

УДК 311:004.9R

ББК 60.6с515

М32

А

Мастецкий, Сергей Эдуардович.

М32 Статистический анализ и визуализация данных с помощью R / С. Э. Мастецкий, В. К. Шитиков. — 2-е изд., эл. — 1 файл pdf : 497 с. — Москва : ДМК Пресс, 2023. — Систем. требования: Adobe Reader XI либо Adobe Digital Editions 4.5 ; экран 10". — Текст : электронный.

ISBN 978-5-89818-601-2

Сегодня язык R является безусловным лидером среди свободно распространяемых систем статистического анализа. Ведущие университеты мира, аналитики крупнейших компаний и исследовательских центров регулярно используют R при проведении научно-технических расчетов и создании крупных информационных проектов. Широкое преподавание статистики на базе этой системы и всемерная поддержка научным сообществом обусловили то, что приведение скриптов кода на языке R постепенно становится общепризнанным стандартом как в журнальных публикациях, так и при неформальном общении ученых всего мира. Настоящая книга дополняет небольшую (пока) коллекцию работ по R на русском языке, обобщая и значительно расширяя совокупность методических сообщений, опубликованных ранее одним из авторов в блоге «R: Анализ и визуализация данных» (r-analytics.blogspot.com).

Книга адресована студентам, аспирантам, а также молодым и состоявшимся ученым, желающим освоить классические и современные методы анализа данных с использованием языка R.

УДК 311:004.9R

ББК 60.6с515

Электронное издание на основе печатного издания: Статистический анализ и визуализация данных с помощью R / С. Э. Мастецкий, В. К. Шитиков. — Москва : ДМК Пресс, 2015. — 496 с. — ISBN 978-5-97060-301-7. — 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 978-5-89818-601-2

© Мастецкий С. Э., Шитиков В. К., 2015

© Оформление, издание, ДМК Пресс, 2015

А

Содержание

Предисловие	10
--------------------------	-----------

Глава 1. Основные компоненты статистической среды R	13
--	-----------

1.1. История возникновения и основные принципы организации среды R.....	13
1.2. Работа с командной консолью	17
1.3. Работа с меню R Commander.....	20
1.4. Объекты, пакеты, функции, устройства	24

Глава 2. Описание языка R	31
--	-----------

2.1. Типы данных	31
2.2. Векторы и матрицы	32
2.3. Факторы	38
2.4. Списки и таблицы данных	40
Заполнение пустых значений.....	45
Сортировка таблиц	46
Объединение таблиц	46
2.5. Импортирование данных в R.....	47
2.6. Представление дат и времени. Временные ряды	51
Форматы представления дат и времени	51
Вычисления с датами и временем.....	52
Преобразование текстовых переменных в машинный формат времени	53
Временные ряды	54
2.7. Организация вычислений: функции, ветвления, циклы	56
Написание собственных функций.....	57
Условия и циклы.....	59
2.8. Векторизованные вычисления в R.....	61

Глава 3. Базовые графические возможности R	70
---	-----------

3.1. Функция plot() и ее параметры	70
Управляющие параметры функции plot()	73
Общие аргументы графических функций	74
3.2. Гистограммы, функции ядерной плотности и функция cdplot()	79
3.3. Диаграммы размахов.....	87
3.4. Круговые и столбиковые диаграммы	91

3.5. Диаграммы Кливленда и одномерные диаграммы рассеяния	99
3.6. Категоризованные графики	107

Глава 4. Описательная статистика, подгонка распределений и смежные задачи 114

4.1. Базовые функции для расчета параметров описательной статистики.....	114
4.2. <code>summary()</code> и функции из дополнительных пакетов	118
4.3. Анализ выбросов	121
4.4. Заполнение пропущенных значений в таблицах данных	125
4.5. Воспроизводимость результатов при использовании генератора случайных чисел	131
4.6. Законы распределения вероятностей, реализованные в R	134
4.7. Подбор закона и параметров распределения в R	136
4.8. Проверка на нормальность распределения	144
Графические способы.....	145
Формальные тесты.....	148

