с непрерывными или приблизительно непрерывными значениями выхода относятся доли голосов на предстоящих выборах, будущие продажи продукта и состояние здоровья, отраженное в медицинском исследовании. К примерам дискретных или категориальных результатов (также называемых классификацией) относятся диагностика заболеваний, предсказание победы или поражения в спортивном соревновании и индивидуального выбора при голосовании;
- *изучение взаимосвязей:* формирование заключения о том, насколько сильно одна или несколько входных переменных оказывают влияние на выход. В качестве примеров можно назвать выявление факторов риска заболевания, определение внешних факторов, влияющих на результат голосования и характеристик в резюме, которые повышают вероятность успешной карьеры. В более общем плане можно использовать такую модель для изучения ассоциаций, стратификаций или структурных отношений между переменными. К примерам относятся взаимосвязь между количеством загрязняющих веществ и заболеваемостью, разная частота задержаний подозреваемых полицией по этнической принадлежности и скорость роста различных частей тела;
- *экстраполяция:* внесение поправки на известные различия между *выборкой* (т. е. данными наблюдений) и представляющей интерес совокупностью. Знакомый всем пример – социальный опрос: выборки людей из реального мира не являются полностью репрезентативными, поэтому необходимо выполнить некоторую корректировку для экстраполяции на генеральную совокупность. Другой пример – использование данных, которые извлекли из ограниченной выборки школ, отобранных исследователями, чтобы сделать выводы обо всех школах в штате. Еще одним примером может служить использование экспериментальных данных об испытаниях лекарственного средства вместе с фоновыми характеристиками всего населения для оценки среднего эффекта воздействия препарата на популяцию;

- *причинно-следственный вывод:* возможно, наиболее важным применением регрессии является оценка *отклика* на воздействие (эффекта эксперимента). Мы дадим более строгое определение причинно-следственного вывода в части V этой книги; а пока просто поговорим о сравнении выборок экспериментальной и контрольной группы или экспериментальных групп с разным уровнем воздействия. Например, в исследовании качества образования откликом могут быть баллы по стандартизированному тесту, роль контрольной группы могут играть студенты, изучавшие стандартный курс, а экспериментальным воздействием может служить новый учебный курс. В области здравоохранения откликом может служить заболеваемость астмой, а продолжительное воздействие может быть связано с наличием какого-либо загрязнителя воздуха. Ключевой проблемой причинно-следственного вывода является обеспечение максимального сходства между экспериментальной и контрольной группой в среднем до начала воздействия или корректировка различий между этими группами.

Во всех этих случаях крайне важно, чтобы регрессионная модель имела достаточную сложность для восприятия необходимой информации. Например, если большинство участников испытания лекарства здоровы и моложе 70 лет, но исследователи заинтересованы в предсказании влияния препарата на пожилое население в целом, то важно включить в регрессионную модель в качестве предикторов возраст и предыдущее состояние здоровья. Если эти предикторы отсутствуют, модели просто не хватит информации для необходимой экстраполяции.

1.3. Несколько примеров регрессии

Чтобы дать вам представление о трудностях, связанных с применением регрессии на практике, мы приведем несколько примеров, иллюстрирующих выборку, прогнозирование и причинный вывод.

Оценка общественного мнения на основе добровольного интернет-опроса

Пример: опрос на платформе Xbox

В исследовательском проекте с коллегами из Microsoft Research мы использовали регрессионную модель для корректировки выборки, чтобы получить качественный мониторинг общественного мнения, более точно привязанный ко времени и с меньшими затратами, чем традиционные методы опроса. Данные были взяты из нового и крайне нерепрезентативного набора данных опроса: серии ежедневных опросов избирателей на президентских выборах 2012 года, проводимых на игровой платформе Xbox, с общим размером выборки 750 148 интервью с 345 858 уникальными респондентами. Это характерная проблема больших данных: очень большая выборка, относительно низкие затраты, но при этом отсутствие прямой репрезентативности для большей генеральной совокупности. После корректировки результатов опроса на платформе Xbox с помощью многоуровневой регрессии и постстра-

тификации (multilevel regression and poststratification, MRP) мы получили оценки, соответствующие прогнозам ведущих аналитиков, основанных на агрегировании сотен традиционных опросов, проведенных в течение избирательного цикла.

Целью проекта Xbox не были ни прогнозирование индивидуальных ответов на опрос, ни выявление важных предикторов или причинно-следственных выводов. Истинная цель состояла в том, чтобы выявить общенациональные тенденции в общественном мнении, а регрессия позволила нам скорректировать различия между выборкой и генеральной совокупностью, о чем пойдет речь в разделе 17.1; для этого нам потребовалась экстраполяция.

Рандомизированный эксперимент по изучению влияния образовательной телепрограммы

Пример: эксперимент Electric Company

Примерно в 1970 году было проведено исследование с целью измерить влияние новой образовательной телепрограммы Electric Company на развитие навыка чтения у детей. Эксперимент проводился на детях 1-4-х классов в двух небольших городах США. Для каждого города и класса экспериментаторы выбрали от 10 до 20 школ, в каждой из которых определили по два класса, чьи средние результаты тестов по чтению были самыми низкими. В каждой паре один класс случайным образом выбрали для продолжения обычного курса обучения чтению, а дети из другого класса регулярно смотрели телепрограмму. Каждый ученик прошел предварительное тестирование в начале учебного года и окончательное тестирование в конце.

На рис. 1.2 показаны данные результатов тестирования контрольных и экспериментальных классов[1]. Сравнивая верхний и нижний ряд графиков, можно предположить, что максимальный положительный эффект от просмотра телепрограммы наблюдается в 1-х и 2-х классах, а в более старших классах эффект уменьшается. Эти результаты выглядят правдоподобными, если учесть, что большинство детей в 3-х и 4-х классах уже хорошо умеют читать.

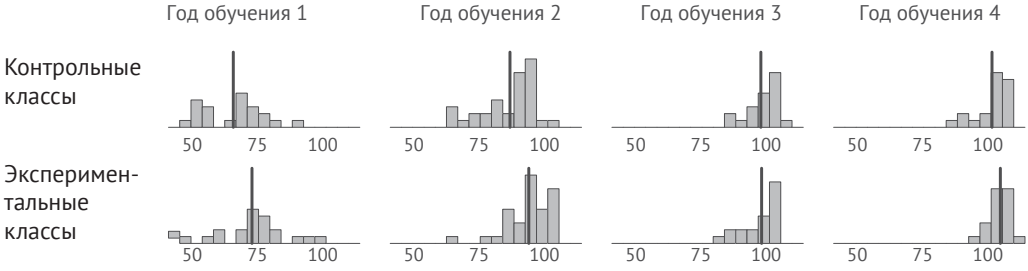


Рис. 1.2. Средние результаты тестов в классе после эксперимента по измерению влияния образовательной телепрограммы на способность детей к чтению. Темная вертикальная линия на каждой гистограмме показывает среднее значение для соответствующей группы классов

[1] Данные и код этого примера находятся в папке ElectricCompany.

Чтобы скорректировать различия между предварительными тестами экспериментальной и контрольной группы, а также для оценки погрешности результатов необходим дальнейший статистический анализ. Мы вернемся к этому примеру в разделе 19.2.

Оценка влияния деятельности Организации Объединенных Наций по поддержанию мира с использованием переменных, предшествующих воздействию, для корректировки различий между экспериментальной и контрольной группой

Пример: миротворцы ООН

Несколько лет назад политолог Пейдж Фортна провела исследование эффективности международных миротворческих сил. Она проанализировала данные из стран, которые были вовлечены в гражданские войны в период с 1989 по 1999 год, сравнивая страны, в которых проводились миротворческие операции Организации Объединенных Наций, со странами, в которых таких операций не было. Конечным критерием служил тот факт, возобновилась ли в стране гражданская война, и если да, то сколько времени продержалось перемирие. Сбор данных закончился в 2004 году, поэтому любые страны, в которых к концу того года не возобновилась гражданская война, были охарактеризованы как находящиеся в состоянии мира. Подмножество обобщенных данных содержит 96 случаев прекращения огня, соответствующих 64 различным войнам[1].

Беглое сравнение стран свидетельствовало о положительном эффекте миротворческой деятельности: в 56 % стран сохранился мир, по сравнению с 34 % стран, в которые не входили миротворцы. Если гражданская война все-таки возобновлялась, обычно это случалось вскоре после перемирия: средний промежуток между прекращением огня и возобновлением боевых действий составлял 17 месяцев в присутствии миротворческих сил и 18 месяцев без них. На рис. 1.3 показаны результаты.

В данном случае возникают большие опасения по поводу *предвзятости отбора*: возможно, миротворцы выбрали легкие случаи. Может быть, по-настоящему жестокие гражданские войны были настолько опасны, что миротворцы не рискнули войти в такие страны, и это объясняет разницу в результатах.

Проще говоря, в этом исследовании «экспериментальное воздействие» – поддержание мира – не было назначено случайным образом. Говоря языком статистики, Фортна проводила *наблюдательное исследование*, а не *эксперимент*, и в этом случае мы должны сделать все возможное, чтобы учесть различия между экспериментальной и контрольной группой, существовавшие до воздействия.

[1] Данные и код для этого примера находятся в папке Peacekeeping.

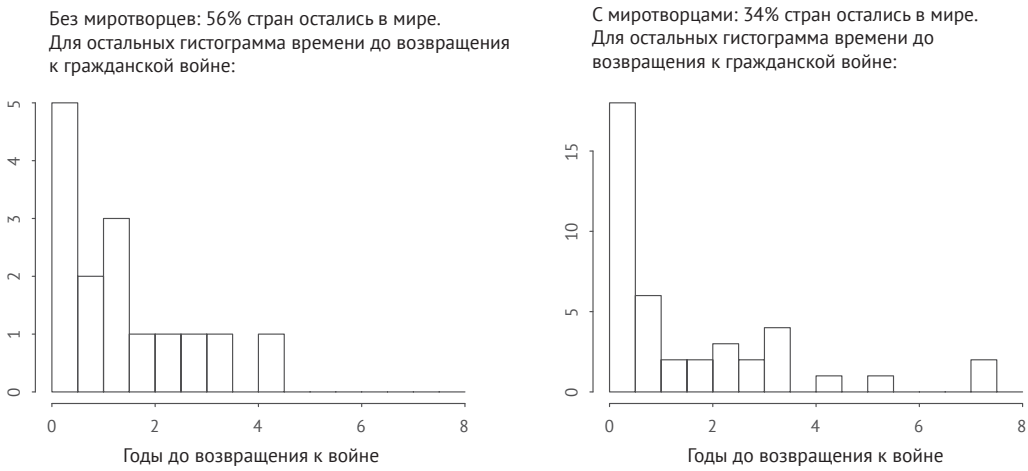


Рис. 1.3. Результаты прекращения гражданской войны в странах с миротворческой деятельностью ООН и без нее. Страны под наблюдением миротворцев ООН с большей вероятностью сохраняли мир, и в среднем требовалось примерно столько же времени на возврат к войне, если это случалось. Однако есть опасения, что страны с миротворческой деятельностью и без нее могут различаться по предшествующим условиям; подробнее на рис. 1.4

Фортна скорректировала начальные условия относительно того, насколько плохим было состояние страны до того, как было принято то или иное решение о вводе миротворческих сил, используя некоторые объективные показатели условий жизни в стране. Анализ еще больше усложнялся тем, что в некоторых странах мы знаем время до возврата к гражданской войне, тогда как относительно других стран мы можем лишь сказать, что гражданская война еще не возобновилась в период сбора данных. В статистике такой вид обработки неполных данных называется *цензурой*, но это не означает, что кто-то отказался предоставить данные. Как правило, некоторые данные не могут быть получены просто в силу особенностей процедуры сбора данных: в данном случае продолжительность времени до возобновления гражданской войны по своей сути не может быть известна для стран, которые сохранили мир к дате завершения сбора данных. Фортна решила эту проблему, используя «модель выживания», в детали которой мы здесь не будем углубляться и просто отметим, что в итоге она суммирует комбинацию предикторов до воздействия как скалярную «оценку плохого состояния». Эта оценка колеблется от 1,9 для гражданской войны в Йемене в 1994 году и 2,0 для восстания сикхов в Индии в 1993 году до случаев с наивысшими оценками негативной ситуации: 6,9 для Анголы в 1991 году и 6,5 для Либерии в 1990 году.

На рис. 1.4 показаны результаты для «экспериментальных» и контрольных стран в зависимости от оценки негативной ситуации, с некоторыми пропусками, когда для такой оценки не хватало некоторых факторов. Согласно этим данным, миротворцы фактически работали

даже в более жестких условиях, чем предполагалось. В результате поправка на плохие исходные условия (если признать, что эта поправка достоверна настолько, насколько достоверны данные и модель, использованные для ее получения) увеличивает предполагаемый положительный эффект миротворческой деятельности, по крайней мере, в период данного исследования.

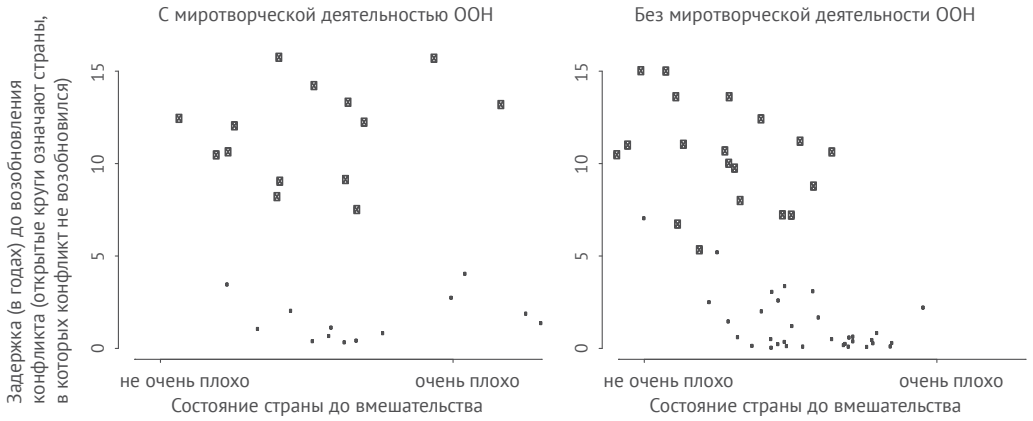


Рис. 1.4. Показатели прекращения гражданской войны в странах с миротворческой деятельностью ООН и без нее в сопоставлении с показателем того, насколько плохой была ситуация в стране. После предварительной корректировки с учетом новой переменной миротворчество по-прежнему ассоциируется с более длительными периодами мира

Оценка эффективности законов об оружии и сложности регрессионного вывода при большом количестве предикторов

Пример: политика контроля над огнестрельным оружием

Ведущий медицинский журнал опубликовал статью, в которой попытался оценить эффективность ряда мер по контролю над огнестрельным оружием:

> Из 25 законов об огнестрельном оружии девять были связаны со снижением смертности от огнестрельного оружия, девять были связаны с повышенной смертностью от огнестрельного оружия, а семь имели неубедительную связь… Прогнозируемое внедрение на федеральном уровне универсальных проверок анкетных данных при покупке огнестрельного оружия могло бы снизить смертность от огнестрельного оружия с 10,35 до 4,46 смертей на 100 000 человек, проверки анкетных данных при покупке боеприпасов могли бы снизить ее до 1,99 на 100 000, а идентификация огнестрельного оружия – до 1,81 на 100 000 человек.

В этом исследовании предпринята попытка причинно-следственного вывода с использованием регрессии факторов воздействия с поправкой на фоновые факторы для коррекции различий между экспериментальной и контрольной группами. Модель настроили таким образом, чтобы сделать прогнозы зависимыми от значений предикторов, соответствующих различным гипотетическим ситуациям в политике.

Но мы считаем, что эти результаты фактически бесполезны по двум причинам: во-первых, в такой регрессии – с 50 точками данных и 30 предикторами и без априорной информации для вывода – коэффициенты модели будут безнадежно зашумлены и скомпрометированы перекрестной зависимостью между предикторами. Во-вторых, исследование было наблюдательным, а не экспериментальным. Другими словами, существуют систематические различия между ситуациями, которые сопровождают различные подходы к контролю над оружием, т. е. различия, которые не будут отражены в других предикторах модели (ковариатах уровня состояния или основных переменных), и нет оснований полагать, что большие различия между штатами в количестве смертей, связанных с огнестрельным оружием, в первую очередь объясняются именно разными законами по контролю над оружием.

Сравнение исследований о деятельности миротворцев и контроле над оружием

Почему мы удовлетворены выводами исследования о деятельности миротворцев, и раскритиковали исследование о контроле над огнестрельным оружием? В обоих случаях политические выводы были сделаны на основе данных наблюдений с использованием регрессионного моделирования для корректировки различий между экспериментальной и контрольной группами. Так чем же различаются эти два проекта?

