

黄湘云
©

现代应用统计与 R 语言

黄湘云

2021-08-07

目录

欢迎	1
本书风格	1
本书定位	10
内容概要	11
致谢名单	11
授权说明	12
运行信息	12
第一章 前言	14
1.1 语言抉择	14
1.2 数据科学	18
1.3 获取帮助	18
1.4 写作环境	18
1.5 记号约定	20
1.6 复现环境	21
1.7 如何发问	22
1.8 作者简介	23
第二章 符号说明	24
第三章 文件操作	28
第四章 数据结构	29
4.1 字符	29
4.2 向量	29
4.3 矩阵	29
4.4 数组	29
4.5 列表	29



4.6 日期	30
第五章 数据操作	33
5.1 查看数据	34
5.2 提取子集	36
5.3 数据重塑	49
5.4 数据转换	54
5.5 按列排序	59
5.6 数据拆分	73
5.7 数据合并	87
5.8 数据去重	90
5.9 数据缺失	91
5.10 数据聚合	93
5.11 表格统计	106
5.12 索引访问	111
5.13 多维数组	111
5.14 其它操作	113
5.14.1 列表属性	113
5.14.2 堆叠向量	114
5.14.3 属性转化	117
5.14.4 绑定环境	118
5.14.5 数据环境	118
5.15 apply 族	123
5.16 with 选项	128
5.17 分组聚合	131
5.18 合并操作	133
5.19 长宽转换	134
5.20 对符合条件的列操作	136
5.21 CASE WHEN 和 fcase	138
5.22 数据操作实战	138
5.23 高频数据操作	138
5.23.1 循环合并	139
5.23.2 分组计数	139
5.23.3 分组抽样	140
5.23.4 分组排序	141



6.1	Spark 与 R 语言	147
6.1.1	sparklyr	147
6.1.2	SparkR	152
6.2	数据库与 R 语言	153
6.3	批量读取 csv 文件	154
6.4	批量导出 xlsx 文件	156
第七章 数据可视化		158
7.1	元素	161
7.1.1	标签	161
7.1.2	注释	161
7.1.3	主题	163
7.2	字体	164
7.2.1	系统字体	164
7.2.2	思源字体	171
7.2.3	数学字体	172
7.2.4	TikZ 设备	176
7.2.5	漫画字体	179
7.2.6	表情字体	180
7.3	配色	181
7.3.1	调色板	184
7.3.2	颜色模式	198
7.3.3	LaTeX 配色	201
7.3.4	ggplot2 配色	201
7.4	图库	203
7.4.1	饼图	203
7.4.2	地图	205
7.4.3	热图	210
7.4.4	条形图	210
7.4.5	函数图	215
7.4.6	密度图	215
7.4.7	提琴图	217
7.4.8	蜂群图	219
7.4.9	瓦片图	220
7.4.10	日历图	221
7.4.11	岭线图	225
7.4.12	椭圆图	226



7.4.13 包络图	229
7.4.14 拟合图	230
7.4.15 地形图	232
7.4.16 树状图	235
7.4.17 留存图	239
7.4.18 瀑布图	240
7.4.19 桑基图	241
7.4.20 马赛克图	243
7.4.21 凹凸图	243
7.4.22 水流图	244
7.4.23 时间线	245
7.4.24 三元图	248
7.4.25 四象限图	248
7.4.26 韦恩图	249
7.4.27 龙卷风图	249
7.4.28 聚类图	250
7.4.29 主成分图	252
7.4.30 组合图	253
7.4.31 动态图	254
第八章 动态文档	260
8.1 文档元素	262
8.1.1 控制选项	262
8.1.2 表格	264
8.1.3 流程图	264
8.2 便携式文档	265
8.2.1 文档汉化	265
8.2.2 添加水印	265
8.2.3 双栏排版	265
8.2.4 参数化报告	266
8.2.5 学术幻灯片	266
8.2.6 文档模版	267
8.2.7 引用文献	267
8.2.8 自定义块	268
8.3 网页文档	269
8.4 编写书籍	269
8.5 个人网站	269



8.6	微软文档	269
8.7	发送邮件	270
8.8	工作流	272
8.9	运行环境	272
第九章 交互图形		274
9.1	散点图	278
9.2	条形图	279
9.3	折线图	280
9.4	双轴图	280
9.5	直方图	282
9.6	箱线图	282
9.7	提琴图	284
9.8	气泡图	285
9.9	曲线图	286
9.10	堆积图	286
9.11	热力图	287
9.12	地图 I	287
9.13	拟合图	289
9.14	轨迹图	290
9.15	三维图	292
9.16	甘特图	292
9.17	帕雷托图	294
9.18	时间线	296
9.19	漏斗图	296
9.20	雷达图	297
9.21	瀑布图	298
9.22	树状图	299
9.23	旭日图	299
9.24	调色板	300
9.25	面积图	301
9.26	动画 I	307
9.27	时序图	312
9.28	图形导出	313
9.29	地图 II	313
9.30	动画 II	317
9.31	网络图	320



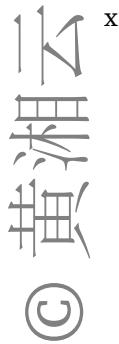
9.31.1 networkD3	320
9.31.2 visNetwork	321
9.31.3 r2d3	322
9.32 Python 交互图形	323
9.33 运行环境	323
第十章 交互表格	326
10.1 DT 和 reactable	327
10.2 gt 和 kableExtra	332
10.3 运行环境	335
第十一章 交互报表	337
11.1 开发流程	340
11.2 开发工具	340
11.3 基础知识	341
11.4 基础组件	341
11.4.1 书签	341
11.4.2 表格	342
11.5 高级主题	347
11.6 部署应用	349
11.7 最佳实践	349
11.8 仪表盘	349
11.9 交互式数据报表 dash	355
11.10 运行环境	355
第十二章 字符串操作	357
12.1 字符串加密	358
第十三章 正则表达式	359
第十四章 文本分析	360
第十五章 抽样分布	361
15.1 正态分布	361
15.2 指数族	362
第十六章 参数估计	366
16.1 点估计	366
16.1.1 矩估计	367



16.1.2 最小二乘估计	367
16.1.3 极大似然估计	371
16.2 区间估计	372
16.2.1 正态分布	372
16.2.2 0-1 分布	373
16.2.3 置信区间和信仰区间	376
16.3 最小角回归	385
16.4 刀切法	385
16.5 重抽样	385
16.6 Delta 方法	385
 第十七章 假设检验	 386
17.1 Ansari-Bradley 检验 ansari.test	389
17.2 Bartlett 检验 bartlett.test	390
17.3 二项检验 binom.test	390
17.4 时间序列独立性检验 Box.test	392
17.5 皮尔逊卡方检验 chisq.test	392
17.6 费舍尔精确检验 fisher.test	392
17.7 方差齐性检验 fligner.test	392
17.8 Friedman 秩和检验 friedman.test	393
17.9 Kruskal-Wallis 秩和检验 kruskal.test	393
17.10 同分布检验 ks.test	394
17.11 Cochran-Mantel-Haenszel 卡方检验 mantelhaen.test	394
17.12 Mauchly 球形检验 mauchly.test	394
17.13 McNemar 卡方检验 mcnemar.test	395
17.14 Mood 方差检验 mood.test	395
17.15 单因素多重比较 oneway.test	396
17.16 配对样本的检验	398
17.16.1 配对比例检验 pairwise.prop.test	398
17.16.2 配对 t 检验 pairwise.t.test	398
17.16.3 配对 Wilcoxon 检验 pairwise.wilcox.test	399
17.16.4 配对样本相关性检验 cor.test	399
17.17 精确泊松检验 poisson.test	399
17.18 单位根检验 PP.test	400
17.19 比例检验 prop.test	400
17.19.1 两个独立二项总体等价性检验	402
17.19.2 不同页面的点击率问题	402



17.19.3 比例齐性检验	404
17.20 比例趋势检验 <code>prop.trend.test</code>	406
17.21 Quade 检验 <code>quade.test</code>	406
17.22 正态性检验 <code>shapiro.test</code>	407
17.23 正态性检验 Epps-Pully 检验	407
17.24 学生 t 检验 <code>t.test</code>	408
17.24.1 正态总体两样本的均值之差的检验	408
17.24.2 办公软件里的 T 检验	415
17.25 方差比检验 <code>var.test</code>	416
17.26 Wilcoxon 秩和检验 <code>wilcox.test</code>	416
17.26.1 ROC 曲线和 <code>wilcox.test</code> 检验的关系	417
17.27 3 + 1 统计检验	419
17.28 经典案例	420
17.28.1 1973 年加州大学伯克利分校的学生招生	420
17.28.2 1976~1977 年美国佛罗里达州的凶杀案件中被告肤色和死刑判决的关系	421
17.28.3 统计专业学生的头发和眼睛的颜色	421
17.29 运行环境	422
 第十八章 功效分析	 425
18.1 方差分析检验的功效	426
18.2 比例检验的功效	426
18.3 t 检验的功效	429
18.4 运行环境	433
 第十九章 试验设计	 435
19.1 学生睡眠质量	436
19.2 驱虫喷雾的效果	436
19.3 重复数不等的多重比较	444
19.4 不同地区的草类植物吸收二氧化碳的情况	445
19.5 果园喷雾剂的效力	445
19.6 验证孟德尔的豌豆实验结果	446
 第二十章 线性模型	 447
20.1 方差分析	447
20.2 单因素方差分析	447
20.3 双因素方差分析	452



20.4 多因素方差分析	452
20.5 核学习	453
20.6 通用机器学习	453
20.7 理论基础	454
20.8 多重多元线性回归	454
20.9 回归诊断	455
20.10 1977 年美国人口普查	458
20.11 石油岩石样品的测量	459
20.12 1888 年瑞士生育率分析	460
20.13 Intercountry Life-Cycle Savings Data 1960-1970	463
20.14 Longley's Economic Regression Data 1947-1962	463
20.15 甲醛的测定	463
20.16 迈克尔逊光速数据分析	464
20.17 不同喂食方式对小鸡体重的影响 I	466
20.18 不同喂食方式对小鸡体重的影响 II	466
20.19 酶的酶联免疫吸附测定	468
20.20 婴儿的体重随年龄的变化情况	469
20.21 火炬松树的生长情况	475
20.22 酶促反应的反应速率	477
20.23 茶碱的药代动力学	478
20.24 本章总结	481
20.25 运行环境	481
第二十一章 广义线性模型	483
21.1 介绍	484
21.2 理论基础	484
21.2.1 岭回归	484
21.2.2 Lasso	485
21.2.3 最优子集回归	485
21.2.4 偏最小二乘回归	485
21.3 吸烟喝酒和食道癌的关系	485
21.4 自然流产和人工流产后的不育	488
21.5 细菌数据集	491
21.6 研究婴儿出生体重低的相关危险因素	493
21.7 哥本哈根住房状况调查	499
21.8 癫痫病发作次数	501
21.9 对数线性模型	502



21.10 泊松回归模型	502
第二十二章 案例研究	511
22.1 统计学家生平	512
22.2 R 语言发展历史	512
22.3 不同实验条件下植物生长情况	512
22.4 橘树生长情况	522
第二十三章 数据探索	526
第二十四章 生存分析	527
24.1 急性粒细胞白血病生存数据	527
第二十五章 时序分析	529
25.1 时序数据	530
25.2 时序图	533
25.3 基本概念	535
25.4 时序检验	539
25.5 指数平滑	539
25.6 Holt-Winters	539
25.7 1749-2013 年太阳黑子数据	541
25.8 1991-1998 年欧洲主要股票市场日闭市价格指数	544
25.9 自回归模型	545
25.10 移动平均模型	545
25.11 自回归移动平均模型	545
25.12 自回归条件异方差模型	546
25.13 广义自回归条件异方差模型	546
25.14 其它特征的时间序列	546
25.15 港股走势	546
25.16 美股走势	548
25.17 51Talk 股价走势	548
25.18 运行环境	549
第二十六章 空间分析	552
26.1 冈比亚儿童疟疾	553
26.2 运行环境	556
第二十七章 空间建模	558



27.1 西非眼线虫病	558
27.2 运行环境	558
第二十八章 贝叶斯模型	561
28.1 软件配置	562
28.2 正态分布	562
28.3 高斯过程	565
28.4 分层正态模型	567
28.4.1 schools 数据	567
28.4.2 rats 数据	572
28.5 非线性模型	574
28.5.1 mcycle 数据	574
第二十九章 梯度提升机	576
29.1 XGBoost	576
第三十章 神经网络	577
30.1 mxnet	577
30.2 运行环境	578
第三十一章 矩阵运算	580
31.1 矩阵乘法	582
31.2 Hadamard 积	583
31.3 矩阵转置	583
31.4 矩阵外积	584
31.5 矩阵乘方	585
31.6 矩阵求幂	587
31.7 矩阵交叉积	587
31.8 矩阵行列式	587
31.9 矩阵条件数	588
31.10 矩阵求逆	588
31.11 矩阵伴随	590
31.12 矩阵范数	590
31.13 矩阵求秩	591
31.14 矩阵求迹	591
31.15 单位矩阵	592
31.16 对角矩阵	592
31.17 上/下三角矩阵	593



31.18 稀疏矩阵	594
31.19 三对角矩阵	595
31.20 LU 分解	595
31.21 Schur 分解	595
31.22 Cholesky 分解	595
31.23 特征值分解	596
31.24 SVD 分解	596
31.25 QR 分解	597
31.26 Jordan 分解	599
31.27 Givens 旋转	599
31.28 特殊函数	600
31.28.1 阶乘	600
31.28.2 伽马函数	600
31.28.3 贝塔函数	602
31.28.4 贝塞尔函数	602
第三十二章 符号计算	605
第三十三章 数值优化	610
33.1 线性规划	612
33.2 整数规划	614
33.2.1 一般整数规划	614
33.2.2 0-1 整数规划	614
33.2.3 混合整数规划	615
33.3 二次规划	617
33.3.1 凸二次规划	617
33.3.2 半正定二次优化	621
33.4 非线性规划	623
33.4.1 一元非线性优化	623
33.4.2 多元非线性无约束优化	624
33.4.3 多元非线性约束优化	645
33.5 非线性方程	666
33.5.1 一元非线性方程	666
33.5.2 非线性方程组	666
33.6 多目标规划	671
33.7 经典优化问题	672
33.8 回归与优化	672



33.9 对数似然	674
33.10 微分方程	675
33.10.1 常微分方程	677
33.10.2 偏微分方程	678
33.10.3 延迟微分方程	685
33.10.4 随机微分方程	685
33.11 运行环境	686
附录 A 命令行操作	688
A.1 查看文件	688
A.2 创建文件夹	690
A.3 移动文件	690
A.4 查看文件大小	691
A.5 终端模拟器	691
A.6 压缩和解压缩	691
A.7 从源码安装 R	693
A.8 安装软件	694
A.9 安装 R 包	695
A.10 软件包管理器	699
A.10.1 dnf	699
A.10.2 apt	701
附录 B 其它软件	703
B.1 文本编辑器	703
B.2 代码编辑器	704
B.3 集成开发环境	705
B.3.1 RStudio 桌面版	705
B.3.2 RStudio 服务器版	706
B.3.3 Shiny 服务器版	708
B.3.4 Eclipse + StatET	708
B.3.5 Emacs + ESS	709
B.3.6 Nvim-R	709
B.4 Git 版本控制	709
B.4.1 安装配置	711
B.4.2 追踪文件	712
B.4.3 合并上流	712
B.4.4 大文件支持	713



B.4.5 新建分支	713
B.4.6 创建 Github Pages 站点	714
B.4.7 博客主题	714
B.4.8 修改远程仓库的位置	715
B.4.9 统计代码仓库的提交量	716
B.4.10 账户共存	716
B.4.11 回车换行	717
B.4.12 子模块	718
B.4.13 克隆项目	718
B.4.14 创建 PR	718
B.4.15 修改 PR	718
B.5 Pandoc 文档处理	719
B.6 Calibre 书籍管理	720
B.7 ImageMagick 图像处理	720
B.8 OptiPNG 图片优化	721
B.9 PDFCrop 裁剪边空	722
B.10 PhantomJS 网页截图	722
B.11 Inkscape 矢量绘图	724
B.12 QPDF PDF 文件操作	724
B.13 UML 标准建模图	725
B.14 Graphviz 流程图	726
B.15 LaTeX 排版工具	728
B.15.1 TinyTeX 发行版	728
B.15.2 安装和更新	729
B.15.3 查询和搜索	729
B.15.4 TikZ 绘图工具	730
B.16 Octave 科学计算	730
B.17 Python 环境配置	732
B.18 Python 基础绘图	733
B.19 Python 基础操作	736
B.20 VBox 虚拟机	747
B.20.1 从命令行启动虚拟机	747
B.21 Docker 虚拟环境	748
B.22 安装的 R 包	753
附录 C 混合编程	773
C.1 函数源码	773



C.2 命名约定	775
C.3 R 与 JavaScripts	775
C.4 R 与 Python	776
C.5 R 与 C	776
C.6 R 与 C++	777
C.7 R 与 LaTeX	778
C.8 运行环境	779
附录 D 面向对象编程	782
D.1 环境	782
D.2 引用	783
D.3 调用栈	783
D.4 闭包	784
D.5 递归	784
D.6 异常	785
D.7 对象	785
D.8 泛型	785
D.9 除虫	787
D.10 性能	787
D.11 质量	787

插图

1	现代统计建模的三重境界	3
2	二项分布参数 p 的置信带	4
3	二项分布 $B(n, \theta)$ 成功概率 θ , 固定样本量 $n = 10$, 分不同的分位点 $q = 2, 4, 6$ 绘制概率随成功概率 θ 的变化	6
4	上帝在掷骰子吗?	8
5	散点图: faithful 数据集	9
1.1	R 语言扩展包生态系统	16
1.2	书籍项目架构图	19
5.1	Tidyverse 和 Base R 的关系	33
5.2	连续型变量分组统计	84
5.3	连续型变量分组统计	84
5.4	太阳黑子的频谱	124
5.5	lapply 函数	126
5.6	1945-1974 美国总统的支持率	127
5.7	with 操作	130
5.8	管道连接数据操作和可视化	137
7.1	简洁美观	160
7.2	美国总统支持率: 自 1945 年第一季度至 1974 年第四季度	162
7.3	文本注释	163
7.4	少量点的情况下可以全部注释, 且可以解决注释重叠的问题	164
7.5	调用系统字体绘图	166
7.6	在 ggplot2 绘图系统中设置中英文字体	168
7.7	调用 hrbrthemes 包设置字体主题	169
7.8	默认字体 Arial Narrow	170
7.9	showtext 包处理图里的中文	172



7.10	斐济地震带	173
7.11	fontcm 处理数学公式	175
7.12	嵌入数学字体	176
7.13	线性回归模型	178
7.14	漫画风格的字体方案	180
7.15	表情字体	181
7.16	R 3.6.0 以前的调色板	182
(a)	terrain.colors 调色板	182
(b)	heat.colors 调色板	182
(c)	topo.colors 调色板	182
(d)	cm.colors 调色板	182
7.17	R 3.6.0 以后的调色板	183
(a)	Grays 调色板	183
(b)	YlOrRd 调色板	183
(c)	Purples 3 调色板	183
(d)	Viridis 调色板	183
7.18	灰度调色板	184
7.19	提取 10 种灰色做调色板	185
7.20	桃色至梨色的渐变	186
7.21	colorRampPalette 自制调色板	187
7.22	RColorBrewer 调色板	188
7.23	grDevices 调色板	189
7.24	grDevices 调色板	191
7.25	colorspace 调色板	193
7.26	源起	195
7.27	Spectral 调色板	196
7.28	美国黄石国家公园的老忠实泉	197
7.29	几种不同的箱线图	202
(a)	简单箱线图	202
(b)	ggplot2 绘制的箱线图	202
(c)	ggplot2 调用默认调色板	202
(d)	ggplot2 调用 Google 调色板	202
7.30	饼图	204
7.31	1975 年美国各州犯罪事件	206
7.32	中国及其周边	206
7.33	画地图的正确姿势	208



(a) 墨卡托投影	208
(b) 北极观察	208
(c) 正交投影	208
(d) 正交投影北极观察	208
7.34 Google 地图示例	209
7.35 条形图的四种常见形态	211
7.36 添加注释到条形图	213
7.37 二维密度图	216
(a) 默认调色板	216
(b) viridis 调色板	216
7.38 几种不同的提琴图	218
(a) 简单箱线图	218
(b) vioplot 绘制的提琴图	218
(c) ggplot2 绘制的提琴图	218
(d) beanplot 绘制的提琴图	218
7.39 正态分布的概率密度曲线	219
7.40 蜜蜂图可视化效果比抖动图好	220
7.41 1949-1960 年国际航线乘客数量的月度趋势	221
7.42 1973 年 5 月至 9 月纽约的气温变化	222
7.43 《现代统计图形》的活跃情况	225
7.44 2016 年在内布拉斯加州林肯市的天气变化	226
7.45 比较数据的分布	227
7.46 几种不同的椭圆图	228
(a) 简单椭圆图	228
(b) 正态和 t 分布	228
(c) 填充几何图形	228
7.47 包络图	229
7.48 自定义样条函数	231
7.49 平滑方法	232
7.50 image 图形	233
7.51 锡安国家公园的高程栅格数据	234
7.52 矩阵树图	236
7.53 世界主要经济体 G20 的人口和经济信息	238
7.54 瀑布图	241
7.55 桑基图	242
7.56 UCBAdmissions 马赛克图	243



7.57 凹凸图	245
7.58 堆积区域图	246
7.59 数据科学的时间轴	247
7.60 四象限图	250
7.61 龙卷风图展示变量重要性	251
7.62 主成分分析	254
7.63 组合图形	255
7.64 药物在人体中的代谢情况	256
7.65 添加过渡效果	259
8.1 R Markdown 极其周边生态	260
8.2 rmarkdown 支持的输出格式	262
8.3 rmarkdown 生态系统	262
8.4 R Markdown 概念图	263
9.1 轨迹数据	291
9.2 斐济地震带	315
9.3 斐济地震带热力图	316
9.4 插入图片	324
11.1 Shiny 生态系统	338
11.2 开发 Shiny 应用扩展的组织	339
15.1 二维正态分布	364
15.2 二维正态分布	365
(a) 等高线图	365
(b) 透视图	365
16.1 μ 的置信水平为 0.95 的置信区间	374
17.1 草类植物吸收 CO ₂	397
17.2 学生睡眠数据 sleep	413
17.3 MacOS 的办公软件 Numbers 做两样本 T 检验	415
17.4 UCBAdmissions 马赛克图	421
17.5 头发、眼睛颜色和性别的比例	423
18.1 t 检验的功效	430
19.1 不同杀虫剂的效果	437



19.2 成对显著性水平	444
19.3 草类植物吸收二氧化碳的量	445
20.1 模型诊断很重要	457
20.2 线性模型可能在欺骗你	458
20.3 1888 年瑞士生育率和社会经济指标的关系	460
20.4 1879 年迈克尔逊光速实验数据	465
20.5 不同喂食方式对小鸡的影响	466
20.6 不同火炬树的生长情况	476
21.1 吸烟喝酒和食道癌的关系	487
22.1 模型	511
22.2 植物干重	513
22.3 橘树生长模型	524
24.1 急性粒细胞白血病生存数据	528
25.1 美国纽黑文的年平均气温，单位：华氏温度	534
25.2 时序图：太阳黑子月均数量	541
25.3 太阳黑子数量年平均时序图	543
25.4 月均太阳黑子数	543
25.5 1991-1998 年间欧洲主要股票市场日闭市价格指数图德国 DAX (Ibis), Switzerland SMI, 法国 CAC 和英国 FTSE	544
25.6 时间序列：非平稳、周期性、非线性	547
(a) 1960-1980 年强生公司每股季度收益	547
(b) 1949-1960 年月均航班乘客数量	547
(c) 1920-1939 年诺丁汉月均气温	547
(d) 1821-1934 年加拿大山猫陷阱数量	547
25.7 51Talk 公司上市以来的股价走势	549
25.8 CEO 股价变化趋势	550
26.1 冈比亚地形海拔数据	554
28.1 参数 μ, σ 的迭代轨迹图和后验分布图	566
(a) 参数的轨迹图	566
(b) 参数的后验分布图	566
32.1 Tetrachoric 函数	607



33.1 可行域	618
33.2 无约束和有约束条件下的解	621
33.3 二分类问题	622
33.4 Himmelblau 函数图像	626
33.5 香蕉函数图像	628
33.6 二维 Ackley 函数图像	631
33.7 Radistrigin 函数	634
33.8 Schaffer 函数	636
33.9 Schaffer 函数	637
33.10 Hölder 函数	639
33.11 Trid 函数	641
33.12 函数图像	643
33.13 局部放大函数图像	644
33.14 正态分布参数的负对数似然函数	676
33.15 洛伦兹曲线	678
33.16 Auckland Maunga Whau 火山地形图 $10m \times 10m$ 。火山的实况地形图 https://en.wikipedia.org/wiki/Maungawhau#/Mount_Eden	680
33.17 一维热传导方程的数值解热力图	681
33.18 解析解的二维图像	683
33.19 解析解的三维透视图像	684
A.1 RStudio IDE 集成的 Zsh 终端模拟器	692
B.1 Typora 主题	704
(a) 默认的主题	704
(b) Vue 主题	704
B.2 开源桌面版 RStudio 集成开发环境	705
B.3 虚拟机里的 RStudio	706
B.4 容器里的 RStudio	707
B.5 基于 Eclipse 的 R 集成开发环境 StatET	708
B.6 Git 代码版本管理	710
B.7 Git 日志查看器	711
B.8 没有优化	721
B.9 优化	722
B.10 图片制作、合成、优化、转换等常用工具	725
B.11 数据分析流程图	726
B.12 matplotlib 示例	735

B.13 matplotlib 复制示例	742
C.1 Python 图形	776
C.2 TeX 系统	780

表格

1 致谢名单	11
2.1 数学符号表	25
4.1 日期表格	32
5.1 <code>apply</code> 函数	123
5.2 不同生长环境下植物的干重	135
5.3 <code>iris</code> 数据集原顺序（左）和新顺序（右）	145
7.1 Windows 系统上四款字体的替代品	165
7.2 数值和比例组合呈现	212
7.3 吲哚美辛在人体中的代谢情况	256
9.1 散点图类型	278
9.2 图层	285
9.3 图形种类	301
10.1 自定义表格样式	334
17.2 伯克利大学各个院系的录取人数	420
18.1 函数 <code>power.t.test()</code> 的参数表	430
20.1 R 包之间的不一致性，计算预测分类的概率的语法	453
20.2 迈克尔逊光速数据	465
20.3 火炬松树的高度（英尺）随时间（年）的变化	476
22.1 不同生长环境下植物的干重	513
22.2 线性回归的输出	517



22.3 躯干周长（毫米）随时间（天）的变化	524
33.1 ROI 插件按优化问题分类	612
A.1 R 和 Shell 命令的等价表示	688
A.2 devtools 的系统依赖	696
B.2 依赖的 R 包	761
B.3 本书使用的 R 包	770
D.1 泛型函数	786

欢迎

警告

Book in early development. Planned release in 202X.

