Оглавление

Предисловие редактора 12
Добро пожаловать в Книгу рецептов R 13
Благодарности 18

Глава 1. Начало работы и получение справочной информации 19
1.1. Загрузка и установка R 20
1.2. Установка RStudio 22
1.3. Запуск RStudio 23
1.4. Ввод команд 24
1.5. Выход из RStudio 26
1.6. Прерывание R 28
1.7. Просмотр прилагаемой документации 28
1.8. Получение справки по функции 30
1.9. Поиск в прилагаемой документации 31
1.10. Получение справки по пакету 33
1.11. Поиск справки в интернете 34
1.12. Поиск соответствующих функций и пакетов 36
1.13. Поиск в списках рассылки 37
1.14. Отправка вопросов в Stack Overflow или в другое место в сообществе 38

Глава 2. Немного основ 42
2.1. Вывод на экран 42
2.2. Установка переменных 44
2.3. Перечисление переменных 45
2.4. Удаление переменных 46
2.5. Создание вектора 48
2.6. Базовая статистика 49
2.7. Создание последовательностей 51
2.8. Сравнение векторов 53
2.9. Выбор элементов вектора 54
2.10. Векторная арифметика 57
2.11. Разбираемся с приоритетом оператора 58
2.12. Меньше печатать и больше делать 60
2.13. Создание конвейера вызовов функций 61
2.14. Как избежать некоторых распространенных ошибок 64

Глава 3. Навигация по программному обеспечению 69
3.1. Получение и настройка рабочего каталога 69
3.2. Создание нового проекта RStudio 70
3.3. Сохранение своего рабочего пространства 72
3.4. Просмотр истории команд 73
3.5. Сохранение результата предыдущей команды 74

3.6. Отображение загруженных пакетов через путь поиска 75
3.7. Просмотр списка установленных пакетов 76
3.8. Доступ к функциям в пакете 77
3.9. Доступ к встроенным наборам данных 78
3.10. Установка пакетов из CRAN 80
3.11. Установка пакета из GitHub 81
3.12. Установка или изменение зеркала CRAN по умолчанию 82
3.13. Запуск сценария 84
3.14. Запуск пакетного сценария 85
3.15. Поиск домашнего каталога R 87
3.16. Настройка запуска R 88
3.17. Использование R и RStudio в облаке 91

Глава 4. Ввод и вывод 93

4.1. Ввод данных с клавиатуры 93
4.2. Вывод меньшего числа цифр (или большего) 94
4.3. Перенаправление вывода в файл 96
4.4. Список файлов 97
4.5. Не удается открыть файл в Windows – что делать? 99
4.6. Чтение записей фиксированной ширины 100
4.7. Чтение файлов табличных данных 102
4.8. Чтение из файлов CSV 105
4.9. Запись в файлы CSV 107
4.10. Чтение табличных или CSV-данных из интернета 109
4.11. Чтение данных из Excel 110
4.12. Запись таблицы данных в Excel 111
4.13. Чтение данных из файла SAS 113
4.14. Чтение данных из таблиц HTML 115
4.15. Чтение файлов со сложной структурой 117
4.16. Чтение из баз данных MySQL 121
4.17. Доступ к базе данных с помощью dbplyr 124
4.18. Сохранение и транспортировка объектов 125

Глава 5. Структуры данных 128

5.1. Добавление данных в вектор 135
5.2. Вставка данных в вектор 136
5.3. Правило повторного использования 137
5.4. Создание фактора (категориальной переменной) 139
5.5. Объединение нескольких векторов в один вектор и фактор 140
5.6. Создание списка 141
5.7. Выбор элементов по месту в списке 143
5.8. Выбор элементов списка по имени 144
5.9. Создание ассоциативного списка «имя/значение» 146
5.10. Удаление элемента из списка 147
5.11. Преобразование списка в вектор 148
5.12. Удаление элементов NULL из списка 149
5.13. Удаление элементов списка с использованием условия 150
5.14. Инициализация матрицы 152
5.15. Операции с матрицами 153
5.16. Задание описательных имен для строк
и столбцов матрицы 154

5.17. Выбор одной строки или столбца из матрицы ... 155
5.18. Инициализация таблицы данных из данных столбца ... 156
5.19. Инициализация таблицы данных из данных строки ... 157
5.20. Добавление строк в таблицу данных ... 160
5.21. Выбор столбцов таблицы по их месту в таблице данных ... 162
5.22. Выбор столбцов таблицы данных по имени ... 165
5.23. Изменение имен столбцов таблицы данных ... 167
5.24. Удаление значений NA из таблицы данных ... 168
5.25. Исключение столбцов по имени ... 169
5.26. Объединение двух таблиц данных ... 170
5.27. Объединение таблиц данных по общему столбцу ... 171
5.28. Приведение атомарных типов данных ... 173
5.29. Приведение структурированных типов данных ... 174

Глава 6. Преобразование данных ... 177

6.1. Применение функции ко всем элементам списка ... 177
6.2. Применение функции к каждой строке таблицы данных ... 179
6.3. Применение функции к каждой строке матрицы ... 180
6.4. Применение функции к каждому столбцу ... 182
6.5. Применение скалярной функции к векторам или спискам ... 183
6.6. Применение функции к группам данных ... 186
6.7. Создание нового столбца по условию ... 187

