

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	<i>Регрессия</i>	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	<i>Дисперсионный анализ</i>	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>pwg</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 *Продвинутые 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 *Продвинутые 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 <i>Создание динамических отчетов</i>	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 <i>Создание пакетов</i>	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 <i>Продвинутая графика с использованием пакета lattice</i>	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

Предисловие от издательства

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 рецензии». 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 следующие 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. Издательство «ДМК Пресс» очень серьезно относится 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 dmkpress@gmail.com.

Мы высоко ценим любую помощь по защите наших авторов, благодаря которой мы можем предоставлять вам качественные материалы.

Предисловие

Что толку в книжке, если в ней нет ни картинок, ни разговоров?

Алиса. *«Алиса в Стране чудес»*¹

Оно чудесно и наделено сокровищами, способными удовлетворить всех от мала до велика, но не предназначено для робких духом.

Кью. *Сериал «Звездный путь: следующее поколение»*

Когда я начал писать эту книгу, я потратил довольно много времени на выбор хорошего эпиграфа. В итоге я остановился на этих двух. R – это потрясающе гибкая платформа и язык для исследования, визуализации и интерпретации данных. Я выбрал цитату из «Алисы в Стране чудес», чтобы передать суть современного статистического анализа – интерактивного процесса, состоящего из исследования, визуализации и интерпретации.

Вторая цитата отражает широко распространенное мнение о том, что R сложен в изучении. Я надеюсь показать вам, что это не так. R обладает настолько широкими возможностями и предлагает такое огромное число аналитических и графических функций (по последним подсчетам их более 50 000), что в одинаковой степени может вызывать бессознательный страх и у новичков, и у опытных пользователей. Однако в этом кажущемся безумии есть своя логика и поэзия. Вооружившись руководствами и инструкциями, вы сможете сориентироваться в огромном разнообразии возможностей и выбрать те инструменты, которые нужны для эффективного и элегантного решения вашей задачи.

Первое мое знакомство с R состоялось несколько лет назад, когда я подал заявление о приеме на должность консультанта по статистике. На встрече перед собеседованием будущий работодатель

¹ Перевод Н. Демуровой.

спросил меня, владею ли я языком R. Следуя стандартным советам специалистов по подбору персонала, я немедленно сказал «да» и приступил к его изучению. Я был опытным статистиком и исследователем с 25-летним опытом программирования в SAS и SPSS, свободно владел несколькими языками программирования. Что тут может быть сложного? Знаменитые последние слова.

Стремясь выучить этот язык программирования (как можно быстрее, ведь день собеседования приближался с угрожающей быстротой), я находил или тома, посвященные внутренней структуре языка, или многочисленные трактаты об отдельных продвинутых статистических методах, написанных специалистами в данной области для своих коллег. Встроенная справка была слишком лаконичной и служила скорее справочником, чем учебным пособием. Каждый раз, когда мне казалось, что я освоил общую логику и возможности R, обнаруживалось что-то новое, заставлявшее почувствовать себя невежественным и ничтожным.

Взявшись осваивать R, я подошел к процессу с позиции исследователя данных. Я пытался понять, что нужно сделать, чтобы успешно обработать, проанализировать и интерпретировать данные, и выделил следующие важные аспекты:

- доступ к данным (получение данных из разных источников);
- очистка данных (замена или удаление пропущенных значений, преобразование признаков в более удобный для обработки формат);
- аннотирование данных (чтобы можно было вспомнить, что представляет каждый их фрагмент);
- обобщение данных (вычисление описательных статистик, помогающих характеризовать данные);
- визуализация данных (потому что картинка на самом деле стоит тысячи слов);
- моделирование данных (выявление зависимостей и проверка гипотез);
- оформление результатов (подготовка таблиц и диаграмм достаточного для публикации качества).