本书风格

可以说，点估计、区间估计、假设检验、统计功效是每一个学数理统计的学生都绕不过去的坎，离开学校从事数据相关的工作，它们仍然是必备的工具。所以，本书会覆盖相关内容，但是和高校的教材最大的区别是更加注重它们之间的区别和联系，毕竟每一个统计概念都是经过了千锤百炼，而我们的主流教材始终如一地遵循的一个基本套路，就是突然给出一大堆定义、命题或定理，紧接着冗长的证明过程，然后给出一些难以找到实际应用背景的例子。三板斧抡完后就是给学生布置大量的习题，这种教学方式无论对于立志从事理论工作的还是将来投身于工业界的学生都是不合适的。

极大似然估计最初由德国数据学家 Gauss 于 1821 年提出，但未得到重视，后来，R. A. Fisher 在 1922 年再次提出极大似然的思想，探讨了它的性质，使它得到广泛的研究和应用。[\[茆诗松 et al., 2006\]](#)

这是国内某著名数理统计教材在极大似然估计开篇第一段的内容，后面是各种定义、定理、公式推导。教材简短一句话，这里面有很多信息值得发散，一个数学家提出了统计学领域极其重要的一个核心思想，他是在研究什么的时候提出了这个想法，为什么后来没有得到重视，整整 100 年以后，Fisher 又是怎么提出这一思想的呢？他做了什么使得这个思想被广泛接受和应用？虽然这可能有点离题，但是读者可以获得很多别的启迪，要知道统计领域核心概念的形成绝不是一蹴而就的，这一点也绝不局限于统计科学，任何一门科学都是这样的，比如物理学之于光的波粒二象性。历史上，各门各派的学者历经多年的思想碰撞才最终沉淀出现在现在的结晶。笔者认为，学校要想培养出有原创理论创新的人才，在对待前辈



的成果上，我们要不吝笔墨和口水，传道不等于满堂灌和刷分机，用寥寥数节课或者数页纸来梳理学者们几十年乃至上百年的智慧结晶是非常值得的，我们甚至可以从当时的社会、人文去剖析。非常欣赏有人在收集关于统计学历史的材料，读者不妨去看看 https://github.com/sctyner/history_of_statistics。另一个不得不提的人是 Allison Horst，她以风趣幽默的漫画形式，以画龙点睛之手法勾勒出基本的统计概念和思想，详见 <https://github.com/allisonhorst/stats-illustrations>，是我见过最好的科普读物。

Bradley Efron 在他的课程中谈及现代统计的研究层次，第一层次是基于正态分布假设的，这种类型已经研究的很清楚了，往往可以得到精确的结果，第二层次是将正态分布推广到指数族，这种类型的也研究的比较多了，常见的情况都研究的比较清楚，罕见的情况也是大量存在的，特别是在实际应用当中，总的来说只能得到部分准确的结果，第三层次对分布没有任何限定，只要满足成为一个统计分布的条件，这种情况下就只能求助于一般的数学工具和渐进理论。

下面以区间估计为例，希望能为传道做一点事情。区间估计的意义是解决点估计可靠性问题，它用置信系数解决了对估计结果的信心问题，弥补了点估计的不足。置信系数是最大的置信水平。

1934 年 C. J. Clopper 和 E. S. Pearson 给出二项分布 $B(n, p)$ 参数 p 的置信带 [Clopper and Pearson, 1934]，图 2 提炼了文章的主要结果。

区间半径这么长，区间估计的意义何在？增加样本量可以使得半径更短，那么至少应该有多少样本量才可以让估计变得有意义呢？就是说用估计比不用估计更好呢？答案是 39 个，留给读者思考一下为什么？读者可能已经注意到，置信带是关于点 $(5, 0.5)$ 中心对称的，这又是为什么，并且两头窄中间胖，像个酒桶？

提示

Base R 提供的 `uniroot()` 函数只能求取一元非线性方程的一个根，而 **root-Solve** 包提供的 `uniroot.all()` 函数可以求取所有的根。在给定分位点下，我们需要满足方程的最小的概率值。

Base R 提供的 `binom.test()` 函数可以精确计算置信区间，而 `prop.test()` 函数可近似计算置信区间。

```
# 近似计算 Wilson 区间
prop.test(x = 2, n = 10, p = 0.95, conf.level = 0.95, correct = TRUE)
## Warning in prop.test(x = 2, n = 10, p = 0.95, conf.level = 0.95, correct =
## TRUE): Chi-squared approximation may be incorrect
##
```

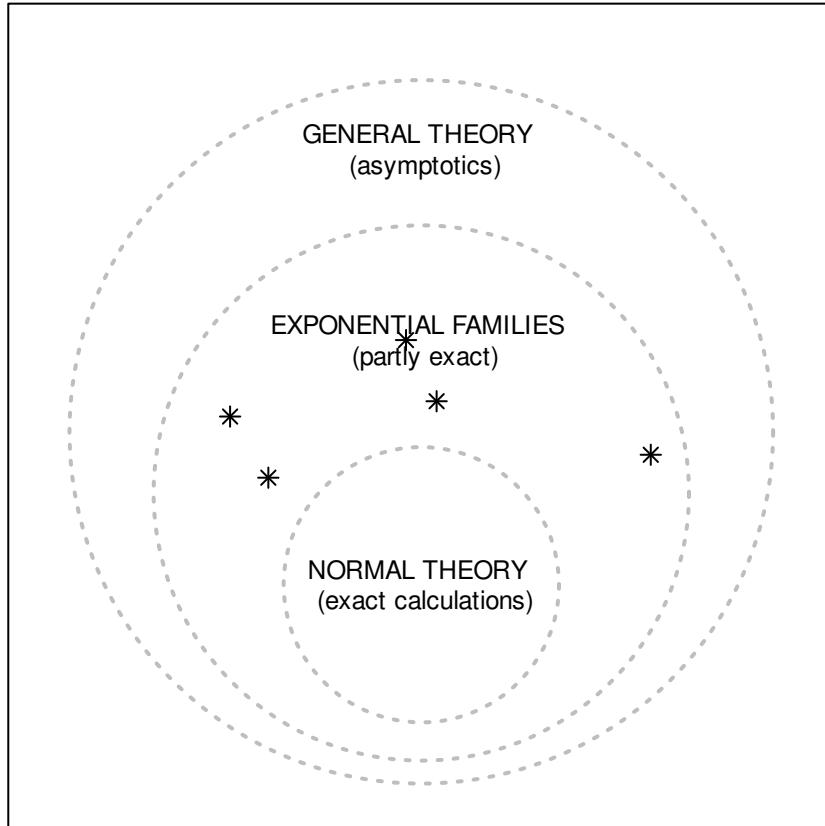


图1: 现代统计建模的三重境界: 修改自 2019 年冬季 Bradley Efron 的课程笔记(第一部分) http://statweb.stanford.edu/~ckirby/brad/STATS305B_Part-1_corrected-2.pdf

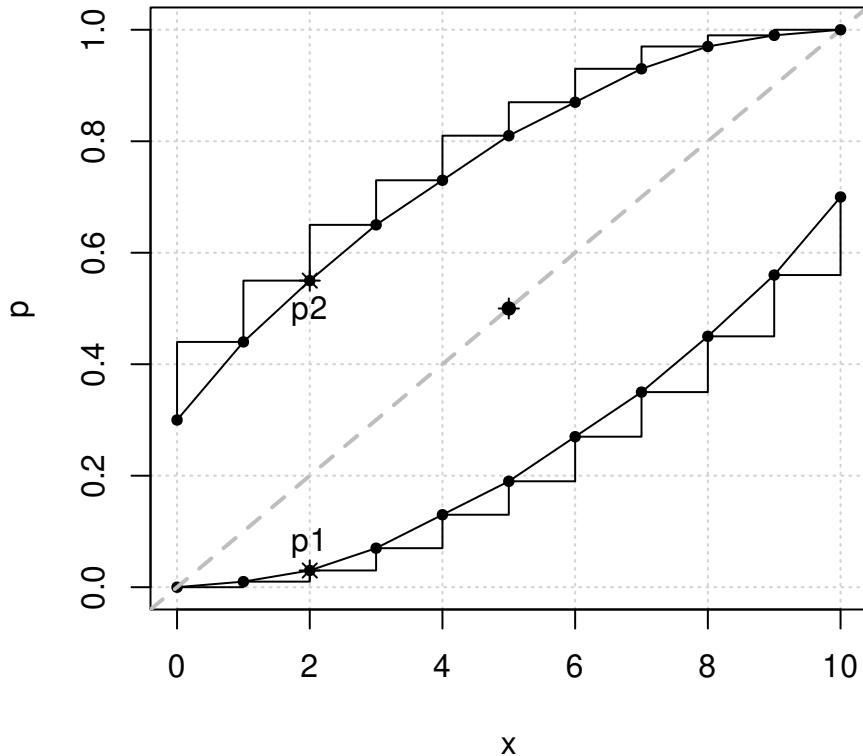


图 2: 给定置信系数 $1 - \alpha = 0.95$ 和样本量 $n = 10$ 的情况下，二项分布参数 p 的置信带。样本量为 10，正面朝上的次数为 2，置信水平为 0.95 的情况下，参数 p 的精确区间估计为 $(p_1, p_2) = (0.03, 0.55)$ 。



```

## 1-sample proportions test with continuity correction
##
## data: 2 out of 10, null probability 0.95
## X-squared = 103, df = 1, p-value <2e-16
## alternative hypothesis: true p is not equal to 0.95
## 95 percent confidence interval:
## 0.03543 0.55782
## sample estimates:
## p
## 0.2
# 精确计算
binom.test(x = 2, n = 10, p = 0.95, conf.level = 0.95)
##
## Exact binomial test
##
## data: 2 and 10
## number of successes = 2, number of trials = 10, p-value = 2e-09
## alternative hypothesis: true probability of success is not equal to 0.95
## 95 percent confidence interval:
## 0.02521 0.55610
## sample estimates:
## probability of success
## 0.2

```

实际达到的置信度水平随真实的未知参数值和样本量的变化而剧烈波动，这意味着这种参数估计方法在实际应用中不可靠、真实场景中参数真值是永远未知的，样本量是可控的，并且是可以变化的。根本原因在于这类分布是离散的，比如这里的二项分布。当数据 x 是离散的情况，置信区间的端点 $\ell(x)$ 和 $u(x)$ 也是离散的。这种缺陷是无法避免的，清晰的置信区间和离散的数据之间存在无法调和的冲突。

覆盖概率 $P_\theta(X = x)$ 和参数真值 θ 的关系 [Brown et al., 2001, Geyer and Meeden, 2005]

比如总体为二项分布 $B(n, \theta)$ 其中 $n=10$ ，在置信水平 $\alpha = 0.95$ 下，问参数 θ 的覆盖概率是多少？随参数 θ 的变化情况如何？<https://d.cosx.org/d/421502-coverage-probability>

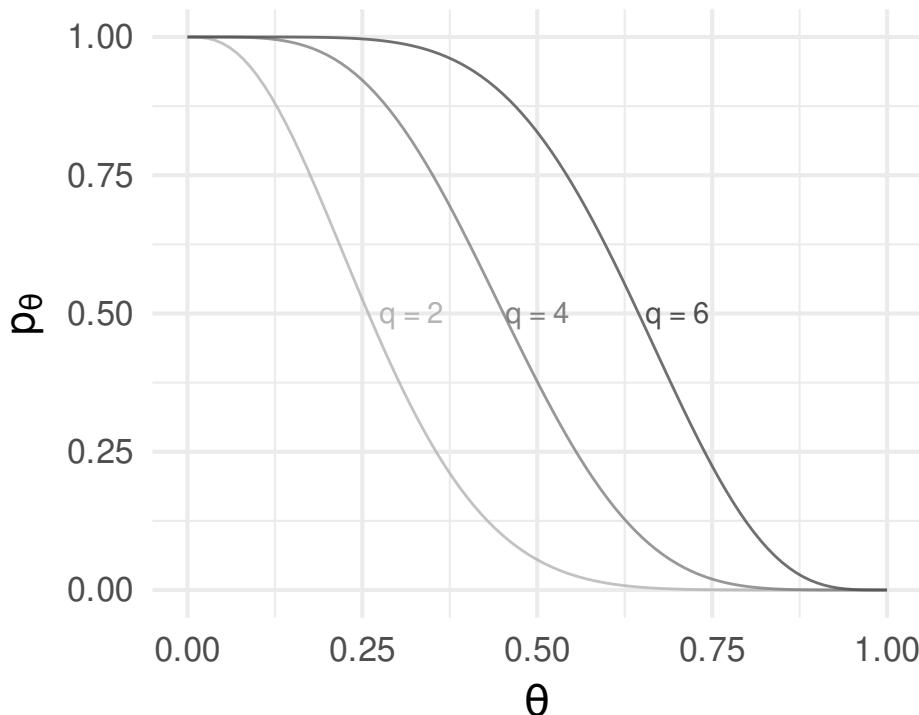


图 3: 二项分布 $B(n, \theta)$ 成功概率 θ , 固定样本量 $n = 10$, 分不同的分位点 $q = 2, 4, 6$ 绘制概率随成功概率 θ 的变化



还是以抛硬币的为例，我来做这个实验，抛 10 次，获得 7 次正面向上，他做这个实验，10 次中 4 次正面，每个人来做这个实验可能都会有所不同，实际上有 $2^{10} = 1024$ 个结果（含位置变化），每个结果都可以用来估计未知的参数 p 及其置信区间，和相应的覆盖概率。

假设参数的真值是 0.7，做一次实验，得到正面朝上的结果，有 6 次

```
set.seed(2019)
rbinom(n = 1, size = 10, prob = 0.7)
```

```
## [1] 6
```

这个检验的原假设是 $p = 0.7$ ，样本落在拒绝域的概率是 $0.4997 > 0.05$ 即不能拒绝原假设。

```
binom.test(x = 6, n = 10, p = 0.7, conf.level = 0.95)
```

```
##
## Exact binomial test
##
## data: 6 and 10
## number of successes = 6, number of trials = 10, p-value = 0.5
## alternative hypothesis: true probability of success is not equal to 0.7
## 95 percent confidence interval:
## 0.2624 0.8784
## sample estimates:
## probability of success
## 0.6
```

比例的真实值 p 落在区间 $(\hat{p} - Z_{1-\alpha/2} * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + Z_{1-\alpha/2} * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}})$ 的概率是 0.95。

```
c(
  0.6 - qnorm(p = 1 - 0.05 / 2, mean = 0, sd = 1) * sqrt(0.6 * (1 - 0.6) / 10),
  0.6 + qnorm(p = 1 - 0.05 / 2, mean = 0, sd = 1) * sqrt(0.6 * (1 - 0.6) / 10)
)
```

```
## [1] 0.2964 0.9036
```

TODO: 多重比较与检验

多重比较 `p.adjust()` 函数 Adjust P-values for Multiple Comparisons 单因素多重比较 `oneway.test()`

```
set.seed(123)
x <- rnorm(50, mean = c(rep(0, 25), rep(3, 25)))
p <- 2 * pnorm(sort(-abs(x)))
# ?p.adjust
round(p, 3)

## [1] 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.001 0.002
## [13] 0.003 0.004 0.005 0.007 0.007 0.009 0.009 0.011 0.021 0.049 0.061 0.063
## [25] 0.074 0.083 0.086 0.119 0.189 0.206 0.221 0.286 0.305 0.466 0.483 0.492
## [37] 0.532 0.575 0.578 0.619 0.636 0.645 0.656 0.689 0.719 0.818 0.827 0.897
## [49] 0.912 0.944

# round(p.adjust(p), 3)
# round(p.adjust(p, "BH"), 3)
```

TODO: 混合正态分布的参数估计

$$y = f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

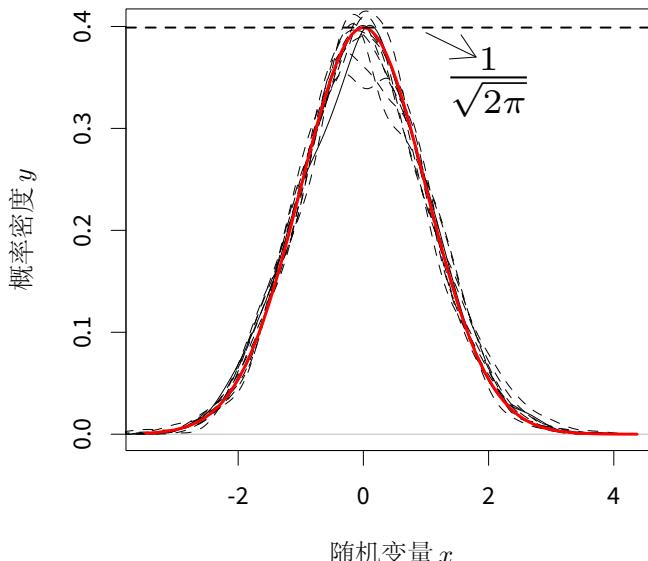


图 4: 上帝在掷骰子吗?

两个二元正态分布的碰撞，点的密度估计值代表概率密度值，

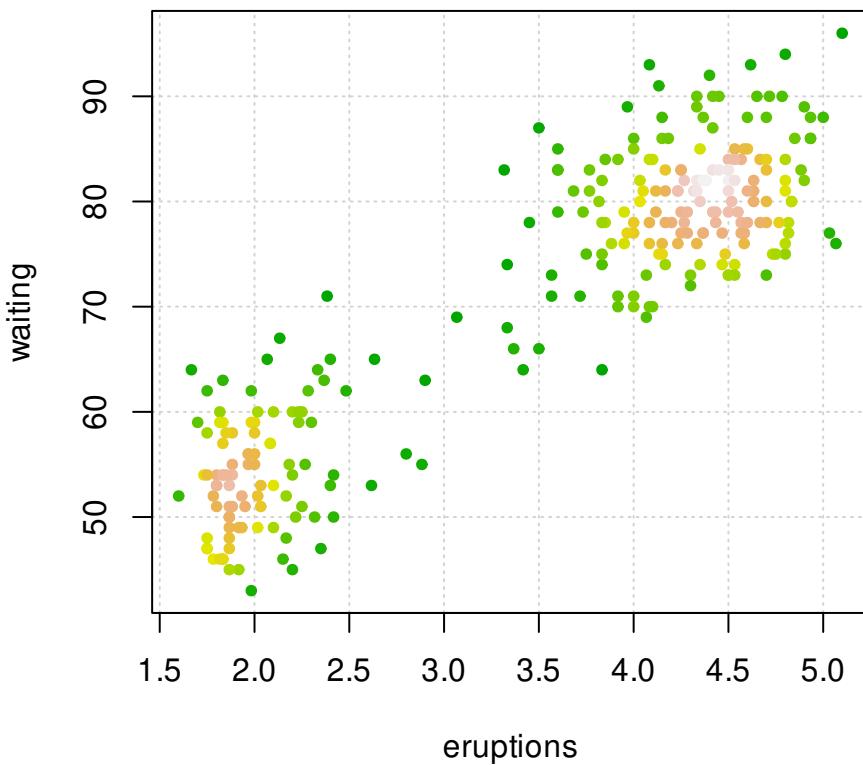
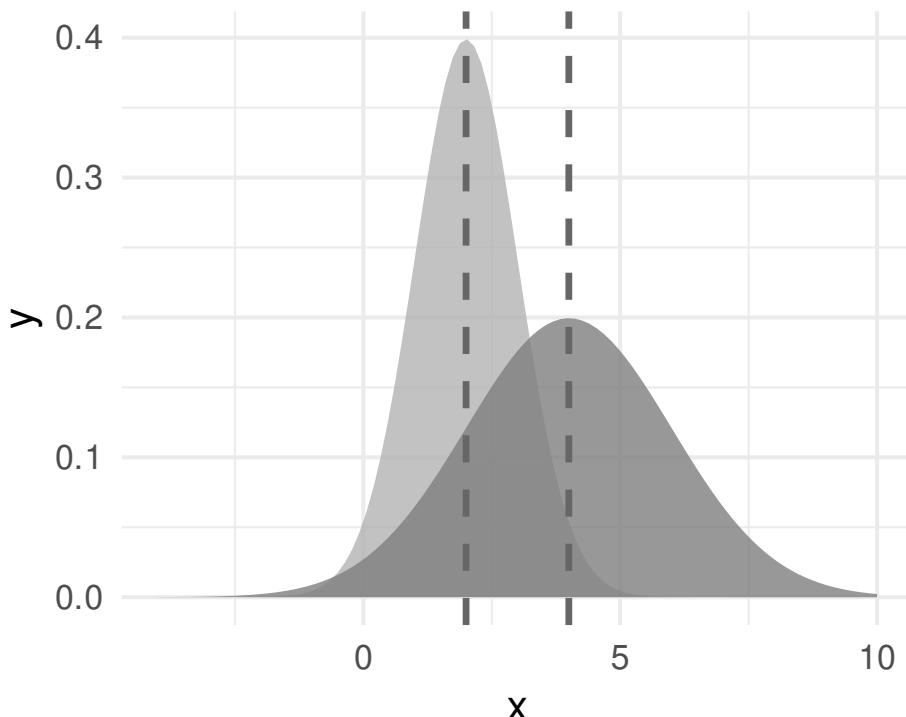


图 5: 散点图: faithful 数据集

TODO: 统计检验，决策风险，显著性水平



Charles J. Geyer 的文章 Fuzzy and Randomized Confidence Intervals and P-Values [Geyer and Meeden, 2005] 文章中的图 1 名义覆盖概率的计算见 [Blyth and Hutchinson, 1960]

本书定位

学习本书需要读者具备基本的概率、统计知识，比如上过一学期的概率论和数理统计学，也需要读者接触过编程知识，比如至少上过一学期的 C 语言、Python 语言或 Matlab 语言。了解基本的线性代数，比如矩阵的加、减、乘、逆四则运算、线性子空间、矩阵的 LU、SVD、Eigen 等分解。



内容概要

第一 章介绍本书的写作背景、语言环境、全书的记号约定、如何获取帮助、作者简介等信息。

第二 章介绍全书的数学公式符号。

第三 章介绍文件操作。

第四 章介绍 R 语言的数据结构。

第五 章介绍数据操作，包括 Base R、**data.table** 和 **magrittr**。

第六 章介绍数据导入导出，**data.table** 之于 csv 文件，**openxlsx** 之于 xlsx 文件。

第七 章介绍数据可视化，分四个部分，基础元素、常用图形、字体和颜色设置。

第八 章介绍动态文档，即 R Markdown 及其生态系统。

第九 章介绍交互图形，以常用的 **plotly** 和 **highcharter** 为主，重点介绍 R 和 JavaScript 库的对应关系。

第十 章介绍交互表格，分两节介绍交互式的 **DT** 和 **reactable**，静态的 **gt** 和 **kableExtra**，掌握这几个 R 包足以应付日常工作。

第十一 章介绍交互报表开发，符合工业标准的最佳实践。

致谢名单

特别感谢 XX，还有很多人通过提交 PR 或 Issues 的方式参与了本书的创作过程，没有这一点一滴的持续改进，本书不会达到现在的样子，所以我将他们列在致谢名单中，详见表 1，人名按照提交量（commit 的个数）降序排列。

表 1：致谢名单

贡献者	提交量
Yadong Liu	1
Yihui Xie	1

黄湘云
于北京

授权说明

警告

本书采用 [知识共享署名-非商业性使用-禁止演绎 4.0 国际许可协议](#) 许可，
请君自重，别没事儿拿去传个什么新浪爱问、百度文库以及 XX 经济论坛，
项目中代码使用 [MIT 协议](#) 开源



运行信息

书籍在 R version 4.1.0 (2021-05-18) 下编译，Pandoc 版本 2.14.1，最新一次编译发生在 2021-08-07 15:36:57。

```
xfun::session_info(packages = c(
  "knitr", "rmarkdown", "bookdown", "equatiomatic",
  "data.table", "DT", "kableExtra", "reactable",
  "patchwork", "plotly", "shiny",
  "ggplot2", "dplyr", "tidyverse"
), dependencies = FALSE)

## R version 4.1.0 (2021-05-18)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 20.04.2 LTS
##
## Locale:
##   LC_CTYPE=en_US.UTF-8          LC_NUMERIC=C
##   LC_TIME=en_US.UTF-8           LC_COLLATE=en_US.UTF-8
##   LC_MONETARY=en_US.UTF-8       LC_MESSAGES=en_US.UTF-8
##   LC_PAPER=en_US.UTF-8          LC_NAME=C
##   LC_ADDRESS=C                  LC_TELEPHONE=C
##   LC_MEASUREMENT=en_US.UTF-8   LC_IDENTIFICATION=C
##
## Package version:
##   bookdown_0.22      data.table_1.14.0  dplyr_1.0.7      DT_0.18
##   equatiomatic_0.2.0 ggplot2_3.3.5     kableExtra_1.3.4  knitr_1.33
```

```
## patchwork_1.1.1   plotly_4.9.4.1    reactable_0.2.3   rmarkdown_2.9  
## shiny_1.6.0       tidyverse_1.3.1  
##  
## Pandoc version: 2.14.1
```





第一章 前言

荃者所以在鱼，得鱼而忘荃；蹄者所以在兔，得兔而忘蹄；言者所以在意，得意而忘言。吾安得夫忘言之人而与之言哉！

— 摘自《庄子·杂篇·物》

庄子谈学习，余深以为然，故引之。

The fish trap exists because of the fish; once you've gotten the fish, you can forget the trap. The rabbit snare exists because of the rabbit; once you've gotten the rabbit, you can forget the snare. Words exist because of meaning; once you've gotten the meaning, you can forget the words. Where can I find a man who has forgotten words so I can have a word with him ? ¹

— Chuang Tzu

1.1 语言抉择

行业内可以做统计分析和建模的软件汗牛充栋，比较顶级的收费产品有 SAS 和 SPSS，在科学计算领域的 Matlab 和 Mathematica 也有相当强的统计功能，而用户基数最大的是微软 Excel，抛开微软公司的商业手段不说，Excel 的市场份额却是既成事实。Brian D. Ripley 20 多年前的一句话很有意思，放在当下也是适用的。

Let's not kid ourselves: the most widely used piece of software for statistics is Excel.