Глава 5. Классические методы статистики 151

5.1. Гипотеза о равенстве средних двух генеральных совокупностей.....	151
Одновыборочный t-критерий	151
Сравнение двух независимых выборок	153
Сравнение двух зависимых выборок.....	155
5.2. Ранговый критерий Уилкоксона-Манна-Уитни	157
Одновыборочный критерий Уилкоксона	157
Сравнение двух независимых выборок	158
Сравнение двух зависимых выборок.....	159
5.3. Рандомизация, бутстреп и оценка статистической мощности (на примере двухвыборочного t-критерия).....	161
5.4. Гипотеза об однородности дисперсий	168
Проверка однородности дисперсии в двух группах	168
Проверка однородности дисперсии в нескольких группах	169
5.5. Введение в дисперсионный анализ	171
Постановка задачи.....	171
Две оценки генеральной дисперсии в дисперсионном анализе	174
Выполнение дисперсионного анализа в R.....	176
Двухфакторный дисперсионный анализ.....	176
5.6. Оценка корреляции двух случайных величин.....	180
5.7. Критерий хи-квадрат	184
Критерий хи-квадрат для таблиц сопряженности размером 2×2	184

Критерий хи-квадрат для таблиц сопряженности размером больше 2×2	187
5.8. Точный тест Фишера. Критерии Мак-Немара и Кохрана-Мантеля-Хензеля.....	187
Точный тест Фишера.....	187
Критерий Мак-Немара	190
Критерий Кохрана-Мантеля-Хензеля для таблиц сопряженности размером $2 \times 2 \times K$	193
5.9. Оценка статистической мощности при сравнении частот.....	197

Глава 6. Дисперсионный анализ.....203

6.1. Протокол разведочного анализа данных	203
Выявление точек-выбросов	204
Проверка однородности групповых дисперсий.....	205
Проверка на нормальность распределения	206
Выявление избыточного числа нулевых значений	207
Выявление коллинеарности	207
Выявление формы связи между переменными	210
Выявление взаимодействий между предикторами	212
Влияние пространственно-временных факторов на анализируемую переменную.....	216
6.2. Дисперсионный анализ как линейная модель	219
6.3. Структура модельных объектов дисперсионного анализа	227
6.4. Оценка адекватности модели дисперсионного анализа	230
Проверка исходных предположений общей линейной модели	230
Проверка условия нормальности распределения	231
Проверка условия однородности групповых дисперсий.....	234
Что делать, когда однофакторный дисперсионный анализ неприменим?....	237
6.5. Дисперсионный анализ по Краскелу-Уоллису	239
6.6. Модели двух- и многофакторного дисперсионного анализа	241
Синтаксис объекта «формула»	242
Выполнение двухфакторного дисперсионного анализа при помощи функции <code>lm()</code>	244
Порядок перечисления предикторов в формуле модели	246
Многофакторный дисперсионный анализ	248
6.7. Контрасты в линейных моделях, содержащих категориальные предикторы.....	249
Основные понятия	250
Контрасты комбинаций условий (treatment contrasts).....	252
Контрасты сумм (sum contrasts)	254
Контрасты Хелмерта	255

Контрасты, задаваемые пользователем	257
6.8. Проблема множественных проверок статистических гипотез	258
Поправка Бонферрони.....	261
Метод Холма	262
Метод Беньямини-Хохберга.....	263
Метод Беньямини-Йекутили	266
6.9. Апостериорные сравнения групповых средних.....	267
Критерий Тьюки	268
Методы множественных проверок гипотез, реализованные в пакете multcomp	271

