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.

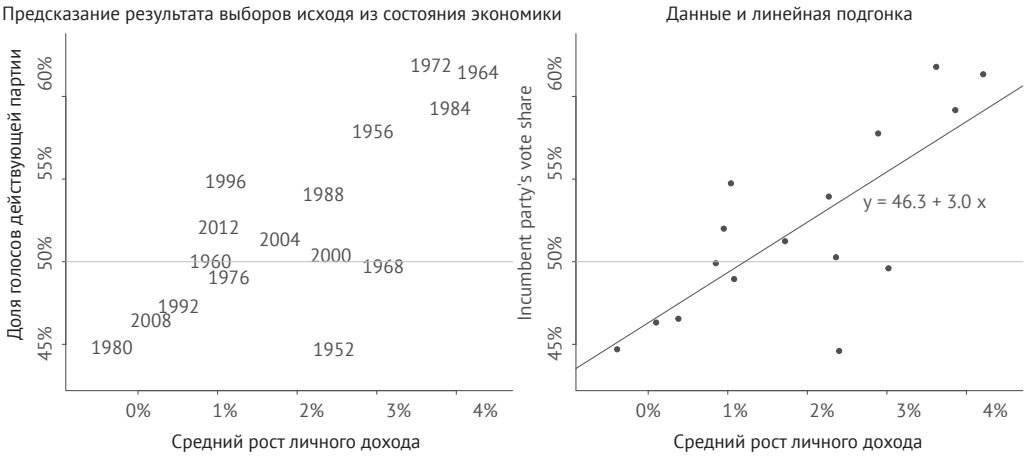

Рис. 1.1. Прогнозирование результата выборов на основе экономических данных: (а) данные, (б) аппроксимация прямой y = 46,3 + 3,0x

[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.