Основное отличие состоит в том, что исследование эффекта от деятельности миротворцев сфокусировано, тогда как исследование, посвященное контролю над огнестрельным оружием, таковым не является. Когда цель одна, удобнее выполнять корректировку. В частности, по поводу исследования миротворческой деятельности высказывалась особая озабоченность, что ООН с большей вероятностью вмешается в более безопасные ситуации. Анализ данных показал обратное: операции по поддержанию мира в среднем выполнялись в несколько худших условиях. Этот вывод не окончательный – в частности, мера плохих условий в стране опирается на определенные измеряемые переменные, и поэтому возможно, что существуют важные неизмеряемые характеристики, которые заставят корректировку пойти другим путем. Тем не менее модель, которую мы видим на основе наблюдаемых переменных, в целом рассказывает более убедительную историю.

Напротив, модель контроля над огнестрельным оружием трудно признать убедительной по двум причинам. Во-первых, модель корректируется сразу для многих потенциальных причинных переменных: влияние каждого закона оценивается при условии, что все остальные законы остаются неизменными, что нереально, так как несколько законов могут быть изменены одновременно, и нет причины предполагать, что их влияние складывается простым способом. Во-вторых, сравнения проводятся между состояниями, но состояния различаются по многим системным критериям, и совсем не ясно, может ли простая модель надеяться на корректировку соответствующих различий.

Да, сравнения в миротворческом проекте зависят от страны, но разработанный критерий плохих условий жизни кажется более уместным для вопроса, задаваемого в этом исследовании.

Мы не хотим делать слишком большое различие между этими исследованиями, которые в конечном итоге оценивают степень, а не характер явления. На самом деле нам действительно нужно оценивать политику в области поддержания мира, контроля над оружием и других областях, и имеет смысл использовать данные и статистический анализ для помощи в принятии решений. Мы рассматриваем исследование деятельности миротворцев, несмотря на все его потенциальные недостатки, как хороший пример, поскольку оно начинается с прямого сравнения данных, а затем целенаправленно устраняет угрозу корректности выводов. Наоборот, в исследовании по контролю над огнестрельным оружием корректировки переменных до воздействия кажутся менее убедительными и фатально зависят от неправдоподобных предположений модели, что часто случается с неструктурированными данными.

Во втором случае статистические методы сами по себе становятся проблемой, поскольку заявления о контроле над огнестрельным оружием никогда не были бы опубликованы без ложного чувства уверенности, основанного на регрессионном анализе. В этом анализе пришлось применить наивный подход, чтобы иметь возможность отслеживать варианты и давать достоверные причинно-следственные выводы из данных наблюдений; а статистическая значимость и доверительные интервалы были столь же наивно подобраны таким образом, чтобы иметь возможность отсеивать шум и предоставлять воспроизводимые утверждения о мире за пределами имеющихся данных. Сложите эти два наивных подхода вместе, и в результате уважаемому медицинскому журналу приходится публиковать внешне убедительные, но плохо обоснованные выводы, сделанные на основе беспорядочного набора агрегированных данных о тенденциях.

1.4. Проблемы построения и интерпретации регрессий

Мы можем выделить два основных способа использования регрессии для причинного вывода: *оценку взаимосвязи* и *корректировку основных переменных*.

Применение регрессии для оценки интересующей нас взаимосвязи

Начнем с простейшего сценария сопоставимости экспериментальной и контрольной групп. Выполнения этого условия можно добиться с помощью рандомизации – подхода, при котором люди (или, в более общем плане, некие объекты эксперимента) случайным образом распределяются по экспериментальным и контрольным группам, или

с помощью более сложной схемы, которая обеспечивает баланс между группами. В части V этой книги мы подробно рассматриваем связи между экспериментальным воздействием, балансом и статистическим анализом. А пока просто отметим, что существуют различные способы достижения приблизительной сопоставимости экспериментальной и контрольной групп, а также корректировки известных или смоделированных различий между группами.

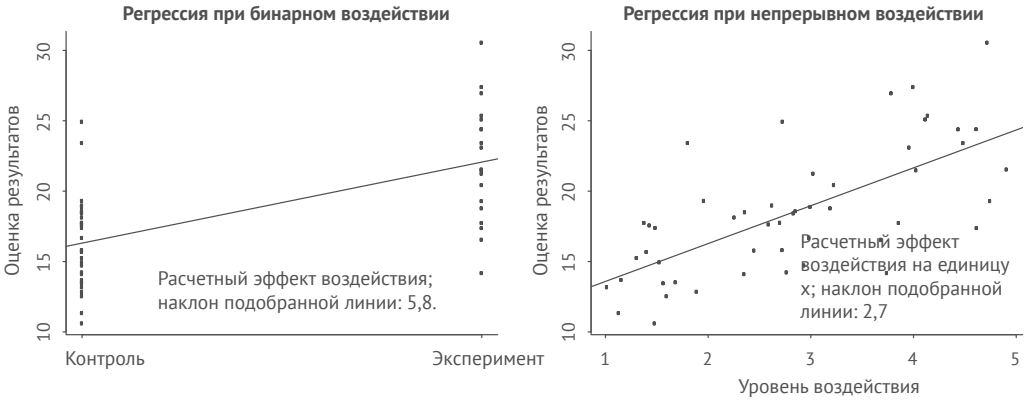


Рис. 1.5. Регрессия для оценки причинно-следственного эффекта с (а) простым сравнением экспериментальной и контрольной группы или (б) в пределах диапазона уровней воздействия

Если нас интересует влияние некоторого воздействия *x* на выход *y* и наши данные получены из рандомизированного или иным образом сбалансированного эксперимента, мы можем применить регрессию, т. е. модель, которая предсказывает *y* по *x* с учетом погрешности.

Если параметр *x* является *бинарным* (*x* = 0 для контроля или *x* = 1 для воздействия), то регрессия особенно проста и выглядит, как показано на рис. 1.5а. Но та же идея верна и для непрерывного предсказателя, который показан на рис. 1.5б[1].

Пример: гипотетические линейные и нелинейные модели

В этой главе мы предполагаем сопоставимость групп, предназначенных для разных видов воздействия, так что регрессионный анализ, прогнозирующий результат данного воздействия, дает нам прямую оценку причинного эффекта. Опять же, мы откладываем до части V обсуждение того, какие допущения – как математические, так и практические – необходимы для того, чтобы эта простая модель давала осмысленный причинный вывод.

Но, если отбросить эти сомнения, мы можем продолжить разработку модели различными способами, чтобы лучше подогнать ее к данным и делать более точные прогнозы. Одним из направлений является рассмотрение нелинейного моделирования эффекта непрерывного воздействия. На рис. 1.5б дана линейная модель, на рис. 1.6а показан пример лежащего в ее основе нелинейного эффекта, а рис. 1.6б ил-

[1] Код для этого примера находится в папке SimpleCasual.

люстрирует, что произойдет, если эту кривую вытянуть в прямую линию. Другое важное направление – моделирование *взаимодействий* – эффектов воздействия, которые варьируются как функция от других предикторов модели. Например, на рис. 1.7 показано предполагаемое влияние радона на уровень заболеваемости раком легких у мужчин. Радон вызывает рак (или, точнее, увеличивает вероятность рака), причем этот эффект сильнее среди курильщиков, чем среди некурящих. В этой модели (которая представляет собой обобщение данных из научных статей и не является результатом подгонки к какому-либо одному набору данных) действие радона предполагается линейным, но взаимодействующим с курением.

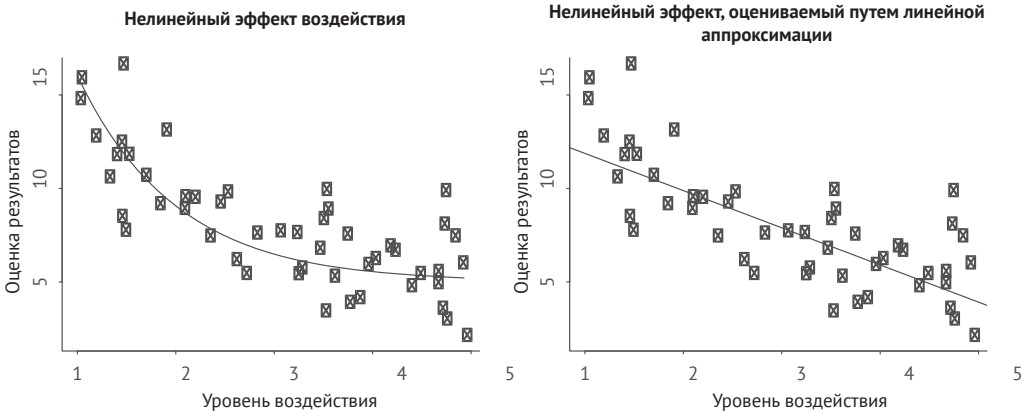


Рис. 1.6. (а) Гипотетические данные, в которых причинный эффект является нелинейной функцией уровня воздействия; (б) те же данные с оцененным линейным эффектом. Всегда можно оценить линейную модель, даже если она не соответствует данным

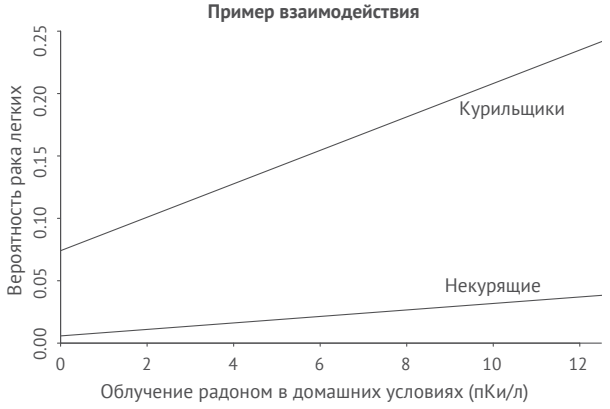


Рис. 1.7. Дополнительный риск возникновения рака легких в течение жизни для мужчин в зависимости от среднего облучения радоном в пикокюри на литр (пКи/л). Связь между заболеваемостью раком и радоном различается для курильщиков и некурящих

Пример: радон, курение и рак легких

Взаимодействие может иметь большое значение, и мы обсуждаем его на протяжении всей книги. Если нас интересует эффект воздействия, нас должно интересовать и то, как этот эффект меняется. Такие изменения могут быть важны по практическим соображениям – например, при принятии решения о том, имеет ли смысл назначать какую-либо дорогостоящую медицинскую процедуру, или кто подвергается наибольшему риску из-за проблем с экологией, – или для научных целей.

Применение регрессии для корректировки различий между экспериментальной и контрольной группами

В большинстве реальных задач причинно-следственного вывода существуют систематические различия между объектами эксперимента и контрольными объектами. Может случиться так, что подопытная группа в среднем была хуже, чем контрольная. Или, возможно, в рамках образовательного эксперимента в классах, в которых применялся новый метод обучения, были более мотивированные учителя, чем в тех, которые придерживались старой программы. В таких условиях важно учитывать различия между группами до воздействия, и для этого мы можем использовать регрессию.

Пример: корректировка гипотетической причинно-следственной связи

На рис. 1.8 показаны некие гипотетические данные с подогнанной линейной регрессией[1]. Ключевое отличие от рис. 1.5 и 1.6 состоит в том, что в этом случае на оси *x* представлены *значения предиктора до воздействия*, а не уровень этого воздействия.

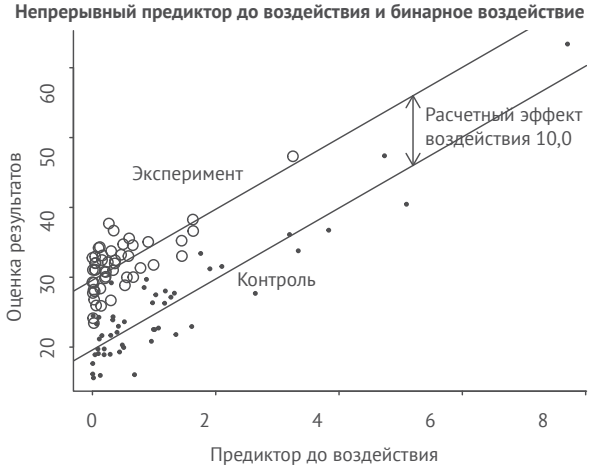


Рис. 1.8. Гипотетические данные результатов бинарного воздействия с непрерывным значением параметра до воздействия. На диаграмме рассеяния экспериментальные объекты обозначены кружками, а контрольные объекты – точками. На них наложена линия регрессии, прогнозирующая результат для данного воздействия и основной переменной, при этом предполагаемый эффект воздействия представляет собой разницу между двумя линиями

[1] Код для этого примера находится в папке SimpleCausal.

Корректировка основных переменных особенно важна при наличии *дисбаланса*, когда между экспериментальной и контрольной группой до воздействия существует различие по ключевым предикторам. Такая корректировка будет зависеть от модели – в примере на рис. 1.8 основными предположениями являются линейность и аддитивность – и после обстоятельного анализа должно следовать четкое объяснение последствий любых корректировок.

Например, гипотетический анализ рис. 1.8 можно резюмировать следующим образом:

> В среднем объекты, подвергшиеся воздействию, расположены на 4,8 балла выше, чем контрольные; $\bar{y} = 31{,}7$ для экспериментальной группы и $\bar{y} = 25{,}5$ для контрольной. Но две группы различались по своему предиктору до воздействия: $\bar{x} = 0{,}4$ для экспериментальной группы и $\bar{x} = 1{,}2$ для контрольной. После корректировки этой разницы мы получили расчетный эффект воздействия 10,0.

Для вычисления эффекта воздействия необходимо наличие модели, но суть этого примера заключается в том, что при наличии дисбаланса между экспериментальным и контрольным показателем ключевого предиктора необходимо выполнить определенную корректировку.

Интерпретация коэффициентов прогнозной модели

Пример: заработок и рост человека

При интерпретации регрессионных моделей могут возникнуть проблемы даже в простейшем случае чистого прогнозирования. Рассмотрим следующую модель, подогнанную под данные опроса: *заработок* = 11 000 + 1500 ∗ (*рост* – 60) + *погрешность*, где годовой заработок измеряется в долларах, рост – в дюймах, а ошибка лежит в диапазоне 22 000 (говоря математическим языком, ошибка имеет среднее значение 0 и среднеквадратическое отклонение 22 000). Эта прогнозная модель на самом деле не годится для *прогнозирования*, потому что погрешность модели очень велика: вряд ли кому-то интересен прогноз заработка 25 000 долл. с возможной погрешностью 22 000. Однако регрессия в некоторой степени полезна для *исследования ассоциаций* в том смысле, что она демонстрирует положительный наклон линии (с учетом среднеквадратической ошибки, передающей неопределенность этого наклона). Коэффициенты регрессии можно интерпретировать в качестве вывода о выборке только в той степени, в которой люди в опросе представляют собой репрезентативную выборку изучаемого населения (взрослые жители США в 1990 году); в ином случае лучше всего включить в модель дополнительные предикторы, чтобы преодолеть разрыв между выборкой и *генеральной совокупностью*. Причинная интерпретация регрессии в этом примере кажется очевидной – каждый дополнительный дюйм роста добавляет вам 1500 долл. дохода в год. Но такая интерпретация сомнительна, потому что высокие и низкие люди могут различаться во многих других отношениях: рост не является случайно применяемым экспериментальным воздействием. Более того, рост – это проблемная переменная, причинно-следственные связи которой

следует изучать другими способами, о которых мы поговорим позже. Для этого примера лучше всего подходит категория *наводящих ассоциаций* (exploring assotiations). Наблюдение за закономерностями в данных может навести исследователя на мысль провести дополнительное исследование, чтобы изучить *истинные* причины, по которым высокие люди зарабатывают больше, чем люди меньшего роста.

Построение, интерпретация и проверка регрессионных моделей

Статистический анализ состоит из четырех этапов. Это:

- *построение модели*, начиная с простых линейных моделей вида $y = a + bx + error$, и последующее расширение за счет дополнительных предикторов, взаимодействий и преобразований;
- *подгонка модели*, которая включает манипулирование данными, программирование и использование алгоритмов для оценки коэффициентов регрессии и их погрешностей, а также для создания вероятностных прогнозов;
- *проверка соответствия модели данным*, что включает в себя построение графиков, дополнительное программирование и активное исследование (несовершенных) связей между измерениями, параметрами и лежащими в их основе объектами исследования;
- *критика модели*, которая заключается не только в выявлении недостатков и сомнительных предположений, но также в рассмотрении направлений улучшения моделей. Или, по крайней мере, ограничение разумными рамками утверждений, которые могли бы быть сделаны при буквальном толковании выводов соответствующей модели.