Затем я постарался понять, как можно использовать R, чтобы выполнить каждую из этих задач. Поскольку я лучше всего учусь, обучая других, со временем я создал сайт (www.statmethods.net), на котором рассказываю все, что узнал сам.

Затем, спустя год, Марьян Бейс (Marjan Base) из издательства Manning позвонила и спросила, не хочу ли я написать книгу про R. К этому времени у меня уже было 50 статей в научных журналах, четыре технических руководства, многочисленные главы в книгах и целая книга по методологии исследований, и что тут может быть сложного? Рискую повториться – знаменитые последние слова.

Первое издание вышло в 2011 году, а второе – в 2015-м. Над третьим изданием я начал работать два с половиной года назад. Описание R всегда было непростой задачей, но за последние несколько лет произошла почти что революция, обусловленная ростом популярности больших данных, широким внедрением программного обеспечения tidyverse (tidyverse.org), быстрой разработкой новых подходов к прогнозной аналитике и машинному обучению, а также появлением новых и более мощных технологий визуализации данных. Я хотел отразить все эти важные изменения в третьем издании.

Книгу, которую вы держите в руках, я мечтал иметь много лет назад. Я постарался написать для вас путеводитель по R, который позволит быстро овладеть всеми возможностями этого уникального продукта с открытым исходным кодом, не испытав разочарований и раздражения, которые пришлось испытать мне. Надеюсь, вам понравится.

P.S. Мне предложили ту должность, но я отказался. Однако знакомство с R развернуло мою карьеру в совершенно неожиданном направлении. Жизнь может быть забавной штукой.

Благодарности

Многие люди приложили значительные усилия, чтобы сделать эту книгу лучше:

- в первую очередь это Марьян Бейс (Marjan Base), глава издательства Manning, которая предложила мне написать эту книгу;
- Себастьян Стирлинг (Sebastian Stirling), Дженифер Стоут (Jennifer Stout) и Карен Миллер (Karen Miller), редакторы-консультанты по аудитории (development editor) первого, второго и третьего изданий этой книги. Они провели многие часы в телефонных беседах со мной, помогая организовать материал, прояснить основные идеи и в целом сделать текст более интересным;
- Майк Шепард (Mike Shepard), научный редактор, который помог выявить непонятные места и представил свое экспертное мнение о тестировании кода. Я мог смело положиться на его подробные отзывы и суждения;
- Алекс Драгосавлевич (Aleks Dragosavljević), редактор-рецензент (review editor), помог найти рецензентов и координировал процесс рецензирования;
- Дейдре Хайам, помогавшая следить за процессом подготовки книги к печати, и ее команда: Сюзанна Дж. Фокс (Suzanne G. Fox), мой редактор, и Кэти Теннант (Katie Tennant), корректор;
- рецензенты, которые потратили много времени на внимательное чтение текста, находили опечатки и делали ценные замечания: Ален Ломпо (Alain Lompo), Алессандро Пуциелли (Alessandro Puzielli), Арав Агарвал (Arav Agarwal), Эшли Пол Итли (Ashley Paul Eatly), Клеменс Баадер (Clemens Baader), Дэниел Си Догерти (Daniel C Daugherty), Дэниел Кенни-Юнг (Daniel Kenney-Jung), Эрико Лендзиан (Erico Lenzian),

Джеймс Фронхофер (James Frohnhofner), Жан-Франсуа Морин (Jean-François Morin), Дженис Том (Jenice Tom), Джим Фронхофер (Jim Frohnhofner), Кей Энгельхардт (Kay Engelhardt), Келвин Микс (Kelvin Meeks), Кришна Шреста (Krishna Shrestha), Луис Фелипе Медейро Алвес (Luis Felipe Medeiros Alves), Марио Гизель (Mario Giesel), Мартин Перри (Martin Perry), Ник Дрозд (Nick Drozd), Николь Кенигштейн (Nicole Koenigstein), Роберт Самохил (Robert Samohyl), Тиклу Гангули (Tiklu Ganguly), Том Джеффрис (Tom Jeffries), Ульрих Гюгер (Ulrich Gueger), Вишал Сингх (Vishal Singh);

- многие участники программы раннего доступа издательства Manning (Manning Early Access Program, MEAP), купившие книгу до того, как она была закончена, задавали великолепные вопросы, указывали на ошибки и давали ценные подсказки.