Глава 7. Строки и даты ... 189

7.1. Получение длины строки ... 191
7.2. Конкатенация строк ... 191
7.3. Извлечение подстрок ... 192
7.4. Разбиение строки по разделителю ... 193
7.5. Замена подстрок ... 195
7.6. Генерация всех попарных комбинаций строк ... 195
7.7. Получение текущей даты ... 197
7.8. Преобразование строки в дату ... 197
7.9. Преобразование даты в строку ... 198
7.10. Преобразование года, месяца и дня в объект Date ... 199
7.11. Получение даты по юлианскому календарю ... 200
7.12. Извлечение частей даты ... 201
7.13. Создание последовательности дат ... 202

Глава 8. Вероятность ... 204

8.1. Подсчет количества комбинаций ... 206
8.2. Генерация комбинаций ... 206
8.3. Генерация случайных чисел ... 207
8.4. Генерация воспроизводимых случайных чисел ... 209
8.5. Генерация случайной выборки ... 210
8.6. Генерация случайных последовательностей ... 212
8.7. Случайная перестановка вектора ... 213
8.8. Расчет вероятностей для дискретных распределений ... 213
8.9. Расчет вероятностей для непрерывных распределений ... 215
8.10. Преобразование вероятностей в квантили ... 216
8.11. Построение графика функции плотности ... 217

Глава 9. Общая статистика 222

9.1. Получение сводки данных 224
9.2. Расчет относительных частот 226
9.3. Представление факторов в виде таблицы и создание таблиц сопряженности 227
9.4. Проверка категориальных переменных на независимость 228
9.5. Расчет квантилей (и квартилей) набора данных 229
9.6. Инвертирование квантиля 230
9.7. Преобразование данных в *z*-оценки 230
9.8. Проверка среднего значения выборки (t-критерий) 231
9.9. Формирование доверительного интервала для среднего значения 233
9.10. Формирование доверительного интервала для медианы 234
9.11. Тестирование доли выборки 235
9.12. Формирование доверительного интервала для доли 236
9.13. Проверка на нормальность 237
9.14. Тест последовательностей 238
9.15. Сравнение средних значений двух выборок 239
9.16. Непараметрическое сравнение местоположений двух выборок 241
9.17. Проверка значимости корреляции 242
9.18. Проверка групп на предмет наличия равных пропорций 244
9.19. Парные сравнения между средними значениями групп 245
9.20. Проверка двух выборок, чтобы определить, принадлежат ли они одному закону распределения 246

Глава 10. Графики 248

10.1. Создание точечной диаграммы 252
10.2. Добавление заголовка и меток 253
10.3. Добавление (или удаление) координатной сетки 254
10.4. Применение темы к графику ggplot 258
10.5. Создание точечной диаграммы 261
10.6. Добавление (или удаление) условных обозначений 263
10.7. Построение регрессионной линии точечной диаграммы 267
10.8. Построение точечных диаграмм 270
10.9. Создание по одной точечной диаграмме 272
10.10. Создание гистограммы 274
10.11. Добавление доверительных интервалов в гистограмму 276
10.12. Раскраска гистограммы 279
10.13. Построение линии из точек *x* и *y* 281
10.14. Изменение типа, ширины или цвета линии 282
10.15. Построение нескольких наборов данных 285
10.16. Добавление вертикальных или горизонтальных линий 286
10.17. Создание диаграммы размаха 288
10.18. Создание диаграммы размаха для каждого уровня фактора 290
10.19. Создание гистограммы 292
10.20. Добавление оценки плотности к гистограмме 293
10.21. Создание графика квантиль-квантиль 295
10.22. Создание других графиков квантиль-квантиль 297
10.23. Построение переменной в нескольких цветах 300
10.24. График функции 302
10.25. Отображение нескольких графиков на одной странице 304

Глава 11. Линейная регрессия и дисперсионный анализ 309
11.1. Простая линейная регрессия 311
11.2. Множественная линейная регрессия 313
11.3. Получение регрессионной статистики 314
11.4. Общая информация о регрессии 318
11.5. Линейная регрессия без свободного члена 321
11.6. Регрессия только тех переменных, которые сильно коррелируют с вашей зависимой переменной 322
11.7. Линейная регрессия с эффектами взаимодействия 325
11.8. Выбор наиболее подходящих переменных регрессии 327
11.9. Регрессия для подмножества данных 331
11.10. Использование выражения в формуле регрессии 333
11.11. Полиномиальная регрессия 334
11.12. Регрессия на преобразованных данных 335
11.13. Поиск наиболее подходящего степенного преобразования (тест Бокса–Кокса) 337
11.14. Формирование доверительных интервалов для коэффициентов регрессии 342
11.15. Построение невязок регрессии 342
11.16. Диагностика линейной регрессии 344
11.17. Обнаружение влиятельных наблюдений 348
11.18. Тестирование невязок на наличие автокорреляции (критерий Дарбина–Уотсона) 349
11.19. Предсказываем новые значения 351
11.20. Формирование интервалов предсказаний 352
11.21. Однофакторный дисперсионный анализ 352
11.22. Создание диаграммы взаимодействия 354
11.23. Находим различия между средними значениями групп 356
11.24. Устойчивый дисперсионный анализ (критерий Краскела–Уоллиса) 358
11.25. Сравнение моделей с использованием функции anova 360