— Brian D. Ripley [Ripley, 2002]

¹译文摘自 Eric D. Kolaczyk



有鉴于 Excel 在人文、社会、经济和管理等领域的影响力，熟悉 R 语言的人把它看作超级收费版的 Excel，这实在是一点也不过分。事实上，我司就是一个很好的明证，一个在线教育类的互联网公司，各大业务部门都在使用 Excel 作为主要的数据分析工具。然而，Excel 的不足也十分突出，工作过程无法保存和重复利用，Excel 也不是数据库，数据集稍大，操作起来愈发困难，对于复杂的展示，需要借助内嵌的 VBA，由于缺乏版本控制，随着时间的推移，几乎不可维护。所以，我们还是放弃 Excel 吧，Jenny Bryan 更在 2016 年国际 R 语言大会上的直截了当地喊出了这句话²。Nathan Stephens 对 Excel 的缺陷不足做了全面的总结³。

Some people familiar with R describe it as a supercharged version of Microsoft's Excel spreadsheet software.

— Ashlee Vance ⁴

另一方面，我们谈谈开源领域的佼佼者 — R (<https://cran.r-project.org/>)，Python (<https://www.python.org/>) 和 Octave (<http://www.gnu.org/software/octave/>)。Python 号称万能的胶水语言，从系统运维到深度学习都有它的广泛存在，它被各大主流 Linux 系统内置，语言风格上更接近于基数庞大的开发人员，形成了强大的生态平台。Octave 号称是可以替代 Matlab 的科学计算软件，在兼容 Matlab 的方面确实做的很不错，然而，根据 Julia 官网给出的各大编程语言的测试 <https://julialang.org/benchmarks/>，性能上不能相提并论。

R 提供了丰富的图形接口，包括 Tcl/Tk, Gtk, Shiny 等，以及基于它们的衍生品 rattle (RGtk2)、Rcmdr (tcl/tk)、radian (shiny)。更多底层介绍，见 John Chamber 的著作《Extending R》。

TikZ 在绘制示意图方面有很大优势，特别是示意图里包含数学公式，这更是 LaTeX 所擅长的方面

JASP <https://jasp-stats.org> 是一款免费的统计软件，源代码托管在 Github 上 <https://github.com/jasp-stats/jasp-desktop>，主要由阿姆斯特丹大学 E. J. Wagenmakers 教授 <https://www.ejwagenmakers.com/> 领导的团队维护开发，实现了很多贝叶斯和频率统计方法，相似的图形用户界面使得 JASP 可以作为 SPSS 的替代，目前实现的功能见 <https://jasp-stats.org/current-functionality/>，统计方法见博客 <https://www.bayesianspectacles.org/>。

国内可视化分析平台，比如 hiplot 基于 R 语言实现可视化分析，各类图形的介

²<https://channel9.msdn.com/Events/useR-international-R-User-conference/useR2016/jailbreak-Get-out-of-Excel-free>

³<https://resources.rstudio.com/wistia-rstudio-essentials-2/how-to-excel-without-using-excel>

⁴<https://www.nytimes.com/2009/01/07/technology/business-computing/07program.html>

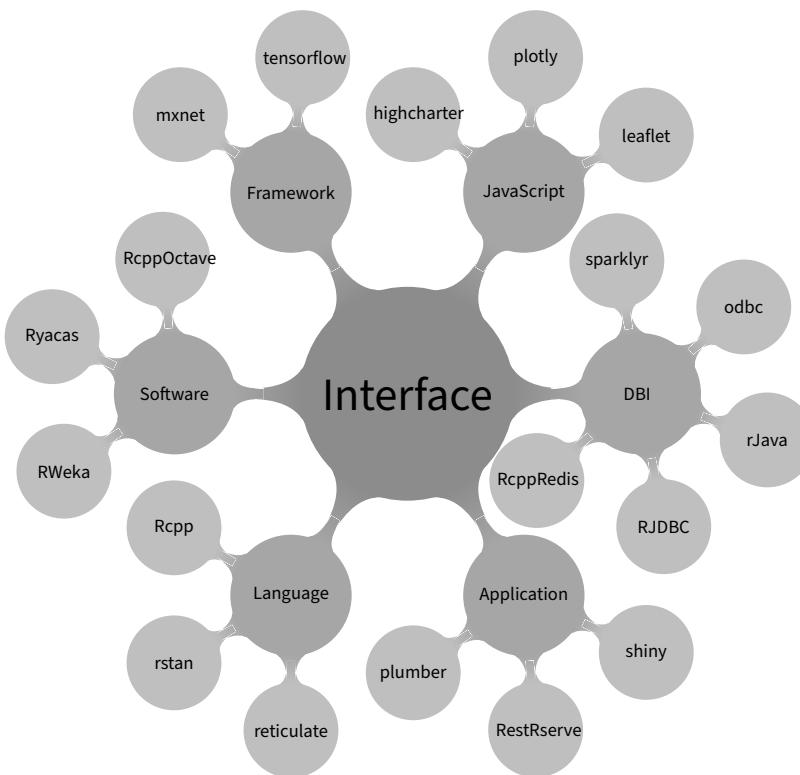


图 1.1: R 语言扩展包生态系统



绍见[文档](#)，极大地降低数据分析人员探索分析的门槛，节省了时间，同时非专业内的人也可借助其完成分析探索的过程，只需明白各类图形的含义即可。美团也建设了自己的可视化分析平台帮助运营人员，详见[文档](#)

[Patrick Burns](#) 收集整理了 R 语言中奇葩的现象，写成 [The R Inferno](#) 直译过来就是《R 之炼狱》。这些奇葩的怪现象可以看做是 R 风格的一部分，对于编程人员来说就是一些建议和技巧，参考之可以避开某些坑。Paul E. Johnson 整理了一份真正的 R 语言建议，记录了他自己从 SAS 转换到 R 的过程中遇到的各种问题 <http://pj.freefaculty.org/R/Rtips.html>。Michail Tsagris 和 Manos Papadakis 也收集了 70 多条 R 编程的技巧和建议，力求以更加 R 范地将语言特性发挥到极致 [[Tsagris and Papadakis, 2018](#)]，Martin Mächler 提供了一份 [Good Practices in R Programming](#)。Python 社区广泛流传着 Tim Peters 的《Python 之禅》，它已经整合进每一版 Python 软件中，只需在 Python 控制台里执行 `import this` 可以获得。

1. Beautiful is better than ugly.
2. Explicit is better than implicit.
3. Simple is better than complex.
4. Complex is better than complicated.
5. Flat is better than nested.
6. Sparse is better than dense.
7. Readability counts.
8. Special cases aren't special enough to break the rules.
9. Although practicality beats purity.
10. Errors should never pass silently.
11. Unless explicitly silenced.
12. In the face of ambiguity, refuse the temptation to guess.
13. There should be one- and preferably only one -obvious way to do it.
14. Although that way may not be obvious at first unless you're Dutch.
15. Now is better than never.
16. Although never is often better than *right* now.
17. If the implementation is hard to explain, it's a bad idea.
18. If the implementation is easy to explain, it may be a good idea.
19. Namespaces are one honking great idea – let's do more of those!

— The Zen of Python

总之，编程语言到一定境界都是殊途同归的，对美的认识也是趋同的，道理更



是相通的，Python 社区的 Pandas <https://github.com/pandas-dev/pandas> 和 Matplotlib <https://github.com/matplotlib/matplotlib> 也有数据框和图形语法的影子。Pandas <https://github.com/pandas-dev/pandas> 明确说了要提供与 `data.frame` 类似的数据结构和对应统计函数等，而 Matplotlib 偷了 ggplot2 绘图样式 https://matplotlib.org/3.2.2/gallery/style_sheets/ggplot.html。

1.2 数据科学

John M. Chambers 谈了数据科学的源起以及和 S、R 语言的渊源 [Chambers, 2020]。

1.3 获取帮助

R 社区提供了丰富的帮助资源，可以在 R 官网搜集的高频问题 <https://cran.r-project.org/faqs.html> 中查找，也可在线搜索 <https://cran.r-project.org/search.html> 或 <https://rseek.org/>，更多获取帮助方式见 <https://www.r-project.org/help.html>。爆栈网问题以标签分类，比如 `r-plotly`、`r-markdown`、`data.table` 和 `ggplot2`，还可以关注一些活跃的社区大佬，比如 [谢益辉](#)。

1.4 写作环境

本书 R Markdown 源文件托管在 Github 仓库里，本地使用 RStudio IDE 编辑，`bookdown` 组织各个章节的 `Rmd` 文件和输出格式，使用 `Git` 进行版本控制。每次提交修改到 Github 上都会触发 `Travis` 自动编译书籍，将一系列 `Rmd` 文件经 `knitr` 调用 R 解释器执行里面的代码块，并将输出结果返回，`Pandoc` 将 `Rmd` 文件转化为 `md`、`html` 或者 `tex` 文件。若想输出 `pdf` 文件，还需要准备 `TeX` 排版环境，最后使用 `Netlify` 托管书籍网站，和 `Travis` 一起实现连续部署，使得每次修改都会同步到网站。最近一次编译时间 2021 年 08 月 07 日 15 时 36 分 58 秒，本书用 R version 4.1.0 (2021-05-18) 编译，完整运行环境如下：

```
xfun:::session_info(packages = c(
  "knitr", "rmarkdown", "bookdown"
), dependencies = FALSE)
```

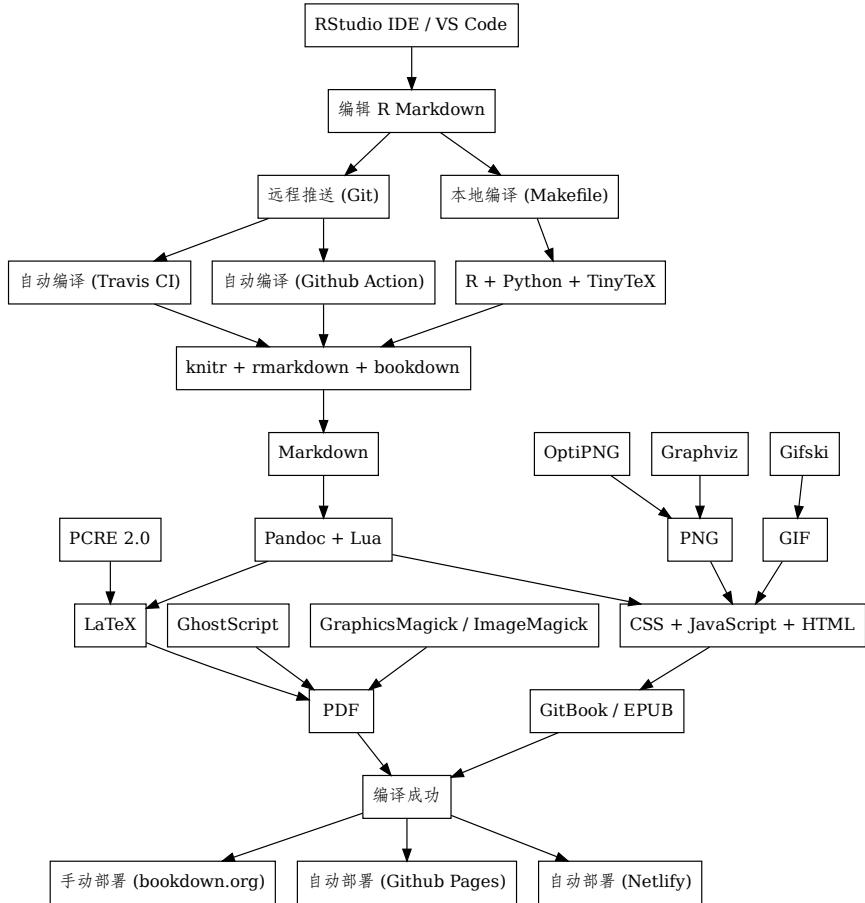


图 1.2: 书籍项目架构图



```
## R version 4.1.0 (2021-05-18)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 20.04.2 LTS
##
## Locale:
##  LC_CTYPE=en_US.UTF-8          LC_NUMERIC=C
##  LC_TIME=en_US.UTF-8           LC_COLLATE=en_US.UTF-8
##  LC_MONETARY=en_US.UTF-8       LC_MESSAGES=en_US.UTF-8
##  LC_PAPER=en_US.UTF-8          LC_NAME=C
##  LC_ADDRESS=C                  LC_TELEPHONE=C
##  LC_MEASUREMENT=en_US.UTF-8   LC_IDENTIFICATION=C
##
## Package version:
##  bookdown_0.22 knitr_1.33      rmarkdown_2.9
##
## Pandoc version: 2.14.1
```

借助 **bookdown** [Xie, 2016] 可以将 Rmd 文件组织起来, **rmarkdown** [Allaire et al., 2021] 和 **knitr** [Xie, 2015] 将源文件编译成 Markdown 文件, **Pandoc** 将 Markdown 文件转化成 HTML 和 TeX 文件, **TinyTeX** [Xie, 2019] 可以将 TeX 文件进一步编译成 PDF 文档, 书中大量的图形在用 **ggplot2** 包制作 [Wickham, 2016], 而统计理论相关的示意图用 Base R 创作。

最后，本书在三个位置提供网页版，网站 [Github Pages](#) 发布最近一次在 [Travis](#) 构建成功的版本 <https://xiangyunhuang.github.io/masr/>，网站 [Bookdown](#) 发布本地手动创建的版本 <https://bookdown.org/xiangyun/masr/>，网站 [Netlify](#) 发布最新的开发版 <https://masr.netlify.app/>。

1.5 记号约定

正文中的代码、函数、参数及参数值以等宽正体表示,如`data(list = c('iris', 'BOD'))`,其中函数名称`data()`,参数及参数值`list = c('iris', 'BOD')`,R程序包用粗体表示,如`graphics`。

ruler()

-----+---1---+---2---+---3---+---4---+---5---+---6---+---7---+---8
1234567890123456789012345678901234567890123456789012345678901234567890



1.6 复现环境

构建容器

本书借助 Github Action 从 Dockerfile 构建容器镜像，然后将镜像文件推送到 Github Package。完成这些操作首先需要从 <https://github.com/settings/tokens> 新建拥有 GitHub Package⁵ 读写删的权限的 TOKEN（俗称访问令牌或密钥），命名为 GITHUB_PKG，并将此令牌保存到本地 TOKEN.txt 文件中，以备后用。

镜像内默认暴露 8181 端口供外部连接使用，进入容器后，默认工作路径是 /home/docker/。创建好镜像后，要先登陆 GitHub Package 然后才有权限将镜像拉取下来

```
# 登陆 GitHub Package
cat ~/TOKEN.txt | docker login https://docker.pkg.github.com -u XiangyunHuang --password @jFJZDf
# 拉取镜像
docker pull docker.pkg.github.com/xiangyunhuang/masr/masr-book:devel
```

读者可以先查看下容器内的信息

```
docker run --rm -u root -v "${PWD}:/home/docker/" \
  docker.pkg.github.com/xiangyunhuang/masr/masr-book:devel \
  bash -c "locale; fc-list | sort"
```

运行容器

下面从镜像创建一个叫 masr-book 的容器，并让它在后台运行，允许以真正的 root 账户权限交互式执行命令，停止容器后，自动销毁容器。此处，不再介绍 Docker 容器的使用，用容器打包本书所有软件环境仅供读者完整复现本书之用，感兴趣的读者可以去参考书籍[Docker 从入门到实践](#)。

```
docker run -itd -p 8282:8787 --rm --name=masr-book --privileged=true \
  docker.pkg.github.com/xiangyunhuang/masr/masr-book:devel /sbin/init
```

接着登陆进入 masr-book 容器，

```
docker exec -it masr-book bash
```

一番骚操作后，用户退出容器，然后停止容器。

⁵<https://docs.github.com/en/packages/using-github-packages-with-your-projects-ecosystem/configuring-docker-for-use-with-github-packages>



```
docker stop masr-book
```



使用 RStudio Server

启动容器后，接着获取容器 masr-book 的 IP 地址，然后依据端口号 8282 从网页进入 RStudio Sever，比如 <http://192.168.100.23:8282>



```
docker inspect --format='{{.NetworkSettings.IPAddress}}' masr-book
```

1.7 如何发问

The phrase “does not work” is not very helpful, it can mean quite a few things including:

- Your computer exploded.
- No explosion, but smoke is pouring out the back and microsoft’s “NoSmoke” utility is not compatible with your power supply.
- The computer stopped working.
- The computer sits around on the couch all day eating chips and watching talk shows.
- The computer has started picketing your house shouting catchy slogans and demanding better working conditions and an increase in memory.
- Everything went dark and you cannot check the cables on the back of the computer because the lights are off due to the power outage.
- R crashed, but the other programs are still working.
- R gave an error message and stopped processing your code after running for a while.
- R gave an error message without running any of your code (and is waiting for your next command).
- R is still running your code and the time has exceeded your patience so you think it has hung.
- R completed and returned a result, but also gave warnings.
- R completed your command, but gave an incorrect answer.
- R completed your command but the answer is different from what you expect (but is correct according to the documentation).



There are probably others. Running your code I think the answer is the last one.

— Greg Snow⁶

1.8 作者简介

热心开源事业，统计之都副主编，经常混迹于统计之都论坛、Github 和客栈网。

个人主页 <https://xiangyun.rbind.io/>

⁶来自 R 社区论坛收集的智语 `fortunes::fortune(324)`。



第二章 符号说明

Fabio Mulazzani: I need to obtain all the 9.somethingExp157 permutations that can be given from the numbers from 1 to 100.

Ted Harding: To an adequate approximation there are 10^{158} of them. Simply to obtain them all (at a rate of 10^{10} per second, which is faster than the CPU frequency of most desktop computers) would take 10^{148} seconds, or slightly longer than 3×10^{140} years. Current estimates of the age of the Universe are of the order of 1.5×10^{10} years, so the Universe will have to last about 2×10^{130} times as long as it has already existed, before the task could be finished.

So: why do you want to do this?

— Fabio Mulazzani and Ted Harding¹

数学符号约定参考花书 https://github.com/goodfeli/dlbook_notation

[Flexible Imputation of Missing Data](#) 的 [符号约定章节](#)

矩阵、向量用粗体大写表示，特殊情况下，Y 只有一列

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Y 叫做因变量或者响应变量 **response variables**, X 叫做自变量、协变量 **covariate** 或者预报变量 **predictor variables**

线性回归模型

$$y = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

¹<https://stat.ethz.ch/pipermail/r-help/2008-November/180820.html>



其中 y 是 $n \times 1$ 的观测向量, X 为 $n \times p$ 的设计矩阵, β 为未知参数向量, β_0 为常数项, $\beta_1, \dots, \beta_{p-1}$ 为回归系数, ϵ 为 $n \times 1$ 随机误差向量, 其均值为 0, 即 $E(\epsilon_i) = 0$

模型假设

1. 误差项方差齐性, 即

$$\text{Var}(\epsilon_i) = \sigma^2, \quad i = 1, 2, \dots, n$$

2. 误差项彼此不相关, 即

$$\text{Cov}(\epsilon_i, \epsilon_j) = 0, \quad i \neq j, \quad i, j = 1, \dots, n$$

线性模型中线性二字实质上是指 y 关于未知参数 β_i 的关系是线性的。

A, B, C, D 斜体表示普通的集合, X, Y, Z 表示矩阵, a, b, c, d 表示常数, $\alpha, \beta, \theta, \phi, \kappa$ 表示模型或者分布函数的参数, Θ 表示参数空间, $\mathbb{R}^n, \mathbb{C}^n$ 表示特殊的 n 维实(复)数域, $\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D}$ 表示一般的数域, $\mathcal{S}, \mathcal{P}, \mathcal{G}$ 分别表示随机过程、概率空间和图

表 2.1: 数学符号表

符号	含义	符号	含义
A	粗体	Ω	全集
A	集合	\mathbb{R}, \mathbb{C}	实(复)数集
\mathcal{A}	集族	\emptyset	空集
A	矩阵	A^-	矩阵的广义逆
A^\top	矩阵转置	\bar{x}	平均值
A^{-1}	矩阵的逆	$ a $	标量绝对值
A^*	伴随矩阵	$\text{diag}(A)$	矩阵的对角
$\ A\ _1$	矩阵的 1 范数	I	单位矩阵
$\ A\ _2$	矩阵的 2 范数	I_n	n 阶单位矩阵
$\ A\ _\infty$	矩阵的无穷范数	1	全 1 矩阵
$\ X\ _1$	向量的 1 范数	1_n	n 阶全 1 矩阵
$\ X\ _2$	向量的 2 范数	$\ X\ _\infty$	向量的无穷范数
$\langle X, Y \rangle$	向量的内积	$f(X)$	随机变量的函数
$X \wedge Y$	向量的外积	∇X	向量微分或梯度
β	模型系数	θ	模型或分布参数



符号	含义	符号	含义
α	模型截距	Θ	参数空间
$\hat{\beta}_{ls}$	模型系数的 LS 估计	$f(x)$	标量值函数
$\hat{\beta}_{mle}$	模型系数的 MLE 估计	$f(\mathbf{X})$	向量的函数
$\hat{\beta}_{bayes}$	模型系数的 Bayes 估计	\mathcal{X}	概率空间
ρ	相关系数	κ	贝塞尔函数的阶
ϕ	尺度参数	u	距离 $\ \mathbf{x}_1 - \mathbf{x}_2\ $
\mathbb{R}^2	二维实数域	$S(x)$	空间过程
\mathcal{S}	$\mathcal{S} = \{S(x) : x \in \mathbb{R}^2\}$	\mathcal{S}^*	随机过程 \mathcal{S} 的近似
\triangleq	定义为或记为	$\hat{\beta}_{ridge}$	β 的岭回归估计
$A \geq 0$	矩阵 A 半正定	$\hat{\beta}_{lar}$	β 的最小角回归估计
$A > 0$	矩阵 A 正定	$\hat{\beta}_{subset}$	β 的最优子集回归估计
$A \otimes B$	矩阵 A 与 B 的 Kronecker 积	MSE 均方误差	$\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$
$\mathcal{M}(A)$	矩阵 A 的列向量张成的子空间	RMSE 均方根误差	$\sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$
$\ A\ $	矩阵 A 的范数	MAE 平均绝对误差	$\frac{1}{n} \sum_{i=1}^n y_i - \hat{y}_i $
$ A $	矩阵 A 的行列式	LSE	最小二乘估计
$rk(A)$	矩阵 A 的秩	BLUE	最佳线性无偏估计
$tr(A)$	方阵 A 的迹	MVUE	最小方差无偏估计
A^{-1}	矩阵 A 的逆	UMVUE	一致最小方差无偏估计
A^-	矩阵 A 的广义逆	MINQUE	最小范数二次无偏估计
$\hat{\beta}_{ols}$	β 的普通最小二乘估计	OLS	普通最小二乘估计
$\hat{\beta}_{pca}$	β 的主成分分析估计	PLS	偏最小二乘估计
$\hat{\beta}_{pls}$	β 的偏最小二乘估计	GLS	广义最小二乘估计
$\hat{\beta}_{svm}$	β 的支持向量机估计	WLS	带权最小二乘估计
$\hat{\beta}_{lasso}$	β 的 Lasso 估计	-	-

多元统计分析高惠璇矩阵符号表示，深度学习符号表示 <https://github.com/XiangyunHuang/dlbook>

举例，线性模型的表示，此处 Y 为 $n \times 1$ 列向量， X 为 $p \times n$ 的矩阵， β 为 $p \times 1$ 的列向量， ϵ 为 $n \times 1$ 列向量

$$Y = X'\beta + \epsilon$$

$$\mathbf{A} = \boldsymbol{\Gamma}^\top \boldsymbol{\Lambda} \boldsymbol{\Gamma}$$



$$\mathbf{u} = (u_1, u_2, \dots, u_n)$$

$$\mathbf{x} = (x_1, x_2, \dots, x_n)$$

期望 \mathbb{E} 正态分布 $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ 对数 log 协方差 Cov, Var

矩阵

$$\mathbf{Y} = (\mathbf{y}^{(1)}, \mathbf{y}^{(n)}, \dots, \mathbf{y}^{(n)})$$

其中 $\mathbf{y}^{(i)} = (y_{1i}, y_{2i}, \dots, y_{ni})$ 表示第 i 列

梅隆函数 (Matern function) 是描述空间相关性的常用函数，它带有两参数 κ 和 ϕ ，具体形式如下：

$$\rho(u) = \{2^{\kappa-1}\Gamma(\kappa)\}^{-1}(u/\phi)^\kappa K_\kappa(u/\phi)$$

