

УДК 311:004.9R
ББК 60.6с515
К12

Кабаков, Роберт И.

К12 R в действии. Анализ и визуализация данных на языке R / Р. И. Кабаков ; пер. с англ. П. А. Волковой. — 2-е изд., эл. — 1 файл pdf : 590 с. — Москва : ДМК Пресс, 2023. — Систем. требования: Adobe Reader XI либо Adobe Digital Editions 4.5 ; экран 10". — Текст : электронный.

ISBN 978-5-89818-347-9

R — это мощный язык для статистических вычислений и графики, который может справиться поистине с любой задачей в области обработки данных. Он работает во всех важных операционных системах и поддерживает тысячи специализированных модулей и утилит. Все это делает R замечательным средством для извлечения полезной информации из гор сырых данных.

«R в действии» — это руководство по обучению этому языку с особым вниманием к практическим задачам. В данной книге представлены полезные примеры статистической обработки данных и описаны изящные методы работы с запутанными и неполными данными, а также с данными, распределение которых отлично от нормального и с которыми трудно справиться обычными методами. Статистический анализ — это только одна сторона дела. Вы также овладеете обширными графическими возможностями для визуального исследования и представления данных.

УДК 311:004.9R
ББК 60.6с515

Электронное издание на основе печатного издания: R в действии. Анализ и визуализация данных на языке R / Р. И. Кабаков ; пер. с англ. П. А. Волковой. — Москва : ДМК Пресс, 2014. — 588 с. — ISBN 978-5-97060-077-1. — Текст : непосредственный.

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.

Материал, изложенный в данной книге, многократно проверен. Но поскольку вероятность технических ошибок все равно существует, издательство не может гарантировать абсолютную точность и правильность приводимых сведений. В связи с этим издательство не несет ответственности за возможные ошибки, связанные с использованием книги.

В соответствии со ст. 1299 и 1301 ГК РФ при устранении ограничений, установленных техническими средствами защиты авторских прав, правообладатель вправе требовать от нарушителя возмещения убытков или выплаты компенсации.

ISBN 978-5-89818-347-9

© 2012 by Manning Publications Co.
© Оформление, перевод на русский язык
ДМК Пресс, 2014



ОГЛАВЛЕНИЕ

От переводчика	5
Предисловие	15
Благодарности	18
Об этой книге	20
Об иллюстрации на обложке	26

ЧАСТЬ I.

Начало работы	27
---------------------	----

Глава 1. Знакомство с R

1.1. Зачем использовать R?	32
1.2. Получение и установка R	35
1.3. Работа в R	35
1.3.1. Начало работы	36
1.3.2. Как получить помощь	39
1.3.3. Рабочее пространство	40
1.3.4. Ввод и вывод	43
1.4. Пакеты	44
1.4.1. Что такое пакеты?	44
1.4.2. Установка пакета	46
1.4.3. Загрузка пакета	46
1.4.4. Получение информации о пакете	46
1.5. Пакетная обработка	47
1.6. Использование вывода в качестве ввода – повторное использование результатов	48
1.7. Работа с большими массивами данных	49
1.8. Учимся на примере	49
1.9. Резюме	51

Глава 2. Создание набора данных

2.1. Что такое набор данных?	53
2.2. Структуры данных	54
2.2.1. Векторы	55
2.2.2. Матрицы	56
2.2.3. Массивы данных	58

2.2.4. Таблицы данных	59
2.2.5. Факторы	63
2.2.6. Списки	65
2.3. Ввод данных	67
2.3.1. Ввод данных с клавиатуры	68
2.3.2. Импорт данных из текстового файла с разделителями	69
2.3.3. Импорт данных из Excel	71
2.3.4. Импорт данных из XML-файлов	72
2.3.5. Извлечение данных из веб-страниц	72
2.3.6. Импорт данных из SPSS	72
2.3.7. Импорт данных из SAS	73
2.3.8. Импорт данных из Stata	73
2.3.9. Импорт данных из netCDF	74
2.3.10. Импорт данных из HDF5	74
2.3.11. Импорт данных из систем управления базами данных	75
2.3.12. Импорт данных при помощи Stat/Transfer	77
2.4. Аннотирование наборов данных	77
2.4.1. Подписи для переменных	78
2.4.2. Пояснение значений переменных	78
2.5. Полезные функции для работы с объектами	79
2.6. Резюме	80
Глава 3. Начало работы с диаграммами	81
3.1. Работа с диаграммами	82
3.2. Простой пример	84
3.3. Графические параметры	86
3.3.1. Символы и линии	87
3.3.2. Цвета	88
3.3.3. Характеристики текста	90
3.3.4. Размеры диаграммы и полей	93
3.4. Добавление текста, настройка параметров осей и условных обозначений	95
3.4.1. Заголовки	95
3.4.2. Оси	96
3.4.3. Опорные линии	99
3.4.4. Легенда	100
3.4.5. Аннотации	102
3.5. Объединение диаграмм	105
3.5.1. Полный контроль над расположением диаграмм	110
3.9. Резюме	112
Глава 4. Основы управления данными	113
4.1. Рабочий пример	113
4.2. Создание новых переменных	116
4.3. Перекодировка переменных	117
4.4. Переименование переменных	119