Глава 12. Полезные хитрости 362
12.1. Просмотр данных 362
12.2. Вывод на экран результата присваивания 364
12.3. Суммирование строк и столбцов 365
12.4. Вывод данных в столбцах 366
12.5. Объединение данных 367
12.6. Поиск положения определенного значения 368
12.7. Выбор каждого *n*-го элемента вектора 368
12.8. Поиск минимумов или максимумов 369
12.9. Генерация всех комбинаций нескольких переменных 371
12.10. Преобразование таблицы данных 372
12.11. Сортировка таблицы данных 373
12.12. Удаление атрибутов из переменной 374
12.13. Раскрываем структуру объекта 375
12.14. Определяем время выполнения кода 378
12.15. Избавляемся от предупреждений и сообщений об ошибках 379
12.16. Извлечение аргументов функции из списка 380
12.17. Определение собственных бинарных операторов 382
12.18. Избавляемся от сообщения о запуске 383
12.19. Получение и настройка переменных среды 384

12.20. Разбиение кода на секции 385
12.21. Локальная параллелизация выполнения кода 386
12.22. Удалённая параллелизация выполнения кода 388

Глава 13. За пределами основных цифр и статистики 392
13.1. Минимизация или максимизация однопараметрической функции 392
13.2. Минимизация или максимизация многопараметрической функции 393
13.4. Метод главных компонент 396
13.5. Простая ортогональная регрессия 397
13.6. Поиск кластеров в данных 399
13.7. Прогнозирование бинарной переменной (логистическая регрессия) 402
13.8. Бутстрэппинг 404
13.9. Факторный анализ 406

Глава 14. Анализ временных рядов 411
14.1. Представление данных временного ряда 412
14.2. Построение данных временных рядов 415
14.3. Извлечение самых старых или самых последних наблюдений 417
14.4. Выбор элементов из временного ряда 419
14.5. Объединение нескольких временных рядов 421
14.6. Заполнение временного ряда 423
14.7. Смещение временного ряда 425
14.8. Вычисление последовательных различий 427
14.9. Выполнение расчетов по временным рядам 428
14.10. Вычисление скользящей средней 429
14.11. Применение функции по календарному периоду 431
14.12. Применение функции rollapply 433
14.13. Построение функции автокорреляции 434
14.14. Тестирование временного ряда на наличие автокорреляций 436
14.15. Построение функции частичной автокорреляции 437
14.16. Поиск корреляций с временным лагом между двумя временными рядами 439
14.17. Удаление тренда из временного ряда 440
14.18. Подгонка модели ARIMA 443
14.19. Удаление незначимых коэффициентов ARIMA 446
14.20. Выполнение диагностики для модели ARIMA 448
14.21. Прогнозирование по модели ARIMA 450
14.22. Построение прогноза 451
14.23. Тестирование на наличие возвращения к среднему 452
14.24. Сглаживание временного ряда 455

Глава 15. Простое программирование 457
15.1. Выбор из двух альтернатив: if/else 458
15.2. Итерация с помощью цикла 460
15.3. Определение функции 461
15.4. Создание локальной переменной 462
15.5. Выбор из нескольких альтернатив: функция switch 463
15.6. Определение значений по умолчанию для параметров функции 464
15.7. Подать сигнал с помощью сообщения об ошибке 465

15.8. Защита от ошибок ... 466
15.9. Создание анонимной функции ... 467
15.10. Создание коллекции многократно используемых функций ... 468
15.11. Автоматическое форматирование кода ... 469

Глава 16. R Markdown и публикации ... 471
16.1. Создание нового документа ... 472
16.2. Добавление заголовка, автора или даты ... 474
16.3. Форматирование текста документа ... 476
16.4. Вставка заголовков документов ... 476
16.5. Вставка списка ... 477
16.6. Вывод результатов из кода R ... 478
16.7. Контролируем, какой код и результаты отображаются ... 480
16.8. Вставка графика ... 481
16.9. Вставка таблицы ... 484
16.10. Вставка таблицы данных ... 485
16.11. Вставка математических уравнений ... 488
16.12. Генерация вывода HTML ... 488
16.13. Генерация вывода в формате PDF ... 490
16.14. Генерация вывода в формате Microsoft Word ... 492
16.15. Генерация выходных данных презентации ... 498
16.16. Создание параметризованного отчета ... 500
16.17. Организация рабочего процесса в R Markdown ... 503

Об авторах ... 506

Колофон ... 506

Предметный указатель ... 507

От редакции ... 508

Предисловие редактора

Более двадцати лет назад, ещё в прошлом тысячелетии, в далёкой, далёкой, почти сказочной Новой Зеландии два статистика Джентельмен и Айхэка из Оклендского университета породили ещё один язык программирования. Немного подумали и назвали его R.

С тех пор, для того, чтобы вы могли легко и непринуждённо посчитать медиану, построить линейную модель, найти кластеры или, страшно подумать, размножить ваш набор данных для целей бутстрэпа, десятки, а возможно, и сотни тысяч квалифицированных и разных любителей статистической обработки данных положили на алтарь с буквой R самое ценное. Да, да – своё время! Пользуйтесь этой программой с уважением.

