

УДК 004.04

ББК 32.372

К12

Роберт И. Кабаков

К12 R в действии / пер. с англ. А. Н. Киселева. 3-е изд. – М.: ДМК Пресс, 2023. – 768 с.: ил.

ISBN 978-5-93700-173-3

R – золотой стандарт, ежедневно используемый исследователями по всему миру для самых разных вычислений и статистического анализа данных. Этот свободно распространяемый язык с открытым исходным кодом включает огромное количество пакетов самой разной направленности, от расширенной визуализации данных до глубокого обучения. Чрезвычайно удобный для пользователей с математическим складом ума, R легко решает практические задачи, не заставляя думать о них с точки зрения программиста.

Данная книга научит вас выполнять статистический анализ и визуализировать результаты с помощью R и его популярных пакетов; решать такие практические задачи, как прогнозирование, интеллектуальный анализ данных и разработка динамических отчетов. В третье издание добавлены новые сведения о построении диаграмм с помощью пакета ggplot2, а также приводятся примеры из области машинного обучения, такие как кластеризация, классификация и анализ временных рядов.

Издание предназначено для широкого круга специалистов по обработке данных.

УДК 004.04

ББК 32.372

Authorized translation of the English edition ©2022 Manning Publications. This translation is published and sold by permission of Manning Publications, the owner of all rights to publish and sell the same.

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.

ISBN (анг.) 978-1-61729-605-5

ISBN (рус.) 978-5-93700-173-3

© 2022 by Manning Publications Co.

© Оформление, издание, перевод,
ДМК Пресс, 2023

Краткое оглавление

ЧАСТЬ I. НАЧАЛО РАБОТЫ..... 35

- 1 ■ Знакомство с R..... 37
- 2 ■ Создание набора данных 58
- 3 ■ Основы управления данными 88
- 4 ■ Начало работы с диаграммами 114
- 5 ■ Дополнительные приемы управления данными 136

ЧАСТЬ II. БАЗОВЫЕ МЕТОДЫ 169

- 6 ■ Базовые диаграммы 171
- 7 ■ Основные методы статистической обработки данных..... 205

ЧАСТЬ III. МЕТОДЫ СРЕДНЕЙ СЛОЖНОСТИ..... 241

- 8 ■ Регрессия..... 243
- 9 ■ Дисперсионный анализ..... 293
- 10 ■ Анализ мощности 327
- 11 ■ Диаграммы средней сложности..... 346
- 12 ■ Статистика повторных выборок и бутстреп-анализ 378

ЧАСТЬ IV. МЕТОДЫ ПОВЫШЕННОЙ СЛОЖНОСТИ 401

- 13 ■ Обобщенные линейные модели 403
- 14 ■ Метод главных компонент и факторный анализ..... 425
- 15 ■ Временные ряды 451
- 16 ■ Кластерный анализ..... 486
- 17 ■ Классификация 512
- 18 ■ Продвинутое методы работы с пропущенными данными 542

ЧАСТЬ V. РАСШИРЕНИЕ ВОЗМОЖНОСТЕЙ..... 569

- 19 ■ Продвинутое методы работы с диаграммами 571
- 20 ■ Продвинутое приемы программирования..... 608
- 21 ■ Создание динамических отчетов 647
- 22 ■ Создание пакетов 667
- 23 ■ Продвинутое графика с использованием пакета lattice..... 696