4.5. Пропущенные значения	121
4.5.1. Перекодировка значений в отсутствующие	122
4.5.2. Исключение пропущенных значений из анализа	122
4.6. Календарные даты как данные	124
4.6.1. Преобразование дат в текстовые переменные	126
4.6.2. Получение дальнейшей информации	126
4.7. Преобразования данных из одного типа в другой	127
4.8. Сортировка данных	128
4.9. Объединение наборов данных	129
4.9.1. Добавление столбцов	129
4.9.2. Добавление строк	130
4.10. Разделение наборов данных на составляющие	130
4.10.1. Выбор переменных	130
4.10.2. Исключение переменных	131
4.10.3. Выбор наблюдений	132
4.10.4. Функция <code>subset()</code>	133
4.10.5. Случайные выборки	134
4.11. Использование команд SQL для преобразования таблиц данных	135
4.12. Резюме	136

Глава 5. Более сложные способы управления данными 137

5.1. Задача по управлению данными, которую нужно решить	138
5.2. Числовые и текстовые функции	139
5.2.1. Математические функции	139
5.2.2. Статистические функции	140
5.2.3. Функции распределения	143
5.2.4. Текстовые функции	148
5.2.5. Другие полезные функции	149
5.2.6. Применение функций к матрицам и таблицам данных	151
5.3. Решение нашей задачи по управлению данными	152
5.4. Управление выполнением команд	157
5.4.1. Повторение и циклы	158
5.4.2. Выполнение при условии	159
5.5. Функции, написанные пользователем	160
5.6. Агрегирование и изменение структуры данных	163
5.6.1. Транспонирование	163
5.6.2. Агрегирование данных	164
5.6.3. Пакет <code>reshape</code>	165
5.7. Резюме	167

ЧАСТЬ II. Базовые методы 169

Глава 6. Базовые диаграммы	171
6.1. Столбчатые диаграммы	172
6.1.1. Простые столбчатые диаграммы	172
6.1.2. Столбчатые диаграммы: составные и с группировкой	174
6.1.3. Столбчатые диаграммы для средних значений	175
6.1.4. Оптимизация столбчатых диаграмм	177
6.1.5. Спинограммы	178
6.2. Круговые диаграммы	179
6.3. Гистограммы	182
6.4. Диаграммы ядерной оценки функции плотности	185
6.5. Диаграммы размахов	188
6.5.1. Использование диаграмм размахов для сравнения групп между собой	189
6.5.2. Скрипичные диаграммы	193
6.6. Точечные диаграммы	194
6.7. Резюме	197

Глава 7. Основные методы статистической обработки данных	198
7.1. Описательные статистики	199
7.1.1. Калейдоскоп методов	200
7.1.2. Вычисление описательных статистик для групп данных	204
7.1.3. Визуализация результатов	208
7.2. Таблицы частот и таблицы сопряженности	208
7.2.1. Создание таблиц частот	209
7.2.2. Тесты на независимость	216
7.2.3. Показатели взаимосвязи	218
7.2.4. Визуализация результатов	219
7.2.5. Преобразование таблиц в неструктурированные файлы	219
7.3. Корреляции	221
7.3.1. Типы корреляций	222
7.3.2. Проверка статистической значимости корреляций	225
7.3.3. Визуализация корреляций	228
7.4. Тесты Стьюдента	228
7.4.1. Тест Стьюдента для независимых выборок	229
7.4.2. Тест Стьюдента для зависимых выборок	230
7.4.3. Когда имеется больше двух групп	231
7.5. Непараметрические тесты межгрупповых различий	231
7.5.1. Сравнение двух групп	231
7.5.2. Сравнение более двух групп	233
7.6. Визуализация групповых различий	236
7.7. Резюме	236

ЧАСТЬ III.