Когда мы (группа едва знакомых по интернету) чуть больше десяти лет назад начали цикл статей «Анализ данных с R» в ещё живой и бумажный журнал Linux Format, никакой печатной продукции, да и вообще хоть сколько-нибудь исчерпывающей информации на русскоязычном пространстве по этой теме просто не было. Хотя во всём остальном мире статистики R захватывал университет за университетом. Чуть позже благодаря усилиям лидера группы биолога Алексея Шипунова из этого цикла родилась первая значительная (мы честно на тот момент так думали) R-книга на русском языке «Наглядная статистика. Используем R!», выпущенная издательством «ДМК Пресс» в 2012 году и находящаяся сейчас в открытом доступе на CRAN для всех желающих.

Поэтому отрадно сейчас видеть, что мировая литература в области R-анализа уже достаточно регулярно, хоть и с задержкой, добирается до наших краёв. R – это просто. По сути, сборник рецептов является самым замечательным форматом для его практического применения. Другое дело – сам сборник рецептов – это только начало, потому что возможности R безграничны! В области обработки статистики, естественно.

Евгений Балдин, к. ф.-м. н.

Добро пожаловать в книгу рецептов R

Язык R представляет собой мощный инструмент для статистики, графики и статистического программирования. Он ежедневно используется десятками тысяч людей для проведения серьезного статистического анализа. Это бесплатная система с открытым исходным кодом, реализация которой является коллективным достижением большого числа умных, трудолюбивых людей. Для этого языка существует свыше 10 000 доступных пакетов дополнений. R является серьезным конкурентом всех коммерческих пакетов для статической обработки данных.

Но R может расстраивать, и может быть не ясно, как выполнить множество задач, даже простых. Простые задачи становятся легкими, когда вы знаете, как это сделать, но при этом понимаете, что это слово «как» может свести с ума.

Эта книга полна практических рецептов, каждый из которых решает определенную проблему. Каждый рецепт включает в себя краткое введение в решение, а затем обсуждение, целью которого является объяснить решение и дать вам представление относительно того, как оно работает. Мы знаем, что эти рецепты полезны, и знаем, что они работают, потому что сами используем их.

Диапазон рецептов обширен. Он начинается с основных задач, а затем переходит к вводу и выводу, общей статистике, графике и линейной регрессии. Любая значительная работа с R будет включать в себя большинство этих областей или их все.

Если вы новичок, то эта книга поможет вам быстрее сориентироваться. Если вы промежуточный пользователь, эта книга будет полезна, чтобы расширить ваш кругозор и покопаться в памяти («Как снова выполнить этот тест Колмогорова–Смирнова?»).

Данная книга не является учебником по языку R, хотя вы и узнаете что-то, изучая рецепты. Это не справочное руководство, но в ней содержится много полезной информации. Это не книга по программированию на R, хотя многие рецепты полезны в сценариях, написанных на этом языке.

Наконец, эта книга не является введением в статистику. Многие рецепты предполагают, что вы знакомы с базовой статистической процедурой, если таковая имеется, и просто хотите знать, как это делается в R.

Рецепты

В большинстве рецептов используется одна или две функции R для решения конкретной задачи. Важно помнить, что мы не описываем функции подробно; скорее, мы приводим достаточно описания, чтобы решить насущную проблему. Почти каждая такая функция имеет дополнительные возможности, помимо описанных здесь, а некоторые из них обладают удивительными возможностями.

Мы настоятельно рекомендуем вам прочитать страницы справки по функциям. Скорее всего, вы узнаете что-нибудь ценное.

Каждый рецепт предлагает один из способов решения конкретной задачи. Конечно, для каждой задачи существует несколько разумных решений. Когда нам было известно несколько вариантов решения, мы обычно выбирали самое простое. Для любой задачи вы, вероятно, сможете найти несколько альтернативных решений самостоятельно. Это книга рецептов, а не Библия.

В частности, R имеет буквально тысячи дополнительных пакетов, доступных для скачивания, многие из которых реализуют альтернативные алгоритмы и статистические методы. Эта книга концентрируется на основных функциях, доступных через базовый дистрибутив, в сочетании с несколькими важными пакетами, известными под общим названием *tidyverse*.

Наиболее краткое определение этого термина дает Хэдли Уикхем (https://blog.rstudio.com/2016/09/15/tidyverse-1-0-0/), его создатель и один из его основных разработчиков:

> *Tidyverse* – это набор пакетов, которые работают в гармонии, потому что они имеют общие представления данных и дизайн API. Пакет tidyverse разработан, чтобы упростить установку и загрузку основных пакетов из *tidyverse* одной командой. Лучшее место, чтобы познакомиться со всеми пакетами в tidyverse и их совместимостью, – это книга R for Data Science (https://r4ds.had.co.nz).

Примечание относительно терминологии

Цель каждого рецепта – решить задачу и сделать это быстро. Вместо того чтобы работать в утомительной прозе, мы иногда упорядочиваем описание с помощью терминологии, которая является правильной, но не точной. В качестве неплохого примера можно привести термин *обобщенная функция*. Мы называем `print(x)` и `plot(x)` обобщенными функциями, потому что они работают для многих видов `x`, соответственно обрабатывая каждый вид. Специалист, работающий в области теории вычислительных машин и систем, вздрогнет при виде нашей терминологии, потому что, строго говоря, это не просто «функции»; это полиморфные методы с динамической диспетчеризацией. Но если бы мы тщательно расшифровывали все эти подробности, основные решения были бы погребены под техническими деталями. Поэтому мы называем их просто функциями, что, как нам кажется, более читабельно.