Оглавление

Предисловие от издательства	17
Предисловие	19
Благодарности	22
Об этой книге	24
Об авторе	33
Об иллюстрации на обложке	34
ЧАСТЬ I. НАЧАЛО РАБОТЫ.....	35
1 Знакомство с R	37
1.1. Зачем использовать R?	39
1.2. Получение и установка R	42
1.3. Работа в R	42
1.3.1. Начало работы	43
1.3.2. Использование RStudio	45
1.3.3. Как получить помощь.....	48
1.3.4. Рабочее пространство	50
1.3.5. Проекты	51
1.4. Пакеты	51
1.4.1. Что такое пакеты?	52
1.4.2. Установка пакета.....	52
1.4.3. Загрузка пакета.....	53
1.4.4. Получение информации о пакете.....	53
1.5. Передача вывода на ввод: повторное использование результатов	54
1.6. Работа с большими массивами данных	55
1.7. Учимся на примере	55
Итоги	57
2 Создание набора данных.....	58
2.1. Что такое набор данных?	59
2.2. Структуры данных.....	60
2.2.1. Векторы.....	61
2.2.2. Матрицы	62
2.2.3. Массивы	64
2.2.4. Таблицы данных.....	64
2.2.5. Факторы	67
2.2.6. Списки	70
2.2.7. Усовершенствованные таблицы данных.....	71

2.3. Ввод данных	73
2.3.1. Ввод данных с клавиатуры	74
2.3.2. Импорт данных из текстового файла с разделителями	76
2.3.3. Импорт данных из Excel	80
2.3.4. Импорт данных из JSON-файлов	81
2.3.5. Извлечение данных из веб-страниц	81
2.3.6. Импорт данных из SPSS	82
2.3.7. Импорт данных из SAS	82
2.3.8. Импорт данных из Stata	82
2.3.9. Импорт данных из баз данных	83
2.3.10. Импорт данных при помощи Stat/Transfer	84
2.4. Аннотирование наборов данных	85
2.4.1. Подписи для переменных	86
2.4.2. Подписи для значений переменных	86
2.5. Полезные функции для работы с объектами	86
Итоги	87

3 Основы управления данными

3.1. Рабочий пример	89
3.2. Создание новых переменных	91
3.3. Перекодирование переменных	92
3.4. Переименование переменных	94
3.5. Пропущенные значения	95
3.5.1. Перекодирование значений в отсутствующие	96
3.5.2. Исключение пропущенных значений из анализа	96
3.6. Календарные даты	98
3.6.1. Преобразование дат в текстовые переменные	100
3.6.2. Получение дополнительной информации	100
3.7. Преобразования данных из одного типа в другой	100
3.8. Сортировка данных	101
3.9. Объединение наборов данных	102
3.9.1. Добавление столбцов	102
3.9.2. Добавление строк	103
3.10. Разделение наборов данных на составляющие	103
3.10.1. Выбор переменных	103
3.10.2. Исключение переменных из выборки	104
3.10.3. Выборка наблюдений	105
3.10.4. Функция subset()	106
3.10.5. Выборка случайных наблюдений	107
3.11. Использование dplyr для работы с таблицами данных	107
3.11.1. Основные функции из пакета dplyr	108

3.11.2. Объединение инструкций с помощью оператора конвейера	111
3.12. Использование инструкций SQL для работы с таблицами данных	112
Итоги	113
4 Начало работы с диаграммами	114
4.1. Создание диаграмм с помощью пакета ggplot2	116
4.1.1. ggplot	116
4.1.2. Геометрические объекты	117
4.1.3. Группировка	121
4.1.4. Масштабирование	123
4.1.5. Категоризованные диаграммы	125
4.1.6. Метки	127
4.1.7. Темы	128
4.2. Особенности пакета ggplot2	130
4.2.1. Параметры с данными и настройками визуального представления	130
4.2.2. Диаграммы как объекты	132
4.2.3. Сохранение диаграмм	133
4.2.4. Типичные ошибки	134
Итоги	135
5 Дополнительные приемы управления данными	136
5.1. Задача по управлению данными	137
5.2. Числовые и текстовые функции	138
5.2.1. Математические функции	138
5.2.2. Статистические функции	139
5.2.3. Функции распределения вероятности	142
5.2.4. Текстовые функции	146
5.2.5. Другие полезные функции	148
5.2.6. Применение функций к матрицам и таблицам данных	149
5.2.7. Решение задачи по управлению данными	150
5.3. Управление потоком выполнения	155
5.3.1. Циклы	156
5.3.2. Выполнение по условию	157
5.4. Пользовательские функции	158
5.5. Агрегирование и реструктуризация данных	160
5.5.1. Транспонирование	161
5.5.2. Преобразование широкого набора данных в длинный и обратно	162
5.6. Агрегирование данных	164
Итоги	167