Следующим шагом является возвращение к этапу построения модели, возможно, с добавлением новых данных и уточнением модели.

Проблема серьезной прикладной работы состоит в том, *как быть критичным, не будучи нигилистом*, и признать, что мы способны сделать выводы из статистического анализа – мы можем делать обобщения от выборки к генеральной совокупности, от экспериментальной группы к контрольной и от наблюдаемых измерений к исследуемым конструктам, – даже если эти выводы вызывают опасения.

Ключевым шагом в критике выводов исследования – и понимании границ такой критики – является выполнение действий, которые связывают более обобщающие утверждения с данными и статистическим анализом. Одним из слабых мест исследования влияния законов по контролю оборота огнестрельного оружия является тот факт, что были сделаны далеко идущие выводы относительно предлагаемых изменений в законах, но сравнения проводились между разными штатами без учета данных о том, какие законы там были приняты или отвергнуты. И наоборот, анализ взаимосвязи роста и заработка был более описательным и не предполагал зависимости от политических изменений. Еще одна проблема, связанная с исследованием законов об огнестрель-

ном оружии, заключалась в том, что предполагаемые эффекты были чересчур значительными и вели к сокращению смертности в пять раз. Это признак чрезмерной интерпретации зашумленных данных. В данном случае исследовали наблюдали существующие различия между состояниями и слишком охотно связывали их с доступными факторами. С таким же успехом можно попытаться сопоставить смертность от огнестрельного оружия с различными законами, регулирующими переработку куриного мяса, и найти подходящие корреляции, из которых вывести причинно-следственную связь. В отличие от этих сомнительных примеров исследование миротворческой деятельности является более строгим – рассмотрение только одного воздействия, а не попытка рассмотреть сразу 25 факторов – и более открытым в отношении вариаций. Смысл рис. 1.4 состоит не в том, чтобы утверждать, что миротворчество дает какой-то конкретный эффект, а в том, чтобы показать, что оно связано с задержкой возврата к гражданской войне по сравнению с аналогичными ситуациями в странах, в которые не вмешивались миротворцы ООН.

Ни одно исследование не идеально. В анализе Xbox мы использовали нерепрезентативную выборку, чтобы сделать вывод о генеральной совокупности избирателей. Исследование Electric Company было контролируемым экспериментом, поэтому нас мало беспокоят различия между экспериментальной и контрольной группой, но следует задуматься о целесообразности обобщения результатов эксперимента с телешоу на всю страну. Речь идет о том, что мы должны уметь распознавать проблемы при экстраполяции, а затем работать над их устранением. В случае исследования Xbox мы использовали регрессию для моделирования мнения как функции демографических переменных, таких как возраст, пол и образование, где выборка отличалась от популяции; данные компании Electric были проанализированы отдельно для каждого года обучения, что дает дополнительное представление о различиях в эффекте воздействия.

1.5. Классический и байесовский вывод

Как статистики, мы тратим большую часть наших усилий на подгонку моделей к данным и использование этих моделей для составления прогнозов. Эти шаги можно выполнять в рамках различных методологических и философских подходов. Общими для всех этих подходов являются три ключевых аспекта: (1) какая *информация* используется в процессе построения модели, (2) какие *предположения и допущения* мы при этом делаем и (3) какой *подход к интерпретации* вывода модели мы применяем – классический или байесовский. Рассмотрим их по очереди.

Информация

Отправной точкой для любой задачи регрессии являются данные о выходной переменной y и одном или нескольких предикторах x. Когда данные являются непрерывными и имеется один прогнозирующий

фактор, они могут отображаться в виде диаграммы рассеяния, как на рис. 1.5 и 1.6. Когда есть один непрерывный и один бинарный предиктор, данные могут быть представлены в виде диаграммы рассеяния с двумя разными символами, как на рис. 1.8. В более общем случае не всегда возможно представить все данные на одном графике.

Помимо самих данных, мы обычно что-то знаем и о том, как они были собраны. Например, анализируя данные социологического опроса, мы можем посмотреть на вопросы и получить представление о том, как их задавали, где и когда проводились интервью. Если данные представляют собой лабораторные анализы, мы можем оценить систематические ошибки и изменчивость результатов измерений и т. д.

Также должна быть доступна информация о том, какие данные вообще наблюдались. В опросе респонденты могут быть случайной выборкой из четко определенной совокупности (например, сформированной путем извлечения случайных имен из списка) или они могут быть так называемой *удобной выборкой*[1], и в этом случае мы должны иметь некоторое представление о том, какие типы людей вошли в эту выборку с большей или меньшей степенью вероятности. В эксперименте воздействие может применяться к объектам случайно или по определенному правилу, и в этом случае у нас обычно будет информация о том, как применялось воздействие. Например, если врачи выбирают, какие методы лечения назначать отдельным пациентам, нам будет полезно знать, какие методы рассматривались каждым врачом и какие характеристики пациентов повлияли на решение о выборе лечения.

Наконец, у нас обычно есть предварительные знания, полученные из источников, отличных от источника имеющихся данных, на основе опыта предыдущих аналогичных исследований. Мы должны использовать такую информацию с осторожностью. Например, опубликованные научные отчеты склонны преувеличивать размеры эффекта, поскольку существует определенная научная предвзятость, из-за которой исследователи вынуждены публиковать в основном большие и «статистически значимые» результаты; об этом пойдет речь в разделе 4.5. Однако бывают ситуации, когда локальных данных недостаточно, и было бы глупо делать выводы, не используя предварительных знаний. В разделе 9.4 мы приводим пример связи между параметрами родителей и полом их детей.

Предположения и допущения

Есть три вида предположений, которые лежат в основе любой регрессионной модели, генерирующей выход y исходя из предикторов x. Во-первых, это форма связующей функции между x и y: обычно мы предполагаем линейную функцию, но эта связь более гибкая, чем может показаться, поскольку мы можем выполнять преобразования пре-

[1] Выборка, сформированная из соображений удобства исследования. Вопреки расхожему мнению, удобная выборка не является синонимом нерепрезентативной выборки. – *Прим. перев.*

дикторов или выхода, а также комбинировать предикторы линейным или нелинейным образом (об этом говорится в главе 12). Тем не менее выбор преобразований, а также выбор переменных, которые следует включить в модель в первую очередь, опираются на предположения о взаимосвязях между различными изучаемыми переменными.

Второй набор предположений включает в себя источник данных: какие потенциальные наблюдения видны, а какие нет, кого опрашивают, а кого пропускают, кто подвергается экспериментальному воздействию и т. д. Эти предположения могут быть сильными – например, строгое предположение о случайной выборке или случайном применении воздействия – или более гибкими, например допускающими, чтобы вероятность участия в опросе была различной для мужчин и женщин и варьировалась в зависимости от этнической принадлежности и образования, или допускающими разную вероятность экспериментального воздействия в зависимости от возраста и предыдущего состояния здоровья. Самые сильные предположения, такие как случайное распределение, обычно просты и понятны, тогда как слабые предположения, будучи более общими, обычно оказываются более сложными.

Третья разновидность предположений относится к актуальности измеренных данных в реальном мире: точны ли ответы на опрос, можно ли обобщить поведение объектов в лабораторном эксперименте на внешний мир, предсказывают ли сегодняшние измерения то, что может произойти завтра? Эти вопросы можно изучать статистически, сравнивая стабильность наблюдений, проводимых разными способами или в разное время, но в контексте регрессии они обычно принимаются как само собой разумеющееся. Интерпретация регрессии y от x зависит также от отношения между измеренным x и фактическими предикторами, на которые мы обратили внимание, а также от отношения между измеренным y и фактическим результатом.

Классический вывод

Традиционный подход к статистическому анализу основан на обобщении информации в данных без использования априорной информации. Вместо нее используют оценки и прогнозы, которые имеют хорошо понятные статистические свойства, низкую систематическую ошибку и низкую дисперсию. Такой подход иногда называют «частотным», поскольку классический исследователь или аналитик интересуется долгосрочными ожиданиями своих методов – оценки должны быть в среднем правильными (*несмещенность*), истинное значение параметра должно входить в доверительный интервал не реже, чем в 95 % случаев (*охват*). Важным принципом классической статистики является *консерватизм*: иногда данные оказываются слабыми, и мы не можем делать убедительных заявлений, но мы хотели бы иметь возможность сказать, по крайней мере приблизительно, что наши оценки беспристрастны и наши интервалы имеют заявленный охват. В классической статистике должен быть четкий и однозначный («объективный») путь от данных к выводам,

который в свою очередь должен быть проверяемым, по крайней мере теоретически, на основе частотных свойств этих данных.

Классическая статистика может многое предложить, и есть искушение обобщать информацию только на основе данных. Слабые стороны классического подхода проявляются, когда исследования небольшие, а данные косвенные или сильно изменчивые. Проиллюстрируем это на примере.

Пример: влияние на развитие детей на Ямайке

В 2013 году группа экономистов опубликовала исследование, в котором сообщалось о «большом влиянии на последующие заработки участников рандомизированного эксперимента, направленного на психосоциальную стимуляцию отстающих в развитии ямайских детей из бедных семей. Экспериментальное воздействие заключалось в одночасовых еженедельных визитах местных ямайских социальных работников в течение 2 лет. Мы повторно опросили участников исследования спустя 20 лет после воздействия». По оценкам исследователей, воздействие увеличило доход участников эксперимента на 42 % при 95%-м доверительном интервале эффекта от воздействия, который мы запишем как [+2 %, +98 %]. То есть оценка, основанная только на данных, состоит в том, что внешнее влияние умножает средний заработок на коэффициент 1,42 с 95%-м доверительным интервалом [1,02, 1,98] для данного коэффициента (упражнение 3.8).

Погрешность здесь весьма велика, что неизбежно, поскольку оценка основана на сравнении доходов только 127 детей, зарплаты которых, когда они выросли, сильно различаются. С точки зрения классического вывода нет ничего плохого в таком широком интервале – если бы эта же статистическая процедура применялась снова и снова ко многим различным задачам, полученные доверительные интервалы содержали бы истинные значения параметров в 95 % случаев (охватывая любые неточности в данных и протоколах экспериментов). Однако мы реалисты и понимаем, что эти интервалы с большей вероятностью попадут в отчеты, если они исключают ноль, и поэтому мы *не* ожидаем, что они будут охватывать 95 % случаев в реальном мире (упражнения 5.8 и 5.9). Но, возможно, более важно то, что некоторые значения в этом интервале гораздо более *правдоподобны*, чем другие: визиты соцработника вполне могут давать отдачу 2 или даже 0 %, но очень маловероятно, что они часто приносят отдачу в 98 % и фактически удваивают заработки людей. Мы можем утверждать это исходя из предшествующих знаний или общего понимания того, как устроено общество. Откровенно говоря, мы не доверяем оценке в 42 %: если бы исследование собирались повторить и нам предложили сделать ставку на то, будет ли результат больше или меньше 42 %, мы бы с уверенностью сделали ставку на «меньше». Это не означает, что исследование бесполезно, просто на основании этих данных мало что можно узнать об эффекте визитов соцработника в раннем детстве.

Байесовский вывод

Байесовский вывод – это подход к статистике, который включает в вывод априорную информацию, выходя за рамки простого обобщения существующих данных. Например, в случае влияния визитов соцработ-

ника на развитие младенцев можно было бы начать с предположения, что воздействие может иметь значение, но что средний эффект, скорее всего, будет менее 10 % в положительном или отрицательном направлении. Мы можем использовать эту информацию в качестве *априорного* распределения вероятностей, предположив, что эффект мультипликативного воздействия будет лежать в диапазоне [0,9, 1,1]; комбинируя это с данными и используя правила байесовского вывода, мы получаем 95%-ный *апостериорный* интервал [0,92, 1,28] – вариацию от 8 % отрицательного эффекта до 28 % положительного; подробности вы найдете в упражнении 9.6. Исходя из этого байесовского анализа, наше лучшее предположение о наблюдаемой разнице в будущем репликационном исследовании будет намного ниже 42 %.

Этот простой пример иллюстрирует как сильные, так и слабые стороны байесовского вывода. С одной стороны, байесовский анализ дает более разумные результаты и может использоваться для прямого прогнозирования результатов будущих экспериментов. С другой стороны, требуется дополнительная информация – некое «предварительное распределение», которое в данном случае основано на спорном утверждении о том, что влияние экспериментального воздействия на будущие доходы составляет менее 10 %. Но в любом случае этой субъективности не избежать: в байесовском выводе априорное распределение представляет собой арену, на которой будут критически оцениваться любые прогнозы. В мире, в котором визиты соцработника способны удвоить будущий средний заработок, грубая оценка коэффициента 1,42 и интервал [1,02, 1,98] дают разумные прогнозы. Но в мире, где такие огромные последствия маловероятны, мы должны соответствующим образом скорректировать наши ожидания и прогнозы.

Итак, у нас есть выбор: классический вывод, ведущий к чистым сводкам данных, которые могут иметь ограниченную ценность в качестве прогнозов; или байесовский вывод, который теоретически может давать достоверные прогнозы даже при слабых данных, но основан на дополнительных предположениях. Здесь нет универсально правильного ответа; мы должны уметь трезво оценивать свои возможности.

Существует также практическое преимущество байесовского подхода, состоящее в том, что все его выводы являются вероятностными и, следовательно, могут быть представлены симулятором вероятностей. По этой причине всякий раз, когда мы хотим суммировать погрешность оценок, лежащих за пределами простых доверительных интервалов, и всякий раз, когда хотим использовать регрессионные модели для прогнозов, мы используем байесовский подход. Как будет показано в главе 9, мы можем выполнить байесовский вывод с использованием неинформативных или малоинформативных априорных вероятностей и получить результаты, аналогичные классическим оценкам, наряду с результатами моделирования, которые можно использовать для выражения прогнозной погрешности, или мы можем использовать информативные априорные значения, когда это необходимо.

Если у нас есть подходящая информация, которой *нет в модели* (например, осведомленность о предвзятости, неизмеренные характеристики объекта, априорная информация об отклике на воздействие и т. д.), то мы обязаны учитывать ее так же тщательно, как и при интерпретации наших статистических сводок данных.

1.6. Вычисление наименьших квадратов и байесовской регрессии

Для построения графиков и вычисления сводок данных, подгонки статистических моделей и моделирования демонстрационных данных из теоретических или подогнанных моделей мы используем код на языке R. Мы вводим этот код в книгу по мере необходимости, а познакомиться с основами R вы можете в приложении А.

В целом мы рекомендуем использовать для регрессии байесовский вывод: если доступна априорная информация, вы можете ее использовать, а в противном случае байесовская регрессия со слабо информативными априорными значениями по умолчанию все равно имеет преимущество, поскольку дает стабильные оценки и позволяет строить модели, дающие логический вывод с прогнозной погрешностью (т. е. оценки с погрешностями и вероятностными предсказаниями или прогнозами). Например, в модели выборов, представленной в разделе 1.2, к которой мы вернемся в главе 7, моделирование на основе подогнанной байесовской модели выявляет погрешность расчетных коэффициентов регрессии и позволяет нам вычислять вероятностные прогнозы для будущих выборов, обусловленные предположениями о состоянии экономики в год выборов.

Вы можете выполнить подгонку байесовской регрессии в R, используя команды вида

```
fit <- stan_glm(y ~ x, data=mydata)
```

Но некоторым пользователям статистических методов может быть незнаком или неудобен байесовский вывод. Если вы один из них или вам приходится общаться с людьми, которым более удобна классическая статистика, можете выполнить подгонку модели по методу наименьших квадратов:

```
fit <- lm(y ~ x, data=mydata)
```

Наконец, еще одна проблема, связанная с методом `stan_glm`, заключается в том, что он может медленно работать с большими объемами данных. Мы можем заставить его работать быстрее, запустив в режиме оптимизации:

```
fit <- stan_glm(y ~ x, data=mydata, algorithm="optimizing")
```

Наборы данных для примеров из этой книги небольшие, и скорость на самом деле не имеет значения, но полезно помнить об этой опции,

приступая к решению более сложных задач. При запуске в режиме оптимизации `stan_glm` выполняет приблизительную подгонку, но попрежнему поддерживает имитационный подход, применяемый для суммирования логической и прогнозной погрешности.

Если вы предпочитаете избегать байесовского вывода, можете заменить большинство экземпляров `stan_glm` в этой книге на `lm` для линейной регрессии или `glm` для логистических и обобщенных линейных моделей и получить почти идентичные результаты. Заметные различия проявляются в примере в разделе 9.5 с сильным априорным распределением, в примерах логистической и упорядоченной логистической регрессии с полным разделением в разделах 14.6 и 15.5, в нашей реализации перекрестной проверки в разделе 11.8 и в различных примерах по всей книге, где мы используем имитацию данных для внесения погрешности в оценки или прогнозы. Также вы можете подогнать регрессию методом наименьших квадратов и получить байесовские погрешности, запустив `stan_glm` с при плоских априорных распределениях (flat prior distributions), которые мы обсудим в разделе 8.4.