Еще один пример, взятый из статистики, – это сложность, связанная с семантикой проверки статистических гипотез. Использование строгого языка теории вероятностей затруднит практическое применение некоторых тестов, поэтому мы используем более разговорный язык при описании всех статистических тестов. *См.* введение к главе 9 для получения дополнительной информации о том, как проверка гипотез представлена в рецептах.

Наша цель состоит в том, чтобы сделать мощь языка R доступной для широкой аудитории, сделав книгу читабельной, а не формальной. Мы надеемся, что эксперты в соответствующих областях поймут, является ли наша терминология неформальной.

Примечания по программному обеспечению и платформам

Базовый дистрибутив R имеет частые и запланированные выпуски, но определение языка и реализация ядра стабильны. Рецепты, приведенные в этой книге, должны работать с любым последним выпуском базового дистрибутива.

Некоторые рецепты имеют специфические для платформы соображения, и мы тщательно их отметили. Эти рецепты в основном касаются проблем, связанных с программным обеспечением, таких как установка и настройка. Насколько нам известно, все остальные рецепты будут работать на всех трех основных платформах для R: Windows, macOS и Linux/Unix.

Дополнительные источники

Ниже приводится ряд советов для получения дополнительной информации.

В сети

Своего рода «основа» для всего, что связано с R, – сайт проекта R (https://www.r-project.org). Отсюда вы можете скачать R для своей платформы, дополнительные пакеты, документацию и исходный код, а также множество других ресурсов.

Помимо сайта проекта R, мы рекомендуем использовать R-поисковик, например RSeek (https://rseek.org), созданный Сашей Гудманом. Вы можете использовать общую поисковую систему, такую как Google, но при применении «R» в качестве поискового запроса выдается слишком много посторонних вещей. *См.* рецепт 1.11 для получения дополнительной информации о поиске в интернете.

Чтение блогов – отличный способ узнать о R и быть в курсе самых передовых разработок. Таких блогов на удивление много, поэтому мы рекомендуем два из них: R-bloggers (https://www.r-bloggers.com), созданный Талом Галили, и PlanetR (http://planetr.stderr.org). Подписавшись на их RSS-каналы, вы будете получать уведомления об интересных и полезных статьях с десятков сайтов.

Книги

Есть много, много книг об обучении и использовании R. Перечисленные здесь несколько изданий мы нашли полезными. Обратите внимание, что сайт проекта R содержит обширную библиографию книг, связанных с R (https://www.r-project.org/doc/bib/R-books.html).

R for Data Science (http://shop.oreilly.com/product/0636920034407.do) Хэдли Уикхэма и Гарретта Гроулмунда (издательство *O'Reilly*) – отличное введение в пакеты *tidyverse*, особенно в том, что касается их использования в анализе данных и статистике. Данное издание также доступно онлайн (https://r4ds.had.co.nz).

Мы полагаем, что книга *R Graphics Cookbook* (http://shop.oreilly.com/product/0636920063704.do) Уинстона Чанга (издательство *O'Reilly*) незаменима для создания графики. Книга *ggplot2: Elegant Graphics for Data Analysis* Хэдли Уикхэма (издательство *Springer*) представляет собой наиболее полный справочник по графическому пакету ggplot2, который мы используем в этой книге. Любому, кто занимается серьезной графикой в R, понадобится книга *R Graphics* Пола Мёррелла (издательство *Chapman & Hall / CRC*).

R in a Nutshell (http://shop.oreilly.com/product/9780596801717.do) Джозефа Адлера

(издательство *O'Reilly*) – это краткое руководство и справочник, который всегда будет рядом с вами. Оно охватывает гораздо больше тем, нежели эта книга рецептов. Новые книги по программированию на языке R появляются регулярно. Мы предлагаем *Hands On Programming with R* (https://www.oreilly.com/library/view/hands-on-programming-with/9781449359089/) Гарретта Гроулмунда (издательство *O'Reilly*) для введения или *The Art of R Programming* Нормала Мэтлоффа (издательство *No Starch Press*). *Advanced R* Хэдли Уикхэма (издательство *Chapman & Hall / CRC*) доступна в печатном виде, а также бесплатно на странице http://adv-r.had.co.nz. В ней еще более подробно представлены продвинутые темы R. *Efficient R Programming* (http://shop.oreilly.com/product/0636920047995.do) Колина Гиллеспи и Робина Ловеласа (издательство *O'Reilly*) – еще одно хорошее руководство для изучения более глубоких концепций программирования на языке R.

Modern Applied Statistics with S, 4-е изд., Уильяма Венейблса и Брайана Рипли (Springer) использует R для иллюстрации множества передовых статистических методов. Функции и наборы данных книги доступны в пакете MASS, который входит в стандартный дистрибутив R.

Серьезные фанаты могут скачать *R Language Definition* (https://cran.r-project.org/doc/manuals/R-lang.pdf) от R Core Team. *R Language Definition* находится в стадии разработки, но может ответить на многие ваши подробные вопросы о R как о языке программирования.

