

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

Все права защищены. Любая часть этой книги не может быть воспроизведена в какой бы то ни было форме и какими бы то ни было средствами без письменного разрешения владельцев авторских прав.

This translation of Regression and Other Stories is published by arrangement with Cambridge University Press.

ISBN (англ.) 978-1-107-02398-7

ISBN (рус.) 978-5-97060-987-3

© Andrew Gelman, Jennifer Hill, Aki Vehtari 2021

© Оформление, издание, перевод, ДМК Пресс, 2022

Оглавление

Предисловие	13
Благодарности	14
Краткое содержание книги	15
Более увлекательные названия глав	16
Скачивание исходного кода примеров	18
Максимально эффективное использование книги	18
В помощь преподавателю:	
возможная структура курсов	18
Типографские соглашения, принятые в книге	19
Отзывы и пожелания	19
Список опечаток	20
Нарушение авторских прав	20
 ЧАСТЬ I. ОСНОВЫ	 21
Глава 1. Обзор темы и знакомство с регрессией	22
1.1. Три задачи статистики	22
1.2. Зачем изучать регрессию?	24
1.3. Несколько примеров регрессии	26
1.4. Проблемы построения и интерпретации регрессий	32
1.5. Классический и байесовский вывод	38
1.6. Вычисление наименьших квадратов и байесовской регрессии	43
1.7. Упражнения	44
 Глава 2. Данные и показатели	 48
2.1. Проверка происхождения данных	48
2.2. Достоверность и надежность	51
2.3. Все графики служат для сравнения	54
2.4. Данные и корректировка: тенденции в уровнях смертности	63
2.5. Упражнения	66

Глава 3. Обзор основных методов математики и теории вероятностей..... 68

3.1. Средневзвешенные значения	68
3.2. Векторы и матрицы	69
3.3. Построение линии	71
3.4. Экспоненциальный и степенной рост и спад, логарифмические отношения	72
3.5. Распределения вероятностей.....	76
3.6. Вероятностное моделирование	83
3.7. Упражнения	86

Глава 4. Статистический вывод..... 88

4.1. Выборочные распределения и генеративные модели	88
4.2. Оценки, стандартные ошибки и доверительные интервалы	90
4.3. Предвзятость и немоделируемая погрешность	98
4.4. Статистическая значимость, проверка гипотез и статистические ошибки	101
4.5. Проблемы с концепцией статистической значимости	106
4.6. Пример проверки гипотезы: 55 000 жителей нуждаются в вашей помощи!	111
4.7. Выход за рамки проверки гипотез.....	115
4.8. Упражнения	117

Глава 5. Моделирование случайных величин..... 120

5.1. Моделирование дискретных вероятностей	120
5.2. Моделирование непрерывных и смешанных дискретно-непрерывных вероятностей	123
5.3. Вычисление сводных показателей моделей с использованием среднего и среднего абсолютного отклонения	125
5.4. Моделирование выборочного распределения с помощью бутстрапа ..	126
5.5. Моделирование имитационных данных как образ жизни	130
5.6. Упражнения	130

ЧАСТЬ II. ЛИНЕЙНАЯ РЕГРЕССИЯ..... 135

Глава 6. Основы регрессионного моделирования..... 136

6.1. Регрессионные модели	136
6.2. Подгонка простой регрессии к смоделированным данным	137
6.3. Интерпретируйте коэффициенты как сравнения, а не как эффекты ..	140
6.4. Историческое происхождение регрессии.....	142
6.5. Парадокс регрессии к среднему.....	145
6.6. Упражнения	149

Глава 7. Линейная регрессия с одним предиктором152

7.1. Пример: прогнозирование итога президентских выборов по экономической ситуации.....	152
7.2. Проверка подгонки модели с помощью моделирования данных	157
7.3. Сравнения как частный случай регрессионных моделей	160
7.4. Упражнения	164

Глава 8. Подгонка регрессионных моделей166

8.1. Наименьшие квадраты, максимальное правдоподобие и байесовский вывод	166
8.2. Влияние отдельных точек в подогнанной регрессии	173
8.3. Наклон линии в методе наименьших квадратов как средневзвешенное значение наклонов пар	174
8.4. Сравнение подгоночных функций <code>lm</code> и <code>stan_glm</code>	175
8.5. Упражнения	178