Все перечисленные помогли сделать эту книгу лучше и полнее.

Я также хотел бы поблагодарить многочисленных разработчиков, сделавших R такой мощной платформой для анализа данных. Этот список включает не только основную команду разработчиков, но и многочисленных сторонников, создавших и поддерживающих дополнительные пакеты, значительно расширяющие возможности R. В приложении E перечислены авторы всех пакетов, упомянутых в этой книге. Отдельно я хотел бы упомянуть Джона Фокса (John Fox), Хадли Викхама (Hadley Wickham), Франка Е. Харрела младшего (Frank E. Harrell Jr.), Дипаяна Саркара (Deeprayan Sarkar) и Вильяма Ревилла (William Revelle), работами которых я восхищаюсь. Я старался как следует отразить их вклад, а ответственность за все ошибки и искажения, непреднамеренно допущенные в этой книге, лежит исключительно на мне.

На самом деле мне следовало бы начать эту книгу с благодарности моей жене и другу Кэрол Линн (Carol Lynn). Она не особенно интересуется статистикой или программированием, но неоднократно прочитала каждую главу и внесла множество исправлений и предложений. Никаким другим способом нельзя выразить свою любовь к другому человеку лучше, чем прочесть ради него текст по многомерной статистике. Она проявила необычайное терпение в вечера и выходные, которые я проводил в работе над этой книгой, выражая свою поддержку и проявляя такт и сочувствие. И за что это мне так повезло?

Есть еще два человека, которых я хочу поблагодарить. Один из них – мой отец, любовь которого к науке вдохновляла меня и помогла понять ценность данных. Другой человек – Гари К. Бургер (Gary K. Burger), мой руководитель в магистратуре. Гари заинтересовал меня статистикой и преподаванием, в то время как я собирался стать врачом. Это все он.

Об этой книге

Если вы выбрали эту книгу, скорее всего, у вас есть какие-то данные, которые нужно собрать, обобщить, преобразовать, исследовать, смоделировать, визуализировать или представить коллегам. Если это так, то R создан для вас! R стал всемирно известным языком программирования для статистического анализа и визуализации данных. В нем реализовано множество методов анализа данных, от самых простых до самых сложных и современных.

Как проект с открытым кодом он доступен для многих платформ, включая Windows, Mac OS X и Linux. Он постоянно развивается, и ежедневно появляются новые процедуры. Кроме того, R поддерживается большим и многоликим сообществом ученых и программистов, которые охотно помогут новичку советами.

Платформа R больше, пожалуй, известна за способность создавать красивые и сложные диаграммы, она может справиться с любой статистической задачей. Базовая версия содержит сотни функций для статистического анализа, управления данными и построения диаграмм. Однако некоторые особенно мощные методы реализованы в дополнительных пакетах, созданных независимыми авторами.

Эта широта возможностей имеет свою цену. Новичкам порой сложно понять, что такое R и как работать с этим языком. Даже самые опытные пользователи R с удивлением обнаруживают какие-то возможности, о которых не подозревали.

Третье издание «R в действии» – это руководство-путеводитель по R, знакомящее с самой платформой и ее возможностями. В книге описаны наиболее полезные функции базовой версии и более 90 наиболее часто используемых дополнительных пакетов. Основной упор в книге делается на практическое применение – на то, чтобы вы, руководствуясь прочитанным, могли проанализировать

Конец ознакомительного фрагмента.

Приобрести книгу можно

в интернет-магазине

«Электронный универс»

e-Univers.ru