ЧАСТЬ II. БАЗОВЫЕ МЕТОДЫ 169

6	<i>Базовые диаграммы</i>	171
6.1.	Столбиковые диаграммы	172
6.1.1.	Простые столбиковые диаграммы.....	172
6.1.2.	Столбиковые диаграммы: составные, с группировкой и спинограммы.....	173
6.1.3.	Столбиковые диаграммы средних значений	175
6.1.4.	Настройка столбиковых диаграмм	178
6.2.	Круговые диаграммы	183
6.3.	Диаграммы «плоское дерево»	186
6.3.	Гистограммы	189
6.5.	Диаграммы ядерной оценки функции плотности	192
6.6.	Коробчатые диаграммы	196
6.6.1.	Использование коробчатых диаграмм для сравнения групп.....	197
6.6.2.	Скрипичные диаграммы.....	200
6.7.	Точечные диаграммы.....	202
	Итоги.....	204
7	<i>Основные методы статистической обработки данных</i>	205
7.1.	Описательные статистики.....	206
7.1.1.	Калейдоскоп методов.....	207
7.1.2.	Дополнительные возможности.....	208
7.1.3.	Вычисление описательных статистик для групп данных	211
7.1.4.	Получение описательных статистик в интерактивном режиме с помощью dplyr	213
7.1.5.	Визуализация результатов.....	215
7.2.	Таблицы частот и таблицы сопряженности	215
7.2.1.	Создание таблиц частот	216
7.2.2.	Критерии независимости	223
7.2.3.	Меры тесноты связи	225
7.2.4.	Визуализация результатов.....	225
7.3.	Корреляция	226
7.3.1.	Типы корреляций	226
7.3.2.	Проверка статистической значимости корреляций	229
7.3.3.	Визуализация корреляций	231
7.4.	Критерий Стьюдента.....	232
7.4.1.	Критерий Стьюдента для независимых выборок	232
7.4.2.	Критерий Стьюдента для зависимых выборок	233
7.4.3.	Когда имеется больше двух групп	234
7.5.	Непараметрические критерии межгрупповых различий	235
7.5.1.	Сравнение двух групп	235
7.5.2.	Сравнение более двух групп	236
7.6.	Визуализация групповых различий.....	239
	Итоги.....	239

ЧАСТЬ III. МЕТОДЫ СРЕДНЕЙ СЛОЖНОСТИ 241

8 Регрессия	243
8.1. Многоликая регрессия	245
8.1.1. Когда используется МНК-регрессия.....	246
8.1.2. Что нужно знать.....	247
8.2. МНК-регрессия.....	247
8.2.1. Подгонка регрессионных моделей при помощи <code>lm()</code>	248
8.2.2. Простая линейная регрессия	250
8.2.3. Полиномиальная регрессия.....	253
8.2.4. Множественная линейная регрессия.....	255
8.2.5. Множественная линейная регрессия с учетом взаимосвязей	258
8.3. Диагностика регрессионных моделей.....	260
8.3.1. Стандартный подход.....	261
8.3.2. Усовершенствованный подход.....	264
8.3.3. Мультиколлинеарность	270
8.4. Необычные наблюдения.....	271
8.4.1. Выбросы.....	271
8.4.2. Точки высокой напряженности	271
8.4.3. Влиятельные наблюдения	273
8.5. Способы корректировки.....	276
8.5.1. Удаление наблюдений.....	277
8.5.2. Преобразование переменных	277
8.5.3. Добавление или удаление переменных.....	279
8.5.4. Применение другого подхода.....	280
8.6. Выбор «лучшей» регрессионной модели	280
8.6.1. Сравнение моделей	281
8.6.2. Выбор переменных	282
8.7. Продолжение анализа	286
8.7.1. Перекрестная проверка.....	286
8.7.2. Относительная важность	288
Итоги.....	292
9 Дисперсионный анализ	293
9.1. Краткий обзор терминологии	294
9.2. Подгонка ANOVA-моделей.....	297
9.2.1. Функция <code>aov()</code>	298
9.2.2. Порядок членов в формуле	299
9.3. Однофакторный дисперсионный анализ	300
9.3.1. Множественное сравнение	303
9.3.2. Проверка справедливости предположений.....	306
9.4. Однофакторный ковариационный анализ	308
9.4.1. Проверка справедливости предположений.....	310
9.4.2. Визуализация результатов.....	311
9.5. Двухфакторный дисперсионный анализ.....	312