Глава 9. Прогнозирование и байесовский вывод.....182

9.1. Распространение погрешности вывода с помощью апостериорного моделирования	182
9.2. Прогноз и погрешность: <code>predict</code> , <code>posterior_linpred</code> и <code>posterior_predict</code>	185
9.3. Априорная информация и байесовский синтез	191
9.4. Пример байесовского вывода: соотношение привлекательности и пола.....	194
9.5. Равномерные, малоинформативные и информативные априорные значения в регрессии	197
9.6. Упражнения	204

Глава 10. Линейная регрессия с несколькими предикторами208

10.1. Добавление предикторов в модель.....	208
10.2. Интерпретация коэффициентов регрессии	212
10.3. Взаимодействия	213
10.4. Индикаторные переменные.....	215
10.5. Построение плана парного и группового эксперимента как задача регрессии	220
10.6. Погрешность прогнозирования выборов в Конгресс	222
10.7. Математические обозначения и статистический вывод.....	228
10.8. Взвешенная регрессия	232
10.9. Подгонка одной модели ко многим наборам данных.....	234
10.10. Упражнения	235

Глава 11. Предположения, диагностика и оценка модели 240

11.1. Предположения регрессионного анализа	240
11.2. Построение графика данных и подогнанной модели	245
11.3. Графики остатков	251
11.4. Сравнение данных с репликациями из подогнанной модели.....	255
11.5. Прогнозное моделирование для проверки подгонки модели временного ряда.....	258
11.6. Остаточное стандартное отклонение σ и объясненная дисперсия R^2	262
11.7. Внешняя валидация: проверка подогнанной модели на новых данных	267
11.8. Перекрестная проверка	268
11.9. Упражнения	280

Глава 12. Регрессия и преобразования данных 283

12.1. Линейные преобразования	283
12.2. Центрирование и стандартизация моделей с взаимодействиями.....	286
12.3. Корреляция и регрессия к среднему.....	289
12.4. Логарифмические преобразования.....	292
12.5. Другие преобразования.....	301
12.6. Создание и сравнение регрессионных моделей для прогнозирования.....	306
12.7. Модели с большим количеством предикторов	317
12.8. Упражнения	324

ЧАСТЬ III. ОБОБЩЕННЫЕ ЛИНЕЙНЫЕ МОДЕЛИ..... 329

Глава 13. Логистическая регрессия..... 330

13.1. Логистическая регрессия с одним предиктором	330
13.2. Интерпретация коэффициентов логистической регрессии и правило деления на 4.....	334
13.3. Прогнозы и сравнения.....	338
13.4. Интерпретация регрессии через скрытые данные.....	343
13.5. Максимальное правдоподобие и байесовский вывод для логистической регрессии.....	346
13.6. Перекрестная проверка и логарифмическая оценка для логистической регрессии	350
13.7. Построение модели логистической регрессии: колодцы в Бангладеш	353
13.8. Упражнения	360

Глава 14. Продолжаем работу с логистической регрессией.... 365

14.1. Графическое представление логистической регрессии и двоичных данных	365
--	-----

14.2. Логистическая регрессия с взаимодействиями	367
14.3. Прогностическое извлечение имитационных данных	374
14.4. Средние прогностические сравнения по шкале вероятности	376
14.5. Остатки регрессии дискретных данных	382
14.6. Идентификация и разделение	387
14.7. Упражнения	392

Глава 15. Другие обобщенные линейные модели396

15.1. Определение и обозначения	396
15.2. Регрессия Пуассона и отрицательная биномиальная регрессия	398
15.3. Логистически-биномиальная модель.....	407
15.4. Пробит-регрессия: нормально распределенные скрытые данные	409
15.5. Упорядоченная и неупорядоченная категориальная регрессия.....	411
15.6. Робастная регрессия с использованием t-модели	418
15.7. Модели конструктивного выбора.....	420
15.8. Выходим за рамки обобщенных линейных моделей	425
15.9. Упражнения	429

ЧАСТЬ IV. ДО И ПОСЛЕ ПОДГОНКИ РЕГРЕССИИ435

Глава 16. План исследования и размер выборки436

16.1. Проблема статистической мощности	436
16.2. Общие принципы разработки исследования на примере оценки долей	439
16.3. Размер выборки и расчет плана для непрерывных результатов.....	445
16.4. Взаимодействия труднее оценить, чем основные эффекты.....	452
16.5. Расчет эксперимента после сбора данных	458
16.6. Анализ эксперимента с использованием имитационных данных	461
16.7. Упражнения	467