其中， $K_\kappa(\cdot)$ 表示 κ 阶修正的贝塞尔函数

第三章 文件操作

第四章 数据结构

网站 <https://r-coder.com/> 主要介绍 Base R，特点是全面细致，排版精美

4.1 字符

4.2 向量

4.3 矩阵

4.4 数组

4.5 列表

```
x <- list(a = 1, b = 2, c = list(d = c(1, 2, 3), e = "hello"))
print(x)

## $a
## [1] 1
##
## $b
## [1] 2
##
## $c
## $c$d
```

```
## [1] 1 2 3
##
## $c$e
## [1] "hello"
base::print.simple.list(x)

##      -
## a    1
## b    2
## c.d1 1
## c.d2 2
## c.d3 3
## c.e  hello
```

4.6 日期

上个季度最后一天

```
# https://d.cosx.org/d/421162/16
as.Date(cut(as.Date(c("2020-02-01", "2020-05-02")), "quarter")) - 1
```

```
## [1] "2019-12-31" "2020-03-31"
```

本季度第一天

```
as.Date(cut(as.Date(c("2020-02-01", "2020-05-02")), "quarter"))
```

```
## [1] "2020-01-01" "2020-04-01"
```

类似地，本月第一天和上月最后一天

```
# 本月第一天
as.Date(cut(as.Date(c("2020-02-01", "2020-05-02")), "month"))
```

```
## [1] "2020-02-01" "2020-05-01"
```

上月最后一天

```
as.Date(cut(as.Date(c("2020-02-01", "2020-05-02")), "month")) - 1
```

```
## [1] "2020-01-31" "2020-04-30"
```



timeDate 提供了很多日期计算函数，比如季初、季末、月初、月末等

```
library(timeDate)
# 季初
as.Date(format(timeFirstDayInQuarter(charvec = c("2020-02-01", "2020-05-02"))), format)
# 季末
as.Date(format(timeLastDayInQuarter(charvec = c("2020-02-01", "2020-05-02"))), format)
# 月初
as.Date(format(timeFirstDayInMonth(charvec = c("2020-02-01", "2020-05-02"))), format)
# 月末
as.Date(format(timeLastDayInMonth(charvec = c("2020-02-01", "2020-05-02"))), format)
```

`cut.Date()` 是一个泛型函数，查看它的所有 S3 方法

```
methods(cut)

## [1] cut.Date      cut.default    cut.dendrogram* cut.POSIXt
## see '?methods' for accessing help and source code
```

格式化输出日期类型数据

```
formatC(round(runif(1, 1e8, 1e9)), digits = 10, big.mark = ",")  
## [1] "550,688,898"  
  
# Sys.setlocale(locale = "C") # 如果是 Windows 系统，必须先设置，否则转化结果是 NA  
as.Date(paste("1990-January", 1, sep = "-"), format = "%Y-%B-%d")  
## [1] "1990-01-01"
```

获取当日零点

```
format(as.POSIXlt(Sys.Date()), "%Y-%m-%d %H:%M:%S")  
## [1] "2021-08-07 00:00:00"
```

从 `POSIXt` 数据对象中，抽取小时和分钟部分，返回字符串

```
strftime(x = Sys.time(), format = "%H:%M")  
## [1] "15:36"
```



表 4.1: 日期表格

代 码	含 义	代 码	含 义
%a	Abbreviated weekday	%A	Full weekday
%b	Abbreviated month	%B	Full month
%c	Locale-specific date and time	%d	Decimal date
%H	Decimal hours (24 hour)	%I	Decimal hours (12 hour)
%j	Decimal day of the year	%m	Decimal month
%M	Decimal minute	%p	Locale-specific AM/PM
%S	Decimal second	%U	Decimal week of the year (starting on Sunday)
%w	Decimal Weekday (0=Sunday)	%W	Decimal week of the year (starting on Monday)
%x	Locale-specific Date	%X	Locale-specific Time
%y	2-digit year	%Y	4-digit year
%z	Offset from GMT	%Z	Time zone (character)

本节介绍了 R 本身提供的基础日期操作，第二十五章着重介绍一般的时间序列类型的数据对象及其操作。



第五章 数据操作

`data.table` 大大加强了 Base R 提供的数据操作，`poorman` 提供最常用的数据操作，但是不依赖 `dplyr`, `fst`, `arrow` 和 `feather` 提供更加高效的数据读写性能。

`collapse` 提供一系列高级和快速的数据操作，支持 Base R、`dplyr`、`tibble`、`data.table`、`plm` 和 `sf` 数据框结构类型。关键的特点有：1. 高级的统计编程，提供一系列统计函数支持在向量、矩阵和数据框上做分组和带权计算。`fastverse` 提供丰富的数据操作和统计计算功能，意图打造一个 `tidyverse` 替代品。

更多参考材料见[A data.table and dplyr tour](#), [Big Data in Economics: Data cleaning and wrangling](#) 和 [DataCamp's data.table cheatsheet](#), 关于采用 Base R 还是 tidyverse 做数据操作的[讨论](#)，数据操作的动画展示参考 <https://github.com/gadenbuie/tidyexplain>。

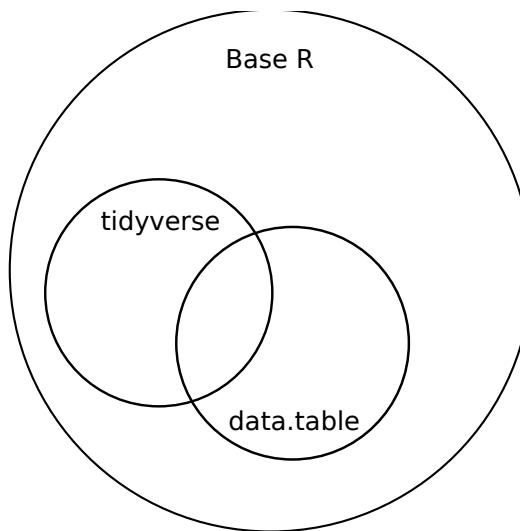


图 5.1: Tidyverse 和 Base R 的关系



什么是 Base R? Base R 指的是 R 语言/软件的核心组件，由 R Core Team 维护

```
Pkgs <- sapply(list.files(R.home("library")), function(x)
  packageDescription(pkg = x, fields = "Priority"))
names(Pkgs[Pkgs == "base" & !is.na(Pkgs)])  
  
## [1] "base"      "compiler"   "datasets"   "graphics"   "grDevices"  "grid"  
## [7] "methods"    "parallel"   "splines"    "stats"     "stats4"    "tcltk"  
## [13] "tools"     "utils"  
  
names(Pkgs[Pkgs == "recommended" & !is.na(Pkgs)])  
  
## [1] "boot"       "class"      "cluster"    "codetools"   "foreign"  
## [6] "KernSmooth" "lattice"    "MASS"       "Matrix"     "mgcv"  
## [11] "nlme"       "nnet"       "rpart"     "spatial"    "survival"
```

数据变形，分组统计聚合等，用以作为模型的输入，绘图的对象，操作的数据对象是数据框 (`data.frame`) 类型的，而且如果没有特别说明，文中出现的数据集都是 Base R 内置的，第三方 R 包或者来源于网上的数据集都会加以说明。

```
# 给定一个/些 R 包名，返回该 R 包存放的位置
sapply(.libPaths(), function(pkg_path) {
  c("survival", "ggplot2") %in% .packages(T, lib.loc = pkg_path)
})  
  
##          /home/runner/work/_temp/Library /opt/R/4.1.0/lib/R/library
## [1,]                FALSE                 TRUE
## [2,]                TRUE                 FALSE
```

5.1 查看数据

查看属性

```
str(iris)  
  
## 'data.frame': 150 obs. of 5 variables:  
## $ Sepal.Length: num 5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...  
## $ Sepal.Width : num 3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...  
## $ Petal.Length: num 1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...  
## $ Petal.Width : num 0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
```



```
## $ Species : Factor w/ 3 levels "setosa","versicolor",...: 1 1 1 1 1 1 1 1 1 ...
```

查看部分数据集

```
head(iris, 5)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1         5.1      3.5         1.4       0.2  setosa
## 2         4.9      3.0         1.4       0.2  setosa
## 3         4.7      3.2         1.3       0.2  setosa
## 4         4.6      3.1         1.5       0.2  setosa
## 5         5.0      3.6         1.4       0.2  setosa
```

```
tail(iris, 5)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 146        6.7      3.0         5.2      2.3 virginica
## 147        6.3      2.5         5.0      1.9 virginica
## 148        6.5      3.0         5.2      2.0 virginica
## 149        6.2      3.4         5.4      2.3 virginica
## 150        5.9      3.0         5.1      1.8 virginica
```

查看文件前（后）5行

```
head -n 5 test.csv
```

```
tail -n 5 test.csv
```

对象的类型，存储方式

```
class(iris)
```

```
## [1] "data.frame"
```

```
mode(iris)
```

```
## [1] "list"
```

```
typeof(iris)
```

```
## [1] "list"
```

查看对象在 R 环境中所占空间的大小

```
object.size(iris)
```

```
## 7256 bytes
```



```
object.size(letters)
## 1712 bytes
object.size(ls)
## 89880 bytes
format(object.size(library), units = "auto")
## [1] "1.8 Mb"
```

5.2 提取子集

```
subset(x, subset, select, drop = FALSE, ...)
```

参数 `subset` 代表行操作，`select` 代表列操作，函数 `subset` 从数据框中提取部分数据

```
subset(iris, Species == "virginica")
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
## 101	6.3	3.3	6.0	2.5	virginica
## 102	5.8	2.7	5.1	1.9	virginica
## 103	7.1	3.0	5.9	2.1	virginica
## 104	6.3	2.9	5.6	1.8	virginica
## 105	6.5	3.0	5.8	2.2	virginica
## 106	7.6	3.0	6.6	2.1	virginica
## 107	4.9	2.5	4.5	1.7	virginica
## 108	7.3	2.9	6.3	1.8	virginica
## 109	6.7	2.5	5.8	1.8	virginica
## 110	7.2	3.6	6.1	2.5	virginica
## 111	6.5	3.2	5.1	2.0	virginica
## 112	6.4	2.7	5.3	1.9	virginica
## 113	6.8	3.0	5.5	2.1	virginica
## 114	5.7	2.5	5.0	2.0	virginica
## 115	5.8	2.8	5.1	2.4	virginica
## 116	6.4	3.2	5.3	2.3	virginica
## 117	6.5	3.0	5.5	1.8	virginica



```
## 118      7.7      3.8      6.7      2.2 virginica
## 119      7.7      2.6      6.9      2.3 virginica
## 120      6.0      2.2      5.0      1.5 virginica
## 121      6.9      3.2      5.7      2.3 virginica
## 122      5.6      2.8      4.9      2.0 virginica
## 123      7.7      2.8      6.7      2.0 virginica
## 124      6.3      2.7      4.9      1.8 virginica
## 125      6.7      3.3      5.7      2.1 virginica
## 126      7.2      3.2      6.0      1.8 virginica
## 127      6.2      2.8      4.8      1.8 virginica
## 128      6.1      3.0      4.9      1.8 virginica
## 129      6.4      2.8      5.6      2.1 virginica
## 130      7.2      3.0      5.8      1.6 virginica
## 131      7.4      2.8      6.1      1.9 virginica
## 132      7.9      3.8      6.4      2.0 virginica
## 133      6.4      2.8      5.6      2.2 virginica
## 134      6.3      2.8      5.1      1.5 virginica
## 135      6.1      2.6      5.6      1.4 virginica
## 136      7.7      3.0      6.1      2.3 virginica
## 137      6.3      3.4      5.6      2.4 virginica
## 138      6.4      3.1      5.5      1.8 virginica
## 139      6.0      3.0      4.8      1.8 virginica
## 140      6.9      3.1      5.4      2.1 virginica
## 141      6.7      3.1      5.6      2.4 virginica
## 142      6.9      3.1      5.1      2.3 virginica
## 143      5.8      2.7      5.1      1.9 virginica
## 144      6.8      3.2      5.9      2.3 virginica
## 145      6.7      3.3      5.7      2.5 virginica
## 146      6.7      3.0      5.2      2.3 virginica
## 147      6.3      2.5      5.0      1.9 virginica
## 148      6.5      3.0      5.2      2.0 virginica
## 149      6.2      3.4      5.4      2.3 virginica
## 150      5.9      3.0      5.1      1.8 virginica

# summary(iris$Sepal.Length)  mean(iris$Sepal.Length)
# 且的逻辑
# subset(iris, Species == "virginica" & Sepal.Length > 5.84333)
```

云
湘
黄
C

```
subset(iris, Species == "virginica" &  
Sepal.Length > mean(Sepal.Length))
```

	## Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
## 101	6.3	3.3	6.0	2.5	virginica
## 103	7.1	3.0	5.9	2.1	virginica
## 104	6.3	2.9	5.6	1.8	virginica
## 105	6.5	3.0	5.8	2.2	virginica
## 106	7.6	3.0	6.6	2.1	virginica
## 108	7.3	2.9	6.3	1.8	virginica
## 109	6.7	2.5	5.8	1.8	virginica
## 110	7.2	3.6	6.1	2.5	virginica
## 111	6.5	3.2	5.1	2.0	virginica
## 112	6.4	2.7	5.3	1.9	virginica
## 113	6.8	3.0	5.5	2.1	virginica
## 116	6.4	3.2	5.3	2.3	virginica
## 117	6.5	3.0	5.5	1.8	virginica
## 118	7.7	3.8	6.7	2.2	virginica
## 119	7.7	2.6	6.9	2.3	virginica
## 120	6.0	2.2	5.0	1.5	virginica
## 121	6.9	3.2	5.7	2.3	virginica
## 123	7.7	2.8	6.7	2.0	virginica
## 124	6.3	2.7	4.9	1.8	virginica
## 125	6.7	3.3	5.7	2.1	virginica
## 126	7.2	3.2	6.0	1.8	virginica
## 127	6.2	2.8	4.8	1.8	virginica
## 128	6.1	3.0	4.9	1.8	virginica
## 129	6.4	2.8	5.6	2.1	virginica
## 130	7.2	3.0	5.8	1.6	virginica
## 131	7.4	2.8	6.1	1.9	virginica
## 132	7.9	3.8	6.4	2.0	virginica
## 133	6.4	2.8	5.6	2.2	virginica
## 134	6.3	2.8	5.1	1.5	virginica
## 135	6.1	2.6	5.6	1.4	virginica
## 136	7.7	3.0	6.1	2.3	virginica
## 137	6.3	3.4	5.6	2.4	virginica

```
## 138      6.4      3.1      5.5      1.8 virginica
## 139      6.0      3.0      4.8      1.8 virginica
## 140      6.9      3.1      5.4      2.1 virginica
## 141      6.7      3.1      5.6      2.4 virginica
## 142      6.9      3.1      5.1      2.3 virginica
## 144      6.8      3.2      5.9      2.3 virginica
## 145      6.7      3.3      5.7      2.5 virginica
## 146      6.7      3.0      5.2      2.3 virginica
## 147      6.3      2.5      5.0      1.9 virginica
## 148      6.5      3.0      5.2      2.0 virginica
## 149      6.2      3.4      5.4      2.3 virginica
## 150      5.9      3.0      5.1      1.8 virginica
```

在行的子集范围内

```
subset(iris, Species %in% c("virginica", "versicolor") &
       Sepal.Length > mean(Sepal.Length))
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1      7.0      3.2      4.7      1.4 versicolor
## 2      6.4      3.2      4.5      1.5 versicolor
## 3      6.9      3.1      4.9      1.5 versicolor
## 4      6.5      2.8      4.6      1.5 versicolor
## 5      6.3      3.3      4.7      1.6 versicolor
## 6      6.6      2.9      4.6      1.3 versicolor
## 7      5.9      3.0      4.2      1.5 versicolor
## 8      6.0      2.2      4.0      1.0 versicolor
## 9      6.1      2.9      4.7      1.4 versicolor
## 10     6.7      3.1      4.4      1.4 versicolor
## 11     6.9      2.2      4.5      1.5 versicolor
## 12     5.9      3.2      4.8      1.8 versicolor
## 13     6.1      2.8      4.0      1.3 versicolor
## 14     6.3      2.5      4.9      1.5 versicolor
## 15     6.1      2.8      4.7      1.2 versicolor
## 16     6.4      2.9      4.3      1.3 versicolor
## 17     6.6      3.0      4.4      1.4 versicolor
## 18     6.8      2.8      4.8      1.4 versicolor
## 19     6.7      3.0      5.0      1.7 versicolor
```

云
湘
黄
C

## 79	6.0	2.9	4.5	1.5 versicolor
## 84	6.0	2.7	5.1	1.6 versicolor
## 86	6.0	3.4	4.5	1.6 versicolor
## 87	6.7	3.1	4.7	1.5 versicolor
## 88	6.3	2.3	4.4	1.3 versicolor
## 92	6.1	3.0	4.6	1.4 versicolor
## 98	6.2	2.9	4.3	1.3 versicolor
## 101	6.3	3.3	6.0	2.5 virginica
## 103	7.1	3.0	5.9	2.1 virginica
## 104	6.3	2.9	5.6	1.8 virginica
## 105	6.5	3.0	5.8	2.2 virginica
## 106	7.6	3.0	6.6	2.1 virginica
## 108	7.3	2.9	6.3	1.8 virginica
## 109	6.7	2.5	5.8	1.8 virginica
## 110	7.2	3.6	6.1	2.5 virginica
## 111	6.5	3.2	5.1	2.0 virginica
## 112	6.4	2.7	5.3	1.9 virginica
## 113	6.8	3.0	5.5	2.1 virginica
## 116	6.4	3.2	5.3	2.3 virginica
## 117	6.5	3.0	5.5	1.8 virginica
## 118	7.7	3.8	6.7	2.2 virginica
## 119	7.7	2.6	6.9	2.3 virginica
## 120	6.0	2.2	5.0	1.5 virginica
## 121	6.9	3.2	5.7	2.3 virginica
## 123	7.7	2.8	6.7	2.0 virginica
## 124	6.3	2.7	4.9	1.8 virginica
## 125	6.7	3.3	5.7	2.1 virginica
## 126	7.2	3.2	6.0	1.8 virginica
## 127	6.2	2.8	4.8	1.8 virginica
## 128	6.1	3.0	4.9	1.8 virginica
## 129	6.4	2.8	5.6	2.1 virginica
## 130	7.2	3.0	5.8	1.6 virginica
## 131	7.4	2.8	6.1	1.9 virginica
## 132	7.9	3.8	6.4	2.0 virginica
## 133	6.4	2.8	5.6	2.2 virginica
## 134	6.3	2.8	5.1	1.5 virginica

```
## 135      6.1      2.6      5.6      1.4  virginica
## 136      7.7      3.0      6.1      2.3  virginica
## 137      6.3      3.4      5.6      2.4  virginica
## 138      6.4      3.1      5.5      1.8  virginica
## 139      6.0      3.0      4.8      1.8  virginica
## 140      6.9      3.1      5.4      2.1  virginica
## 141      6.7      3.1      5.6      2.4  virginica
## 142      6.9      3.1      5.1      2.3  virginica
## 144      6.8      3.2      5.9      2.3  virginica
## 145      6.7      3.3      5.7      2.5  virginica
## 146      6.7      3.0      5.2      2.3  virginica
## 147      6.3      2.5      5.0      1.9  virginica
## 148      6.5      3.0      5.2      2.0  virginica
## 149      6.2      3.4      5.4      2.3  virginica
## 150      5.9      3.0      5.1      1.8  virginica
```

```
# 在列的子集中 先选中列
subset(iris, Sepal.Length > mean(Sepal.Length),
       select = c("Sepal.Length", "Species")
     )
```

```
##      Sepal.Length   Species
## 51      7.0 versicolor
## 52      6.4 versicolor
## 53      6.9 versicolor
## 55      6.5 versicolor
## 57      6.3 versicolor
## 59      6.6 versicolor
## 62      5.9 versicolor
## 63      6.0 versicolor
## 64      6.1 versicolor
## 66      6.7 versicolor
## 69      6.2 versicolor
## 71      5.9 versicolor
## 72      6.1 versicolor
## 73      6.3 versicolor
## 74      6.1 versicolor
```

云
湘
黄
©

```
## 75      6.4 versicolor
## 76      6.6 versicolor
## 77      6.8 versicolor
## 78      6.7 versicolor
## 79      6.0 versicolor
## 84      6.0 versicolor
## 86      6.0 versicolor
## 87      6.7 versicolor
## 88      6.3 versicolor
## 92      6.1 versicolor
## 98      6.2 versicolor
## 101     6.3 virginica
## 103     7.1 virginica
## 104     6.3 virginica
## 105     6.5 virginica
## 106     7.6 virginica
## 108     7.3 virginica
## 109     6.7 virginica
## 110     7.2 virginica
## 111     6.5 virginica
## 112     6.4 virginica
## 113     6.8 virginica
## 116     6.4 virginica
## 117     6.5 virginica
## 118     7.7 virginica
## 119     7.7 virginica
## 120     6.0 virginica
## 121     6.9 virginica
## 123     7.7 virginica
## 124     6.3 virginica
## 125     6.7 virginica
## 126     7.2 virginica
## 127     6.2 virginica
## 128     6.1 virginica
## 129     6.4 virginica
## 130     7.2 virginica
```



```
## 131      7.4  virginica
## 132      7.9  virginica
## 133      6.4  virginica
## 134      6.3  virginica
## 135      6.1  virginica
## 136      7.7  virginica
## 137      6.3  virginica
## 138      6.4  virginica
## 139      6.0  virginica
## 140      6.9  virginica
## 141      6.7  virginica
## 142      6.9  virginica
## 144      6.8  virginica
## 145      6.7  virginica
## 146      6.7  virginica
## 147      6.3  virginica
## 148      6.5  virginica
## 149      6.2  virginica
## 150      5.9  virginica
```

高级操作：加入正则表达式筛选

```
## sometimes requiring a logical 'subset' argument is a nuisance
nm <- rownames(state.x77)
start_with_M <- nm %in% grep("^M", nm, value = TRUE)
subset(state.x77, start_with_M, Illiteracy:Murder)
```

	Illiteracy	Life	Exp	Murder
## Maine	0.7	70.39	2.7	
## Maryland	0.9	70.22	8.5	
## Massachusetts	1.1	71.83	3.3	
## Michigan	0.9	70.63	11.1	
## Minnesota	0.6	72.96	2.3	
## Mississippi	2.4	68.09	12.5	
## Missouri	0.8	70.69	9.3	
## Montana	0.6	70.56	5.0	



简化

```
subset(state.x77, subset = grepl("^M", rownames(state.x77)), select = Illiteracy:Murder)

##          Illiteracy Life Exp Murder
## Maine           0.7   70.39   2.7
## Maryland        0.9   70.22   8.5
## Massachusetts   1.1   71.83   3.3
## Michigan         0.9   70.63  11.1
## Minnesota       0.6   72.96   2.3
## Mississippi     2.4   68.09  12.5
## Missouri        0.8   70.69   9.3
## Montana         0.6   70.56   5.0
```

继续简化

```
subset(state.x77, grepl("^M", rownames(state.x77)), Illiteracy:Murder)
```

```
##          Illiteracy Life Exp Murder
## Maine           0.7   70.39   2.7
## Maryland        0.9   70.22   8.5
## Massachusetts   1.1   71.83   3.3
## Michigan         0.9   70.63  11.1
## Minnesota       0.6   72.96   2.3
## Mississippi     2.4   68.09  12.5
## Missouri        0.8   70.69   9.3
## Montana         0.6   70.56   5.0
```

注意

警告：这是一个为了交互使用打造的便捷函数。对于编程，最好使用标准的子集函数，如 [，特别地，参数 `subset` 的非标准计算 (non-standard evaluation)^a 可能带来意想不到的后果。

^aThomas Lumley (2003) Standard nonstandard evaluation rules. <https://developer.r-project.org/nonstandard-eval.pdf>

使用索引 [

```
iris[iris$Species == "virginica", ]
```

```
##      Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 101         6.3        3.3         6.0        2.5 virginica
```

## 102	5.8	2.7	5.1	1.9 virginica
## 103	7.1	3.0	5.9	2.1 virginica
## 104	6.3	2.9	5.6	1.8 virginica
## 105	6.5	3.0	5.8	2.2 virginica
## 106	7.6	3.0	6.6	2.1 virginica
## 107	4.9	2.5	4.5	1.7 virginica
## 108	7.3	2.9	6.3	1.8 virginica
## 109	6.7	2.5	5.8	1.8 virginica
## 110	7.2	3.6	6.1	2.5 virginica
## 111	6.5	3.2	5.1	2.0 virginica
## 112	6.4	2.7	5.3	1.9 virginica
## 113	6.8	3.0	5.5	2.1 virginica
## 114	5.7	2.5	5.0	2.0 virginica
## 115	5.8	2.8	5.1	2.4 virginica
## 116	6.4	3.2	5.3	2.3 virginica
## 117	6.5	3.0	5.5	1.8 virginica
## 118	7.7	3.8	6.7	2.2 virginica
## 119	7.7	2.6	6.9	2.3 virginica
## 120	6.0	2.2	5.0	1.5 virginica
## 121	6.9	3.2	5.7	2.3 virginica
## 122	5.6	2.8	4.9	2.0 virginica
## 123	7.7	2.8	6.7	2.0 virginica
## 124	6.3	2.7	4.9	1.8 virginica
## 125	6.7	3.3	5.7	2.1 virginica
## 126	7.2	3.2	6.0	1.8 virginica
## 127	6.2	2.8	4.8	1.8 virginica
## 128	6.1	3.0	4.9	1.8 virginica
## 129	6.4	2.8	5.6	2.1 virginica
## 130	7.2	3.0	5.8	1.6 virginica
## 131	7.4	2.8	6.1	1.9 virginica
## 132	7.9	3.8	6.4	2.0 virginica
## 133	6.4	2.8	5.6	2.2 virginica
## 134	6.3	2.8	5.1	1.5 virginica
## 135	6.1	2.6	5.6	1.4 virginica
## 136	7.7	3.0	6.1	2.3 virginica
## 137	6.3	3.4	5.6	2.4 virginica

云
湘
黄
◎