Методы обработки данных средней сложности ...	237
--	------------

Глава 8. Регрессия	239
8.1. Многоликая регрессия	241
8.1.1. Ситуации, в которых используется МНК-регрессия	242
8.1.2. Что вам нужно знать	244
8.2. МНК-регрессия	244
8.2.1. Подгонка регрессионных моделей при помощи команды <code>lm()</code>	245
8.2.2. Простая линейная регрессия	247
8.2.3. Полиномиальная регрессия	250
8.2.4. Множественная линейная регрессия	253
8.2.5. Множественная линейная регрессия со взаимодействиями	257
8.3. Диагностика регрессионных моделей	259
8.3.1. Стандартный подход	260
8.3.2. Усовершенствованный подход	264
8.3.3. Общая проверка выполнения требований, предъявляемых к линейным моделям	272
8.3.4. Мультиколлинеарность	273
8.4. Необычные наблюдения	274
8.4.1. Выбросы	275
8.4.2. Точки высокой напряженности	275
8.4.3. Влиятельные наблюдения	277
8.5. Способы корректировки	281
8.5.1. Удаление наблюдений	281
8.5.2. Преобразование переменных	281
8.5.3. Добавление или удаление переменных	284
8.5.4. Попытка применить другой подход	284
8.6. Выбор «лучшей» регрессионной модели	285
8.6.1. Сравнение моделей	285
8.6.2. Выбор переменных	286
8.7. Продолжение анализа	291
8.7.1. Кросс-валидация	292
8.7.2. Относительная важность	294
8.8. Резюме	298
Глава 9. Дисперсионный анализ	299
9.1. Ускоренный курс терминологии	300
9.2. Подгонка ANOVA-моделей	304
9.2.1. Функция <code>aov()</code>	304
9.2.2. Порядок членов в формуле	305
9.3. Однофакторный дисперсионный анализ	307
9.3.1. Множественные сравнения	308
9.3.2. Проверка справедливости допущений, лежащих в основе теста	312
9.4. Однофакторный ковариационный анализ	314
9.4.1. Проверка допущений, лежащих в основе теста	316

9.4.2. Визуализация результатов	317
9.5. Двухфакторный дисперсионный анализ	318
9.6. Дисперсионный анализ для повторных измерений	323
9.7. Многомерный дисперсионный анализ	326
9.7.1. Проверка предположений, лежащих в основе теста	328
9.7.2. Устойчивый многомерный дисперсионный анализ	330
9.8. Дисперсионный анализ как регрессия	331
9.9. Резюме	333
Глава 10. Анализ мощности	335
10.1. Краткий обзор процедуры проверки гипотез	336
10.2. Проведение анализа мощности при помощи пакета <code>pwr</code>	339
10.2.1. Тесты Стьюдента	340
10.2.2. Дисперсионный анализ	342
10.2.3. Корреляции	343
10.2.4. Линейные модели	344
10.2.5. Сравнение пропорций	345
10.2.6. Тесты хи-квадрат	346
10.2.7. Выбор подходящего размера эффекта в незнакомых ситуациях	348
10.3. Графический анализ мощности	350
10.4. Другие пакеты	352
10.5. Резюме	354
Глава 11. Диаграммы средней сложности	356
11.1. Диаграммы рассеяния	357
11.1.1. Матрицы диаграмм рассеяния	361
11.1.2. Диаграммы рассеяния высокой плотности	367
11.1.3. Трехмерные диаграммы рассеяния	370
11.1.4. Пузырьковые диаграммы	375
11.2. Линейные графики	377
11.3. Кореллограммы	382
11.4. Мозаичные диаграммы	388
11.5. Резюме	391
Глава 12. Статистика повторных выборок и бутстреп-анализ	392
12.1. Перестановочные тесты	393
12.2. Перестановочные тесты в пакете <code>coin</code>	395
12.2.1. Тесты на независимость для двух и k выборок	397
12.2.2. Независимость в таблицах сопряженности	399
12.2.3. Независимость между числовыми переменными	400
12.2.4. Тесты для двух и k зависимых выборок	400
12.2.5. Дополнительная информация	401

12.3. Перестановочные тесты, реализованные в пакете <code>ImPerm</code>	401
12.3.1. Простая и полиномиальная регрессия	402
12.3.2. Множественная регрессия	403
12.3.3. Однофакторные дисперсионный и ковариационный анализы	404
12.3.4. Двухфакторный дисперсионный анализ	405
12.4. Дополнительные замечания о перестановочных тестах	407
12.5. Бутстреп-анализ	408
12.6. Бутстреп-анализ при помощи пакета <code>boot</code>	409
12.6.1. Бутстреп-анализ для одной статистики	411
12.6.2. Бутстреп-анализ для нескольких статистик	413
12.7. Резюме	416

ЧАСТЬ IV.