Значение байесовского и имитационного подхода возрастает при подгонке регуляризованной регрессии и многоуровневых моделей. Эти темы выходят за рамки данной книги, но как только вы научитесь использовать имитационный метод для преодоления погрешности, у вас будут хорошие возможности для изучения этих более продвинутых моделей и работы с ними.

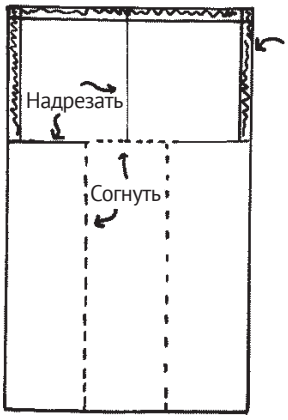
Мы вернемся к примеру прогнозирования выборов в главе 7, а также подробно обсудим регрессию роста–заработка, опрос Xbox и эксперимент Electric Company в главах 6, 17 и 19 соответственно.

1.7. Упражнения

Данные для примеров и упражнений в этой и других главах можно найти на сайте по адресу https://avehtari.github.io/ROS-Examples/. В приложении A представлен краткий обзор языка R и программного обеспечения, которое вы будете использовать для вычислений.

Пример: конструирование вертолета

1.1. *От проектирования до принятия решения*. На рис. 1.9 показан прототип бумажного «вертолета». Цель этого задания – спроектировать вертолет, которому требуется как можно больше времени, чтобы достичь пола при падении с фиксированной высоты, например 2,5 м. Вертолеты должны иметь только одинаковую базовую конструкцию, показанную на эскизе. Никаких дополнительных складок, изгибов или перфораций не допускается. Длина корпуса и ширина лопасти вертолета – только эти два конструктивных параметра вертолета можно менять. Ширина и длина основания должны оставаться одинаковыми для всех вертолетов. К нижней части вертолета прикреплена металлическая скрепка.



Helicopter_ID	width	length	time
1	1.8	3.2	1.64
1	1.8	3.2	1.74
1	1.8	3.2	1.68
1	1.8	3.2	1.62
1	1.8	3.2	1.68
2	1.8	3.2	1.62
2	1.8	3.2	1.65
2	1.8	3.2	1.66
2	1.8	3.2	1.63
2	1.8	3.2	1.66

Рис. 1.9. (а) Схема изготовления «вертолета» из половины листа бумаги и скрепки. Длинные сегменты слева и справа загибаются к середине, и полученная длинная трехслойная полоса скрепляется скрепкой. Один из двух верхних сегментов загибается вперед, а другой – назад. При падении вертолет вращается в воздухе. (б) Файл данных, показывающий время полета в секундах для 5 полетов каждого из двух идентичных вертолетов с шириной лопасти 1,8 дюйма (~4,6 см) и длиной лопасти 3,2 дюйма (8,1 см), сброшенных с высоты примерно 2,5 м. Авторы Гельман и Нолан (2017).

Вот несколько комментариев студентов, которые выполняли это задание.

> Рич слишком сильно согнул лопасти, и вертолеты падали камнем, переворачивались вверх тормашками, боком и т. д.

> Эти вертолеты сильно реагируют на увеличение длины лопасти. И похоже, что слишком большая ширина делает вертолет неустойчивым. Он переворачивается во время полета.

> Энди предлагает сделать жесткий шаблон для складывания основания втрое.

> После небольшой тренировки мы решили сменить подход. Лучше всего получилось, когда Юй давал отсчет, а Джон отпускал вертолет. 3 - 2 - 1 - ПУСК.

Раздайте каждой группе студентов 25 половинок листа бумаги и 2 скрепки. Ширина корпуса будет составлять одну треть ширины листа, поэтому ширина лопасти может составлять от 1/6 до 1/2 ширины корпуса, как показано на рис. 1.9а. Длина корпуса определяется преподавателем.

Например, если листы бумаги А4 имеют стандартный для США размер 8,5×5,5 дюймов (21,59×13,97 см), а длина корпуса составляет 7,62 см, то ширина лопасти может составлять от 2,31 до 6,98 см, а длина – от 0 до 13,97 см.

В этом задании вы можете поэкспериментировать с 25 листами и 10 скрепками. Из каждой половинки листа можно сделать только один вертолет. Но вы можете последовательно подбирать параметры конструкции, задав для каждого вертолета ширину лопасти и длину корпуса с учетом уже записанных вами данных. Сделайте несколько

измерений для каждого вертолета, сбрасывая его с постоянной высоты и рассчитывая время, необходимое для приземления.

(а) Запишите ширину лопасти и длину корпуса для каждого из ваших 25 вертолетов вместе с вашими измерениями времени в один файл, в котором каждое наблюдение находится в отдельной строке в соответствии с шаблоном Helicopters.txt в папке Helicopters, также показанном на рис. 1.9б.

(б) Графически представьте данные так, как сочтете нужным.

(в) Исходя из полученных результатов предложите такую конструкцию (ширину и длину лопасти), которая, по вашему мнению, позволит достичь максимального ожидаемого времени полета вертолета. Вам не обязательно использовать здесь формализованную регрессионную модель, но вы должны использовать общий регрессионный подход.

Приведенное выше описание адаптировано из работы Гельмана и Нолана (2017, раздел 20.4).

1.2. *Построение регрессионной модели и данных.* На рис. 1.1б показаны данные, соответствующие подогнанной линии регрессии $y = 46{,}3 + 3{,}0x$ с остаточным стандартным отклонением 3,9 и значениями x в диапазоне примерно от 0 до 4 %.

(а) Приблизительно изобразите на графике гипотетические данные с тем же диапазоном x, но соответствующие линии $y = 30 + 10x$ с остаточным стандартным отклонением 3,9.

(b) Приблизительно изобразите на графике гипотетические данные с тем же диапазоном x, но соответствующие линии $y = 30 + 10x$ с остаточным стандартным отклонением 10.

1.3. *Цели регрессии.* Загрузите какие-либо данные по интересующей вас теме. Не изображая эти данные на графиках и не выполняя статистический анализ, обсудите, как вы могли бы использовать эти данные для следующих целей.

(а) Подбор регрессии для оценки интересующей взаимосвязи.

(б) Использование регрессии для корректировки различий между экспериментальной и контрольной группами.

(в) Использование регрессии для прогнозирования.

1.4. *Задачи статистики.* Приведите примеры применения прикладной статистики в интересующей вас области, в которых приходится решать следующие задачи.

(а) Обобщение выборки на генеральную совокупность.

(б) Обобщение вывода экспериментальной группы на контрольную.

(в) Обобщение наблюдаемых измерений на неявные исследуемые конструкты.

Объясните свои ответы.

1.5. *Цели регрессии.* Приведите примеры применения прикладной статистики в интересующей вас области, в которых цели заключаются в следующем.

(а) Прогнозирование/классификация.
(б) Обнаружение взаимосвязей.
(в) Экстраполяция.
(г) Причинный вывод.

Объясните свои ответы.

1.6. *Причинно-следственный вывод.* Найдите реальный и заслуживающий внимания пример с экспериментальной группой, контрольной группой, предиктором до воздействия и предиктором после воздействия. Постройте график, подобный рис. 1.8, используя данные из этого примера.

1.7. *Статистика как обобщение.* Найдите опубликованную статью по интересующей вас теме, в которой, по вашему мнению, уделено недостаточно внимания:

(а) обобщению выборки на генеральную совокупность;
(б) обобщению экспериментальной группы на контрольную;
(в) обобщению наблюдаемых измерений на неявные конструкты.

Объясните свои ответы.

1.8. *Статистика как обобщение.* Найдите опубликованную статью по интересующей вас теме, в которой, по вашему мнению, были хорошо освещены следующие вопросы:

(а) обобщение выборки на генеральную совокупность;
(б) обобщение экспериментальной группы на контрольную;
(в) обобщение наблюдаемых измерений на неявные конструкты.

Объясните свои ответы.

1.9. *Проблема линейных моделей.* Вернемся к эксперименту по проектированию вертолета в упражнении 1.1. Предположим, вы должны изготовить 25 вертолетов, измерить время их падения, подогнать линейную модель, предсказывающую этот результат с учетом ширины крыла и длины корпуса:

$$время = \beta_0 + \beta_1 * ширина + \beta_2 * длина + погрешность,$$

а затем использовать подогнанную модель $время = \beta_0 + \beta_1 * ширина + \beta_2 * длина$, чтобы оценить значения ширины лопасти и длины корпуса, которые максимально увеличивают ожидаемое время полета.

(а) Почему этот подход потерпит неудачу?
(б) Предложите более подходящую модель, в которой нет этой проблемы.

1.10. *Работа над собственным примером.* Найдите и скачайте или соберите данные по интересующей вас теме. Вы будете использовать этот пример для работы с идеями и методами, описанными на протяжении всей книги, поэтому пример стоит потраченного времени и должен быть достаточно сложным. Это задание повторяется на протяжении всей книги в качестве заключительного упражнения каждой главы. В этом первом упражнении обсудите свои прикладные цели при изучении этого примера и то, как данные могут помочь в достижении этих целей.

Глава 2

Данные и показатели

В этой книге мы будем подгонять прямые линии (и некоторые кривые) к данным, выполнять сравнение и прогнозирование и оценивать степень погрешности полученных выводов. Мы обсудим предположения, лежащие в основе регрессионных моделей, методы проверки этих предположений и направления улучшения подогнанных моделей. Разберем проблемы экстраполяции имеющихся данных на причинно-следственные выводы и прогнозы для новых данных и будем использовать имитационное моделирование, чтобы внести погрешности в наших оценки и прогнозы.

Однако, прежде чем подгонять модель, следует разобраться, откуда берутся ваши числа. В этой главе на примерах показано, как использовать графические инструменты для изучения и понимания данных и показателей.

2.1. Проверка происхождения данных

Пример: Индекс человеческого развития

Рисунок 2.1 несколько лет назад приобрел в интернете вирусную популярность. На карте сравниваются 50 штатов и Вашингтон, округ Колумбия, по так называемому Индексу человеческого развития (ИЧР), который ранее использовался для сравнения различных стран по ряду социально-экономических показателей. У этой карты довольно нелепая визуализация: к штатам с тремя самыми низкими значениями относятся Луизиана (0,801), Западная Вирджиния (0,800) и Миссисипи (0,799), но алгоритм закрашивания выделяет только Миссисипи.

Но у нас есть к карте и более серьезные вопросы, чем раскраска. Неужели Аляска так развита? А что насчет Вашингтона, который, согласно отчету, занимает 4-е место после Коннектикута, Массачусетса и Нью-Джерси?

Пора присмотреться к числам. Согласно опубликованному отчету, ИЧР объединяет три основных показателя:

- *ожидаемую продолжительность жизни* при рождении как показатель здоровья и долголетия населения;
- *знания и образование*, измеряемые уровнем грамотности взрослого населения (весовой коэффициент 2/3) и комбинированным валовым коэффициентом охвата начальным, средним и высшим образованием (весовой коэффициент 1/3);
- *уровень жизни*, измеряемый натуральным логарифмом валового внутреннего продукта (ВВП) на душу населения по паритету покупательной способности (ППС) в долларах США.

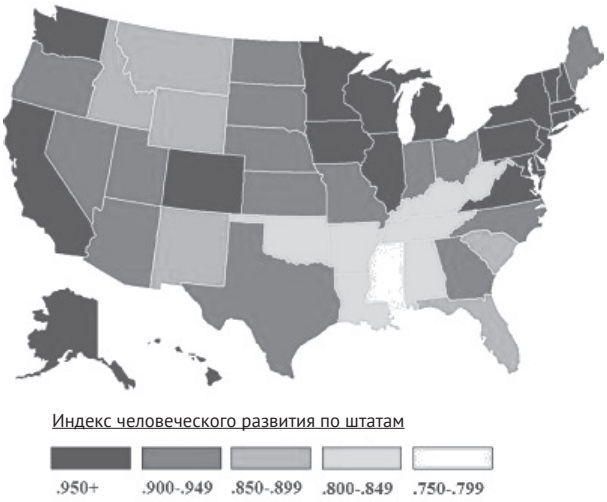


Рис. 2.1. Карта распределения Индекса человеческого развития, представляющая 50 штатов и Вашингтон, округ Колумбия. Источник: PlatypeanArchcow (2009)

Теперь мы можем видеть, что происходит. Ожидаемая продолжительность жизни, грамотность или посещаемость школы не сильно различаются в зависимости от штата. Конечно, гавайцы живут на несколько лет дольше, чем жители Миссисипи, и есть некоторые различия в том, кто продолжает учиться в школе, но, безусловно, самые большие различия между штатами, исходя из этих показателей, заключаются в ВВП. Средний доход в Коннектикуте вдвое больше, чем в Миссисипи. А Вашингтон, округ Колумбия, занимает высокое место, потому что его жители имеют высокий средний доход.

Чтобы проверить связь между ИЧР и доходом, мы загрузили табличные данные ИЧР и нанесли их на график в сравнении с историческими данными о среднем доходе по штатам[1]. На рис. 2.2а показан результат. Картина распределения выраженная, но нелинейная. На рис. 2.2б показаны рейтинги штатов и выявляется четкая закономерность: большинство штатов попадает прямо на 45-градусную линию, и между этими двумя рейтингами наблюдается высокая корреляция. Мы были удивлены, что корреляция не выше – и удивились, что первая диаграмма рассеяния настолько нелинейна, – но, опять же, мы используем доход штата, а не ВВП, и, возможно, с этими данными что-то не так. Нет, дело не в логарифмическом преобразовании, по крайней мере, если вы регистрируете доход, как указано в отчете. Логарифмирование немного расширяет нижнюю границу шкалы, но не меняет общей картины графика. Значения дохода имеют не настолько широкий динамический диапазон, чтобы логарифмическое представление дало ощутимый эффект.

[1] Данные и код для этого примера находятся в папке HDI.

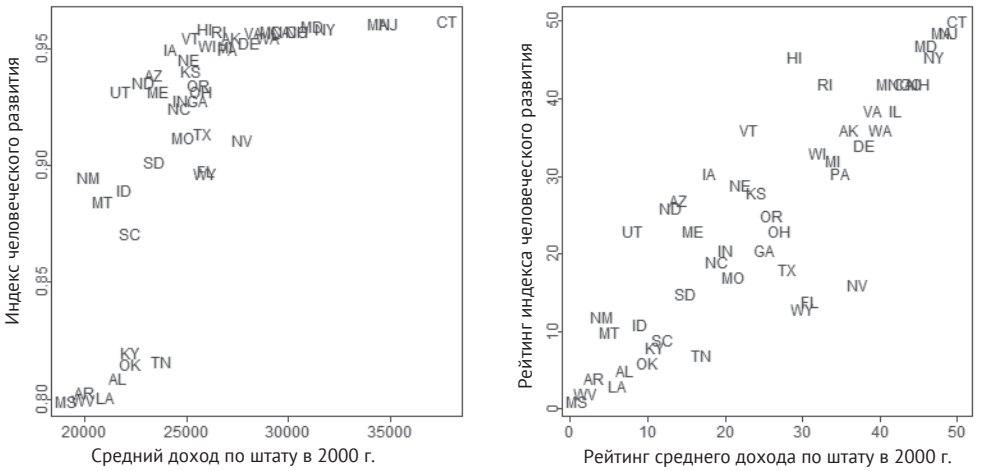


Рис. 2.2. График индекса человеческого развития в сравнении со средним доходом по штатам: (а) диаграмма распределения данных, (б) диаграмма распределения рейтингов

Возможно, с этими данными проделали более серьезные манипуляции, чем мы думаем. Если кому-то интересно проследить за этим, мы предлагаем изучить Южную Каролину (SC) и Кентукки (KY), которые так близки по среднему доходу и так сильно отличаются по уровню ИЧР (рис. 2.2а).

В любом случае карта на рис. 2.1 в значительной степени представляет собой карту распределения доходов штата со странно преобразованными данными и запоминающимся названием. Актуальность этого примера заключается в том, что мы смогли лучше понять данные, построив диаграмму их распределения разными способами.

Особенности показателя могут быть важны

Пример: политическая идеология и принадлежность к партии

В американской политике есть две основные партии, и большинство избирателей попадает в идеологический диапазон от левых (либералов) до правых (консерваторов). Раскол между демократами и республиканцами примерно совпадает с разделением на либералов и консерваторов.

Но эти две шкалы принадлежности к партии и идеологии не идентичны[1]. На рис. 2.3а видно, что доли либералов, умеренных и консерваторов приблизительно равны для всех уровней доходов. Напротив, на рис. 2.3б видна сильная связь между доходом и принадлежностью к республиканской партии, по крайней мере, по состоянию на 2008 год, когда были собраны эти данные опроса. Доли принадлежащих к партии и политической идеологии оценивались по пятибалльной шкале

[1] Данные и код для этого примера находятся в папке Pew.