Книги по статистике

Для изучения статистики прекрасно подойдет *Using R for Introductory Statistics* Джона Верзани (издательство *Chapman & Hall / CRC*). Эта книга обучает статистике и R, давая необходимые навыки работы с компьютером для применения статистических методов.

Вам понадобится хороший учебник по статистике или справочник, чтобы точно интерпретировать статистические тесты, выполненные в R. Подобных хороших книг множество – их слишком много, чтобы можно было рекомендовать какую-либо одну.

Авторы книг по статистике все чаще используют R для иллюстрации своих методов. Если вы работаете в специализированной области, то, вероятно, найдете полезную и актуальную книгу в библиографии проекта R (https://www.r-project.org/doc/bib/R-books.html).

Условные обозначения и соглашения, принятые в книге

В книге используются следующие типографские соглашения.

Курсив используется для смыслового выделения новых терминов, адресов электронной почты, а также имен и расширений файлов.

`Моноширинный шрифт` применяется для листингов программ, а также в обычном тексте для обозначения имен переменных, функций, типов, объектов, баз данных, переменных среды, операторов и ключевых слов.

`Моноширинный полужирный шрифт` используется для обозначения команд или фрагментов текста, которые пользователь должен ввести дословно без изменений.

`Моноширинный курсив` – для обозначения в исходном коде или в командах шаблонных меток-заполнителей, которые должны быть заменены соответствующими контексту реальными значениями.

Этот элемент означает подсказку или предложение.

Этот элемент означает общее примечание.

Этот элемент указывает на предупреждение или предостережение.

Использование примеров кода

Дополнительный материал (примеры кода, исходный код, упражнения и т. д.) доступен для скачивания на странице http://rc2e.com. Аккаунт в Twitter, посвященный этой книге, – @R_cookbook (https://twitter.com/R_cookbook).

Данная книга призвана помочь вам выполнить свою работу. В общем, вы можете использовать код из этой книги в своих программах и документации. Вам не нужно обращаться к нам за разрешением, если вы не воспроизводите значительную часть кода. Например, для написания программы, в которой используется несколько фрагментов кода из этой книги, оно не требуется. Продажа или распространение CD-ROM с примерами из книг O'Reilly требует разрешения. Чтобы ответить на вопрос, сославшись на эту книгу и приведя пример кода, разрешение не требуется. Включение значительного количества примеров кода из этой книги в документацию вашего продукта требует разрешения.

Атрибуция желательна, но не является обязательной. Обычно она включает в себя название книги, автора, издателя и ISBN. Например: «Книга рецептов R, 2-е изд., Дж. Д. Лонг и Пол Титор. Copyright 2019 Дж. Лонг и Пол Титор, 978-1-492-04068-2».

Если вы считаете, что применение примеров кода выходит за рамки добросовестного использования или только что описанного разрешения, свяжитесь с нами по адресу permissions@oreilly.com.

Обучение в режиме онлайн

На протяжении почти 40 лет *O'Reilly Media* (https://www.oreilly.com) предоставляет технологии и бизнес-тренинги, знания и анализ, чтобы помочь компаниям добиваться успеха.

Наша уникальная сеть экспертов и новаторов делится своими знаниями и опытом через книги, статьи, конференции и нашу онлайн-платформу обучения. Платформа онлайн-обучения O'Reilly предоставляет доступ по требованию к курсам обучения в режиме реального времени, углубленным способам обучения, интерактивным средам кодирования и обширной коллекции текстов и видео от O'Reilly и свыше 200 других издательств. Для получения дополнительной информации, пожалуйста, посетите сайт http://oreilly.com.

Как с нами связаться

Пожалуйста, направляйте комментарии и вопросы относительно этой книги издателю:

O'Reilly Media, Inc.
1005 Gravenstein Highway North
Sebastopol, CA 95472
800-998-9938 (в США или Канаде)
707-829-0515 (международный или местный)
707-829-0104 (факс)

У нас есть страница этой книги в сети, где перечислены ошибки, примеры и вся дополнительная информация. Вы можете получить доступ к этой странице по адресу http://bit.ly/RCookbook_2e.

Чтобы оставить комментарий или задать технические вопросы об этой книге, отправьте электронное письмо по адресу bookquestions@oreilly.com.

Для получения дополнительной информации о наших книгах, курсах, конференциях и новостях посетите наш сайт по адресу http://www.oreilly.com.

Ищите нас на Facebook: http://facebook.com/oreilly.

Подпишитесь на нас в Twitter: http://twitter.com/oreillymedia.

Смотрите нас в YouTube: http://www.youtube.com/oreillymedia.

Благодарности

Выражаем благодарность сообществу R в целом и R Core Team в частности. Их бескорыстный вклад огромен. Мир статистики извлекает огромную выгоду из их работы. Участники дискуссии сообщества R Studio очень помогли в разработке идей касательно того, как объяснить многие вещи. А сотрудники и руководство R Studio оказывали поддержку во многих отношениях. Мы в долгу перед ними за все, что они дали сообществу R.