Глава 7. Регрессионные модели зависимостей между количественными переменными279

7.1. О понятии «статистическая модель»	279
Пример простейшей статистической модели	279
Исследование свойств статистических моделей имитационными методами	282
Пример модели с одним количественным предиктором.....	287
Назначение регрессионных моделей.....	289
7.2. Простая линейная регрессия: каков возраст Вселенной?.....	290
Модель для оценки постоянной Хаббла	291
Доверительные интервалы.....	293
Оценка неопределенности в отношении параметров линейной регрессии	295
Оценка «качества» регрессионной модели.....	301
7.3. Стандартные методы диагностики линейных моделей	304
Проверка допущений в отношении остатков модели.....	304
Проверка адекватности структуры систематической части модели	308
Встроенные диагностические графики.....	313
Выявление необычных и влиятельных наблюдений	314
7.4. Модели регрессии при разных видах функции потерь.....	325
Два типа регрессионных моделей	325
Робастные процедуры	329
7.5. Критерии выбора моделей оптимальной сложности.....	331
7.6. Полиномиальные и нелинейные модели регрессии	335
Полиномиальная регрессия	335
Нелинейная регрессия.....	338
7.7. Модель множественной регрессии и выбор ее спецификации	344
Полная модель и обоснование необходимости ее оптимизации	345
Пошаговые алгоритмы селекции переменных.....	347

Построение «всех возможных моделей»	348
Пошаговое включение предикторов в сочетании с перекрестной проверкой	350
7.8. Диагностика моделей множественной регрессии	353
Сравнение нескольких альтернативных моделей	353
Диагностика допущений в отношении остатков модели.....	354
Учет нелинейного характера влияния предикторов на отклик	359
7.9. Регуляризация множественной регрессии.....	361
Гребневая регрессия.....	362
Лассо-регрессия Тибширани	364
7.10. Регрессия на главные компоненты	366
7.11. Сравнение эффективности различных моделей при прогнозировании	372
Формирование исходных данных для построения моделей	372
Общая линейная модель и ее тестирование на проверочной выборке	374
Выбор информативного комплекса предикторов	375
Модели с использованием регуляризации	377
Регрессия на главные компоненты.....	380
Результаты и некоторые выводы	382

Глава 8. Обобщенные, структурные и иные модели регрессии384

8.1. Модели сглаживания	384
Ядерная модель сглаживания	389
Сплаины.....	393
8.2. Обобщенные модели регрессии	395
8.3. Модели пробит- и логит-регрессии	399
Пробит-регрессия для моделирования зависимости «доза–эффект»	400
Логистическая регрессия.....	407
8.4. Пример использования обобщенных моделей для оценки экологической толерантности	411
Модели с нормально распределенным откликом	412
Модели с бинарным откликом.....	416
8.5. Ковариационный анализ.....	419
8.6. Модели со смешанными эффектами для иерархически организованных данных.....	424
Основные идеи	424
Пример с морскими животными: несколько частных моделей.....	426
8.7. Индуктивные модели (метод группового учета аргументов).....	433

8.8. Моделирование структурными уравнениями	440
---	-----

Глава 9. Пространственный анализ и создание картограмм451

9.1. Простая карта: использование растрового рисунка и расчет расстояний.....	451
Использование географических расстояний в статистическом анализе	452
Расчет расстояния между объектами по их географическим координатам	457
9.2. Анализ пространственного размещения точек	460
9.3. Использование сервисов картографической системы Google Maps	466
9.4. Создание картограмм при помощи R	469
Шейп-файлы	470
Функция <code>spplot()</code> из пакета <code>sp</code>	474
Создание картограмм при помощи пакета <code>ggplot2</code>	478

Библиография и интернет-ресурсы484

Основные литературные ссылки по тексту книги	484
Литература по R.....	484
Общеметодическая литература по статистическому анализу	485
Библиографический указатель литературы по R.....	485
Рекомендуемые интернет-ресурсы	494
Русскоязычные ресурсы.....	494
Англоязычные ресурсы	495