слева направо, но, как показывают графики, между двумя переменными есть явные различия.

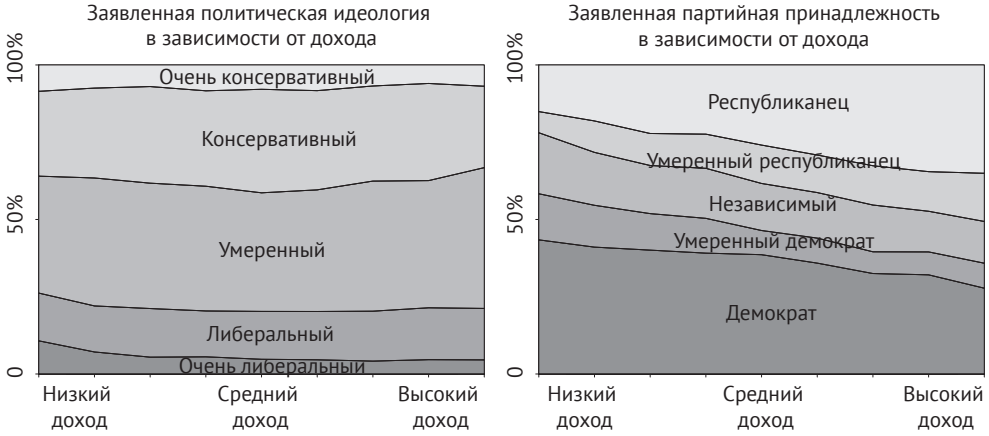


Рис. 2.3. Распределение (а) приверженности к политической идеологии и (б) партийной принадлежности в зависимости от дохода по данным опроса, проведенного во время избирательной кампании в США в 2008 году

Как рис. 2.3 соотносится с общей темой нашей книги? Регрессия – это способ подвести итоги и сделать выводы из данных. Следовательно, выводы из регрессий будут зависеть от качества анализируемых данных и их соответствия задаваемым вопросам. Пример принадлежности к партии и идеологии служит напоминанием о том, что даже очень похожие измерения могут дать разные ответы в зависимости от вопроса.

К сожалению, разрыв между показателем и реальностью – общая проблема научных исследований и коммуникации. Например, медицинская школа Университета Темпл выпустила пресс-релиз под названием «Оливковое масло первого отжима сохраняет память и защищает мозг от болезни Альцгеймера», но фактическое исследование проводилось на мышах и не имело прямой связи с деменцией или болезнью Альцгеймера. Таким образом, утверждение не имеет реального обоснования (раздел 2.2). Такой скачок умозаключений случается постоянно. В некотором смысле это необходимо – лабораторные эксперименты предшествуют клиническим испытаниям, – но мы должны быть честными и отдавать себе отчет в том, что мы на самом деле знаем.

2.2. Достоверность и надежность

Мы обсуждаем важный вопрос показателей по двум причинам. Во-первых, нам нужно понять, что на самом деле означают наши данные. Мы рассмотрели способы визуализации данных и извлечения информации. Но если мы не знаем, что на самом деле представляют данные, то не сможем извлечь нужную информацию.

Анализ данных заходит в тупик, если у нас плохие данные. Существуют проблемы показателей, которые невозможно решить никаки-

ми исправлениями и настройками. В разделе 1.3 мы рассказали о том, как внесли изменения в данные опроса Xbox, чтобы учесть различия между выборкой и генеральной совокупностью. Но если бы мы задали нашим респондентам неправильный вопрос или не записали бы ключевые базовые переменные, которые можно использовать для корректировки, то простого исправления не получилось бы.

Вторая причина заключается в том, что изучение точности, надежности и достоверности данных создаст основу для изучения понятий дисперсии, корреляции и ошибки, которые понадобятся при построении линейных моделей в следующих главах.

Большинство из нас не особо задумывается о повседневных измерениях показателей, в первую очередь потому что мы принимаем как должное показатели, которыми пользуемся, и даже если мы знаем, что есть проблемы с точностью, обычно она достаточно хороша для наших целей. Поэтому у нас не возникает проблем с измерением температуры на улице, веса продуктов, скорости автомобиля и т. д. Мы считаем само собой разумеющимся соответствие между числом и предметом измерения. И обычно нас не волнует точность – в быту нам не нужна температура с точностью до половины градуса или скорость машины с точностью до шести знаков после запятой.

Все это зависит от того, что мы измеряем и каковы предполагаемые выводы. Весы, которые измеряют вес с точностью до 1 кг, отлично подходят для взвешивания слонов, приемлемы для взвешивания взрослых людей и совершенно не подходят для взвешивания драгоценностей. Понятие «достаточной точности» – это сочетание свойств весов и цели, для который мы собираемся их использовать.

В социальных науках способ измерения того, что мы пытаемся измерить, не так очевиден, как в повседневной жизни. Иногда это происходит потому, что объект измерения «реален» и хорошо определен, но плохо поддается подсчету. Примерами могут служить подсчет количества иммигрантов или измерение ежедневного потребления пищи в неконтролируемых условиях.

В других случаях объект измерения на первый взгляд однозначен, но немного нечетко определен, и способы подсчета не очевидны – например, подсчет количества людей в вашем районе, которых вы помните, или знакомых, которым вы доверяете, или подсчет количества слов, которые вы знаете.

Иногда мы пытаемся измерить то, что, по нашему мнению, имеет значение, но это субъективно для каждого человека и не подразумевает наличия объектов, которые можно посчитать или измерить линейкой. Например, это политические убеждения, намерения голосовать и удовлетворенность клиентов. Во всех этих случаях мы хорошо понимаем, о чем идет речь; это понимание глубоко укоренилось в нашем языке, и мы осознаем, что люди имеют свое мнение о вещах и чувствах. Но это глубоко субъективные вещи; их нельзя просто взвесить или измерить ширину. Поэтому для измерения субъективных показателей каждый раз приходится изобретать специальную шкалу, напри-

мер: «Оцените по шкале от 0 до 100, насколько вам понравилось наше приложение?». Ответ на этот вопрос, безусловно, важен, но мы могли бы задать вопрос по шкале от 1 до 3 или даже от 300 до 500. Мы просто надеемся, что люди будут искренне отвечать на вопрос и что они будут использовать шкалу так, как мы задумали. Эти опасения возникают, если вы разрабатываете собственное исследование или анализируете данные, собранные другими.

Часто бывает полезно провести несколько измерений исследуемого неявного конструкта. Например, при исследовании качества образования студентам обычно задают несколько вопросов об уровне преподавателя и курса. А различные состояния здоровья измеряются с помощью стандартных наборов вопросов. Например, «Шкала депрессии» Бека состоит из 21 пункта, каждый из которых дает оценку от 0 до 3, а затем они складываются, чтобы получить общую сумму от 0 до 63.

Показатель может быть полезным для одних целей и неуместным для других. Например, в исследованиях общественного здравоохранения «никогда не куривший» обычно определяется как человек, который выкурил менее 100 сигарет за свою жизнь, что кажется разумным определением при изучении поведения и здоровья взрослых. Но при исследовании подростков было бы ошибкой относить школьника, выкурившего 90 сигарет, к той же категории «никогда не куривших», как и молодого человека, который выкурил ноль, одну или две сигареты.

Достоверность показателей

Показатели достоверны в той степени, в которой они представляет то, что вы пытаетесь измерить. Легко привести отрицательные примеры. Письменный тест не является достоверной оценкой музыкальных способностей. Существует огромный разрыв между доказательствами и тем, о чем мы хотим сделать выводы.

Точно так же вопрос о том, насколько люди удовлетворены какой-либо государственной службой, может не считаться достоверным показателем эффективности этой службы. *Достоверные показатели* – это показатели, для которых существует общее согласие о том, что наблюдения тесно связаны с предполагаемым конструктом.

Мы можем определить достоверность процесса измерения как свойство давать правильный ответ в среднем по широкому диапазону вероятных сценариев. Чтобы изучить достоверность эмпирическим путем, в идеале вам нужны условия, в которых существует наблюдаемое истинное значение и может быть выполнено несколько измерений.

В социальных науках достоверность бывает трудно оценить. Когда истина недоступна, показатели можно сравнить с мнением экспертов или другим «золотым стандартом». Например, набор вопросов, предназначенных для измерения депрессии в новой популяции, можно сравнить с мнением опытного психиатра о группе пациентов, а также с хорошо зарекомендовавшим себя опросником Бека.

Надежность показателей

Надежные показатели – это точные и стабильные показатели. Если мы выполним измерение, а затем у нас будет возможность сделать это снова, мы надеемся, что значение сильно не изменится. Другими словами, мы надеемся, что изменчивость в нашей выборке обусловлена реальными различиями между людьми или предметами, а не случайной ошибкой, возникшей в процессе измерения.

Например, рассмотрим тест, который дважды сдает одна и та же группа студентов. Мы могли бы использовать корреляцию между оценками, полученными в ходе двух тестирований, чтобы понять, насколько *надежно* этот тест измеряет данный конструкт.

Другой подход заключается в том, чтобы разные оценщики применяли одну и ту же меру в одном и том же контексте. Например, мы могли бы сравнить оценки качества обучения от разных инспекторов, которые наблюдали за одним и тем же классом в одно и то же время. Или мы могли бы сравнить судейские оценки качества выполнения гимнастками определенного элемента при условии одновременного просмотра всеми судьями. В данном случае надежность оценок определяется так называемой *межэкспертной надежностью* (inter-rater reliability).

Определение выборки

Еще один критерий качества данных – *определение выборки* (sample selection). Суть в том, что данные, которые вы наблюдаете, могут быть нерепрезентативной выборкой из более широкой совокупности, которая вам не видна. Например, предположим, что вас интересует система общественного транспорта в городе, поэтому вы опрашиваете людей, которые ездят на автобусах и поездах; возможно, вы даже проведете измерения, такие как время в пути или процент времени, проведенного сидя или стоя. Но при этом вы ограничиваете выборку – включаете в нее только тех, кто решил ездить на автобусе или поезде. Среди исключенных окажутся те, кто перестал ездить на автобусе или поезде, потому что им не нравятся эти услуги, но вы упустите их мнение о состоянии общественного транспорта.

Помимо такого рода систематической ошибки отбора, основанной на том, кто включен в набор данных, существуют также ошибки из-за неполучения ответов от субъектов опроса, неполных измерений и неправильного выбора кодирования и интерпретации данных. Мы предпочитаем рассматривать все упомянутые проблемы измерения, включая достоверность, надежность и определение выборки, с точки зрения более крупных моделей, связывающих показатели с исследованием основополагающих взаимосвязей.

2.3. Все графики служат для сравнения

Как мы уже не раз говорили, можно многому научиться, непредвзято глядя на данные. Далее рассмотрим три простых примера. По большому счету, занимаясь статистикой, мы постоянно перемещаемся туда

и обратно между исследованием данных, моделированием, выводом и построением модели, и для каждого шага требуются свои собственные инструменты.

Простые диаграммы рассеяния

Пример: расходы на здравоохранение и продолжительность жизни

На рис. 2.4 показаны некоторые данные о расходах на здравоохранение и ожидаемой продолжительности жизни, из которых видно, что Соединенные Штаты тратят на человека намного больше, чем любая другая страна, не получая при этом какой-либо очевидной выгоды в отношении продолжительности жизни[1].

Код R для построения диаграммы распределения выглядит так:

```
health <- read.table("healthdata.txt", header=TRUE)
country <- rownames(health)
plot(health$spending, health$lifespan, type="n")
text(health$spending, health$lifespan, country)
```

Чтобы придать графику окончательный вид, как показано на рис. 2.4, требуются дополнительные команды, и они доступны на нашем вебсайте, а приведенный здесь код делает основную работу.

График показывает исключительное положение Соединенных Штатов, а также взаимосвязь между расходами и продолжительностью жизни в других странах.

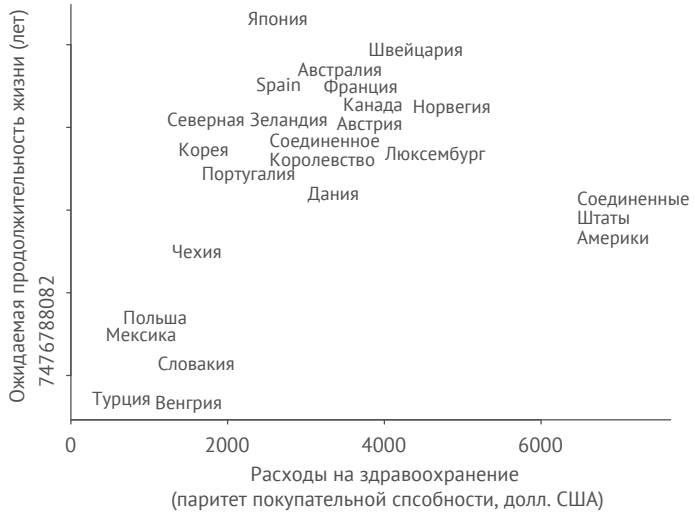


Рис. 2.4. Расходы на здравоохранение и ожидаемая продолжительность жизни в некоторых странах. Эта диаграмма рассеяния показывает две вещи: в целом положительную корреляцию между расходами и продолжительностью жизни и особое положение Соединенных Штатов

[1] Данные и код для этого примера находятся в папке HealthExpenditure.

Отображение дополнительной информации на графике

Вы можете сделать столько графиков, сколько захотите (или сколько позволяет ваше терпение), но полезно немного подумать над каждым графиком, так же как полезно немного подумать о каждой модели, которую вы подгоняете.

Точки на диаграмме рассеяния соответствуют единице анализа в вашем исследовании. По крайней мере теоретически с помощью диаграммы рассеяния вы можете легко отобразить пять переменных: положение *x*, положение *y*, символ, размер символа и цвет символа. Добавление на график сетки позволяет задействовать еще два измерения, в результате чего общее количество потенциально отображаемых переменных достигает семи.

Пример: пересмотр округов и предвзятость

Мы демонстрируем некоторые достоинства подробного визуального описания данных и оценок с помощью рис. 2.5 – графика из наших прикладных исследований, которые сыграли центральную роль в достижении наших ключевых научных результатов. Эта диаграмма рассеяния отображает три переменные, представленные положением *x*, положением *y* и символом, сравнение экспериментальных и контрольных групп с измерениями до и после. В данном случае объекты исследования – это выборы в законодательные органы штата, и график отображает предполагаемую предвзятость (меру того, в какой степени положение границ округа благоприятствует той или иной партии) в течение двух последовательных лет выборов. «Экспериментальные» точки представляют собой различные варианты изменения границ округов, а «контрольные» точки (обозначенные точками на рисунке) – пары последовательных выборов без промежуточного изменения границ избирательных округов. Мы отображаем все данные, а также показываем линии регрессии в том же масштабе. На самом деле мы не занимались подгонкой несовпадающих линий регрессии; только после создания графика и отображения параллельных линий мы поняли, что несовпадающие линии (т. е. взаимодействие между воздействием и измерением «до воздействия») – то, что нам нужно. Взаимодействие имеет решающее значение для интерпретации этих данных: (1) когда границы округов не меняются, предвзятость голосования не изменяется систематически; (2) наибольший эффект от любого изменения границ округов обычно выражается в приближении предвзятости пристрастия к нулю. Линии и точки на одном графике показывают это гораздо яснее, чем любая таблица с числами.

Иногда нам удавалось использовать описательные имена символов, такие как двухбуквенные аббревиатуры штатов. Но, если есть только две или три категории, нам больше нравится брать визуально различимые символы. Например, чтобы различать мужчин и женщин, мы не будем использовать М и Ж или M (male) и F (female). В генеалогических таблицах мужчин и женщин часто обозначают пустыми квадратами и пустыми кружками соответственно, но даже эти символы трудно отличить друг от друга внутри группы. Мы предпочитаем чет-

ко различимые цвета или символы, такие как пустые кружки, закрашенные кружки, крестики и точки на рис. 2.5. Если график состоит из нескольких линий, разместите метки непосредственно на линии, как показано на рис. 1.7.