Мы хотели бы поблагодарить технических рецензентов книги: Дэвида Кёррана, Джастина Ши и майора Дасти Тернера. Их отзывы были очень важны для улучшения качества, точности и полезности этой книги. Наши редакторы, Мелисса Поттер и Рэйчел Монаган, оказали невообразимую помощь, часто не давая нам публично демонстрировать свое невежество. Нашему редактору по производству, Кристен Браун, завидуют все технические авторы благодаря ее скорости и умению работать с Markdown и Git.

Павел хотел бы поблагодарить свою семью за поддержку и терпение при создании этой книги.

Дж. Д. Лонг хотел бы поблагодарить свою жену Мэри Бет и дочь Аду за их терпение в течение всех ранних утренних часов и выходных, которые он провел с ноутбуком, работая над этой книгой.

Глава 1

Начало работы и получение справочной информации

Эта глава закладывает основу для других глав. В ней объяснено, как скачать, установить и запустить R.

Что еще более важно, в ней также объясняется, как получить ответы на ваши вопросы. Сообщество R предоставляет большое количество документации и помощи. Вы не одиноки. Вот несколько распространенных источников справочной информации.

Локальная, установленная документация

Когда вы устанавливаете R на свой компьютер, вместе с ним устанавливается огромное количество документации. Вы можете просматривать локальную документацию (рецепт 1.7) и искать ее (рецепт 1.9). Поразительно, насколько часто ищем в интернете ответ на тот или иной вопрос только для того, чтобы обнаружить, что он уже доступен в установленной документации.

Представления задач

Представление задач (https://cran.r-project.org/web/views/) описывает пакеты, которые относятся к одной из областей статистической работы, такой как эконометрика, медицинская визуализация, психометрия или пространственная статистика. Каждое представление написано и поддерживается экспертом в данной области. Существует свыше 35 таких представлений, поэтому, вероятно, среди них будет один или несколько, имеющих отношение к вашей области интересов. Мы рекомендуем каждому новичку найти и прочитать хотя бы одно из представлений, чтобы познакомиться с возможностями R (рецепт 1.12).

Документация пакетов

Большинство пакетов содержат полезную документацию. Документация хранится вместе с пакетами в репозиториях пакетов, таких как CRAN (https://cran.r-project.org), и автоматически устанавливается на ваш компьютер при установке пакета.

Сайты с вопросами и ответами на них

На сайте, содержащем вопросы и ответы на них, каждый может задать вопрос, на который могут ответить знающие люди. Читатели голосуют за ответы, поэтому со временем будут показаны лучшие ответы. Вся эта информация помечается и архивируется для поиска. Эти сайты представляют собой нечто среднее между списком рассылки и социальной сетью; канонический пример – сайт Stack Overflow (https://stackoverflow.com).

Сеть

В сети много информации о R, и есть специальные инструменты для ее поиска (рецепт 1.11). Информация, содержащаяся в сети, постоянно обновляется, поэтому ищите новые, улучшенные способы организации и поиска информации о языке R.

Списки рассылки

Волонтеры щедро пожертвовали много часов, чтобы ответить на вопросы новичков, которые публикуются в списках рассылок R. Списки архивируются, поэтому вы можете искать в архивах ответы на свои вопросы (рецепт 1.13).

1.1. Загрузка и установка R

Задача

Вам необходимо установить R на свой компьютер.

Решение

Пользователи Windows и macOS могут скачать R на сайте CRAN (Comprehensive R Archive Network). Пользователи Linux и Unix могут установить пакеты R с помощью своего инструмента управления пакетами.

Windows

1. Перейдите на страницу http://www.r-project.org/ в своем браузере.
2. Нажмите **CRAN**. Вы увидите список зеркал сайта, упорядоченных по странам.
3. Выберите ближайший к вам сайт или самый первый указанный в верхней части списка (0-Cloud), который, как правило, хорошо работает для большинства местоположений (https://cloud.r-project.org/).
4. Нажмите **Download R for Windows** (Скачать R для Windows) в разделе **Download and Install R** (Скачать и установить R).
5. Нажмите **base**.
6. Нажмите на ссылку, чтобы скачать последнюю версию R (файл .exe).
7. По завершении загрузки дважды щелкните файл *.exe* и ответьте на привычные вопросы.

macOS

1. Перейдите на страницу http://www.r-project.org/ в своем браузере.
2. Нажмите **CRAN**. Вы увидите список зеркал сайта, упорядоченных по странам.

3. Выберите ближайший к вам сайт или самый первый указанный в верхней части списка (0-Cloud), который, как правило, хорошо работает для большинства местоположений.
4. Нажмите **Download R for (Mac) OS X** (Скачать R для (Mac) OS X).
5. Нажмите на файл *.pkg* для последней версии R в разделе **Latest release** (Последний выпуск), чтобы скачать его.
6. По завершении загрузки дважды щелкните файл *.pkg* и ответьте на привычные вопросы.

Linux или Unix

В основных дистрибутивах Linux есть пакеты для установки R. В табл. 1-1 приведены примеры.

Таблица 1-1. Дистрибутивы Linux

Дистрибутив	Название пакета
Ubuntu или Debian	r-base
Red Hat или Fedora	R.i386
SUSE	R-base

Используйте системный менеджер пакетов для скачивания и установки пакета. Вам понадобится пароль администратора или привилегии sudo; в противном случае попросите системного администратора выполнить установку.