9.6. Дисперсионный анализ повторных измерений	315
9.7. Многомерный дисперсионный анализ.....	319
9.7.1. Проверка справедливости предположений.....	320
9.7.2. Устойчивый многомерный дисперсионный анализ	322
9.8. Дисперсионный анализ как регрессия	323
Итоги.....	325
10 Анализ мощности	327
10.1. Краткий обзор проверки значимости гипотез	328
10.2. Проведение анализа мощности при помощи пакета <code>rwg</code>	331
10.2.1. Критерий Стьюдента	332
10.2.2. Дисперсионный анализ	334
10.2.3. Корреляции	335
10.2.4. Линейные модели	335
10.2.5. Сравнение пропорций.....	337
10.2.6. Критерий хи-квадрат	338
10.2.7. Выбор размера эффекта в незнакомых ситуациях.....	339
10.3. Графический анализ мощности	342
10.4. Другие пакеты.....	344
Итоги.....	345
11 Диаграммы средней сложности	346
11.1. Диаграммы рассеяния	347
11.1.1. Матрицы диаграмм рассеяния	351
11.1.2. Диаграммы рассеяния высокой плотности.....	354
11.1.3. Трехмерные диаграммы рассеяния	357
11.1.4. Вращение трехмерных диаграмм рассеяния	360
11.1.5. Пузырьковые диаграммы	362
11.2. Линейные графики	365
11.3. Кореллограммы	367
11.4. Мозаичные диаграммы.....	373
Итоги.....	376
12 Статистика повторных выборок и бутстреп-анализ....	378
12.1. Критерии перестановок	379
12.2. Критерии перестановок в пакете <code>coin</code>	382
12.2.1. Проверка независимости двух и k выборок	383
12.2.2. Независимость в таблицах сопряженности	385
12.2.3. Независимость между числовыми переменными	386
12.2.4. Критерии перестановок для двух и k зависимых выборок.....	386
12.2.5. Дополнительная информация.....	387
12.3. Критерии перестановок в пакете <code>lmPerm</code>	387
12.3.1. Простая и полиномиальная регрессия	387
12.3.2. Множественная регрессия	389
12.3.3. Однофакторные дисперсионный и ковариационный анализы	390

12.3.4. Двухфакторный дисперсионный анализ	391
12.4. Дополнительные замечания о критериях перестановок	392
12.5. Бутстреп-анализ	392
12.6. Проведение бутстреп-анализа при помощи пакета boot	393
12.6.1. Бутстреп-анализ для одной статистики	395
12.6.2. Бутстреп-анализ для нескольких статистик	397
Итоги	399