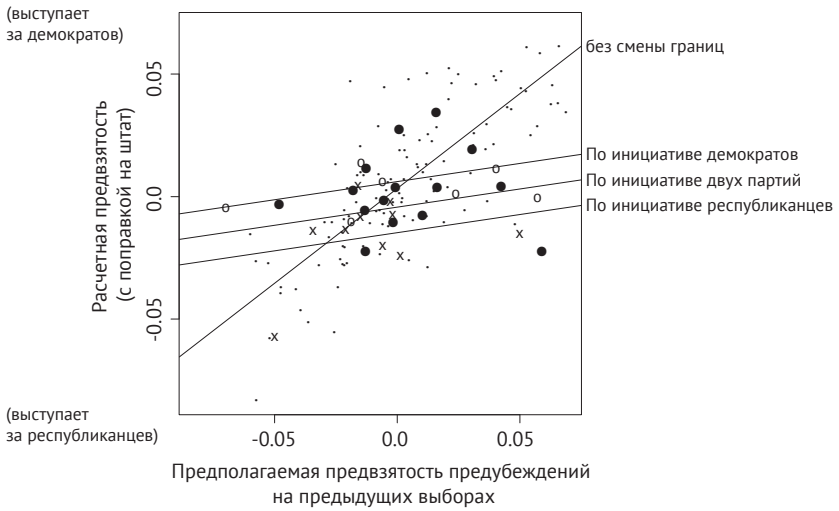


Рис. 2.5. Эффект от изменения границ избирательных округов из-за предвзятого отношения к выборам в законодательные органы штата США. Каждый символ представляет штат и год выборов, с закрашенными кружками, открытыми кружками и крестиками, обозначающими перераспределение избирательных округов по инициативе демократов, двух партий и республиканцев соответственно. Маленькие точки – это контрольные случаи – годы выборов, которые не последовали сразу после перераспределения округов. Линии показывают соответствующие регрессионные модели. Изменение границ избирательных округов, как правило, делает выборы менее предвзятыми, но небольшие партийные предубеждения остаются, и обычно соответствуют партии, инициировавшей изменение границ

Эти предложения основаны на нашем опыте и здравом смысле. Насколько нам известно, они не были подтверждены (или опровергнуты) ни в одном систематическом исследовании.

Несколько распределений

Пример: последние буквы имен

Неожиданный взгляд на данные может привести к открытию. Например, на рис. 2.6 показано распределение последних букв имен мальчиков в Соединенных Штатах в 1906 году. Наиболее распространенными именами в этом году были, например, John, James, George и Edward.

Мы можем изучать данные, помещая несколько связанных графиков рядом. На рис. 2.6 и 2.7 показаны резкие изменения в распределении последних букв в именах мальчиков в течение XX века. В последние годы более трети мальчиков получали имена, оканчивающиеся на «n», наиболее распространенными из которых были Ethan, Jayden, Aiden, Mason и Logan.

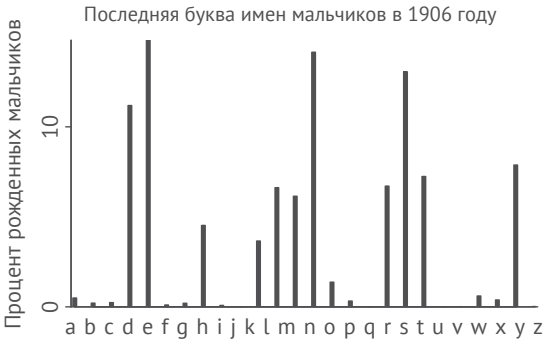


Рис. 2.6. Распределение последних букв имен мальчиков из базы данных американских младенцев 1906 года рождения. Перерисовано с графика Лауры Ваттенберг

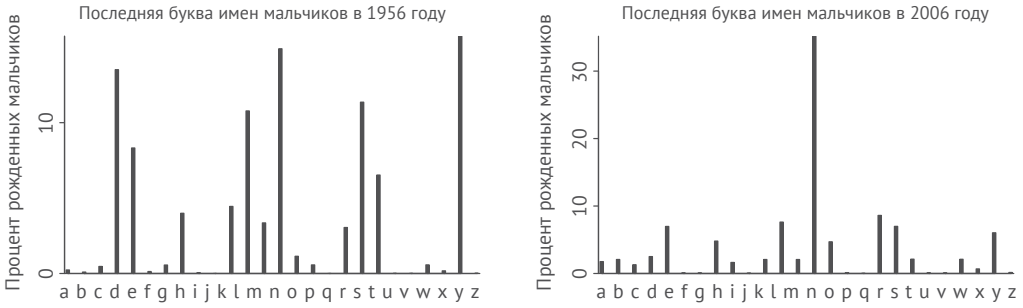


Рис. 2.7. Распределение последних букв имен мальчиков из базы данных американских младенцев, родившихся в 1956 и 2006 годах. Перерисовано с графиков Лауры Ваттенберг. Объединение этих графиков с графиком 1906 года (рис. 2.6) показывает поразительную тенденцию

Не существует единственного лучшего способа отобразить набор данных. Для другого представления вышеупомянутых данных мы создали рис. 2.8, на котором показаны временные ряды регистрируемой каждый год процентной доли частоты последней буквы в именах мальчиков[1]. На графике 26 линий, и мы выделили три из них. Мы испробовали разные представления, но обнаружили, что графики трудно читать, когда выделено более трех строк. За последние 60 лет наблюдается устойчивый рост количества имен мальчиков, оканчивающихся на «n». Если взглянуть на данные об именах с другой стороны, на рис. 2.9 показано соотношение имен мальчиков и девочек, которые каждый год входили в десятку наиболее популярных имен для каждого пола. Традиционно имена мальчиков выбирались из более узкого диапазона, чем имена девочек, причем первые 10 имен представляли 30–40 % всех мальчиков, но в последние годы выбор имен в США стал намного разнообразнее. В этом обширном наборе данных можно найти еще много закономерностей.

[1] Данные и код для этого примера находятся в папке Names.

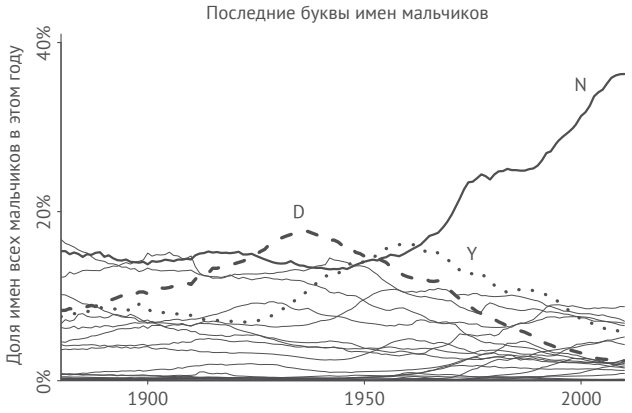


Рис. 2.8. Динамика процентной доли каждой буквы в окончании имен мальчиков. На этом графике 26 линий, причем линии для N, D и Y выделены, чтобы показать различные тенденции для имен, звучащих по-разному. Сравните с рис. 2.6 и 2.7, на которых показаны срезы распределения последней буквы в 1906, 1956 и 2006 годах

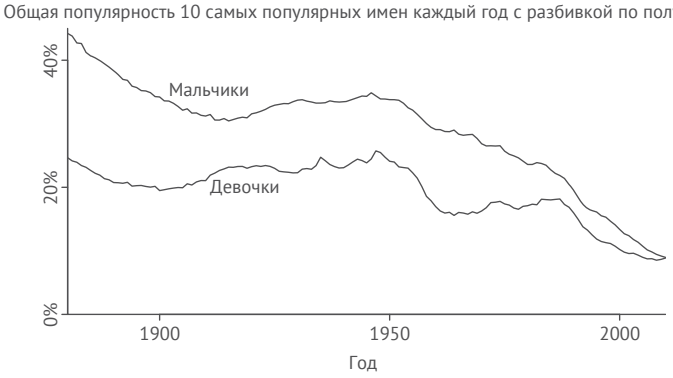


Рис. 2.9. Тенденции изменения количества имен мальчиков и девочек. В конце 1800-х годов, а затем снова в разные периоды с 1950 года наблюдалось резкое снижение доли младенцев, которым давали самые популярные имена, так что теперь 10 наиболее популярных имен каждого пола получают лишь около 10 % детей. Таким образом, хотя звучание имен мальчиков стало более единообразным (на что указывает график последних букв, показанный на рис. 2.6), сами по себе имена стали более разнообразными

Панели диаграмм рассеяния

Диаграмма рассеяния отображает две непрерывные переменные, скажем, y в зависимости от x_1. Окраска точек позволяет нам разместить на диаграмме третью переменную x_2 с небольшим количеством дискретных уровней. На практике бывает трудно прочитать график с более чем двумя цветами. Мы можем включить в представление данных еще две дискретные переменные, построив *двухмерную сетку диаграмм* (*мультипанель*), представляющих дискретные переменные x_3 и x_4. Такой

подход с использование небольших панелей может быть более эффективным, чем попытка втиснуть пять переменных на один график.

На рис. 2.10 представлена мультипанель, относящаяся к результатам выборов в Конгресс США[1]. На каждом графике показана доля демократов от одних выборов к другим относительно доли кандидатов-демократов в голосовании на первых выборах, где каждая точка представляет отдельное место в Палате представителей, серым цветом показаны выборы, на которых претенденты баллотируются на переизбрание, а черным – выборы на свободные места. Каждая строка диаграммы показывает разную пару лет национальных выборов, а четыре столбца показывают данные из разных регионов страны.

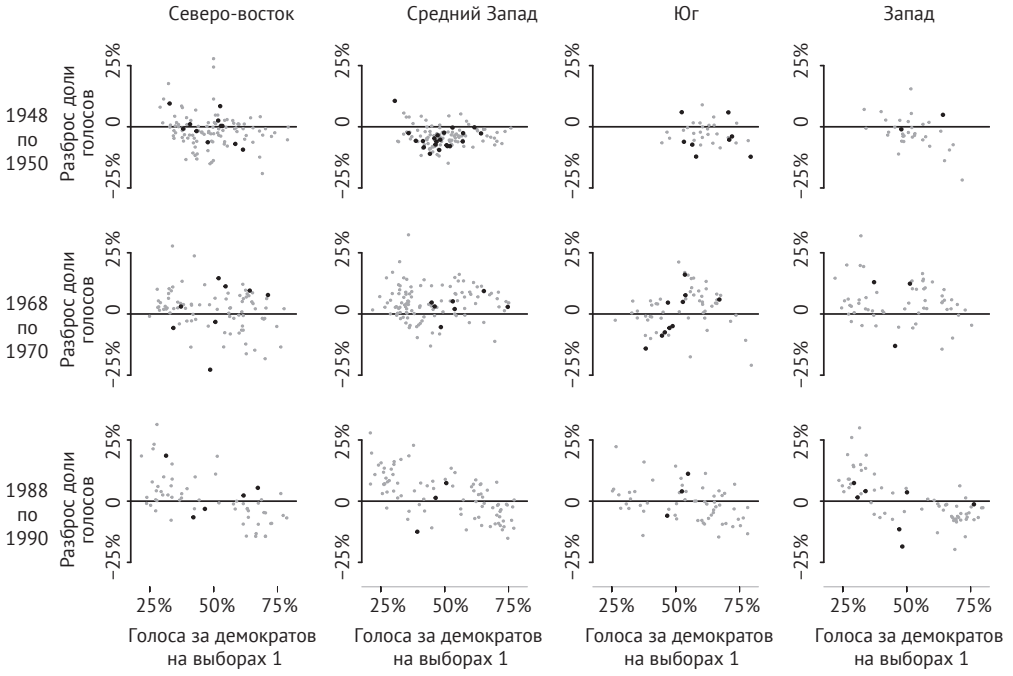


Рис. 2.10. Колебания доли голосов на выборах в Конгресс США за три разных периода. Эта сетка графиков демонстрирует, как мы можем отобразить результат (в данном случае приоритет демократов или республиканцев между двумя выборами в избирательном округе) в зависимости от четырех предикторов: доли голосов от Демократической партии в предыдущем голосовании, статуса кандидата (серый цвет для баллотирующихся действующих депутатов и черный – выборы на освободившиеся места), региона страны и периода времени. Безальтернативные выборы исключены

Разбиение данных по разным панелям позволяет нам увидеть некоторые закономерности, такие как усиление политической поляризации (с 1940-х по 1960-е и далее по 1980-е годы мы видим уменьшение количества выборов с долей голосов около 50 %), увеличение волатильности выборов (более сильные колебания в более поздние периоды)

[1] Данные и код этого примера находятся в папке Congress.

и перемены на Юге, который в 1940-х годах был в подавляющем большинстве демократическим, но к 1980-м годам демонстрировал более симметричный диапазон результатов выборов. Увидеть все это на одной диаграмме очень сложно; кроме того, график можно легко расширить дополнительными строками (данные за большее количество лет) или столбцами (более мелкое разбиение по регионам).

В более общем смысле мы можем построить график непрерывного результата y относительно непрерывного предиктора x_1 и дискретных предикторов x_2, x_3 и x_4. При необходимости мы также можем построить соответствующие линии на каждом графике, показывающие ожидаемое значение y как функцию от x_1 для различных фиксированных значений трех других предикторов.

Дискретные переменные также могут представлять непрерывные величины. Например, чтобы отобразить данные эксперимента с лекарствами от кровяного давления, мы могли бы построить график измерений до и после воздействия с разными цветами для экспериментальной и контрольной группы, где верхняя и нижняя строки отображают данные для мужчин и женщин, а столбцы соответствуют разным возрастным категориям пациентов. Возраст – это непрерывная переменная, но для этого графика ее можно разделить на категории.

Общие принципы коммуникаций и графического представления данных

Представляя данные и результаты анализа, вы всегда должны ставить себя на место читателя отчета. Избегайте перегружать читателя нерелевантным материалом. В качестве простейшего (но все же важного) примера рассмотрим представление числовых результатов отдельно или в виде таблиц.

Не используйте числа со слишком большим количеством десятичных знаков. Абсолютного стандарта для количества значащих цифр не существует; вы должны выбирать точность, которая подчеркивает неопределенность и изменчивость представленных чисел. Например, интервал погрешности [3,276, 6,410] будет более нагляден в записи [3,3, 6,4]. (Исключением является случай, когда нужно сохранить много дополнительных цифр для последующих точных вычислений, например 51,7643 – 51,7581.) Дополнительное соображение заключается в том, что вы часто можете сделать список или таблицу чисел более понятными, сначала вычтя среднее значение (по таблице, строке или столбцу). Подходящее количество значащих цифр зависит от погрешности. Но на практике трех цифр обычно бывает достаточно, потому что, если бы потребовалось больше, мы сначала вычли бы среднее значение.

Основным источником чрезмерного количества значащих цифр обычно является компьютерный вывод по умолчанию. Одно из решений – задать округление в компьютерной программе (например, в R `options(digits = 2)`).

График почти всегда можно сделать меньше, чем вы думаете, но при этом сохранить его читаемость. Благодаря этому остается место для

большего количества графиков в сетке, что позволяет одновременно видеть и сравнивать больше закономерностей.

Никогда не демонстрируйте график, о котором не можете рассказать. Снабдите каждый график развернутой подписью, как мы попытались сделать в этой книге. Это объясняет вам и другим, что вы пытаетесь показать и какие знания вы извлекли из каждого распределения. Избегайте графиков, которые были построены только потому, что их удобно строить.

Графики, облегчающие понимание статистических моделей

Мы выделяем три основных варианта использования графиков в статистическом анализе.

1. Отображение необработанных данных, часто называемое «исследовательским анализом». Такие графики не должны выглядеть красиво; их назначение – показать то, чего вы не ожидали или даже не собирались искать.
2. Графики подогнанных моделей и выводов, иногда с наложением графиков данных, чтобы понять подгонку модели, иногда со структурированием или обобщением выводов для многих параметров, чтобы увидеть более крупную закономерность. Кроме того, мы можем сгенерировать имитационные данные при помощи подогнанной модели, затем нанести их на график и сравнить с сопоставимыми графиками необработанных данных.
3. Графики, представляющие ваши окончательные результаты, – инструмент коммуникации с аудиторией. Часто ваша самая важная аудитория здесь – это вы сами: четко представив все свои результаты на графике, вы внезапно поймете общую картину.

По большому счету назначение любого графика – коммуникация с собой или с другими. Более того, графики – это сопоставления: с нулем, с другими графиками, с горизонтальными линиями и т. д. Мы «читаем» график путем извлечения как ожидаемой информации (например, наклона подогнанной линии регрессии, сравнения серии интервалов погрешности с нулем и друг с другом), так и неожиданной. По нашему опыту главной неожиданностью часто оказывается не «выброс» или отклонение от нормы, а систематическая закономерность в некоторой части данных.

Некоторые из наиболее эффективных графиков просто показывают нам, что делает подогнанная модель. В качестве примера взгляните на рис. 15.6.

Графики как сравнения

Все графики можно рассматривать как сравнения. При построении графика выстраивайте их так, чтобы наиболее важные сравнения были наиболее четкими. Сравнение становится наиболее четким, когда согласованы масштабы. Для эффективного отображения числовых данных может потребоваться творческое мышление, но иногда ваше

творчество может принести лучшие результаты, если предварительно хорошо обдумать свои цели. Процесс подготовки графика похож на написание письма – текст лучше обдумать заранее, и иногда приходится переставлять предложения.