```
## 138      6.4      3.1      5.5      1.8 virginica
## 139      6.0      3.0      4.8      1.8 virginica
## 140      6.9      3.1      5.4      2.1 virginica
## 141      6.7      3.1      5.6      2.4 virginica
## 142      6.9      3.1      5.1      2.3 virginica
## 143      5.8      2.7      5.1      1.9 virginica
## 144      6.8      3.2      5.9      2.3 virginica
## 145      6.7      3.3      5.7      2.5 virginica
## 146      6.7      3.0      5.2      2.3 virginica
## 147      6.3      2.5      5.0      1.9 virginica
## 148      6.5      3.0      5.2      2.0 virginica
## 149      6.2      3.4      5.4      2.3 virginica
## 150      5.9      3.0      5.1      1.8 virginica

iris[iris$Species == "virginica" &
  iris$Sepal.Length > mean(iris$Sepal.Length), ]
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 101      6.3      3.3      6.0      2.5 virginica
## 103      7.1      3.0      5.9      2.1 virginica
## 104      6.3      2.9      5.6      1.8 virginica
## 105      6.5      3.0      5.8      2.2 virginica
## 106      7.6      3.0      6.6      2.1 virginica
## 108      7.3      2.9      6.3      1.8 virginica
## 109      6.7      2.5      5.8      1.8 virginica
## 110      7.2      3.6      6.1      2.5 virginica
## 111      6.5      3.2      5.1      2.0 virginica
## 112      6.4      2.7      5.3      1.9 virginica
## 113      6.8      3.0      5.5      2.1 virginica
## 116      6.4      3.2      5.3      2.3 virginica
## 117      6.5      3.0      5.5      1.8 virginica
## 118      7.7      3.8      6.7      2.2 virginica
## 119      7.7      2.6      6.9      2.3 virginica
## 120      6.0      2.2      5.0      1.5 virginica
## 121      6.9      3.2      5.7      2.3 virginica
## 123      7.7      2.8      6.7      2.0 virginica
## 124      6.3      2.7      4.9      1.8 virginica
```

```
## 125      6.7      3.3      5.7      2.1 virginica
## 126      7.2      3.2      6.0      1.8 virginica
## 127      6.2      2.8      4.8      1.8 virginica
## 128      6.1      3.0      4.9      1.8 virginica
## 129      6.4      2.8      5.6      2.1 virginica
## 130      7.2      3.0      5.8      1.6 virginica
## 131      7.4      2.8      6.1      1.9 virginica
## 132      7.9      3.8      6.4      2.0 virginica
## 133      6.4      2.8      5.6      2.2 virginica
## 134      6.3      2.8      5.1      1.5 virginica
## 135      6.1      2.6      5.6      1.4 virginica
## 136      7.7      3.0      6.1      2.3 virginica
## 137      6.3      3.4      5.6      2.4 virginica
## 138      6.4      3.1      5.5      1.8 virginica
## 139      6.0      3.0      4.8      1.8 virginica
## 140      6.9      3.1      5.4      2.1 virginica
## 141      6.7      3.1      5.6      2.4 virginica
## 142      6.9      3.1      5.1      2.3 virginica
## 144      6.8      3.2      5.9      2.3 virginica
## 145      6.7      3.3      5.7      2.5 virginica
## 146      6.7      3.0      5.2      2.3 virginica
## 147      6.3      2.5      5.0      1.9 virginica
## 148      6.5      3.0      5.2      2.0 virginica
## 149      6.2      3.4      5.4      2.3 virginica
## 150      5.9      3.0      5.1      1.8 virginica
```

```
iris[
  iris$Species == "virginica" &
  iris$Sepal.Length > mean(iris$Sepal.Length),
  c("Sepal.Length", "Species")
]
```

```
##      Sepal.Length  Species
## 101      6.3 virginica
## 103      7.1 virginica
## 104      6.3 virginica
## 105      6.5 virginica
```

云
湘
黄
©

```
## 106      7.6 virginica
## 108      7.3 virginica
## 109      6.7 virginica
## 110      7.2 virginica
## 111      6.5 virginica
## 112      6.4 virginica
## 113      6.8 virginica
## 116      6.4 virginica
## 117      6.5 virginica
## 118      7.7 virginica
## 119      7.7 virginica
## 120      6.0 virginica
## 121      6.9 virginica
## 123      7.7 virginica
## 124      6.3 virginica
## 125      6.7 virginica
## 126      7.2 virginica
## 127      6.2 virginica
## 128      6.1 virginica
## 129      6.4 virginica
## 130      7.2 virginica
## 131      7.4 virginica
## 132      7.9 virginica
## 133      6.4 virginica
## 134      6.3 virginica
## 135      6.1 virginica
## 136      7.7 virginica
## 137      6.3 virginica
## 138      6.4 virginica
## 139      6.0 virginica
## 140      6.9 virginica
## 141      6.7 virginica
## 142      6.9 virginica
## 144      6.8 virginica
## 145      6.7 virginica
## 146      6.7 virginica
```

```
## 147      6.3 virginica
## 148      6.5 virginica
## 149      6.2 virginica
## 150      5.9 virginica

iris[iris$Species == "setosa" & iris$Sepal.Length > 5.5, grepl("Sepal", colnames(
  iris,
  subset = Species == "setosa" & Sepal.Length > 5.5,
  select = grepl("Sepal", colnames(iris))
))

##      Sepal.Length Sepal.Width
## 15      5.8        4.0
## 16      5.7        4.4
## 19      5.7        3.8
```

5.3 数据重塑

重复测量数据的变形 Reshape Grouped Data，将宽格式 wide 的数据框变长格式 long 的，反之也行。reshape 还支持正则表达式

```
str(Indometh)

## Classes 'nfnGroupedData', 'nfGroupedData', 'groupedData' and 'data.frame': 66 ob
## $ Subject: Ord.factor w/ 6 levels "1"<"4"<"2"<"5"<...: 1 1 1 1 1 1 1 1 ...
## $ time   : num  0.25 0.5 0.75 1 1.25 2 3 4 5 6 ...
## $ conc   : num  1.5 0.94 0.78 0.48 0.37 0.19 0.12 0.11 0.08 0.07 ...
## - attr(*, "formula")=Class 'formula' language conc ~ time | Subject
##   .. ..- attr(*, ".Environment")=<environment: R_EmptyEnv>
## - attr(*, "labels")=List of 2
##   ..$ x: chr "Time since drug administration"
##   ..$ y: chr "Indomethacin concentration"
## - attr(*, "units")=List of 2
```

云
湘
黄
④

```
##   ..$ x: chr "(hr)"
##   ..$ y: chr "(mcg/ml)"

summary(Indometh)

##   Subject      time          conc
## 1:11    Min.   :0.250   Min.   :0.0500
## 4:11    1st Qu.:0.750   1st Qu.:0.1100
## 2:11    Median  :2.000   Median  :0.3400
## 5:11    Mean    :2.886   Mean    :0.5918
## 6:11    3rd Qu.:5.000   3rd Qu.:0.8325
## 3:11    Max.    :8.000   Max.    :2.7200
```

长的变宽

```
wide <- reshape(Indometh,
  v.names = "conc", idvar = "Subject",
  timevar = "time", direction = "wide"
)
wide[, 1:6]
```

```
##   Subject conc.0.25 conc.0.5 conc.0.75 conc.1 conc.1.25
## 1       1     1.50    0.94     0.78    0.48    0.37
## 12      2     2.03    1.63     0.71    0.70    0.64
## 23      3     2.72    1.49     1.16    0.80    0.80
## 34      4     1.85    1.39     1.02    0.89    0.59
## 45      5     2.05    1.04     0.81    0.39    0.30
## 56      6     2.31    1.44     1.03    0.84    0.64
```

宽的变长

```
reshape(wide, direction = "long")
```

```
##           Subject time conc
## 1.0.25      1 0.25 1.50
## 2.0.25      2 0.25 2.03
## 3.0.25      3 0.25 2.72
## 4.0.25      4 0.25 1.85
## 5.0.25      5 0.25 2.05
## 6.0.25      6 0.25 2.31
## 1.0.5       1 0.50 0.94
```

```
## 2.0.5      2 0.50 1.63
## 3.0.5      3 0.50 1.49
## 4.0.5      4 0.50 1.39
## 5.0.5      5 0.50 1.04
## 6.0.5      6 0.50 1.44
## 1.0.75     1 0.75 0.78
## 2.0.75     2 0.75 0.71
## 3.0.75     3 0.75 1.16
## 4.0.75     4 0.75 1.02
## 5.0.75     5 0.75 0.81
## 6.0.75     6 0.75 1.03
## 1.1        1 1.00 0.48
## 2.1        2 1.00 0.70
## 3.1        3 1.00 0.80
## 4.1        4 1.00 0.89
## 5.1        5 1.00 0.39
## 6.1        6 1.00 0.84
## 1.1.25    1 1.25 0.37
## 2.1.25    2 1.25 0.64
## 3.1.25    3 1.25 0.80
## 4.1.25    4 1.25 0.59
## 5.1.25    5 1.25 0.30
## 6.1.25    6 1.25 0.64
## 1.2        1 2.00 0.19
## 2.2        2 2.00 0.36
## 3.2        3 2.00 0.39
## 4.2        4 2.00 0.40
## 5.2        5 2.00 0.23
## 6.2        6 2.00 0.42
## 1.3        1 3.00 0.12
## 2.3        2 3.00 0.32
## 3.3        3 3.00 0.22
## 4.3        4 3.00 0.16
## 5.3        5 3.00 0.13
## 6.3        6 3.00 0.24
## 1.4        1 4.00 0.11
```

云 湘 黄

## 2.4	2 4.00 0.20
## 3.4	3 4.00 0.12
## 4.4	4 4.00 0.11
## 5.4	5 4.00 0.11
## 6.4	6 4.00 0.17
## 1.5	1 5.00 0.08
## 2.5	2 5.00 0.25
## 3.5	3 5.00 0.11
## 4.5	4 5.00 0.10
## 5.5	5 5.00 0.08
## 6.5	6 5.00 0.13
## 1.6	1 6.00 0.07
## 2.6	2 6.00 0.12
## 3.6	3 6.00 0.08
## 4.6	4 6.00 0.07
## 5.6	5 6.00 0.10
## 6.6	6 6.00 0.10
## 1.8	1 8.00 0.05
## 2.8	2 8.00 0.08
## 3.8	3 8.00 0.08
## 4.8	4 8.00 0.07
## 5.8	5 8.00 0.06
## 6.8	6 8.00 0.09

宽的格式变成长的格式 <https://stackoverflow.com/questions/2185252> 或者长的格式变成宽的格式 <https://stackoverflow.com/questions/5890584/>

```
set.seed(45)
dat <- data.frame(
  name = rep(c("Orange", "Apple"), each=4),
  numbers = rep(1:4, 2),
  value = rnorm(8))
dat
```

```
##      name numbers      value
## 1 Orange       1  0.3407997
## 2 Orange       2 -0.7033403
## 3 Orange       3 -0.3795377
```

```
## 4 Orange      4 -0.7460474
## 5 Apple       1 -0.8981073
## 6 Apple       2 -0.3347941
## 7 Apple       3 -0.5013782
## 8 Apple       4 -0.1745357

reshape(dat, idvar = "name", timevar = "numbers", direction = "wide")

##      name    value.1    value.2    value.3    value.4
## 1 Orange  0.3407997 -0.7033403 -0.3795377 -0.7460474
## 5 Apple   -0.8981073 -0.3347941 -0.5013782 -0.1745357

## times need not be numeric
df <- data.frame(id = rep(1:4, rep(2,4)),
                  visit = I(rep(c("Before","After"), 4)),
                  x = rnorm(4), y = runif(4))

df

##   id visit        x         y
## 1  1 Before  1.8090374 0.89106978
## 2  1 After   -0.2301050 0.06920426
## 3  2 Before -1.1304182 0.94623103
## 4  2 After   0.2159889 0.74850150
## 5  3 Before  1.8090374 0.89106978
## 6  3 After   -0.2301050 0.06920426
## 7  4 Before -1.1304182 0.94623103
## 8  4 After   0.2159889 0.74850150

reshape(df, timevar = "visit", idvar = "id", direction = "wide")

##   id x.Before y.Before x.After y.After
## 1  1  1.809037 0.8910698 -0.2301050 0.06920426
## 3  2 -1.130418 0.9462310  0.2159889 0.74850150
## 5  3  1.809037 0.8910698 -0.2301050 0.06920426
## 7  4 -1.130418 0.9462310  0.2159889 0.74850150

## warns that y is really varying
reshape(df, timevar = "visit", idvar = "id", direction = "wide", v.names = "x")

## Warning in reshapeWide(data, idvar = idvar, timevar = timevar, varying =
## varying, : some constant variables (y) are really varying
```



```
##   id      y  x.Before  x.After
## 1  1 0.8910698 1.809037 -0.2301050
## 3  2 0.9462310 -1.130418  0.2159889
## 5  3 0.8910698 1.809037 -0.2301050
## 7  4 0.9462310 -1.130418  0.2159889
```

更加复杂的例子，`gambia` 数据集，重塑的效果是使得个体水平的长格式变为村庄水平的宽格式

```
# data(gambia, package = "geoR")
# 在线下载数据集
gambia <- read.table(
  file =
    paste("http://www.leg.ufpr.br/lib/exe/fetch.php",
          "pessoais:paulojus:mbgbook:datasets:gambia.txt",
          sep = "/"),
  header = TRUE
)
head(gambia)

# Building a "village-level" data frame
ind <- paste("x", gambia[, 1], "y", gambia[, 2], sep = "")
village <- gambia[!duplicated(ind), c(1:2, 7:8)]
village$prev <- as.vector(tapply(gambia$pos, ind, mean))
head(village)
```

5.4 数据转换

`transform` 对数据框中的某些列做计算，取对数，将计算的结果单存一列加到数据框中

```
transform(iris, scale.sl = (max(Sepal.Length) - Sepal.Length) / (max(Sepal.Length) - mi
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species  scale.sl
## 1          5.1       3.5        1.4       0.2  setosa 0.77777778
## 2          4.9       3.0        1.4       0.2  setosa 0.83333333
## 3          4.7       3.2        1.3       0.2  setosa 0.88888889
## 4          4.6       3.1        1.5       0.2  setosa 0.91666667
## 5          5.0       3.6        1.4       0.2  setosa 0.80555556
```

```
## 6      5.4    3.9    1.7    0.4    setosa 0.694444444
## 7      4.6    3.4    1.4    0.3    setosa 0.916666667
## 8      5.0    3.4    1.5    0.2    setosa 0.805555556
## 9      4.4    2.9    1.4    0.2    setosa 0.972222222
## 10     4.9    3.1    1.5    0.1    setosa 0.833333333
## 11     5.4    3.7    1.5    0.2    setosa 0.694444444
## 12     4.8    3.4    1.6    0.2    setosa 0.861111111
## 13     4.8    3.0    1.4    0.1    setosa 0.861111111
## 14     4.3    3.0    1.1    0.1    setosa 1.000000000
## 15     5.8    4.0    1.2    0.2    setosa 0.583333333
## 16     5.7    4.4    1.5    0.4    setosa 0.611111111
## 17     5.4    3.9    1.3    0.4    setosa 0.694444444
## 18     5.1    3.5    1.4    0.3    setosa 0.777777778
## 19     5.7    3.8    1.7    0.3    setosa 0.611111111
## 20     5.1    3.8    1.5    0.3    setosa 0.777777778
## 21     5.4    3.4    1.7    0.2    setosa 0.694444444
## 22     5.1    3.7    1.5    0.4    setosa 0.777777778
## 23     4.6    3.6    1.0    0.2    setosa 0.916666667
## 24     5.1    3.3    1.7    0.5    setosa 0.777777778
## 25     4.8    3.4    1.9    0.2    setosa 0.861111111
## 26     5.0    3.0    1.6    0.2    setosa 0.805555556
## 27     5.0    3.4    1.6    0.4    setosa 0.805555556
## 28     5.2    3.5    1.5    0.2    setosa 0.750000000
## 29     5.2    3.4    1.4    0.2    setosa 0.750000000
## 30     4.7    3.2    1.6    0.2    setosa 0.888888889
## 31     4.8    3.1    1.6    0.2    setosa 0.861111111
## 32     5.4    3.4    1.5    0.4    setosa 0.694444444
## 33     5.2    4.1    1.5    0.1    setosa 0.750000000
## 34     5.5    4.2    1.4    0.2    setosa 0.666666667
## 35     4.9    3.1    1.5    0.2    setosa 0.833333333
## 36     5.0    3.2    1.2    0.2    setosa 0.805555556
## 37     5.5    3.5    1.3    0.2    setosa 0.666666667
## 38     4.9    3.6    1.4    0.1    setosa 0.833333333
## 39     4.4    3.0    1.3    0.2    setosa 0.972222222
## 40     5.1    3.4    1.5    0.2    setosa 0.777777778
## 41     5.0    3.5    1.3    0.3    setosa 0.805555556
```

云
湘
黄
◎

## 42	4.5	2.3	1.3	0.3	setosa 0.94444444
## 43	4.4	3.2	1.3	0.2	setosa 0.97222222
## 44	5.0	3.5	1.6	0.6	setosa 0.80555556
## 45	5.1	3.8	1.9	0.4	setosa 0.77777778
## 46	4.8	3.0	1.4	0.3	setosa 0.86111111
## 47	5.1	3.8	1.6	0.2	setosa 0.77777778
## 48	4.6	3.2	1.4	0.2	setosa 0.91666667
## 49	5.3	3.7	1.5	0.2	setosa 0.72222222
## 50	5.0	3.3	1.4	0.2	setosa 0.80555556
## 51	7.0	3.2	4.7	1.4	versicolor 0.25000000
## 52	6.4	3.2	4.5	1.5	versicolor 0.41666667
## 53	6.9	3.1	4.9	1.5	versicolor 0.27777778
## 54	5.5	2.3	4.0	1.3	versicolor 0.66666667
## 55	6.5	2.8	4.6	1.5	versicolor 0.38888889
## 56	5.7	2.8	4.5	1.3	versicolor 0.61111111
## 57	6.3	3.3	4.7	1.6	versicolor 0.44444444
## 58	4.9	2.4	3.3	1.0	versicolor 0.83333333
## 59	6.6	2.9	4.6	1.3	versicolor 0.36111111
## 60	5.2	2.7	3.9	1.4	versicolor 0.75000000
## 61	5.0	2.0	3.5	1.0	versicolor 0.80555556
## 62	5.9	3.0	4.2	1.5	versicolor 0.55555556
## 63	6.0	2.2	4.0	1.0	versicolor 0.52777778
## 64	6.1	2.9	4.7	1.4	versicolor 0.50000000
## 65	5.6	2.9	3.6	1.3	versicolor 0.63888889
## 66	6.7	3.1	4.4	1.4	versicolor 0.33333333
## 67	5.6	3.0	4.5	1.5	versicolor 0.63888889
## 68	5.8	2.7	4.1	1.0	versicolor 0.58333333
## 69	6.2	2.2	4.5	1.5	versicolor 0.47222222
## 70	5.6	2.5	3.9	1.1	versicolor 0.63888889
## 71	5.9	3.2	4.8	1.8	versicolor 0.55555556
## 72	6.1	2.8	4.0	1.3	versicolor 0.50000000
## 73	6.3	2.5	4.9	1.5	versicolor 0.44444444
## 74	6.1	2.8	4.7	1.2	versicolor 0.50000000
## 75	6.4	2.9	4.3	1.3	versicolor 0.41666667
## 76	6.6	3.0	4.4	1.4	versicolor 0.36111111
## 77	6.8	2.8	4.8	1.4	versicolor 0.30555556

## 78	6.7	3.0	5.0	1.7	versicolor 0.33333333
## 79	6.0	2.9	4.5	1.5	versicolor 0.52777778
## 80	5.7	2.6	3.5	1.0	versicolor 0.61111111
## 81	5.5	2.4	3.8	1.1	versicolor 0.66666667
## 82	5.5	2.4	3.7	1.0	versicolor 0.66666667
## 83	5.8	2.7	3.9	1.2	versicolor 0.58333333
## 84	6.0	2.7	5.1	1.6	versicolor 0.52777778
## 85	5.4	3.0	4.5	1.5	versicolor 0.69444444
## 86	6.0	3.4	4.5	1.6	versicolor 0.52777778
## 87	6.7	3.1	4.7	1.5	versicolor 0.33333333
## 88	6.3	2.3	4.4	1.3	versicolor 0.44444444
## 89	5.6	3.0	4.1	1.3	versicolor 0.63888889
## 90	5.5	2.5	4.0	1.3	versicolor 0.66666667
## 91	5.5	2.6	4.4	1.2	versicolor 0.66666667
## 92	6.1	3.0	4.6	1.4	versicolor 0.50000000
## 93	5.8	2.6	4.0	1.2	versicolor 0.58333333
## 94	5.0	2.3	3.3	1.0	versicolor 0.80555556
## 95	5.6	2.7	4.2	1.3	versicolor 0.63888889
## 96	5.7	3.0	4.2	1.2	versicolor 0.61111111
## 97	5.7	2.9	4.2	1.3	versicolor 0.61111111
## 98	6.2	2.9	4.3	1.3	versicolor 0.47222222
## 99	5.1	2.5	3.0	1.1	versicolor 0.77777778
## 100	5.7	2.8	4.1	1.3	versicolor 0.61111111
## 101	6.3	3.3	6.0	2.5	virginica 0.44444444
## 102	5.8	2.7	5.1	1.9	virginica 0.58333333
## 103	7.1	3.0	5.9	2.1	virginica 0.22222222
## 104	6.3	2.9	5.6	1.8	virginica 0.44444444
## 105	6.5	3.0	5.8	2.2	virginica 0.38888889
## 106	7.6	3.0	6.6	2.1	virginica 0.08333333
## 107	4.9	2.5	4.5	1.7	virginica 0.83333333
## 108	7.3	2.9	6.3	1.8	virginica 0.16666667
## 109	6.7	2.5	5.8	1.8	virginica 0.33333333
## 110	7.2	3.6	6.1	2.5	virginica 0.19444444
## 111	6.5	3.2	5.1	2.0	virginica 0.38888889
## 112	6.4	2.7	5.3	1.9	virginica 0.41666667
## 113	6.8	3.0	5.5	2.1	virginica 0.30555556

云
湘
黄
◎

## 114	5.7	2.5	5.0	2.0	virginica	0.61111111
## 115	5.8	2.8	5.1	2.4	virginica	0.58333333
## 116	6.4	3.2	5.3	2.3	virginica	0.41666667
## 117	6.5	3.0	5.5	1.8	virginica	0.38888889
## 118	7.7	3.8	6.7	2.2	virginica	0.05555556
## 119	7.7	2.6	6.9	2.3	virginica	0.05555556
## 120	6.0	2.2	5.0	1.5	virginica	0.52777778
## 121	6.9	3.2	5.7	2.3	virginica	0.27777778
## 122	5.6	2.8	4.9	2.0	virginica	0.63888889
## 123	7.7	2.8	6.7	2.0	virginica	0.05555556
## 124	6.3	2.7	4.9	1.8	virginica	0.44444444
## 125	6.7	3.3	5.7	2.1	virginica	0.33333333
## 126	7.2	3.2	6.0	1.8	virginica	0.19444444
## 127	6.2	2.8	4.8	1.8	virginica	0.47222222
## 128	6.1	3.0	4.9	1.8	virginica	0.50000000
## 129	6.4	2.8	5.6	2.1	virginica	0.41666667
## 130	7.2	3.0	5.8	1.6	virginica	0.19444444
## 131	7.4	2.8	6.1	1.9	virginica	0.13888889
## 132	7.9	3.8	6.4	2.0	virginica	0.00000000
## 133	6.4	2.8	5.6	2.2	virginica	0.41666667
## 134	6.3	2.8	5.1	1.5	virginica	0.44444444
## 135	6.1	2.6	5.6	1.4	virginica	0.50000000
## 136	7.7	3.0	6.1	2.3	virginica	0.05555556
## 137	6.3	3.4	5.6	2.4	virginica	0.44444444
## 138	6.4	3.1	5.5	1.8	virginica	0.41666667
## 139	6.0	3.0	4.8	1.8	virginica	0.52777778
## 140	6.9	3.1	5.4	2.1	virginica	0.27777778
## 141	6.7	3.1	5.6	2.4	virginica	0.33333333
## 142	6.9	3.1	5.1	2.3	virginica	0.27777778
## 143	5.8	2.7	5.1	1.9	virginica	0.58333333
## 144	6.8	3.2	5.9	2.3	virginica	0.30555556
## 145	6.7	3.3	5.7	2.5	virginica	0.33333333
## 146	6.7	3.0	5.2	2.3	virginica	0.33333333
## 147	6.3	2.5	5.0	1.9	virginica	0.44444444
## 148	6.5	3.0	5.2	2.0	virginica	0.38888889
## 149	6.2	3.4	5.4	2.3	virginica	0.47222222

```
## 150      5.9      3.0      5.1      1.8 virginica 0.555555556
```

验证一下 `scale.s1` 变量的第一个值

```
(max(iris$Sepal.Length) - 5.1) / (max(iris$Sepal.Length) - min(iris$Sepal.Length))
```