Продвинутые методы 417

Глава 13. Обобщенные линейные модели 419

13.1. Обобщенные линейные модели и функция <code>glm()</code>	420
13.1.1. Функция <code>glm()</code>	421
13.1.2. Вспомогательные функции	423
13.1.3. Соответствие модели данным и регрессионная диагностика	424
13.2. Логистическая регрессия	425
13.2.1. Интерпретация параметров модели	428
13.2.2. Оценка влияния независимых переменных на вероятность исхода	430
13.2.3. Избыточная дисперсия	431
13.2.4. Дополнительные методы	432
13.3. Пуассоновская регрессия	433
13.3.1. Интерпретация параметров модели	436
13.3.2. Избыточная дисперсия	437
13.3.3. Дополнительные методы	439
13.4. Резюме	442

Глава 14. Главные компоненты и факторный анализ 443

14.1. Выполнение анализа главных компонент и факторного анализа в R	446
14.2. Главные компоненты	447
14.2.1. Выбор необходимого числа компонент	449
14.2.2. Выделение главных компонент	451
14.2.3. Вращение главных компонент	455
14.2.4. Вычисление значений главных компонент	456
14.3. Разведочный факторный анализ	459

14.3.1. Определение числа извлекаемых факторов	460
14.3.2. Выделение общих факторов	462
14.3.3. Вращение факторов	463
14.3.4. Значения факторов	467
14.3.5. Другие пакеты для проведения факторного анализа	468
14.4. Другие модели для латентных переменных	468
14.5. Резюме	470

Глава 15. Продвинутое методы работы

с пропущенными данными 472

15.1. Этапы работы с пропущенными данными	474
15.2. Обнаружение пропущенных значений	476
15.3. Исследование структуры пропущенных данных.....	477
15.3.1. Представление пропущенных значений в виде таблицы	478
15.3.2. Визуальное исследование структуры пропущенных данных.....	479
15.3.3. Использование корреляции для исследования пропущенных значений	482
15.4. Выявление источников пропущенных данных и эффекта от них	484
15.5. Рациональный подход	486
15.6. Анализ полных строк (построчное удаление).....	487
15.7. Метод множественного восстановления пропущенных данных.....	489
15.8. Другие подходы к пропущенным данным	495
15.8.1. Парное удаление	496
15.8.2. Простое (нестохастическое) восстановление данных.....	496
15.9. Резюме	497

Глава 16. Продвинутое графические методы 499

16.1. Четыре графические системы R	500
16.2. Пакет lattice	501
16.2.1. Условные переменные	507
16.2.2. Функции для изменения формата ячеек	509
16.2.3. Группировка переменных	512
16.2.4. Графические параметры	518
16.2.5. Расположение диаграмм на странице	519
16.3. Пакет ggplot2	520
16.4. Интерактивная графика	526
16.4.1. Взаимодействие с диаграммами: идентификация точек	527
16.4.2. Пакет playwith	527
16.4.3. Пакет latticist	529
16.4.4. Создание интерактивной графики при помощи пакета iplots.....	530

16.4.5. Пакет rggobi	532
16.5. Резюме	533
Послесловие: В погоне за кроликом	535
Приложение А.	
Графические пользовательские интерфейсы	539
Приложение В.	
Настройка начальной конфигурации программы ...	543
Приложение С.	
Экспорт данных из R	545
С.1. Текстовый файл с разделителями.....	545
С.2. Таблица Excel.....	545
С.3. Другие статистические программы	546
Приложение D.	
Сохранение результатов в пригодном для публикации качестве	547
D.1. Подготовка отчета типографского качества при помощи пакета Sweave (R + LaTeX)	548
D.2. Объединение сил с OpenOffice при помощи пакета odfWeave	554
D.3. Комментарии.....	557
Приложение Е.	
Матричная алгебра в R	558
Приложение F.	
Пакеты, упомянутые в этой книге	561
Приложение G.	
Работа с большими наборами данных	570
G.1. Эффективное программирование	571
G.2. Хранение данных вне оперативной памяти	572
G.3. Аналитические пакеты для больших объемов данных	573
Приложение H.	
Обновление версии R	574
Список литературы	576
Указатель пакетов и функций.....	581