ЧАСТЬ IV. МЕТОДЫ ПОВЫШЕННОЙ СЛОЖНОСТИ...401

13 <i>Обобщенные линейные модели</i>	403
13.1. Обобщенные линейные модели и функция glm()	404
13.1.1. Функция glm()	405
13.1.2. Вспомогательные функции	407
13.1.3. Соответствие модели фактическим данным и регрессионная диагностика	408
13.2. Логистическая регрессия	409
13.2.1. Интерпретация параметров модели	412
13.2.2. Оценка влияния независимых переменных на вероятность исхода	413
13.2.3. Избыточная дисперсия	414
13.2.4. Дополнительные методы	416
13.3. Пуассоновская регрессия	417
13.3.1. Интерпретация параметров модели	419
13.3.2. Избыточная дисперсия	420
13.3.3. Дополнительные методы	422
Итоги	424
14 <i>Метод главных компонент и факторный анализ</i>	425
14.1. Поддержка метода главных компонент и факторного анализа в R	427
14.2. Главные компоненты	429
14.2.1. Выбор числа главных компонент	430
14.2.2. Выделение главных компонент	432
14.2.3. Вращение главных компонент	436
14.2.4. Вычисление оценок главных компонент	437
14.3. Разведочный факторный анализ	440
14.3.1. Определение числа извлекаемых факторов	441
14.3.2. Выделение общих факторов	442
14.3.3. Вращение факторов	443
14.3.4. Оценки факторов	447
14.3.5. Другие пакеты для проведения факторного анализа	448
14.4. Другие модели скрытых переменных	448
Итоги	449
15 <i>Временные ряды</i>	451
15.1. Создание объекта временного ряда	454

15.2. Сглаживание и сезонная декомпозиция	457
15.2.1 Сглаживание с помощью простых скользящих средних	457
15.2.2. Сезонная декомпозиция	459
15.3. Экспоненциальные модели прогнозирования	466
15.3.1. Простое экспоненциальное сглаживание	467
15.3.2. Экспоненциальное сглаживание Холта и Холта–Уинтерса	470
15.3.3. Функция ets() и автоматизация прогнозирования	473
15.4. Модели прогнозирования ARIMA	475
15.4.1. Основные понятия	475
15.4.2. Модели ARMA и ARIMA	477
15.5. Дополнительная информация	485
Итоги	485
16 Кластерный анализ	486
16.1. Общие этапы кластерного анализа	488
16.2. Вычисление расстояний	490
16.3. Иерархический кластерный анализ	492
16.4. Разделяющие методы кластерного анализа	498
16.4.1. Кластеризация методом k -средних	498
16.4.2. Разделение вокруг медоидов	505
16.5. Исключение несуществующих кластеров	507
16.6. Дополнительная информация	511
Итоги	511
17 Классификация	512
17.1. Подготовка данных	514
17.2. Логистическая регрессия	515
17.3. Деревья решений	517
17.3.1. Классические деревья решений	518
17.3.2. Деревья условного вывода	522
17.4. Случайные леса	523
17.5. Машины опорных векторов	526
17.5.1. Настройка модели SVM	529
17.6. Выбор лучшего прогностического решения	531
17.7. Интерпретация прогнозов черного ящика	535
17.7.1. Графики разбивки	536
17.7.2. График значений Шепли	538
17.8. Дополнительная информация	539
Итоги	541
18 Продвинутое методы работы с пропущенными данными	542
18.1. Этапы работы с пропущенными данными	544
18.2. Идентификация пропущенных значений	546
18.3. Исследование структуры пропущенных данных	547

18.3.1. Представление пропущенных значений в виде таблицы	548
18.3.2. Использование корреляции для исследования пропущенных значений.....	552
18.4. Определение причин отсутствия данных и их влияния.....	554
18.5. Рациональный подход к обработке отсутствующих данных	555
18.6. Удаление пропущенных данных	557
18.6. Анализ полных строк (построчное удаление)	557
18.6.2. Анализ доступных наблюдений (попарное удаление)	559
18.7. Одиночное восстановление пропущенных данных	559
18.7.1. Простое восстановление.....	560
18.7.2. Восстановление методом k -ближайших соседей	560
18.7.3. missForest.....	562
18.8. Множественное восстановление пропущенных данных.....	563
18.9. Другие подходы обработки пропущенных данных	567
Итоги.....	568

ЧАСТЬ V. РАСШИРЕНИЕ ВОЗМОЖНОСТЕЙ 569

19 Продвинутое методы работы с диаграммами..... 571

19.1. Управление отображением осей.....	572
19.1.1. Настройка осей	573
19.1.2. Настройка цветов	579
19.2. Изменение темы оформления	584
19.2.1. Предопределенные темы оформления.....	585
19.2.2. Настройка шрифтов.....	586
19.2.3. Настройка легенды	589
19.2.4. Настройка оформления области диаграммы.....	591
19.3. Добавление аннотаций	593
19.4. Объединение диаграмм.....	601
19.5. Создание интерактивных диаграмм	603
Итоги.....	606