Обсуждение

Установка R в Windows или macOS проста, потому что для этих платформ существуют готовые двоичные файлы (скомпилированные программы). Вам нужно всего лишь следовать предыдущим инструкциям. Страницы CRAN также содержат ссылки на ресурсы, связанные с установкой, например задаваемые вопросы (FAQ) и советы для особых ситуаций («Работает ли R на Windows Vista / 7/8 / Server 2008?»), что может оказаться полезным.

Наиболее подходящий способ установить R в Linux или Unix – использовать диспетчер пакетов дистрибутивов Linux для установки R в виде пакета. Дистрибутивные пакеты значительно упрощают как первоначальную установку, так и последующие обновления.

В Ubuntu или Debian используйте команду `apt-get` для загрузки и установки R. Воспользуйтесь утилитой `sudo`, чтобы получить необходимые привилегии:

```
$ sudo apt-get install r-base
```

В Red Hat или Fedora используйте `yum`:

```
$ sudo yum install R.i386
```

Большинство платформ Linux также имеют графические менеджеры пакетов, которые могут оказаться более удобными.

Помимо базовых пакетов, мы рекомендуем также установить пакеты документации. Нам нравится `r-base-html` (потому что мы любим просматривать документацию с гиперссылками), а также `r-doc-html`, который устанавливает важные руководства по R локально:

```
$ sudo apt-get install r-base-html r-doc-html
```

Некоторые репозитории Linux включают в себя готовые копии пакетов R, доступных в CRAN. Мы не используем их, потому что предпочитаем получать программное обеспечение непосредственно от самого CRAN, где обычно имеются самые свежие версии.

В редких случаях вам может понадобиться выполнить сборку R с нуля. У вас может быть неподдерживаемая версия Unix, или у вас могут быть особые соображения относительно производительности или конфигурации. Процедура сборки в Linux или Unix вполне стандартна. Скачайте архив с домашней страницы вашего зеркала CRAN; он будет называться как-нибудь вроде *R-3.5.1.tar.gz*, за исключением того, что вместо 3.5.1 будет указана последняя версия. Распакуйте архив, найдите файл с именем *INSTALL* и следуйте инструкциям.

См. также

В книге *R in a Nutshell* Джозефа Адлера (издательство *O'Reilly*) содержится более подробная информация о загрузке и установке R, включая инструкции по сборке версий для Windows и macOS. Возможно, самым полным является руководство по установке и администрированию R, доступное в CRAN, где описывается сборка и установка R на различных платформах.

Это рецепт установки базового пакета. *См.* рецепт 3.10 для установки пакетов дополнений из CRAN.

1.2. Установка RStudio

Задача

Вам нужна более комплексная интегрированная среда разработки (IDE), чем та, что есть в R по умолчанию. Другими словами, вы хотите установить RStudio Desktop.

Решение

За последние несколько лет RStudio стал наиболее широко используемой интегрированной средой разработки для R. Мы считаем, что почти вся работа с R должна выполняться в среде разработки RStudio Desktop, если нет веской причины поступить иначе. RStudio производит ряд продуктов, включая RStudio Desktop, RStudio Server и RStudio Shiny Server среди прочих. В этой книге мы будем использовать термин RStudio для обозначения RStudio Desktop, хотя большинство концепций применимо и к RStudio Server.

Чтобы установить RStudio, загрузите последнюю версию установщика для вашей платформы с сайта RStudio (https://rstudio.com/products/rstudio/download/).

Версию RStudio Desktop с открытым исходным кодом можно скачать и использовать бесплатно.

Обсуждение

В этой книге используются RStudio версии 1.2.x и R версии 3.5.x. Новые версии RStudio выпускаются каждые несколько месяцев, поэтому регулярно обновляй-

те их. Обратите внимание, что RStudio работает с любой версией R, которую вы установили, поэтому обновление до последней версии RStudio *не* обновляет вашу версию R. R необходимо обновлять отдельно.

Взаимодействие с R в RStudio немного отличается от встроенного пользовательского интерфейса R. Для этой книги мы решили использовать RStudio для всех примеров.

1.3. Запуск RStudio

Задача

Вы хотите запустить RStudio на своем компьютере.

Решение

Распространенная ошибка, которую допускают новые пользователи R и RStudio, – это случайный запуск R, когда они намеревались запустить RStudio. Самый простой способ убедиться в том, что вы действительно запускаете RStudio, – это найти RStudio на своем рабочем столе, а затем использовать любой метод, который предоставляет ваша ОС для закрепления иконки там, где ее легко можно будет найти позже:

Windows

Нажмите на меню **Пуск** в левом нижнем углу экрана. В поле поиска введите `RStudio`.

macOs

Найдите в панели запуска приложение RStudio или нажмите **Cmd-space** (**Cmd** – это команда или клавиша ⌘)) и введите `RStudio`, чтобы выполнить поиск с помощью Spotlight Search.

Ubuntu

Нажмите сочетание клавиш **Alt-F1** и введите `RStudio` для поиска RStudio.

Обсуждение

На рис. 1-1, видно, что иконки R и RStudio похожи , что может привести вас в замешательство

Рис. 1-1. Иконки R и RStudio в macOS

Конец ознакомительного фрагмента.
Приобрести книгу можно
в интернет-магазине «Электронный универс»
(e-Univers.ru)