Графики подогнанных моделей

Иногда бывает полезно изобразить согласованную модель и данные на одном и том же графике, как мы это делаем на протяжении всей книги. Нам также нравится графически отображать наборы предполагаемых параметров (например, на рис. 10.9). Графики оценок параметров можно рассматривать как протомодели в том смысле, что график подразумевает связь между осью *y* (отображаемые оценки параметров) и осью *x* (часто время или какой-либо другой показатель различных подмножеств данных, по которым подгоняют модель). Эти графики содержат неявную модель или сравнение с неявной моделью точно так же, как любая диаграмма рассеяния содержит начальное значение прогнозной модели.

Еще одно использование графиков с подогнанными моделями – построение предсказанных наборов данных и их визуальное сравнение с фактическими данными, как мы обсуждаем в разделах 11.4 и 11.5. Для структур данных, более сложных, чем простые пакеты или временные ряды, графики могут быть нацелены на конкретные аспекты проверяемых моделей.

2.4. Данные и корректировка: тенденции в уровнях смертности

Даже когда нет сомнений в качестве данных или модели, иногда имеет смысл скорректировать измерения, чтобы ответить на практические вопросы реального мира.

Пример: тенденции в уровне смертности

В конце 2015 года экономисты Энн Кейс и Ангус Дитон опубликовали график, иллюстрирующий «заметное увеличение смертности от всех причин среди белых не латиноамериканских мужчин и женщин среднего возраста в США в период с 1999 по 2013 год». Авторы заявили, что их числовые данные «выбиваются из общего ряда для возрастной группы 45–54». Они рассчитывали уровень смертности каждый год путем деления общего числа смертей в возрастной группе на население в целом и обратили внимание на эту конкретную подгруппу, потому что она выделялась своим ростом: показатели смертности для других возрастных и этнических групп в этот период снижались.

Подозревая наличие систематической ошибки, мы исследовали, может ли увеличение совокупных показателей смертности для этой возрастной группы в значительной степени быть вызвано изменением состава возрастной группы от 45 до 54 лет в период с 1990 по 2013 год. Если бы это было так, изменение уровня смертности в группе с течением времени могло не отражать изменение возрастных коэффициентов

смертности. Поправка на возраст подтвердила это подозрение. Вопреки первоначальному утверждению, основанному на необработанных данных, мы обнаружили, что после поправки на возрастной состав больше не наблюдается устойчивого роста показателей смертности для этой возрастной группы. Напротив, с 1999 по 2005 год наблюдается тенденция к увеличению, а затем – постоянные показатели. Более того, стратификация по полу скорректированных коэффициентов смертности показывает заметный рост только для женщин, а не для мужчин.

Мы демонстрируем необходимость корректировки по возрасту на рис. 2.11[1]. Нескорректированные числа на рис. 2.11а показывают устойчивый рост уровня смертности среди белых нелатиноамериканцев в возрасте от 45 до 54 лет. Однако в этот период средний возраст в этой группе увеличился по мере того, как прошло поколение бэби-бума. На рис. 2.11б показано это увеличение.

Предположим на данный момент, что показатели смертности лиц в этой возрастной группе с 1999 по 2013 год не изменились. В этом случае мы могли бы рассчитать изменение уровня групповой смертности исключительно из-за изменения основного возраста населения. Мы делаем это, беря уровни смертности за 2013 год для каждого возраста и вычисляя средневзвешенный показатель каждый год, используя количество людей в каждой возрастной группе. На рис. 2.11в показан результат. Изменение возрастного состава объясняет примерно половину изменения уровня смертности в этой группе с 1999 года и *все* изменения с 2005 года.

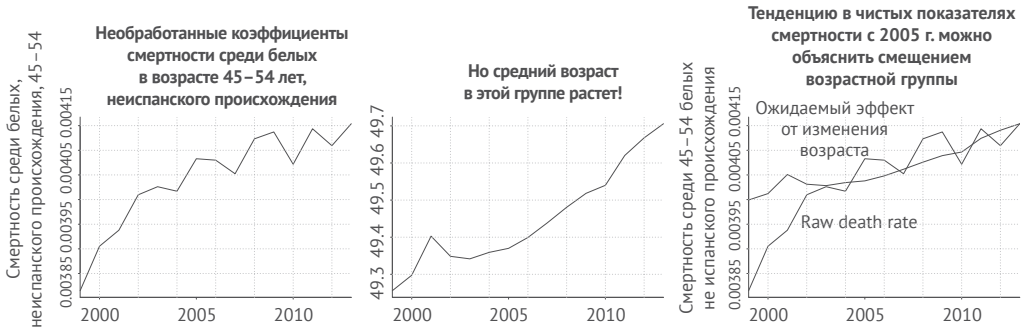


Рис. 2.11. (а) Наблюдаемое увеличение уровня смертности среди белых нелатиноамериканцев в возрасте от 45 до 54 лет без поправки на возраст; (б) увеличение среднего возраста этой группы по мере прохождения поколения бэби-бума; (в) необработанный коэффициент смертности, наряду с тенденцией смертности, обусловленной только изменением возрастного распределения, повсюду имел возрастные коэффициенты смертности на уровне 2013 года

Продемонстрировав важность корректировки по возрасту, мы теперь выполняем корректировку с учетом меняющегося возрастного состава. Мы спрашиваем, как бы выглядели данные, если бы возрастные группы оставались неизменными каждый год и менялись бы только

[1] Данные и код для этого примера находятся в папке AgePeriodCohort.

индивидуальные показатели смертности. На рис. 2.12а показана такая простейшая корректировка, приводящая показатели каждого года к гипотетической равномерно распределенной популяции, в которой количество людей одинаково в каждом возрасте от 45 до 54 лет. То есть мы рассчитываем уровень смертности каждый год путем деления количества смертей для каждого возраста от 45 до 54 лет на численность населения этого возраста с последующим усреднением. Это позволяет нам сравнивать уровни смертности по годам. В соответствии с рис. 2.11а итоговый коэффициент смертности увеличился с 1999 по 2005 год, а затем перестал расти.

Мы могли бы так же легко использовать другое возрастное распределение, чтобы проводить достоверные сравнения по годам. Проверяя, мы обнаруживаем, что скорректированная по возрасту тенденция не чувствительна к возрастному распределению, используемому для нормализации показателей смертности. На рис. 2.12б показаны предполагаемые изменения уровня смертности при трех вариантах: первый предполагает равномерное распределение возрастов 45–54 лет; во втором использовано распределение по возрастам, существовавшее в 1999 году, которое смещено в сторону более молодого начала группы 45–54; и в третьем использовано возрастное распределение на 2013 год, которое смещено в сторону старшего возраста. Общая картина не меняется.

Расчет скорректированных по возрасту коэффициентов отдельно для каждого пола дает важный результат, который мы отобразили на рис. 2.12в. Смертность среди белых нелатиноамериканских женщин выросла с 1999 по 2013 год. Однако среди соответствующей группы мужчин рост смертности с 1999 по 2005 год был почти обратным с 2005 по 2013 год.

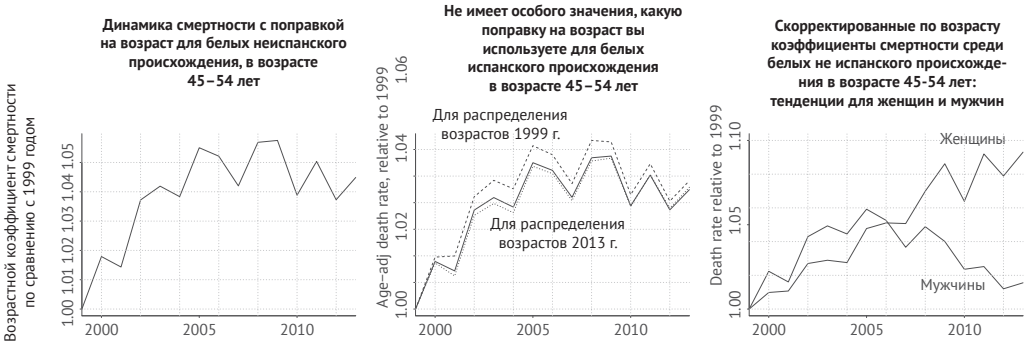


Рис. 2.12. (а) Скорректированные по возрасту коэффициенты смертности среди белых нелатиноамериканцев в возрасте от 45 до 54 лет, демонстрирующие рост с 1999 по 2005 год и стабильную тенденцию с 2005 года; (б) сравнение двух разных поправок на возраст; (в) тенденции скорректированных по возрасту коэффициентов смертности с разбивкой по полу. Три графика представлены в разных масштабах

Следовательно, корректировка возраста – это не просто академическое упражнение. Из-за меняющегося состава возрастной группы от 45 до 54 лет поправка на возраст существенно меняет интерпретацию данных. Это не меняет ключевой результат, который был замечен в не-

скорректированных данных: сравнение белых нелатиноамериканских граждан США среднего возраста с другими странами и другими этническими группами. Эти сравнения остаются в силе после корректировки по возрасту. Ошибка агрегирования в опубликованных нескорректированных цифрах составляет порядка 5 % в тенденции с 1999 по 2003 год, в то время как показатели смертности в других странах и других группах снизились примерно на 20 % за этот период.

Продолжив разбивку данных, можно узнать еще больше. Например, на рис. 2.13 коэффициенты смертности с поправкой на возраст в этой группе разбиты по регионам США. Наиболее заметной тенденцией стал рост смертности женщин на Юге. Напротив, уровень смертности для обоих полов снижается на северо-востоке, в регионе, где показатели смертности были самыми низкими с самого начала. Эти графики демонстрируют ценность такого рода исследования данных.

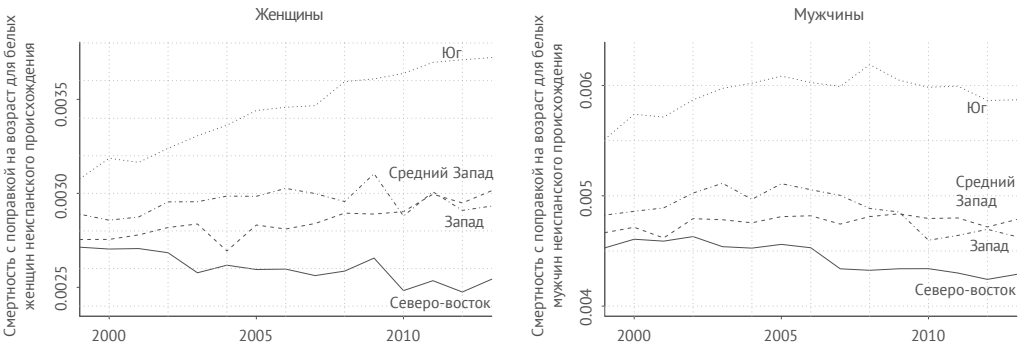


Рис. 2.13. Скорректированные по возрасту коэффициенты смертности среди белых нелатиноамериканских мужчин и женщин в возрасте от 45 до 54 лет в разбивке по регионам страны. Наиболее заметной тенденцией стал рост смертности женщин на Юге. Напротив, уровень смертности для обоих полов на северо-востоке снижается. Графики представлены в разных масштабах; как видно из оси ординат, уровень смертности среди женщин ниже, чем среди мужчин

2.5. Упражнения

2.1. *Составные показатели.* Следуя примеру Индекса человеческого развития в разделе 2.1, найдите составной показатель по интересующей вас теме. Отследите отдельные компоненты показателя и используйте диаграммы рассеяния, чтобы понять, как работает показатель, как это было сделано для указанного примера в книге.

2.2. *Значащие разряды.*

(а) Найдите статью, опубликованную в журнале о статистике или общественных науках, в которой используется слишком много значащих разрядов, т. е. где числа представлены или отображены с излишним уровнем точности. Объясните свой выбор.

(б) Найдите в журнале о статистике или общественных науках опубликованную статью, для которой является нормальным использование большого количества значащих разрядов.

2.3. *Обработка данных:* перейдите в папку `Names` и создайте график, аналогичный рис. 2.8, но для девочек.

2.4. *Визуализация данных.* Выполните любое упражнение по анализу данных из этой книги и представьте необработанные данные несколькими различными способами. Обсудите преимущества и недостатки каждого представления.

2.5. *Визуализация подогнанных моделей.* Выполните любое упражнение по анализу данных из этой книги и представьте подогнанную модель несколькими различными способами. Обсудите преимущества и недостатки каждого представления.

2.6. *Визуализация данных.* Возьмите данные по какой-то интересующей вас проблеме и постройте несколько графиков, чтобы выделить различные аспекты данных, как это было сделано на рис. 2.6–2.8.

2.7. *Надежность и достоверность.*

(а) Приведите пример сценария измерений, который дает достоверные, но недостаточно надежные данные.

(б) Приведите пример сценария измерений, который дает надежные, но не достоверные данные.

2.8. *Надежность и достоверность.* Обсудите достоверность, надежность и отбор данных в контексте измерений по интересующей вас теме. Напишите несколько примеров измерений, чтобы продемонстрировать надежность, напишите примеры истинных и измеренных значений, чтобы продемонстрировать достоверность, а также приведите примеры наблюдаемых и полных данных, чтобы продемонстрировать отбор.

2.9. *Графическое отображение параллельных временных рядов*: данные о смертности из раздела 2.4 доступны на сайте Центра по контролю и профилактике заболеваний США http://wonder.cdc.gov. Загрузите данные о смертности из этого источника, но выберите только одну конкретную причину смерти, а затем сделайте графики, аналогичные приведенным в разделе 2.4, с разбивкой тенденций смертности по возрасту, полу и регионам страны.

2.10. *Работа над собственным примером.* Продолжая пример из упражнения 1.10, нарисуйте свои данные в виде графика и обсудите проблемы достоверности и надежности. Как вы могли бы собрать дополнительные данные, по крайней мере теоретически, для решения этих проблем?

Глава 3

Обзор основных методов математики и теории вероятностей

Знание относительно несложных методов начальной математики и статистики играет три важные роли в регрессионном моделировании. Во-первых, линейная алгебра и простые распределения вероятностей являются строительными блоками для сложных моделей. Во-вторых, полезно понимать основные принципы вывода отдельно от деталей конкретной модели. В-третьих, на практике часто бывает полезно построить быстрые оценки и сравнения для небольших частей задачи – до того, как подогнать под нее сложную модель, или для понимания выхода такой модели. В этой главе дается краткий обзор некоторых основных понятий и методов.

3.1. Средневзвешенные значения

В статистике часто используют взвешенные данные или выводы, чтобы приспособиться к целевой группе населения.

Вот простой пример. В 2010 году в Северной Америке проживало 456 млн человек: 310 млн жителей США, 112 млн мексиканцев и 34 млн канадцев. Средний возраст людей в каждой стране в этом году приведен в табл. 3.1.

Таблица 3.1. Население и средний возраст стран Северной Америки. (Данные из CIA World Factbook 2010.) Средний возраст всех жителей Северной Америки представляет собой средневзвешенное значение среднего возраста в каждой стране

Страта, j	Метка	Население, N_j, млн	Средний возраст, $\bar{y}_j$
1	США	310	36,8
2	Мексика	112	26,7
3	Канада	34	40,7

Средний возраст всех жителей Северной Америки является *средневзвешенным*:

$$\text{средний возраст} = \frac{310000000 * 36,8 + 112000000 * 26,7 + 34000000 * 40,7}{310000000 + 112000000 + 34000000}.$$

Это средневзвешенное, а не простое среднее значение, потому что значения среднего возраста для каждой страны (36,8, 26,7, 40,7) умножаются на «веса», равные численности населения каждой страны. Общая численность населения Северной Америки составляла 310 + 112 + 34 = 456 млн, и мы можем переписать приведенное выше выражение как:

$$\text{средний возраст} = \frac{310000000}{456000000} * 36,8 + \frac{112\,000\,000}{456000\,000} * 26,7 + \frac{34\,000\,000}{456\,000\,000} * 40,7 =$$
$$= 0,6798 * 36,8 + 0,2456 * 26,7 + 0,0746 * 40,7 = 34,6.$$

Дробные коэффициенты 0,6798, 0,2456 и 0,0746 (сумма которых равна 1) представляют собой веса стран в этом средневзвешенном значении.

Мы можем записать средневзвешенное значение в более общем виде, используя обозначение суммы:

$$\text{средневзвешенное} = \frac{\sum_j N_j \bar{y}_j}{\sum_j N_j},$$

где j является индексом страны, а сумма складывается по всем *стратам* или *типическим группам* (в данном случае по трем странам).