20 Продвинутое приемы программирования 608

20.1. Обзор языка	609
20.1.1. Типы данных	609
20.1.2. Структуры управления потоком выполнения.....	617
20.1.3. Создание функций.....	619
20.2. Работа с окружениями.....	622
20.3. Нестандартная оценка	624
20.4. Объектно-ориентированное программирование.....	627
20.4.1. Обобщенные функции.....	627
20.4.2. Ограничения модели S3	629
20.5. Разработка эффективного кода	630
20.5.1. Эффективный ввод данных	630
20.5.2. Векторизация	631

20.5.3. Правильный размер объектов.....	632
20.5.4. Распараллеливание	633
20.6. Отладка	635
20.6.1. Распространенные источники ошибок	635
20.6.2. Инструменты отладки.....	636
20.6.3. Параметры сеанса для поддержки отладки	639
20.6.4. Визуальный отладчик RStudio	643
20.7. Дополнительная информация	645
Итоги.....	646
21 Создание динамических отчетов	647
21.1. Шаблонный подход к отчетам	650
21.2. Создание отчета с помощью R и R Markdown	651
21.3. Создание отчетов на R и LaTeX	657
21.3.1. Создание параметризованного отчета	660
21.4. Преодоление типичных проблем с R Markdown	663
21.5. Дополнительная информация	665
Итоги.....	666
22 Создание пакетов.....	667
22.1. Пакет edatools	668
22.2. Создание пакета	670
22.2.1. Установка средств разработки.....	671
22.2.2. Создание проекта пакета.....	671
22.2.3. Написание функций для пакета	672
22.2.4. Добавление документации с описанием функций	678
22.2.5. Добавление общего файла справки (необязательно)	680
22.2.6. Добавление демонстрационных данных в пакет (необязательно)	681
22.2.7. Добавление виньетки (необязательно)	682
22.2.8. Редактирование файла DESCRIPTION	683
22.2.9. Сборка и установка пакета	685
22.3. Распространение пакета	689
22.3.1. Распространение исходного файла пакета	689
22.3.2. Отправка в CRAN.....	689
22.3.3. Размещение на GitHub.....	690
22.3.4. Создание веб-сайта пакета	692
22.4. Дополнительная информация	694
Итоги.....	694
23 Продвинутая графика с использованием пакета lattice.....	696
23.1. Пакет lattice	697
23.2. Условные переменные.....	702
23.3. Функции для изменения формата ячеек	703
23.4. Группировка переменных	707
23.5. Графические параметры	711

23.6. Настройка планок на диаграммах	713
23.7. Размещение диаграмм на странице.....	714
23.8. Дополнительная информация	717
Послесловие. В погоне за кроликом.....	718
Приложение А. Графические пользовательские интерфейсы.....	721
Приложение В. Начальная настройка окружения	724
Приложение С. Экспорт данных из R	727
С.1. Текстовый файл CSV	727
С.2. Электронная таблица Excel.....	728
С.3. Другие статистические приложения.....	728
Приложение D. Матричная алгебра в R.....	729
Приложение E. Пакеты, использованные в этой книге	731
Приложение F. Работа с большими наборами данных.....	738
F.1. Эффективное программирование.....	739
F.2. Хранение данных вне оперативной памяти	740
F.3. Аналитические пакеты для больших объемов данных.....	740
F.4. Комплексные решения для работы с огромными наборами данных	741
Приложение G. Обновление версии R.....	744
G.1. Автоматизированное обновление R (только для Windows)	744
G.2. Обновление R вручную (для Windows и macOS)	745
G.3. Обновление R в Linux	746
Список литературы.....	747
Предметный указатель.....	752