Глава 17. Постстратификация и внедрение недостающих данных471

17.1. Постстратификация: использование регрессии для обобщения на новую популяцию	471
17.2. Генерация имитационных данных для регрессии и постстратификации.....	482
17.3. Моделирование недостающих данных	485
17.4. Простые подходы к работе с отсутствующими данными.....	488
17.5. Что такое множественная подстановка?	491
17.6. Неисключающие модели отсутствующих данных	501
17.7. Упражнения	502

ЧАСТЬ V. ПРИЧИННЫЙ ВЫВОД..... 507

Глава 18. Причинный вывод и рандомизированные эксперименты..... 508

18.1. Основы причинного вывода	508
18.2. Средние причинные эффекты	514
18.3. Рандомизированные эксперименты	518
18.4. Распределения выборки, распределения рандомизации и систематическая ошибка в оценке.....	520
18.5. Использование дополнительной информации при планировании экспериментов	522
18.6. Свойства, допущения и ограничения рандомизированных экспериментов.....	527
18.7. Упражнения	536

Глава 19. Причинный вывод с использованием регрессии по переменной воздействия..... 544

19.1. Ковариаты до воздействия, методы воздействия и потенциальные результаты	544
19.2. Пример: эффект от показа детям образовательного телешоу	546
19.3. Использование предикторов, известных до воздействия	551
19.4. Различные эффекты воздействия, взаимодействие и постстратификация.....	555
19.5. Проблемы интерпретации коэффициентов регрессии как эффектов воздействия.....	559
19.6. Не применяйте для корректировки модели вторичные переменные	561
19.7. Промежуточные результаты и причинно-следственные связи	564
19.8. Упражнения	569

Глава 20. Наблюдательные исследования со всеми предполагаемыми искажающими факторами..... 574

20.1. Проблема причинного вывода	574
20.2. Использование регрессии для оценки причинного эффекта по данным наблюдений	578
20.3. Допущение о неведении при назначении воздействия в наблюдательном исследовании	581
20.4. Дисбаланс и недостаточное перекрытие	586
20.5. Пример: оценка программы по воспитанию детей	592
20.6. Подклассификация и средние эффекты воздействия	595

20.7. Сопоставление меры склонности в примере ухода за детьми.....	600
20.8. Реструктуризация для создания сбалансированных экспериментальных и контрольных групп	609
20.9. Дополнительные соображения относительно наблюдательных исследований.....	623
20.10. Упражнения	627

Глава 21. Дополнительные соображения о причинном выводе.....634

21.1. Косвенная оценка причинно-следственных связей с использованием инструментальных переменных.....	634
21.2. Инструментальные переменные в регрессионном подходе	643
21.3. Разрывная регрессия: известный механизм назначения, но без перекрытия.....	652
21.4. Идентификация с использованием различий внутри или между группами	663
21.5. Причины следствий и следствия причин	672
21.6. Упражнения	678

ЧАСТЬ VI. ЧТО ДАЛЬШЕ?687

Глава 22. Расширенная регрессия и многоуровневые модели688

22.1. Представление моделей в наиболее обобщенном виде	688
22.2. Неполные данные	689
22.3. Коррелированные ошибки и многомерные модели.....	691
22.4. Регуляризация моделей со многими предикторами.....	692
22.5. Многоуровневые, или иерархические, модели.....	693
22.6. Нелинейные модели – демонстрация с использованием Stan	694
22.7. Непараметрическая регрессия и машинное обучение	699
22.8. Вычислительная эффективность	705
22.9. Упражнения	709

Приложение А. Вычисления в R711

А.1. Загрузка и установка R и Stan.....	711
А.2. Скачивание данных и кода примеров	713
А.3. Основы	713
А.4. Чтение, запись и просмотр данных.....	719
А.5. Создание графиков.....	721
А.6. Работа с неупорядоченными данными.....	725
А.7. Основы программирования на R.....	729
А.8. Работа с объектами rstanarm	732

Приложение В. 10 кратких советов

по регрессионному моделированию 735

В.1. Не забывайте о вариации и репликации	735
В.2. Забудьте о статистической значимости	735
В.3. Изображайте на графике только релевантные данные	736
В.4. Интерпретируйте коэффициенты регрессии как сравнения	737
В.5. Изучайте методы статистики при помощи симуляции данных	737
В.6. Подгоняйте много моделей	738
В.7. Настройте вычислительную часть рабочего процесса	739
В.8. Используйте преобразования	740
В.9. Делайте целенаправленные выводы о причинно-следственных связях	740
В.10. Изучайте методы на живых примерах	741

Предметный указатель 742