Выбор весов зависит от контекста. Например, 51 % американцев – женщины и 49 % – мужчины. Средний возраст американских женщин и мужчин составляет 38,1 и 35,5 лет соответственно. Таким образом, средний возраст всех американцев составляет 0,51 * 38,1 + 0,49 * 35,5 = 36,8 (что согласуется со средним показателем для США на рис. 3.1). Но теперь рассмотрим немного другую задачу: оценку средней зарплаты всех учителей в стране. Согласно переписи 2010 года, в Соединенных Штатах насчитывалось 5 700 000 учителей-женщин и 1 500 000 учителей-мужчин (т. е. *среди учителей* 79 % женщин и 21 % мужчин) со средним доходом 45 865 и 49 207 долл. США соответственно. Средний доход всех учителей составлял 0,79 * 45 865 + 0,21 * 49 207 = 46 567 долл., а *не* 0,51 * 45 865 + 0,49 * 49 207 = 47 503 долл.

3.2. Векторы и матрицы

Список чисел называется *вектором*. Прямоугольный массив чисел называется *матрицей*. Векторы и матрицы применяются в регрессии, когда нужно представить прогнозы для многих случаев с использованием одной модели.

Пример: выборы и экономика

В разделе 1.2 мы представили модель для прогнозирования процента голосов действующей партии на президентских выборах в США исходя из экономических условий в годы, предшествовавшие выборам:

Прогнозируемый процент голосов = 46,3 + 3,0 * (темп роста среднего личного дохода),

которую мы запишем как:

$$\hat{y} = 46,3 + 3,0x,$$

или в еще более общем виде:

$$\hat{y} = \hat{a} + \hat{b}x.$$

Выражения $\hat{a}$ и $\hat{b}$ обозначают оценки – расчетные коэффициенты 46,3 и 3,0 были получены путем подгонки линейной модели к историческим данным, а $\hat{y}$ обозначает прогнозируемое значение. В нашем случае мы будем использовать y для обозначения фактического результата выборов, а $\hat{y}$ – это прогноз модели. Здесь мы используем линейное прогнозирование, поэтому работаем с $\hat{y}$.

Давайте применим эту модель к нескольким частным случаям.

1. $x = -1$. Темп роста –1 % (т. е. спад экономики на 1 %) означает, что доля голосов действующей партии составляет 46,3 + 3,0 ∗ (–1) = 43,3 %.
2. $x = 0$. Если в год, предшествующий президентским выборам, будет нулевой экономический рост, модель предсказывает, что кандидат от действующей партии получит 46,3 + 3,0 ∗ 0 = 46,3 % голосов двух партий; т. е. прогнозируется, что он проиграет выборы.
3. $x = 3$. Экономический рост в 3 % означает, что кандидат действующей партии набирает 46,3 + 3,0 ∗ 3 = 55,3 % голосов.

Мы можем определить x как вектор, включающий эти три случая, т. е. $x = (-1, 0, 3)$.

Мы можем объединить эти три прогноза:

$$\hat{y}_1 = 43,3 = 46,3 + 3,0 * (-1),$$
$$\hat{y}_2 = 46,3 = 46,3 + 3,0 * 0,$$
$$\hat{y}_3 = 55,3 = 46,3 + 3,0 * 3,$$

которые можно записать в виде векторов:

$$\hat{y} = \begin{pmatrix} 43,3 \\ 46,3 \\ 55,3 \end{pmatrix} = \begin{pmatrix} 46,3 + 3,0 * (-1) \\ 46,3 + 3,0 * 0 \\ 46,3 + 3,0 * 3 \end{pmatrix},$$

или в матричной форме:

$$\hat{y} = \begin{pmatrix} 43,3 \\ 46,3 \\ 55,3 \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 1 & 0 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} 46,3 \\ 3,0 \end{pmatrix},$$

или в более общем виде:

$$\hat{y} = X\hat{\beta}.$$

Здесь y и x – векторы длины 3, X – матрица 3×2 со столбцом единиц и столбцом, равным вектору x, а $\hat{\beta}$ = (46,3, 3,0) – вектор расчетных коэффициентов.

3.3. Построение линии

Чтобы эффективно использовать линейную регрессию, вам необходимо понимать алгебру и геометрию прямых линий, которые мы кратко рассмотрим здесь.

На рис. 3.1 показана линия $y = a + bx$. *Пересечение* a – это значение y при $x = 0$; коэффициент b – это *наклон* прямой. Линия направлена вверх, если $b > 0$ (как на рис. 3.1a), направлена вниз, если $b < 0$ (как на рис. 3.1b), и является горизонтальной, если $b = 0$. Чем больше абсолютная величина b, тем круче наклон линии.

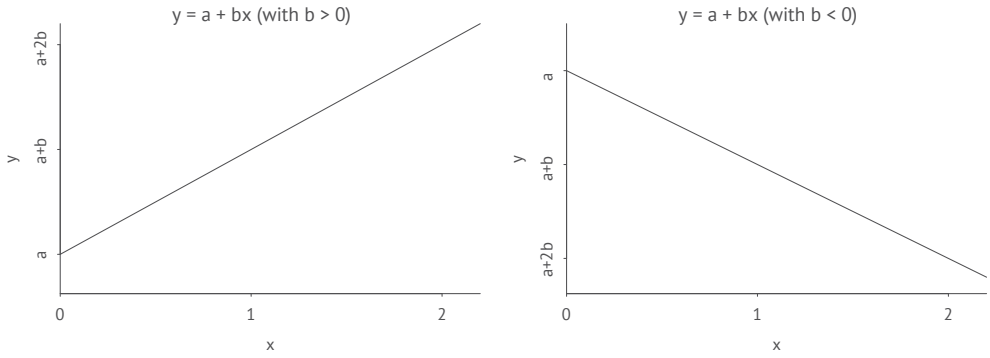


Рис. 3.1. Линии y = a + bx с положительным и отрицательным наклоном

Пример: забег на милю

На рис. 3.2a показан числовой пример: $y = 1007 - 0{,}39x$. Таким образом, $y = 1007$, когда $x = 0$, и y уменьшается на 0,39 на каждую единицу увеличения x. Эта линия аппроксимирует траекторию мирового рекорда времени (в секундах) в забеге на милю с 1900 по 2000 год (рис. A.1 в приложении A). На рис. 3.2б показан график только одной линии в этом масштабе. Эту линию можно легко начертить с помощью R[1]:

```
curve(1007 - 0.393*x, from=1900, to=2000, xlab=»Year», ylab=»Время (секунды)»,
  main=»Прибл. тенденция значения мирового рекорда\nв забеге на милю»)
```

Но как нарисовать эту линию от руки? Мы не можем просто начать с точки пересечения при $x = 0$ и двигаться оттуда, поскольку тогда весь интересующий диапазон с 1900 по 2000 год будет Втиснут в небольшой угол графика. Значение $x = 0$ выходит далеко за пределы диапазона данных. Вместо этого мы используем уравнение для вычисления значения y на двух крайних точках графика:

[1] Данные и код этого примера находятся в папке Mile.

при $x = 1900$ $y = 1007 - 0{,}393 * 1900 = 260$.
при $x = 2000$ $y = 1007 - 0{,}393 * 2000 = 221$.

Это две конечные точки на рис. 3.2, между которыми мы можем провести прямую линию.

Этот пример демонстрирует важность расположения и масштаба. Линии на рис. 3.2а и 3.2б описываются одним и тем же алгебраическим уравнением, но отображаются в разных диапазонах x, и только второй график пригоден для какой-либо прикладной цели.

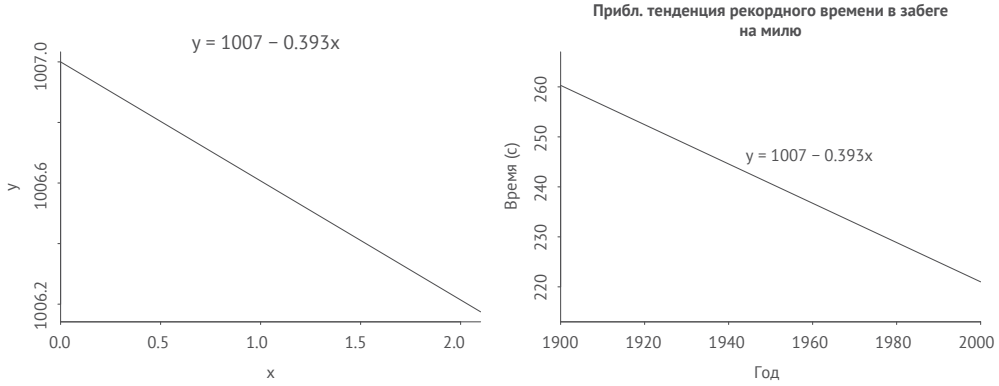


Рис. 3.2. (а) Линия y = 1007 – 0,393x. (б) Для значений x между 1900 и 2000 линия y = 1007 – 0,393x аппроксимирует тенденцию изменения мирового рекорда в забеге на одну милю. Сравните с рис. A.1 в приложении A

Мы также видим сложность интерпретации точки пересечения (значение a в $y = a + bx$) в этом примере. В уравнении $y = 1007 - 0{,}393x$ точка пересечения 1007 с (эквивалентно 16,8 мин) представляет собой прогнозируемое время мирового рекорда в забеге на милю в год 0, что является явно неуместной экстраполяцией. Было бы лучше описать эту модель уравнением $y = 260 - 0{,}393(x - 1900)$ или, возможно, $y = 241 - 0{,}393(x - 1950)$.

Наконец, мы вернемся к интерпретации наклона. Обычное толкование состоит в том, что увеличение на один год приводит к уменьшению мирового рекорда в забеге на милю в среднем на 0,393 с. Однако было бы не очень корректно приписывать времени такую неявную причинную роль. Более уместен описательный подход, когда мы говорим, что при сравнении любых двух соседних лет наблюдаемого периода мировой рекорд текущего года в среднем на 0,393 с меньше, чем в предыдущем году.

3.4. Экспоненциальный и степенной рост и спад, логарифмические отношения

Линию $y = a + bx$ можно использовать для выражения более общего класса отношений, выполняя логарифмические преобразования.

Формула $\log(y) = a + bx$ представляет экспоненциальный рост (если $b > 0$) или снижение (если $b < 0$): $y = Ae^{bx}$, где $A = e^{a}$[1]. Параметр A представляет собой значение y при $x = 0$, а параметр b определяет скорость роста или снижения. Разница в одну единицу в x соответствует аддитивной разности b в $\log(y)$ и, следовательно, мультипликативному коэффициенту e^{b} в y. Вот два примера.

- *Экспоненциальный рост.* Предположим, что население мира начинается с 1,5 млрд в 1900 году и увеличивается экспоненциально, удваиваясь каждые 50 лет (это не точное описание, а лишь грубое приближение). Мы можем записать это как $y = A * 2^{(x-1900)/50}$, где $A = 1{,}5 * 10^{9}$. Это эквивалентно $y = Ae^{(\log(2)/50))(x-1900)} = A^{e0{,}014(x-1900)}$. В статистике мы используем «log» для обозначения *натурального логарифма* (логарифм с основанием e, а не с основанием 10) по причинам, описанным в разделе 12.4.

  Модель $y = Ae^{0{,}014\ (x-1900)}$ представляет собой экспоненциальный рост со скоростью 0,014, что означает, что y увеличивается в $e^{0{,}014} = 1{,}014$ раза в год, или в $e^{0{,}14} = 1{,}15$ за десять лет, или $e^{1{,}4} = 4{,}0$ за сто лет. Мы можем взять логарифм обеих частей уравнения, чтобы получить $\log(y) = 21{,}1 + 0{,}014(x - 1900)$. Здесь $\log(A) = \log(1{,}5 * 10^{9}) = 21{,}1$.

- *Экспоненциальный спад.* Рассмотрим актив, который изначально стоит 1000 долл. и ежегодно падает в цене на 20 %. Тогда его значение в год x можно записать как $y = 1000 * 0{,}8^{x}$ или, что то же самое, $y = 1000e^{\log(0{,}8)x} = 1000e^{-0{,}22x}$. Логарифмирование обеих сторон уравнения дает $\log(y) = \log(1000) - 0{,}22x = 6{,}9 - 0{,}22x$.

Формула $\log(y) = a + b * \log(x)$ представляет степенной рост (если $b > 0$) или спад (если $b < 0$): $y = Ax^{b}$, где $A = e^{a}$. Параметр A представляет собой значение y при $x = 1$, а параметр b определяет скорость роста или снижения. Разница в одну единицу в $\log(x)$ соответствует аддитивной разнице b в $\log(y)$. Вот два примера.

Степенной закон. Пусть y – площадь квадрата, а x – его периметр. Тогда $y = (x/4)^{2}$, и мы можем взять логарифм с обеих сторон, чтобы получить $\log(y) = 2(\log(x) - \log(4)) = -2{,}8 + 2\log(x)$.

Нецелочисленный степенной закон. Пусть y – площадь поверхности куба, а x – его объем. Если L – длина стороны куба, то $y = 6L^{2}$ и $x = L^{3}$, следовательно, соотношение между x и y равно $y = 6x^{2/3}$; таким образом, $\log(y) = \log(6) + \frac{2}{3}\log(x) = 1{,}8 + \frac{2}{3}\log(x)$.

Пример: уровень метаболизма животных

Вот пример того, как интерпретировать степенной закон или *двойную логарифмическую регрессию*[2]. На рис. 3.3 показаны данные логарифмической скорости метаболизма в зависимости от массы тела, указывающие на приблизительную лежащую в основе линейную зависимость. В качестве исходного значения будем считать, что точка

[1] В этой книге авторы используют запись $\log(x)$ для обозначения натурального логарифма. Они объяснят свой выбор позже. – *Прим. перев.*

[2] Код для этого примера находится в папке Metabolic.

с надписью *Человек* соответствует человеку с массой тела 70 кг и метаболизмом 100 Вт; таким образом, группа из 100 человек весом 7000 кг эквивалентна обогревателю на 10 000 Вт. Для сравнения вы можете вычислить количество тепла, выделяемого одним слоном (который весит около 7000 кг согласно графику) или 10 000 крыс (которые вместе также весят около 7000 кг). Ответ заключается в том, что слон выделяет намного меньше тепла, чем группа людей с таким же весом, а крысы выделяют больше тепла, чем люди. Это соответствует наклону менее 1 по двойной логарифмической шкале.

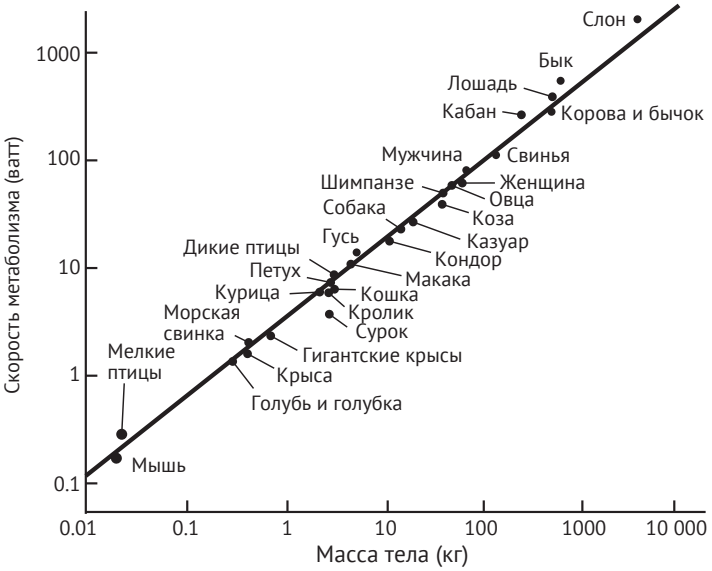


Рис. 3.3. Логарифмическая скорость метаболизма в сравнении с логарифмической массой тела животных по данным Schmidt-Nielsen (1984). Эти данные иллюстрируют двойное логарифмическое преобразование. Подогнанная линия имеет наклон 0,74. Рассмотрите также рис. 3.4

Каково уравнение линии на рис. 3.3? Вопрос не так прост, как кажется, поскольку график имеет логарифмическую шкалу, но оси размечены в исходном (линейном) масштабе. Начнем с нанесения на оси логарифмической шкалы (по основанию e), как показано на рис. 3.4а. Затем мы можем определить уравнение линии, указав две точки, через которые она проходит: например, когда $\log(x) = -4$, $\log(y) = 1{,}6$, и когда $\log(x) = 6$, $\log(y) = 5{,}8$. Таким образом, сравнивая двух животных, у которых $\log(x)$ отличается на 10, мы видим, что средняя разница в $\log(y)$ составляет $5{,}8 - (-1{,}6) = 7{,}4$. Тогда наклон линии равен $7{,}4/10 = 0{,}74$. Поскольку прямая проходит через точку (0, 1,4), ее уравнение можно записать так:

$$\log(y) = 1{,}4 + 0{,}74\log(x).$$