```
## [1] 0.7777778
```

注意

Warning: This is a convenience function intended for use interactively. For programming it is better to use the standard subsetting arithmetic functions, and in particular the non-standard evaluation of argument `transform` can have unanticipated consequences.

5.5 按列排序

在数据框内，根据 (`order`) 某一列或几列对行进行排序 (`sort`)，根据鸢尾花 (`iris`) 的类别 (`Species`) 对萼片 (`sepal`) 的长度进行排序，其余的列随之变化

对萼片的长度排序

```
iris[order(iris$Species, iris$Sepal.Length), ]
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
## 14	4.3	3.0	1.1	0.1	setosa
## 9	4.4	2.9	1.4	0.2	setosa
## 39	4.4	3.0	1.3	0.2	setosa
## 43	4.4	3.2	1.3	0.2	setosa
## 42	4.5	2.3	1.3	0.3	setosa
## 4	4.6	3.1	1.5	0.2	setosa
## 7	4.6	3.4	1.4	0.3	setosa
## 23	4.6	3.6	1.0	0.2	setosa
## 48	4.6	3.2	1.4	0.2	setosa
## 3	4.7	3.2	1.3	0.2	setosa
## 30	4.7	3.2	1.6	0.2	setosa
## 12	4.8	3.4	1.6	0.2	setosa
## 13	4.8	3.0	1.4	0.1	setosa
## 25	4.8	3.4	1.9	0.2	setosa
## 31	4.8	3.1	1.6	0.2	setosa

云 湘 黃 ⓒ

## 46	4.8	3.0	1.4	0.3	setosa
## 2	4.9	3.0	1.4	0.2	setosa
## 10	4.9	3.1	1.5	0.1	setosa
## 35	4.9	3.1	1.5	0.2	setosa
## 38	4.9	3.6	1.4	0.1	setosa
## 5	5.0	3.6	1.4	0.2	setosa
## 8	5.0	3.4	1.5	0.2	setosa
## 26	5.0	3.0	1.6	0.2	setosa
## 27	5.0	3.4	1.6	0.4	setosa
## 36	5.0	3.2	1.2	0.2	setosa
## 41	5.0	3.5	1.3	0.3	setosa
## 44	5.0	3.5	1.6	0.6	setosa
## 50	5.0	3.3	1.4	0.2	setosa
## 1	5.1	3.5	1.4	0.2	setosa
## 18	5.1	3.5	1.4	0.3	setosa
## 20	5.1	3.8	1.5	0.3	setosa
## 22	5.1	3.7	1.5	0.4	setosa
## 24	5.1	3.3	1.7	0.5	setosa
## 40	5.1	3.4	1.5	0.2	setosa
## 45	5.1	3.8	1.9	0.4	setosa
## 47	5.1	3.8	1.6	0.2	setosa
## 28	5.2	3.5	1.5	0.2	setosa
## 29	5.2	3.4	1.4	0.2	setosa
## 33	5.2	4.1	1.5	0.1	setosa
## 49	5.3	3.7	1.5	0.2	setosa
## 6	5.4	3.9	1.7	0.4	setosa
## 11	5.4	3.7	1.5	0.2	setosa
## 17	5.4	3.9	1.3	0.4	setosa
## 21	5.4	3.4	1.7	0.2	setosa
## 32	5.4	3.4	1.5	0.4	setosa
## 34	5.5	4.2	1.4	0.2	setosa
## 37	5.5	3.5	1.3	0.2	setosa
## 16	5.7	4.4	1.5	0.4	setosa
## 19	5.7	3.8	1.7	0.3	setosa
## 15	5.8	4.0	1.2	0.2	setosa
## 58	4.9	2.4	3.3	1.0	versicolor

## 61	5.0	2.0	3.5	1.0 versicolor
## 94	5.0	2.3	3.3	1.0 versicolor
## 99	5.1	2.5	3.0	1.1 versicolor
## 60	5.2	2.7	3.9	1.4 versicolor
## 85	5.4	3.0	4.5	1.5 versicolor
## 54	5.5	2.3	4.0	1.3 versicolor
## 81	5.5	2.4	3.8	1.1 versicolor
## 82	5.5	2.4	3.7	1.0 versicolor
## 90	5.5	2.5	4.0	1.3 versicolor
## 91	5.5	2.6	4.4	1.2 versicolor
## 65	5.6	2.9	3.6	1.3 versicolor
## 67	5.6	3.0	4.5	1.5 versicolor
## 70	5.6	2.5	3.9	1.1 versicolor
## 89	5.6	3.0	4.1	1.3 versicolor
## 95	5.6	2.7	4.2	1.3 versicolor
## 56	5.7	2.8	4.5	1.3 versicolor
## 80	5.7	2.6	3.5	1.0 versicolor
## 96	5.7	3.0	4.2	1.2 versicolor
## 97	5.7	2.9	4.2	1.3 versicolor
## 100	5.7	2.8	4.1	1.3 versicolor
## 68	5.8	2.7	4.1	1.0 versicolor
## 83	5.8	2.7	3.9	1.2 versicolor
## 93	5.8	2.6	4.0	1.2 versicolor
## 62	5.9	3.0	4.2	1.5 versicolor
## 71	5.9	3.2	4.8	1.8 versicolor
## 63	6.0	2.2	4.0	1.0 versicolor
## 79	6.0	2.9	4.5	1.5 versicolor
## 84	6.0	2.7	5.1	1.6 versicolor
## 86	6.0	3.4	4.5	1.6 versicolor
## 64	6.1	2.9	4.7	1.4 versicolor
## 72	6.1	2.8	4.0	1.3 versicolor
## 74	6.1	2.8	4.7	1.2 versicolor
## 92	6.1	3.0	4.6	1.4 versicolor
## 69	6.2	2.2	4.5	1.5 versicolor
## 98	6.2	2.9	4.3	1.3 versicolor
## 57	6.3	3.3	4.7	1.6 versicolor

云
湘
黄
④

## 73	6.3	2.5	4.9	1.5 versicolor
## 88	6.3	2.3	4.4	1.3 versicolor
## 52	6.4	3.2	4.5	1.5 versicolor
## 75	6.4	2.9	4.3	1.3 versicolor
## 55	6.5	2.8	4.6	1.5 versicolor
## 59	6.6	2.9	4.6	1.3 versicolor
## 76	6.6	3.0	4.4	1.4 versicolor
## 66	6.7	3.1	4.4	1.4 versicolor
## 78	6.7	3.0	5.0	1.7 versicolor
## 87	6.7	3.1	4.7	1.5 versicolor
## 77	6.8	2.8	4.8	1.4 versicolor
## 53	6.9	3.1	4.9	1.5 versicolor
## 51	7.0	3.2	4.7	1.4 versicolor
## 107	4.9	2.5	4.5	1.7 virginica
## 122	5.6	2.8	4.9	2.0 virginica
## 114	5.7	2.5	5.0	2.0 virginica
## 102	5.8	2.7	5.1	1.9 virginica
## 115	5.8	2.8	5.1	2.4 virginica
## 143	5.8	2.7	5.1	1.9 virginica
## 150	5.9	3.0	5.1	1.8 virginica
## 120	6.0	2.2	5.0	1.5 virginica
## 139	6.0	3.0	4.8	1.8 virginica
## 128	6.1	3.0	4.9	1.8 virginica
## 135	6.1	2.6	5.6	1.4 virginica
## 127	6.2	2.8	4.8	1.8 virginica
## 149	6.2	3.4	5.4	2.3 virginica
## 101	6.3	3.3	6.0	2.5 virginica
## 104	6.3	2.9	5.6	1.8 virginica
## 124	6.3	2.7	4.9	1.8 virginica
## 134	6.3	2.8	5.1	1.5 virginica
## 137	6.3	3.4	5.6	2.4 virginica
## 147	6.3	2.5	5.0	1.9 virginica
## 112	6.4	2.7	5.3	1.9 virginica
## 116	6.4	3.2	5.3	2.3 virginica
## 129	6.4	2.8	5.6	2.1 virginica
## 133	6.4	2.8	5.6	2.2 virginica

```
## 138      6.4      3.1      5.5      1.8  virginica
## 105      6.5      3.0      5.8      2.2  virginica
## 111      6.5      3.2      5.1      2.0  virginica
## 117      6.5      3.0      5.5      1.8  virginica
## 148      6.5      3.0      5.2      2.0  virginica
## 109      6.7      2.5      5.8      1.8  virginica
## 125      6.7      3.3      5.7      2.1  virginica
## 141      6.7      3.1      5.6      2.4  virginica
## 145      6.7      3.3      5.7      2.5  virginica
## 146      6.7      3.0      5.2      2.3  virginica
## 113      6.8      3.0      5.5      2.1  virginica
## 144      6.8      3.2      5.9      2.3  virginica
## 121      6.9      3.2      5.7      2.3  virginica
## 140      6.9      3.1      5.4      2.1  virginica
## 142      6.9      3.1      5.1      2.3  virginica
## 103      7.1      3.0      5.9      2.1  virginica
## 110      7.2      3.6      6.1      2.5  virginica
## 126      7.2      3.2      6.0      1.8  virginica
## 130      7.2      3.0      5.8      1.6  virginica
## 108      7.3      2.9      6.3      1.8  virginica
## 131      7.4      2.8      6.1      1.9  virginica
## 106      7.6      3.0      6.6      2.1  virginica
## 118      7.7      3.8      6.7      2.2  virginica
## 119      7.7      2.6      6.9      2.3  virginica
## 123      7.7      2.8      6.7      2.0  virginica
## 136      7.7      3.0      6.1      2.3  virginica
## 132      7.9      3.8      6.4      2.0  virginica
```

对花瓣的长度排序

```
iris[order(iris$Species, iris$Petal.Length), ]
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
## 23	4.6	3.6	1.0	0.2	setosa
## 14	4.3	3.0	1.1	0.1	setosa
## 15	5.8	4.0	1.2	0.2	setosa
## 36	5.0	3.2	1.2	0.2	setosa
## 3	4.7	3.2	1.3	0.2	setosa

云
湘
黄
④

## 17	5.4	3.9	1.3	0.4	setosa
## 37	5.5	3.5	1.3	0.2	setosa
## 39	4.4	3.0	1.3	0.2	setosa
## 41	5.0	3.5	1.3	0.3	setosa
## 42	4.5	2.3	1.3	0.3	setosa
## 43	4.4	3.2	1.3	0.2	setosa
## 1	5.1	3.5	1.4	0.2	setosa
## 2	4.9	3.0	1.4	0.2	setosa
## 5	5.0	3.6	1.4	0.2	setosa
## 7	4.6	3.4	1.4	0.3	setosa
## 9	4.4	2.9	1.4	0.2	setosa
## 13	4.8	3.0	1.4	0.1	setosa
## 18	5.1	3.5	1.4	0.3	setosa
## 29	5.2	3.4	1.4	0.2	setosa
## 34	5.5	4.2	1.4	0.2	setosa
## 38	4.9	3.6	1.4	0.1	setosa
## 46	4.8	3.0	1.4	0.3	setosa
## 48	4.6	3.2	1.4	0.2	setosa
## 50	5.0	3.3	1.4	0.2	setosa
## 4	4.6	3.1	1.5	0.2	setosa
## 8	5.0	3.4	1.5	0.2	setosa
## 10	4.9	3.1	1.5	0.1	setosa
## 11	5.4	3.7	1.5	0.2	setosa
## 16	5.7	4.4	1.5	0.4	setosa
## 20	5.1	3.8	1.5	0.3	setosa
## 22	5.1	3.7	1.5	0.4	setosa
## 28	5.2	3.5	1.5	0.2	setosa
## 32	5.4	3.4	1.5	0.4	setosa
## 33	5.2	4.1	1.5	0.1	setosa
## 35	4.9	3.1	1.5	0.2	setosa
## 40	5.1	3.4	1.5	0.2	setosa
## 49	5.3	3.7	1.5	0.2	setosa
## 12	4.8	3.4	1.6	0.2	setosa
## 26	5.0	3.0	1.6	0.2	setosa
## 27	5.0	3.4	1.6	0.4	setosa
## 30	4.7	3.2	1.6	0.2	setosa

## 31	4.8	3.1	1.6	0.2	setosa
## 44	5.0	3.5	1.6	0.6	setosa
## 47	5.1	3.8	1.6	0.2	setosa
## 6	5.4	3.9	1.7	0.4	setosa
## 19	5.7	3.8	1.7	0.3	setosa
## 21	5.4	3.4	1.7	0.2	setosa
## 24	5.1	3.3	1.7	0.5	setosa
## 25	4.8	3.4	1.9	0.2	setosa
## 45	5.1	3.8	1.9	0.4	setosa
## 99	5.1	2.5	3.0	1.1	versicolor
## 58	4.9	2.4	3.3	1.0	versicolor
## 94	5.0	2.3	3.3	1.0	versicolor
## 61	5.0	2.0	3.5	1.0	versicolor
## 80	5.7	2.6	3.5	1.0	versicolor
## 65	5.6	2.9	3.6	1.3	versicolor
## 82	5.5	2.4	3.7	1.0	versicolor
## 81	5.5	2.4	3.8	1.1	versicolor
## 60	5.2	2.7	3.9	1.4	versicolor
## 70	5.6	2.5	3.9	1.1	versicolor
## 83	5.8	2.7	3.9	1.2	versicolor
## 54	5.5	2.3	4.0	1.3	versicolor
## 63	6.0	2.2	4.0	1.0	versicolor
## 72	6.1	2.8	4.0	1.3	versicolor
## 90	5.5	2.5	4.0	1.3	versicolor
## 93	5.8	2.6	4.0	1.2	versicolor
## 68	5.8	2.7	4.1	1.0	versicolor
## 89	5.6	3.0	4.1	1.3	versicolor
## 100	5.7	2.8	4.1	1.3	versicolor
## 62	5.9	3.0	4.2	1.5	versicolor
## 95	5.6	2.7	4.2	1.3	versicolor
## 96	5.7	3.0	4.2	1.2	versicolor
## 97	5.7	2.9	4.2	1.3	versicolor
## 75	6.4	2.9	4.3	1.3	versicolor
## 98	6.2	2.9	4.3	1.3	versicolor
## 66	6.7	3.1	4.4	1.4	versicolor
## 76	6.6	3.0	4.4	1.4	versicolor

云
湘
黄
⑥

## 88	6.3	2.3	4.4	1.3 versicolor
## 91	5.5	2.6	4.4	1.2 versicolor
## 52	6.4	3.2	4.5	1.5 versicolor
## 56	5.7	2.8	4.5	1.3 versicolor
## 67	5.6	3.0	4.5	1.5 versicolor
## 69	6.2	2.2	4.5	1.5 versicolor
## 79	6.0	2.9	4.5	1.5 versicolor
## 85	5.4	3.0	4.5	1.5 versicolor
## 86	6.0	3.4	4.5	1.6 versicolor
## 55	6.5	2.8	4.6	1.5 versicolor
## 59	6.6	2.9	4.6	1.3 versicolor
## 92	6.1	3.0	4.6	1.4 versicolor
## 51	7.0	3.2	4.7	1.4 versicolor
## 57	6.3	3.3	4.7	1.6 versicolor
## 64	6.1	2.9	4.7	1.4 versicolor
## 74	6.1	2.8	4.7	1.2 versicolor
## 87	6.7	3.1	4.7	1.5 versicolor
## 71	5.9	3.2	4.8	1.8 versicolor
## 77	6.8	2.8	4.8	1.4 versicolor
## 53	6.9	3.1	4.9	1.5 versicolor
## 73	6.3	2.5	4.9	1.5 versicolor
## 78	6.7	3.0	5.0	1.7 versicolor
## 84	6.0	2.7	5.1	1.6 versicolor
## 107	4.9	2.5	4.5	1.7 virginica
## 127	6.2	2.8	4.8	1.8 virginica
## 139	6.0	3.0	4.8	1.8 virginica
## 122	5.6	2.8	4.9	2.0 virginica
## 124	6.3	2.7	4.9	1.8 virginica
## 128	6.1	3.0	4.9	1.8 virginica
## 114	5.7	2.5	5.0	2.0 virginica
## 120	6.0	2.2	5.0	1.5 virginica
## 147	6.3	2.5	5.0	1.9 virginica
## 102	5.8	2.7	5.1	1.9 virginica
## 111	6.5	3.2	5.1	2.0 virginica
## 115	5.8	2.8	5.1	2.4 virginica
## 134	6.3	2.8	5.1	1.5 virginica

## 142	6.9	3.1	5.1	2.3	virginica
## 143	5.8	2.7	5.1	1.9	virginica
## 150	5.9	3.0	5.1	1.8	virginica
## 146	6.7	3.0	5.2	2.3	virginica
## 148	6.5	3.0	5.2	2.0	virginica
## 112	6.4	2.7	5.3	1.9	virginica
## 116	6.4	3.2	5.3	2.3	virginica
## 140	6.9	3.1	5.4	2.1	virginica
## 149	6.2	3.4	5.4	2.3	virginica
## 113	6.8	3.0	5.5	2.1	virginica
## 117	6.5	3.0	5.5	1.8	virginica
## 138	6.4	3.1	5.5	1.8	virginica
## 104	6.3	2.9	5.6	1.8	virginica
## 129	6.4	2.8	5.6	2.1	virginica
## 133	6.4	2.8	5.6	2.2	virginica
## 135	6.1	2.6	5.6	1.4	virginica
## 137	6.3	3.4	5.6	2.4	virginica
## 141	6.7	3.1	5.6	2.4	virginica
## 121	6.9	3.2	5.7	2.3	virginica
## 125	6.7	3.3	5.7	2.1	virginica
## 145	6.7	3.3	5.7	2.5	virginica
## 105	6.5	3.0	5.8	2.2	virginica
## 109	6.7	2.5	5.8	1.8	virginica
## 130	7.2	3.0	5.8	1.6	virginica
## 103	7.1	3.0	5.9	2.1	virginica
## 144	6.8	3.2	5.9	2.3	virginica
## 101	6.3	3.3	6.0	2.5	virginica
## 126	7.2	3.2	6.0	1.8	virginica
## 110	7.2	3.6	6.1	2.5	virginica
## 131	7.4	2.8	6.1	1.9	virginica
## 136	7.7	3.0	6.1	2.3	virginica
## 108	7.3	2.9	6.3	1.8	virginica
## 132	7.9	3.8	6.4	2.0	virginica
## 106	7.6	3.0	6.6	2.1	virginica
## 118	7.7	3.8	6.7	2.2	virginica
## 123	7.7	2.8	6.7	2.0	virginica

云
湘
黄

```
## 119      7.7      2.6      6.9      2.3  virginica
```

先对花瓣的宽度排序，再对花瓣的长度排序

```
iris[order(iris$Petal.Width, iris$Petal.Length), ]
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
## 14	4.3	3.0	1.1	0.1	setosa
## 13	4.8	3.0	1.4	0.1	setosa
## 38	4.9	3.6	1.4	0.1	setosa
## 10	4.9	3.1	1.5	0.1	setosa
## 33	5.2	4.1	1.5	0.1	setosa
## 23	4.6	3.6	1.0	0.2	setosa
## 15	5.8	4.0	1.2	0.2	setosa
## 36	5.0	3.2	1.2	0.2	setosa
## 3	4.7	3.2	1.3	0.2	setosa
## 37	5.5	3.5	1.3	0.2	setosa
## 39	4.4	3.0	1.3	0.2	setosa
## 43	4.4	3.2	1.3	0.2	setosa
## 1	5.1	3.5	1.4	0.2	setosa
## 2	4.9	3.0	1.4	0.2	setosa
## 5	5.0	3.6	1.4	0.2	setosa
## 9	4.4	2.9	1.4	0.2	setosa
## 29	5.2	3.4	1.4	0.2	setosa
## 34	5.5	4.2	1.4	0.2	setosa
## 48	4.6	3.2	1.4	0.2	setosa
## 50	5.0	3.3	1.4	0.2	setosa
## 4	4.6	3.1	1.5	0.2	setosa
## 8	5.0	3.4	1.5	0.2	setosa
## 11	5.4	3.7	1.5	0.2	setosa
## 28	5.2	3.5	1.5	0.2	setosa
## 35	4.9	3.1	1.5	0.2	setosa
## 40	5.1	3.4	1.5	0.2	setosa
## 49	5.3	3.7	1.5	0.2	setosa
## 12	4.8	3.4	1.6	0.2	setosa
## 26	5.0	3.0	1.6	0.2	setosa
## 30	4.7	3.2	1.6	0.2	setosa
## 31	4.8	3.1	1.6	0.2	setosa

## 47	5.1	3.8	1.6	0.2	setosa
## 21	5.4	3.4	1.7	0.2	setosa
## 25	4.8	3.4	1.9	0.2	setosa
## 41	5.0	3.5	1.3	0.3	setosa
## 42	4.5	2.3	1.3	0.3	setosa
## 7	4.6	3.4	1.4	0.3	setosa
## 18	5.1	3.5	1.4	0.3	setosa
## 46	4.8	3.0	1.4	0.3	setosa
## 20	5.1	3.8	1.5	0.3	setosa
## 19	5.7	3.8	1.7	0.3	setosa
## 17	5.4	3.9	1.3	0.4	setosa
## 16	5.7	4.4	1.5	0.4	setosa
## 22	5.1	3.7	1.5	0.4	setosa
## 32	5.4	3.4	1.5	0.4	setosa
## 27	5.0	3.4	1.6	0.4	setosa
## 6	5.4	3.9	1.7	0.4	setosa
## 45	5.1	3.8	1.9	0.4	setosa
## 24	5.1	3.3	1.7	0.5	setosa
## 44	5.0	3.5	1.6	0.6	setosa
## 58	4.9	2.4	3.3	1.0	versicolor
## 94	5.0	2.3	3.3	1.0	versicolor
## 61	5.0	2.0	3.5	1.0	versicolor
## 80	5.7	2.6	3.5	1.0	versicolor
## 82	5.5	2.4	3.7	1.0	versicolor
## 63	6.0	2.2	4.0	1.0	versicolor
## 68	5.8	2.7	4.1	1.0	versicolor
## 99	5.1	2.5	3.0	1.1	versicolor
## 81	5.5	2.4	3.8	1.1	versicolor
## 70	5.6	2.5	3.9	1.1	versicolor
## 83	5.8	2.7	3.9	1.2	versicolor
## 93	5.8	2.6	4.0	1.2	versicolor
## 96	5.7	3.0	4.2	1.2	versicolor
## 91	5.5	2.6	4.4	1.2	versicolor
## 74	6.1	2.8	4.7	1.2	versicolor
## 65	5.6	2.9	3.6	1.3	versicolor
## 54	5.5	2.3	4.0	1.3	versicolor

云
湘
黄
①

## 72	6.1	2.8	4.0	1.3 versicolor
## 90	5.5	2.5	4.0	1.3 versicolor
## 89	5.6	3.0	4.1	1.3 versicolor
## 100	5.7	2.8	4.1	1.3 versicolor
## 95	5.6	2.7	4.2	1.3 versicolor
## 97	5.7	2.9	4.2	1.3 versicolor
## 75	6.4	2.9	4.3	1.3 versicolor
## 98	6.2	2.9	4.3	1.3 versicolor
## 88	6.3	2.3	4.4	1.3 versicolor
## 56	5.7	2.8	4.5	1.3 versicolor
## 59	6.6	2.9	4.6	1.3 versicolor
## 60	5.2	2.7	3.9	1.4 versicolor
## 66	6.7	3.1	4.4	1.4 versicolor
## 76	6.6	3.0	4.4	1.4 versicolor
## 92	6.1	3.0	4.6	1.4 versicolor
## 51	7.0	3.2	4.7	1.4 versicolor
## 64	6.1	2.9	4.7	1.4 versicolor
## 77	6.8	2.8	4.8	1.4 versicolor
## 135	6.1	2.6	5.6	1.4 virginica
## 62	5.9	3.0	4.2	1.5 versicolor
## 52	6.4	3.2	4.5	1.5 versicolor
## 67	5.6	3.0	4.5	1.5 versicolor
## 69	6.2	2.2	4.5	1.5 versicolor
## 79	6.0	2.9	4.5	1.5 versicolor
## 85	5.4	3.0	4.5	1.5 versicolor
## 55	6.5	2.8	4.6	1.5 versicolor
## 87	6.7	3.1	4.7	1.5 versicolor
## 53	6.9	3.1	4.9	1.5 versicolor
## 73	6.3	2.5	4.9	1.5 versicolor
## 120	6.0	2.2	5.0	1.5 virginica
## 134	6.3	2.8	5.1	1.5 virginica
## 86	6.0	3.4	4.5	1.6 versicolor
## 57	6.3	3.3	4.7	1.6 versicolor
## 84	6.0	2.7	5.1	1.6 versicolor
## 130	7.2	3.0	5.8	1.6 virginica
## 107	4.9	2.5	4.5	1.7 virginica

## 78	6.7	3.0	5.0	1.7 versicolor
## 71	5.9	3.2	4.8	1.8 versicolor
## 127	6.2	2.8	4.8	1.8 virginica
## 139	6.0	3.0	4.8	1.8 virginica
## 124	6.3	2.7	4.9	1.8 virginica
## 128	6.1	3.0	4.9	1.8 virginica
## 150	5.9	3.0	5.1	1.8 virginica
## 117	6.5	3.0	5.5	1.8 virginica
## 138	6.4	3.1	5.5	1.8 virginica
## 104	6.3	2.9	5.6	1.8 virginica
## 109	6.7	2.5	5.8	1.8 virginica
## 126	7.2	3.2	6.0	1.8 virginica
## 108	7.3	2.9	6.3	1.8 virginica
## 147	6.3	2.5	5.0	1.9 virginica
## 102	5.8	2.7	5.1	1.9 virginica
## 143	5.8	2.7	5.1	1.9 virginica
## 112	6.4	2.7	5.3	1.9 virginica
## 131	7.4	2.8	6.1	1.9 virginica
## 122	5.6	2.8	4.9	2.0 virginica
## 114	5.7	2.5	5.0	2.0 virginica
## 111	6.5	3.2	5.1	2.0 virginica
## 148	6.5	3.0	5.2	2.0 virginica
## 132	7.9	3.8	6.4	2.0 virginica
## 123	7.7	2.8	6.7	2.0 virginica
## 140	6.9	3.1	5.4	2.1 virginica
## 113	6.8	3.0	5.5	2.1 virginica
## 129	6.4	2.8	5.6	2.1 virginica
## 125	6.7	3.3	5.7	2.1 virginica
## 103	7.1	3.0	5.9	2.1 virginica
## 106	7.6	3.0	6.6	2.1 virginica
## 133	6.4	2.8	5.6	2.2 virginica
## 105	6.5	3.0	5.8	2.2 virginica
## 118	7.7	3.8	6.7	2.2 virginica
## 142	6.9	3.1	5.1	2.3 virginica
## 146	6.7	3.0	5.2	2.3 virginica
## 116	6.4	3.2	5.3	2.3 virginica

```
## 149      6.2      3.4      5.4      2.3  virginica
## 121      6.9      3.2      5.7      2.3  virginica
## 144      6.8      3.2      5.9      2.3  virginica
## 136      7.7      3.0      6.1      2.3  virginica
## 119      7.7      2.6      6.9      2.3  virginica
## 115      5.8      2.8      5.1      2.4  virginica
## 137      6.3      3.4      5.6      2.4  virginica
## 141      6.7      3.1      5.6      2.4  virginica
## 145      6.7      3.3      5.7      2.5  virginica
## 101      6.3      3.3      6.0      2.5  virginica
## 110      7.2      3.6      6.1      2.5  virginica
```

sort/ordered 排序， 默认是升序

```
dd <- data.frame(
  b = factor(c("Hi", "Med", "Hi", "Low"),
  levels = c("Low", "Med", "Hi"), ordered = TRUE
),
  x = c("A", "D", "A", "C"), y = c(8, 3, 9, 9),
  z = c(1, 1, 1, 2)
)
str(dd)
```

```
## 'data.frame': 4 obs. of 4 variables:
## $ b: Ord.factor w/ 3 levels "Low" < "Med" < "Hi": 3 2 3 1
## $ x: chr "A" "D" "A" "C"
## $ y: num 8 3 9 9
## $ z: num 1 1 1 2
```

```
dd[order(-dd[,4], dd[,1]), ]
```

```
##      b x y z
## 4 Low C 9 2
## 2 Med D 3 1
## 1 Hi A 8 1
## 3 Hi A 9 1
```

根据变量 z

```
dd[order(dd$z, dd$b), ]  
  
##      b x y z  
## 2 Med D 3 1  
## 1 Hi A 8 1  
## 3 Hi A 9 1  
## 4 Low C 9 2
```

5.6 数据拆分

数据拆分通常是按找某一个分类变量分组，分完组就是计算，计算完就把结果按照原来的分组方式合并

```
## Notice that assignment form is not used since a variable is being added  
g <- airquality$Month  
l <- split(airquality, g) # 分组  
l <- lapply(l, transform, Oz.Z = scale(Ozone)) # 计算：按月对 Ozone 标准化  
aq2 <- unsplit(l, g) # 合并  
head(aq2)
```

```
##   Ozone Solar.R Wind Temp Month Day       Oz.Z  
## 1     41      190  7.4   67      5   1  0.7822293  
## 2     36      118  8.0   72      5   2  0.5572518  
## 3     12      149 12.6   74      5   3 -0.5226399  
## 4     18      313 11.5   62      5   4 -0.2526670  
## 5     NA      NA 14.3   56      5   5      NA  
## 6     28      NA 14.9   66      5   6  0.1972879
```

tapply 自带分组的功能，按月份 Month 对 Ozone 中心标准化，其返回一个列表
with(airquality, tapply(Ozone, Month, scale))

```
## $`5`  
## [1,]  
## [1,]  0.78222929  
## [2,]  0.55725184  
## [3,] -0.52263993  
## [4,] -0.25266698
```

云
湘
黄
⑩

```
## [5,]          NA
## [6,]  0.19728792
## [7,] -0.02768953
## [8,] -0.20767149
## [9,] -0.70262189
## [10,]         NA
## [11,] -0.74761738
## [12,] -0.34265796
## [13,] -0.56763542
## [14,] -0.43264895
## [15,] -0.25266698
## [16,] -0.43264895
## [17,]  0.46726086
## [18,] -0.79261287
## [19,]  0.28727890
## [20,] -0.56763542
## [21,] -1.01759032
## [22,] -0.56763542
## [23,] -0.88260385
## [24,]  0.37726988
## [25,]          NA
## [26,]          NA
## [27,]          NA
## [28,] -0.02768953
## [29,]  0.96221125
## [30,]  4.11189557
## [31,]  0.60224733
## attr(),"scaled:center")
## [1] 23.61538
## attr(),"scaled:scale")
## [1] 22.22445
##
## $`6`
## [,1]
## [1,]          NA
## [2,]          NA
```

```
## [3,]      NA
## [4,]      NA
## [5,]      NA
## [6,]      NA
## [7,] -0.02440942
## [8,]      NA
## [9,]  2.28228109
## [10,]  0.52480260
## [11,]      NA
## [12,]      NA
## [13,] -0.35393664
## [14,]      NA
## [15,]      NA
## [16,] -0.46377904
## [17,]  0.41496020
## [18,] -0.51870025
## [19,] -0.95806987
## [20,] -0.90314867
## [21,]      NA
## [22,]      NA
## [23,]      NA
## [24,]      NA
## [25,]      NA
## [26,]      NA
## [27,]      NA
## [28,]      NA
## [29,]      NA
## [30,]      NA
## attr(),"scaled:center")
## [1] 29.44444
## attr(),"scaled:scale")
## [1] 18.2079
##
## 
## $`7`
##      [,1]
## [1,] 2.398691600
```

云
湘
黄
④

```
## [2,] -0.319744496
## [3,] -0.857109771
## [4,] NA
## [5,] 0.154401335
## [6,] -0.604231995
## [7,] 0.565327721
## [8,] 1.197522162
## [9,] 1.197522162
## [10,] 0.818205498
## [11,] NA
## [12,] -1.552523656
## [13,] -1.015158381
## [14,] NA
## [15,] -1.647352822
## [16,] -0.351354218
## [17,] -0.762280605
## [18,] 0.059572168
## [19,] 0.628547165
## [20,] 0.122791613
## [21,] -1.362865324
## [22,] NA
## [23,] NA
## [24,] 0.660156887
## [25,] 1.545229105
## [26,] -1.236426436
## [27,] -0.224915330
## [28,] 0.723376332
## [29,] -0.288134774
## [30,] 0.154401335
## [31,] -0.003647276
## attr(),"scaled:center")
## [1] 59.11538
## attr(),"scaled:scale")
## [1] 31.63584
##
## $`8`
```

```
## [1] -0.52824846
## [2] -1.28427379
## [3] -1.10786788
## [4]  0.45458446
## [5] -0.62905184
## [6]  0.15217433
## [7]  1.56342160
## [8]  0.73179374
## [9]  1.26101147
## [10] NA
## [11] NA
## [12] -0.40224424
## [13] -0.80545775
## [14]  0.12697348
## [15] NA
## [16] -0.95666281
## [17] -0.02423158
## [18] -0.93146197
## [19] -0.72985522
## [20] -0.40224424
## [21] -0.98186366
## [22] -1.28427379
## [23] NA
## [24] -0.37704340
## [25]  2.72266043
## [26]  0.32858024
## [27] NA
## [28]  0.40418277
## [29]  1.46261822
## [30]  0.60578952
## [31]  0.63099037
## attr(,"scaled:center")
## [1] 59.96154
## attr(,"scaled:scale")
## [1] 39.68121
```

云
湘
黄
④

```
##  
## $`9`  
## [,1]  
## [1,] 2.67385466  
## [2,] 1.92826057  
## [3,] 1.72115110  
## [4,] 2.46674519  
## [5,] 0.64418186  
## [6,] 0.02285346  
## [7,] -0.47420927  
## [8,] -0.34994359  
## [9,] -0.43278737  
## [10,] -0.30852169  
## [11,] 0.51991618  
## [12,] -0.43278737  
## [13,] -0.14283412  
## [14,] -0.92985010  
## [15,] -0.76416252  
## [16,] 0.60275997  
## [17,] -0.55705305  
## [18,] -0.76416252  
## [19,] -0.30852169  
## [20,] -0.63989684  
## [21,] -0.76416252  
## [22,] -0.34994359  
## [23,] 0.18854103  
## [24,] -1.01269388  
## [25,] -0.72274063  
## [26,] -0.05999033  
## [27,] NA  
## [28,] -0.72274063  
## [29,] -0.55705305  
## [30,] -0.47420927  
## attr(),"scaled:center")  
## [1] 31.44828  
## attr(),"scaled:scale")
```

```
## [1] 24.14182
```

上面的过程等价于

```
do.call("rbind", lapply(split(airquality, airquality$Month), transform, Oz.Z = sca
```

	Ozone	Solar.R	Wind	Temp	Month	Day	Oz.Z
## 5.1	41	190	7.4	67	5	1	0.782229293
## 5.2	36	118	8.0	72	5	2	0.557251841
## 5.3	12	149	12.6	74	5	3	-0.522639926
## 5.4	18	313	11.5	62	5	4	-0.252666984
## 5.5	NA	NA	14.3	56	5	5	NA
## 5.6	28	NA	14.9	66	5	6	0.197287919
## 5.7	23	299	8.6	65	5	7	-0.027689532
## 5.8	19	99	13.8	59	5	8	-0.207671494
## 5.9	8	19	20.1	61	5	9	-0.702621887
## 5.10	NA	194	8.6	69	5	10	NA
## 5.11	7	NA	6.9	74	5	11	-0.747617377
## 5.12	16	256	9.7	69	5	12	-0.342657965
## 5.13	11	290	9.2	66	5	13	-0.567635416
## 5.14	14	274	10.9	68	5	14	-0.432648945
## 5.15	18	65	13.2	58	5	15	-0.252666984
## 5.16	14	334	11.5	64	5	16	-0.432648945
## 5.17	34	307	12.0	66	5	17	0.467260861
## 5.18	6	78	18.4	57	5	18	-0.792612867
## 5.19	30	322	11.5	68	5	19	0.287278900
## 5.20	11	44	9.7	62	5	20	-0.567635416
## 5.21	1	8	9.7	59	5	21	-1.017590319
## 5.22	11	320	16.6	73	5	22	-0.567635416
## 5.23	4	25	9.7	61	5	23	-0.882603848
## 5.24	32	92	12.0	61	5	24	0.377269880
## 5.25	NA	66	16.6	57	5	25	NA
## 5.26	NA	266	14.9	58	5	26	NA
## 5.27	NA	NA	8.0	57	5	27	NA
## 5.28	23	13	12.0	67	5	28	-0.027689532
## 5.29	45	252	14.9	81	5	29	0.962211254
## 5.30	115	223	5.7	79	5	30	4.111895575
## 5.31	37	279	7.4	76	5	31	0.602247332

云
湘
黄
①

## 6.32	NA	286	8.6	78	6	1	NA
## 6.33	NA	287	9.7	74	6	2	NA
## 6.34	NA	242	16.1	67	6	3	NA
## 6.35	NA	186	9.2	84	6	4	NA
## 6.36	NA	220	8.6	85	6	5	NA
## 6.37	NA	264	14.3	79	6	6	NA
## 6.38	29	127	9.7	82	6	7	-0.024409423
## 6.39	NA	273	6.9	87	6	8	NA
## 6.40	71	291	13.8	90	6	9	2.282281088
## 6.41	39	323	11.5	87	6	10	0.524802603
## 6.42	NA	259	10.9	93	6	11	NA
## 6.43	NA	250	9.2	92	6	12	NA
## 6.44	23	148	8.0	82	6	13	-0.353936639
## 6.45	NA	332	13.8	80	6	14	NA
## 6.46	NA	322	11.5	79	6	15	NA
## 6.47	21	191	14.9	77	6	16	-0.463779045
## 6.48	37	284	20.7	72	6	17	0.414960198
## 6.49	20	37	9.2	65	6	18	-0.518700247
## 6.50	12	120	11.5	73	6	19	-0.958069868
## 6.51	13	137	10.3	76	6	20	-0.903148666
## 6.52	NA	150	6.3	77	6	21	NA
## 6.53	NA	59	1.7	76	6	22	NA
## 6.54	NA	91	4.6	76	6	23	NA
## 6.55	NA	250	6.3	76	6	24	NA
## 6.56	NA	135	8.0	75	6	25	NA
## 6.57	NA	127	8.0	78	6	26	NA
## 6.58	NA	47	10.3	73	6	27	NA
## 6.59	NA	98	11.5	80	6	28	NA
## 6.60	NA	31	14.9	77	6	29	NA
## 6.61	NA	138	8.0	83	6	30	NA
## 7.62	135	269	4.1	84	7	1	2.398691600
## 7.63	49	248	9.2	85	7	2	-0.319744496
## 7.64	32	236	9.2	81	7	3	-0.857109771
## 7.65	NA	101	10.9	84	7	4	NA
## 7.66	64	175	4.6	83	7	5	0.154401335
## 7.67	40	314	10.9	83	7	6	-0.604231995

## 7.68	77	276	5.1	88	7	7	0.565327721
## 7.69	97	267	6.3	92	7	8	1.197522162
## 7.70	97	272	5.7	92	7	9	1.197522162
## 7.71	85	175	7.4	89	7	10	0.818205498
## 7.72	NA	139	8.6	82	7	11	NA
## 7.73	10	264	14.3	73	7	12	-1.552523656
## 7.74	27	175	14.9	81	7	13	-1.015158381
## 7.75	NA	291	14.9	91	7	14	NA
## 7.76	7	48	14.3	80	7	15	-1.647352822
## 7.77	48	260	6.9	81	7	16	-0.351354218
## 7.78	35	274	10.3	82	7	17	-0.762280605
## 7.79	61	285	6.3	84	7	18	0.059572168
## 7.80	79	187	5.1	87	7	19	0.628547165
## 7.81	63	220	11.5	85	7	20	0.122791613
## 7.82	16	7	6.9	74	7	21	-1.362865324
## 7.83	NA	258	9.7	81	7	22	NA
## 7.84	NA	295	11.5	82	7	23	NA
## 7.85	80	294	8.6	86	7	24	0.660156887
## 7.86	108	223	8.0	85	7	25	1.545229105
## 7.87	20	81	8.6	82	7	26	-1.236426436
## 7.88	52	82	12.0	86	7	27	-0.224915330
## 7.89	82	213	7.4	88	7	28	0.723376332
## 7.90	50	275	7.4	86	7	29	-0.288134774
## 7.91	64	253	7.4	83	7	30	0.154401335
## 7.92	59	254	9.2	81	7	31	-0.003647276
## 8.93	39	83	6.9	81	8	1	-0.528248464
## 8.94	9	24	13.8	81	8	2	-1.284273789
## 8.95	16	77	7.4	82	8	3	-1.107867880
## 8.96	78	NA	6.9	86	8	4	0.454584458
## 8.97	35	NA	7.4	85	8	5	-0.629051841
## 8.98	66	NA	4.6	87	8	6	0.152174328
## 8.99	122	255	4.0	89	8	7	1.563421601
## 8.100	89	229	10.3	90	8	8	0.731793744
## 8.101	110	207	8.0	90	8	9	1.261011471
## 8.102	NA	222	8.6	92	8	10	NA
## 8.103	NA	137	11.5	86	8	11	NA

云
湘
黄
◎

## 8.104	44	192	11.5	86	8	12	-0.402244243
## 8.105	28	273	11.5	82	8	13	-0.805457750
## 8.106	65	157	9.7	80	8	14	0.126973484
## 8.107	NA	64	11.5	79	8	15	NA
## 8.108	22	71	10.3	77	8	16	-0.956662815
## 8.109	59	51	6.3	79	8	17	-0.024231581
## 8.110	23	115	7.4	76	8	18	-0.931461970
## 8.111	31	244	10.9	78	8	19	-0.729855217
## 8.112	44	190	10.3	78	8	20	-0.402244243
## 8.113	21	259	15.5	77	8	21	-0.981863659
## 8.114	9	36	14.3	72	8	22	-1.284273789
## 8.115	NA	255	12.6	75	8	23	NA
## 8.116	45	212	9.7	79	8	24	-0.377043399
## 8.117	168	238	3.4	81	8	25	2.722660432
## 8.118	73	215	8.0	86	8	26	0.328580237
## 8.119	NA	153	5.7	88	8	27	NA
## 8.120	76	203	9.7	97	8	28	0.404182770
## 8.121	118	225	2.3	94	8	29	1.462618224
## 8.122	84	237	6.3	96	8	30	0.605789523
## 8.123	85	188	6.3	94	8	31	0.630990367
## 9.124	96	167	6.9	91	9	1	2.673854658
## 9.125	78	197	5.1	92	9	2	1.928260571
## 9.126	73	183	2.8	93	9	3	1.721151102
## 9.127	91	189	4.6	93	9	4	2.466745189
## 9.128	47	95	7.4	87	9	5	0.644181865
## 9.129	32	92	15.5	84	9	6	0.022853459
## 9.130	20	252	10.9	80	9	7	-0.474209266
## 9.131	23	220	10.3	78	9	8	-0.349943585
## 9.132	21	230	10.9	75	9	9	-0.432787373
## 9.133	24	259	9.7	73	9	10	-0.308521691
## 9.134	44	236	14.9	81	9	11	0.519916184
## 9.135	21	259	15.5	76	9	12	-0.432787373
## 9.136	28	238	6.3	77	9	13	-0.142834116
## 9.137	9	24	10.9	71	9	14	-0.929850097
## 9.138	13	112	11.5	71	9	15	-0.764162523
## 9.139	46	237	6.9	78	9	16	0.602759971



```
## 9.140   18    224 13.8   67    9   17 -0.557053054
## 9.141   13    27 10.3   76    9   18 -0.764162523
## 9.142   24    238 10.3   68    9   19 -0.308521691
## 9.143   16    201 8.0    82    9   20 -0.639896841
## 9.144   13    238 12.6   64    9   21 -0.764162523
## 9.145   23    14 9.2    71    9   22 -0.349943585
## 9.146   36    139 10.3   81    9   23  0.188541034
## 9.147   7     49 10.3   69    9   24 -1.012693885
## 9.148   14    20 16.6   63    9   25 -0.722740629
## 9.149   30    193 6.9    70    9   26 -0.059990329
## 9.150   NA    145 13.2   77    9   27          NA
## 9.151   14    191 14.3   75    9   28 -0.722740629
## 9.152   18    131 8.0    76    9   29 -0.557053054
## 9.153   20    223 11.5   68    9   30 -0.474209266
```

由于上面对 Ozone 正态标准化，所以标准化后的 oz.z 再按月分组计算方差自然每个月都是 1，而均值都是 0。

```
with(aq2, tapply(Oz.Z, Month, sd, na.rm = TRUE))
```

```
## 5 6 7 8 9
## 1 1 1 1 1
```

```
with(aq2, tapply(Oz.Z, Month, mean, na.rm = TRUE))
```

```
##           5           6           7           8           9
## -4.240273e-17 1.052760e-16 5.841432e-17 5.898060e-17 2.571709e-17
```

循着这个思路，我们可以用 tapply 实现分组计算，上面函数 sd 和 mean 完全可以用自定义的更加复杂的函数替代

cut 函数可以将连续型变量划分为分类变量

```
set.seed(2019)
Z <- stats::rnorm(10)
cut(Z, breaks = -6:6)
```

```
## [1] (0,1]  (-1,0]  (-2,-1] (0,1]  (-2,-1] (0,1]  (-1,0]  (0,1]  (-2,-1]
## [10] (-1,0]
## 12 Levels: (-6,-5] (-5,-4] (-4,-3] (-3,-2] (-2,-1] (-1,0] (0,1] (1,2] ... (5,6]
```

```
# labels = FALSE 返回每个数所落的区间位置  
cut(Z, breaks = -6:6, labels = FALSE)
```

```
## [1] 7 6 5 7 5 7 6 7 5 6
```

我们还可以指定参数 `dig.lab` 设置分组的精度，`ordered` 将分组变量看作是有序的，`breaks` 传递单个数时，表示分组数，而不是断点

```
cut(Z, breaks = 3, dig.lab = 4, ordered = TRUE)
```

```
## [1] (0.06396,0.9186] (-0.7881,0.06396] (-1.643,-0.7881] (0.06396,0.9186]  
## [5] (-1.643,-0.7881] (0.06396,0.9186] (-0.7881,0.06396] (0.06396,0.9186]  
## [9] (-1.643,-0.7881] (-0.7881,0.06396]  
## Levels: (-1.643,-0.7881] < (-0.7881,0.06396] < (0.06396,0.9186]
```

此时，统计每组的频数，如图 ??

```
# 条形图  
plot(cut(Z, breaks = -6:6))
```

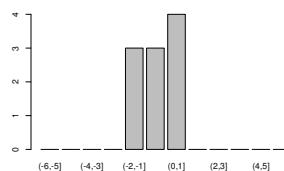


图 5.2: 连续型变量分组统计

```
# 直方图  
hist(Z, breaks = -6:6)
```

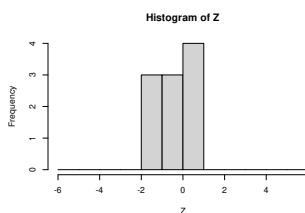


图 5.3: 连续型变量分组统计

在指定分组数的情况下，我们还想获取分组的断点

```
labs <- levels(cut(z, 3))
labs
```

```
## [1] "(-1.64,-0.788]" "(-0.788,0.064]" "(0.064,0.919]"
```

用正则表达式抽取断点

```
cbind(
  lower = as.numeric(sub("\\\\((.+),.*", "\\1", labs)),
  upper = as.numeric(sub("[^,]*,([^,]*)\\\\]", "\\1", labs)))
)
```

```
##      lower   upper
## [1,] -1.640 -0.788
## [2,] -0.788  0.064
## [3,]  0.064  0.919
```

更多相关函数可以参考 `findInterval` 和 `embed`

`tabulate` 和 `table` 有所不同，它表示排列

```
t(combn(8, 4, tabulate, nbins = 8))
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
## [1,]    1    1    1    1    0    0    0    0
## [2,]    1    1    1    0    1    0    0    0
## [3,]    1    1    1    0    0    0    1    0
## [4,]    1    1    1    0    0    0    0    1
## [5,]    1    1    1    0    0    0    0    1
## [6,]    1    1    0    1    1    0    0    0
## [7,]    1    1    0    1    0    1    0    0
## [8,]    1    1    0    1    0    0    1    0
## [9,]    1    1    0    1    0    0    0    1
## [10,]   1    1    0    0    1    1    0    0
## [11,]   1    1    0    0    1    0    1    0
## [12,]   1    1    0    0    1    0    0    1
## [13,]   1    1    0    0    0    1    1    0
## [14,]   1    1    0    0    0    1    0    1
## [15,]   1    1    0    0    0    0    1    1
## [16,]   1    0    1    1    1    0    0    0
## [17,]   1    0    1    1    0    1    0    0
```

云
湘
黄
◎

```
## [18,] 1 0 1 1 0 0 1 0
## [19,] 1 0 1 1 0 0 0 1
## [20,] 1 0 1 0 1 1 0 0
## [21,] 1 0 1 0 1 0 1 0
## [22,] 1 0 1 0 1 0 0 1
## [23,] 1 0 1 0 0 1 1 0
## [24,] 1 0 1 0 0 1 0 1
## [25,] 1 0 1 0 0 0 1 1
## [26,] 1 0 0 1 1 1 0 0
## [27,] 1 0 0 1 1 0 1 0
## [28,] 1 0 0 1 1 0 0 1
## [29,] 1 0 0 1 0 1 1 0
## [30,] 1 0 0 1 0 1 0 1
## [31,] 1 0 0 1 0 0 1 1
## [32,] 1 0 0 0 1 1 1 0
## [33,] 1 0 0 0 1 1 0 1
## [34,] 1 0 0 0 1 0 1 1
## [35,] 1 0 0 0 0 1 1 1
## [36,] 0 1 1 1 1 0 0 0
## [37,] 0 1 1 1 0 1 0 0
## [38,] 0 1 1 1 0 0 1 0
## [39,] 0 1 1 1 0 0 0 1
## [40,] 0 1 1 0 1 1 0 0
## [41,] 0 1 1 0 1 0 1 0
## [42,] 0 1 1 0 1 0 0 1
## [43,] 0 1 1 0 0 1 1 0
## [44,] 0 1 1 0 0 1 0 1
## [45,] 0 1 1 0 0 0 1 1
## [46,] 0 1 0 1 1 1 0 0
## [47,] 0 1 0 1 1 0 1 0
## [48,] 0 1 0 1 1 0 0 1
## [49,] 0 1 0 1 0 1 1 0
## [50,] 0 1 0 1 0 1 0 1
## [51,] 0 1 0 1 0 0 1 1
## [52,] 0 1 0 0 1 1 1 0
## [53,] 0 1 0 0 1 1 0 1
```



```
## [54,] 0 1 0 0 1 0 1 1 1
## [55,] 0 1 0 0 0 1 1 1 1
## [56,] 0 0 1 1 1 1 0 0 0
## [57,] 0 0 1 1 1 0 1 0 0
## [58,] 0 0 1 1 1 0 0 0 1
## [59,] 0 0 1 1 0 1 1 1 0
## [60,] 0 0 1 1 0 1 0 0 1
## [61,] 0 0 1 1 0 0 0 1 1
## [62,] 0 0 1 0 1 1 1 1 0
## [63,] 0 0 1 0 1 1 0 0 1
## [64,] 0 0 1 0 1 0 1 1 1
## [65,] 0 0 1 0 0 0 1 1 1
## [66,] 0 0 0 1 1 1 1 1 0
## [67,] 0 0 0 1 1 1 0 0 1
## [68,] 0 0 0 1 1 0 1 1 1
## [69,] 0 0 0 1 0 1 1 1 1
## [70,] 0 0 0 0 1 1 1 1 1
```

5.7 数据合并

merge 合并两个数据框

```
authors <- data.frame(
  ## I(*) : use character columns of names to get sensible sort order
  surname = I(c("Tukey", "Venables", "Tierney", "Ripley", "McNeil")),
  nationality = c("US", "Australia", "US", "UK", "Australia"),
  deceased = c("yes", rep("no", 4))
)
authorN <- within(authors, {
  name <- surname
  rm(surname)
})
books <- data.frame(
  name = I(c(
    "Tukey", "Venables", "Tierney",
    "Ripley", "Ripley", "McNeil", "R Core"
  ))
)
```

黄湘云

```
)),
  title = c(
    "Exploratory Data Analysis",
    "Modern Applied Statistics ...",
    "LISP-STAT",
    "Spatial Statistics", "Stochastic Simulation",
    "Interactive Data Analysis",
    "An Introduction to R"
),
  other.author = c(
    NA, "Ripley", NA, NA, NA, NA,
    "Venables & Smith"
)
)
```

authors

```
##   surname nationality deceased
## 1 Tukey          US      yes
## 2 Venables       Australia    no
## 3 Tierney        US      no
## 4 Ripley         UK      no
## 5 McNeil         Australia    no
```

authorN

```
##   nationality deceased     name
## 1          US      yes    Tukey
## 2    Australia      no Venables
## 3          US      no  Tierney
## 4          UK      no  Ripley
## 5    Australia      no  McNeil
```

books

```
##      name                  title  other.author
## 1    Tukey  Exploratory Data Analysis      <NA>
## 2  Venables Modern Applied Statistics ...      Ripley
## 3  Tierney            LISP-STAT      <NA>
```



```
## 4 Ripley           Spatial Statistics      <NA>
## 5 Ripley           Stochastic Simulation <NA>
## 6 McNeil          Interactive Data Analysis <NA>
## 7 R Core           An Introduction to R Venables & Smith
```

默认找到同名的列，然后是同名的行合并，多余的没有匹配到的就丢掉

```
merge(authorsN, books)
```

```
##   name nationality deceased           title other.author
## 1 McNeil    Australia       no Interactive Data Analysis <NA>
## 2 Ripley     UK           no      Spatial Statistics <NA>
## 3 Ripley     UK           no      Stochastic Simulation <NA>
## 4 Tierney    US           no        LISP-STAT <NA>
## 5 Tukey      US           yes Exploratory Data Analysis <NA>
## 6 Venables   Australia     no Modern Applied Statistics ... Ripley
```

还可以指定合并的列，先按照 `surname` 合并，留下 `surname`

```
merge(authors, books, by.x = "surname", by.y = "name")
```

```
##   surname nationality deceased           title other.author
## 1 McNeil    Australia       no Interactive Data Analysis <NA>
## 2 Ripley     UK           no      Spatial Statistics <NA>
## 3 Ripley     UK           no      Stochastic Simulation <NA>
## 4 Tierney    US           no        LISP-STAT <NA>
## 5 Tukey      US           yes Exploratory Data Analysis <NA>
## 6 Venables   Australia     no Modern Applied Statistics ... Ripley
```

留下的是 `name`

```
merge(books, authors, by.x = "name", by.y = "surname")
```

```
##   name           title other.author nationality deceased
## 1 McNeil Interactive Data Analysis <NA>    Australia   no
## 2 Ripley    Spatial Statistics <NA>        UK       no
## 3 Ripley    Stochastic Simulation <NA>        UK       no
## 4 Tierney    LISP-STAT <NA>        US       no
## 5 Tukey     Exploratory Data Analysis <NA>        US      yes
## 6 Venables  Modern Applied Statistics ... Ripley  Australia   no
```

为了比较清楚地观察几种合并的区别，这里提供对应的动画展示 <https://github.com>.



com/gadenbuie/tidyexplain

(inner, outer, left, right, cross) join 共 5 种合并方式详情请看 <https://stackoverflow.com/questions/1299871>

cbind 和 rbind 分别是按列和行合并数据框

5.8 数据去重

单个数值型向量去重，此时和 unique 函数作用一样

```
(x <- c(9:20, 1:5, 3:7, 0:8))

## [1]  9 10 11 12 13 14 15 16 17 18 19 20  1  2  3  4  5  3  4  5  6  7  0  1  2
## [26]  3  4  5  6  7  8

## extract unique elements
x[!duplicated(x)]
```

```
## [1]  9 10 11 12 13 14 15 16 17 18 19 20  1  2  3  4  5  6  7  0  8

unique(x)
```

```
## [1]  9 10 11 12 13 14 15 16 17 18 19 20  1  2  3  4  5  6  7  0  8
```

数据框类型数据中，去除重复的行，这个重复可以是多个变量对应的向量

```
set.seed(123)
df <- data.frame(
  x = sample(0:1, 10, replace = T),
  y = sample(0:1, 10, replace = T),
  z = 1:10
)
df
```

```
##     x y z
## 1  0 1 1
## 2  0 1 2
## 3  0 1 3
## 4  1 0 4
## 5  0 1 5
## 6  1 0 6
```



```
## 7 1 1 7  
## 8 1 0 8  
## 9 0 0 9  
## 10 0 0 10  
df[!duplicated(df[, 1:2]), ]  
  
## x y z  
## 1 0 1 1  
## 4 1 0 4  
## 7 1 1 7  
## 9 0 0 9
```

5.9 数据缺失

缺失数据操作

```
data("airquality")  
head(airquality)
```

```
## Ozone Solar.R Wind Temp Month Day  
## 1 41 190 7.4 67 5 1  
## 2 36 118 8.0 72 5 2  
## 3 12 149 12.6 74 5 3  
## 4 18 313 11.5 62 5 4  
## 5 NA NA 14.3 56 5 5  
## 6 28 NA 14.9 66 5 6
```

对缺失值的处理默认是 `na.action = na.omit`

```
# Ozone 最高的那天  
aggregate(data = airquality, Ozone ~ Month, max)  
  
## Month Ozone  
## 1 5 115  
## 2 6 71  
## 3 7 135  
## 4 8 168  
## 5 9 96
```



```
# 每月 Ozone, Solar.R, Wind, Temp 平均值
aggregate(data = airquality, Ozone ~ Month, mean)
```

```
##   Month   Ozone
## 1      5 23.61538
## 2      6 29.44444
## 3      7 59.11538
## 4      8 59.96154
## 5      9 31.44828
```

缺失值处理

```
library(DataExplorer)
plot_missing(airquality)
```

查看包含缺失的记录，不完整的记录

```
airquality[!complete.cases(airquality), ]
```

```
##       Ozone Solar.R Wind Temp Month Day
## 5       NA      NA 14.3  56     5    5
## 6      28      NA 14.9  66     5    6
## 10     NA     194  8.6  69     5   10
## 11     7       NA  6.9  74     5   11
## 25     NA      66 16.6  57     5   25
## 26     NA     266 14.9  58     5   26
## 27     NA      NA  8.0  57     5   27
## 32     NA     286  8.6  78     6    1
## 33     NA     287  9.7  74     6    2
## 34     NA     242 16.1  67     6    3
## 35     NA     186  9.2  84     6    4
## 36     NA     220  8.6  85     6    5
## 37     NA     264 14.3  79     6    6
## 39     NA     273  6.9  87     6    8
## 42     NA     259 10.9  93     6   11
## 43     NA     250  9.2  92     6   12
## 45     NA     332 13.8  80     6   14
## 46     NA     322 11.5  79     6   15
## 52     NA     150  6.3  77     6   21
```



```
## 53     NA      59  1.7   76    6  22
## 54     NA      91  4.6   76    6  23
## 55     NA     250  6.3   76    6  24
## 56     NA     135  8.0   75    6  25
## 57     NA     127  8.0   78    6  26
## 58     NA     47 10.3   73    6  27
## 59     NA     98 11.5   80    6  28
## 60     NA     31 14.9   77    6  29
## 61     NA     138  8.0   83    6  30
## 65     NA     101 10.9   84    7  4
## 72     NA     139  8.6   82    7  11
## 75     NA     291 14.9   91    7  14
## 83     NA     258  9.7   81    7  22
## 84     NA     295 11.5   82    7  23
## 96      78     NA  6.9   86    8  4
## 97      35     NA  7.4   85    8  5
## 98      66     NA  4.6   87    8  6
## 102     NA     222  8.6   92    8  10
## 103     NA     137 11.5   86    8  11
## 107     NA     64 11.5   79    8  15
## 115     NA     255 12.6   75    8  23
## 119     NA     153  5.7   88    8  27
## 150     NA     145 13.2   77    9  27
```

Ozone 和 Solar.R 同时包含缺失值的行

```
airquality[is.na(airquality$Ozone) & is.na(airquality$Solar.R), ]
```

```
##   Ozone Solar.R Wind Temp Month Day
## 5     NA      NA 14.3   56      5   5
## 27    NA      NA  8.0   57      5  27
```

5.10 数据聚合

分组求和 <https://stackoverflow.com/questions/1660124>

主要是分组统计



```

apropos("apply")

## [1] "apply"      "dendrapply" "eapply"      "kernapply"   "lapply"
## [6] "mapply"     "rapply"     "sapply"      "tapply"      "vapply"

# 分组求和 colSums colMeans max
unique(iris$Species)

## [1] setosa      versicolor virginica
## Levels: setosa versicolor virginica

# 分类求和
# colSums(iris[iris$Species == "setosa", -5])
# colSums(iris[iris$Species == "virginica", -5])
colSums(iris[iris$Species == "versicolor", -5])

## Sepal.Length Sepal.Width Petal.Length Petal.Width
##          296.8        138.5       213.0        66.3

# apply(iris[iris$Species == "setosa", -5], 2, sum)
# apply(iris[iris$Species == "setosa", -5], 2, mean)
# apply(iris[iris$Species == "setosa", -5], 2, min)
# apply(iris[iris$Species == "setosa", -5], 2, max)
apply(iris[iris$Species == "setosa", -5], 2, quantile)

##      Sepal.Length Sepal.Width Petal.Length Petal.Width
## 0%            4.3       2.300      1.000       0.1
## 25%           4.8       3.200      1.400       0.2
## 50%           5.0       3.400      1.500       0.2
## 75%           5.2       3.675      1.575       0.3
## 100%          5.8       4.400      1.900       0.6

aggregate: Compute Summary Statistics of Data Subsets

# 按分类变量 Species 分组求和
# aggregate(subset(iris, select = -Species), by = list(iris[, "Species"]), FUN = sum)
aggregate(iris[, -5], list(iris[, 5]), sum)

##      Group.1 Sepal.Length Sepal.Width Petal.Length Petal.Width
## 1      setosa       250.3       171.4       73.1       12.3
## 2  versicolor       296.8       138.5       213.0        66.3

```

```
## 3 virginica      329.4      148.7      277.6      101.3
# 先确定位置，假设有很多分类变量
ind <- which("Species" == colnames(iris))
# 分组统计
aggregate(iris[, -ind], list(iris[, ind]), sum)

##      Group.1 Sepal.Length Sepal.Width Petal.Length Petal.Width
## 1      setosa      250.3      171.4       73.1      12.3
## 2 versicolor      296.8      138.5      213.0      66.3
## 3 virginica      329.4      148.7      277.6      101.3
```

按照 `Species` 划分的类别，分组计算，使用公式表示形式，右边一定是分类变量，否则会报错误或者警告，输出奇怪的结果，请读者尝试运行 `aggregate(Species ~ Sepal.Length, data = iris, mean)`。公式法表示分组计算，~ 左手边可以做加 + 减 - 乘 * 除 / 取余 %% 等数学运算。下面以数据集 `iris` 为例，只对 `Sepal.Length` 按 `Species` 分组计算

```
aggregate(Sepal.Length ~ Species, data = iris, mean)

##      Species Sepal.Length
## 1      setosa      5.006
## 2 versicolor      5.936
## 3 virginica      6.588
```

与上述分组统计结果一样的命令，在大数据集上，与 `aggregate` 相比，`tapply` 要快很多，`by` 是 `tapply` 的包裹，处理速度差不多。读者可以构造伪随机数据集验证。

```
# tapply(iris$Sepal.Length, list(iris$Species), mean)
with(iris, tapply(Sepal.Length, Species, mean))

##      setosa versicolor  virginica
##      5.006      5.936      6.588

by(iris$Sepal.Length, iris$Species, mean)

## iris$Species: setosa
## [1] 5.006
## -----
## iris$Species: versicolor
## [1] 5.936
```



```
## -----
## iris$Species: virginica
## [1] 6.588
```

对所有变量按 Species 分组计算

```
aggregate(. ~ Species, data = iris, mean)
```

```
##      Species Sepal.Length Sepal.Width Petal.Length Petal.Width
## 1      setosa      5.006     3.428      1.462     0.246
## 2 versicolor      5.936     2.770      4.260     1.326
## 3  virginica      6.588     2.974      5.552     2.026
```

对变量 Sepal.Length 和 Sepal.Width 求和后，按 Species 分组计算

```
aggregate(Sepal.Length + Sepal.Width ~ Species, data = iris, mean)
```

```
##      Species Sepal.Length + Sepal.Width
## 1      setosa                  8.434
## 2 versicolor                  8.706
## 3  virginica                  9.562
```

对多个分类变量做分组计算，在数据集 ChickWeight 中 Chick 和 Diet 都是数字编码的分类变量，其中 Chick 是有序的因子变量，Diet 是无序的因子变量，而 Time 是数值型的变量，表示小鸡出生的天数。

查看数据

```
str(ChickWeight)
```

```
## Classes 'nfnGroupedData', 'nfGroupedData', 'groupedData' and 'data.frame': 578 obs. of
##   $ weight: num  42 51 59 64 76 93 106 125 149 171 ...
##   $ Time  : num  0 2 4 6 8 10 12 14 16 18 ...
##   $ Chick : Ord.factor w/ 50 levels "18" < "16" < "15" < ...: 15 15 15 15 15 15 15 15 15 15 ...
##   $ Diet  : Factor w/ 4 levels "1", "2", "3", "4": 1 1 1 1 1 1 1 1 1 1 ...
##   - attr(*, "formula")=Class 'formula' language weight ~ Time | Chick
##   .. ..- attr(*, ".Environment")=<environment: R_EmptyEnv>
##   - attr(*, "outer")=Class 'formula' language ~Diet
##   .. ..- attr(*, ".Environment")=<environment: R_EmptyEnv>
##   - attr(*, "labels")=List of 2
##     ..$ x: chr "Time"
##     ..$ y: chr "Body weight"
```

```
## - attr(*, "units")=List of 2
##   ..$ x: chr "(days)"
##   ..$ y: chr "(gm)"
```

查看数据集 ChickWeight 的前几行

```
head(ChickWeight)
```

```
##   weight Time Chick Diet
## 1     42     0     1     1
## 2     51     2     1     1
## 3     59     4     1     1
## 4     64     6     1     1
## 5     76     8     1     1
## 6     93    10     1     1
```

```
str(ChickWeight)
```

```
## Classes 'nfnGroupedData', 'nfGroupedData', 'groupedData' and 'data.frame':  578 observations
##   $ weight: num  42 51 59 64 76 93 106 125 149 171 ...
##   $ Time   : num  0 2 4 6 8 10 12 14 16 18 ...
##   $ Chick  : Ord.factor w/ 50 levels "18"<"16"<"15"<...: 15 15 15 15 15 15 15 15 15 15 ...
##   $ Diet   : Factor w/ 4 levels "1","2","3","4": 1 1 1 1 1 1 1 1 1 1 ...
##   - attr(*, "formula")=Class 'formula' language weight ~ Time | Chick
##   .. ..- attr(*, ".Environment")=<environment: R_EmptyEnv>
##   - attr(*, "outer")=Class 'formula' language ~Diet
##   .. ..- attr(*, ".Environment")=<environment: R_EmptyEnv>
##   - attr(*, "labels")=List of 2
##     ..$ x: chr "Time"
##     ..$ y: chr "Body weight"
##   - attr(*, "units")=List of 2
##     ..$ x: chr "(days)"
##     ..$ y: chr "(gm)"
```

对于数据集 ChickWeight 中的有序变量 Chick, aggregate 会按照既定顺序返回分组计算的结果

```
aggregate(weight ~ Chick, data = ChickWeight, mean)
```

```
##   Chick      weight
## 1     18 37.00000
```

云
湘
黄
⑩

```
## 2      16 49.71429
## 3      15 60.12500
## 4      13 67.83333
## 5      9 81.16667
## 6      20 78.41667
## 7      10 83.08333
## 8      8 92.00000
## 9      17 92.50000
## 10     19 86.75000
## 11     4 99.33333
## 12     6 113.75000
## 13     11 129.91667
## 14     3 115.83333
## 15     1 111.66667
## 16     12 114.08333
## 17     2 119.91667
## 18     5 126.66667
## 19     14 151.33333
## 20     7 150.00000
## 21     24 66.25000
## 22     30 103.50000
## 23     22 104.25000
## 24     23 111.41667
## 25     27 110.41667
## 26     28 129.91667
## 27     26 131.00000
## 28     25 143.08333
## 29     29 141.83333
## 30     21 184.50000
## 31     33 109.75000
## 32     37 102.50000
## 33     36 134.91667
## 34     31 128.58333
## 35     39 134.25000
## 36     38 142.33333
## 37     32 157.58333
```



```
## 38     40 157.58333
## 39     34 168.83333
## 40     35 193.16667
## 41     44 102.10000
## 42     45 119.58333
## 43     43 143.00000
## 44     41 128.41667
## 45     47 127.91667
## 46     49 137.75000
## 47     46 134.08333
## 48     50 147.50000
## 49     42 149.08333
## 50     48 157.66667

aggregate(weight ~ Diet, data = ChickWeight, mean)

##   Diet    weight
## 1     1 102.6455
## 2     2 122.6167
## 3     3 142.9500
## 4     4 135.2627
```

分类变量没有用数字编码，以 CO2 数据集为例，该数据集描述草植对二氧化碳的吸收情况，Plant 是具有 12 个水平的有序的因子变量，Type 表示植物的源头分别是魁北克 (Quebec) 和密西西比 (Mississippi)，Treatment 表示冷却 (chilled) 和不冷却 (nonchilled) 两种处理方式，conc 表示周围环境中二氧化碳的浓度，uptake 表示植物吸收二氧化碳的速率。

```
# 查看数据集
head(CO2)

##   Plant  Type Treatment conc uptake
## 1 Qn1 Quebec nonchilled  95   16.0
## 2 Qn1 Quebec nonchilled 175   30.4
## 3 Qn1 Quebec nonchilled 250   34.8
## 4 Qn1 Quebec nonchilled 350   37.2
## 5 Qn1 Quebec nonchilled 500   35.3
## 6 Qn1 Quebec nonchilled 675   39.2
```



```
str(CO2)

## Classes 'nfnGroupedData', 'nfGroupedData', 'groupedData' and 'data.frame': 84 obs. of 
## $ Plant    : Ord.factor w/ 12 levels "Qn1"<"Qn2"<"Qn3"<..: 1 1 1 1 1 1 2 2 2 ...
## $ Type     : Factor w/ 2 levels "Quebec","Mississippi": 1 1 1 1 1 1 1 1 1 ...
## $ Treatment: Factor w/ 2 levels "nonchilled","chilled": 1 1 1 1 1 1 1 1 1 ...
## $ conc      : num  95 175 250 350 500 675 1000 95 175 250 ...
## $ uptake    : num  16 30.4 34.8 37.2 35.3 39.2 39.7 13.6 27.3 37.1 ...
## - attr(*, "formula")=Class 'formula' language uptake ~ conc | Plant
##   ... .- attr(*, ".Environment")=<environment: R_EmptyEnv>
## - attr(*, "outer")=Class 'formula' language ~Treatment * Type
##   ... .- attr(*, ".Environment")=<environment: R_EmptyEnv>
## - attr(*, "labels")=List of 2
##   ..$ x: chr "Ambient carbon dioxide concentration"
##   ..$ y: chr "CO2 uptake rate"
## - attr(*, "units")=List of 2
##   ..$ x: chr "(uL/L)"
##   ..$ y: chr "(umol/m^2 s)"
```

对单个变量分组统计

```
aggregate(uptake ~ Plant, data = CO2, mean)
```

```
##     Plant    uptake
## 1     Qn1 33.22857
## 2     Qn2 35.15714
## 3     Qn3 37.61429
## 4     Qc1 29.97143
## 5     Qc3 32.58571
## 6     Qc2 32.70000
## 7     Mn3 24.11429
## 8     Mn2 27.34286
## 9     Mn1 26.40000
## 10    Mc2 12.14286
## 11    Mc3 17.30000
## 12    Mc1 18.00000
```

```
aggregate(uptake ~ Type, data = CO2, mean)
```

```
##           Type   uptake
## 1      Quebec 33.54286
## 2 Mississippi 20.88333
aggregate(uptake ~ Treatment, data = C02, mean)
```

```
##    Treatment   uptake
## 1 nonchilled 30.64286
## 2 chilled    23.78333
```

对多个变量分组统计，查看二氧化碳吸收速率 uptake 随类型 Type 和处理方式 Treatment

```
aggregate(uptake ~ Type + Treatment, data = C02, mean)
```

```
##           Type Treatment   uptake
## 1      Quebec nonchilled 35.33333
## 2 Mississippi nonchilled 25.95238
## 3      Quebec     chilled 31.75238
## 4 Mississippi     chilled 15.81429
```

```
tapply(C02$uptake, list(C02>Type, C02>Treatment), mean)
```

```
##           nonchilled     chilled
## Quebec          35.33333 31.75238
## Mississippi    25.95238 15.81429
```

```
by(C02$uptake, list(C02>Type, C02>Treatment), mean)
```

```
## : Quebec
## : nonchilled
## [1] 35.33333
## -----
## : Mississippi
## : nonchilled
## [1] 25.95238
## -----
## : Quebec
## : chilled
## [1] 31.75238
## -----
```



```
## : Mississippi
## : chilled
## [1] 15.81429
```

在这个例子中 tapply 和 by 的输出结果的表示形式不一样，aggregate 返回一个 data.frame 数据框，tapply 返回一个表格 table，by 返回特殊的数据类型 by。

Function by is an object-oriented wrapper for tapply applied to data frames.

```
# 分组求和
# by(iris[, 1], INDICES = list(iris$Species), FUN = sum)
# by(iris[, 2], INDICES = list(iris$Species), FUN = sum)
by(iris[, 3], INDICES = list(iris$Species), FUN = sum)

## : setosa
## [1] 73.1
## -----
## : versicolor
## [1] 213
## -----
## : virginica
## [1] 277.6

by(iris[1:3], INDICES = list(iris$Species), FUN = sum)

## : setosa
## [1] 494.8
## -----
## : versicolor
## [1] 648.3
## -----
## : virginica
## [1] 755.7

by(iris[1:3], INDICES = list(iris$Species), FUN = summary)

## : setosa
##   Sepal.Length   Sepal.Width   Petal.Length
##   Min.    :4.300   Min.    :2.300   Min.    :1.000
##   1st Qu.:4.800   1st Qu.:3.200   1st Qu.:1.400
##   Median :5.000   Median :3.400   Median :1.500
```

```
## Mean :5.006  Mean :3.428  Mean :1.462
## 3rd Qu.:5.200 3rd Qu.:3.675 3rd Qu.:1.575
## Max. :5.800  Max. :4.400  Max. :1.900
## -----
## : versicolor
##   Sepal.Length   Sepal.Width   Petal.Length
##   Min. :4.900   Min. :2.000   Min. :3.00
##   1st Qu.:5.600 1st Qu.:2.525 1st Qu.:4.00
##   Median :5.900  Median :2.800  Median :4.35
##   Mean    :5.936  Mean    :2.770  Mean    :4.26
##   3rd Qu.:6.300 3rd Qu.:3.000 3rd Qu.:4.60
##   Max.   :7.000  Max.   :3.400  Max.   :5.10
## -----
## : virginica
##   Sepal.Length   Sepal.Width   Petal.Length
##   Min. :4.900   Min. :2.200   Min. :4.500
##   1st Qu.:6.225 1st Qu.:2.800 1st Qu.:5.100
##   Median :6.500  Median :3.000  Median :5.550
##   Mean    :6.588  Mean    :2.974  Mean    :5.552
##   3rd Qu.:6.900 3rd Qu.:3.175 3rd Qu.:5.875
##   Max.   :7.900  Max.   :3.800  Max.   :6.900
by(iris, INDICES = list(iris$Species), FUN = summary)

## : setosa
##   Sepal.Length   Sepal.Width   Petal.Length   Petal.Width
##   Min. :4.300   Min. :2.300   Min. :1.000   Min. :0.100
##   1st Qu.:4.800 1st Qu.:3.200 1st Qu.:1.400 1st Qu.:0.200
##   Median :5.000  Median :3.400  Median :1.500  Median :0.200
##   Mean    :5.006  Mean    :3.428  Mean    :1.462  Mean    :0.246
##   3rd Qu.:5.200 3rd Qu.:3.675 3rd Qu.:1.575 3rd Qu.:0.300
##   Max.   :5.800  Max.   :4.400  Max.   :1.900  Max.   :0.600
##       Species
##   setosa   :50
##   versicolor: 0
##   virginica : 0
##
```

```

## 
## 
## -----
## : versicolor
##   Sepal.Length   Sepal.Width   Petal.Length   Petal.Width      Species
##   Min.  :4.900   Min.  :2.000   Min.  :3.00   Min.  :1.000   setosa    :0
##   1st Qu.:5.600  1st Qu.:2.525  1st Qu.:4.00  1st Qu.:1.200  versicolor:50
##   Median :5.900  Median :2.800  Median :4.35  Median :1.300  virginica :0
##   Mean    :5.936  Mean    :2.770  Mean    :4.26  Mean    :1.326
##   3rd Qu.:6.300  3rd Qu.:3.000  3rd Qu.:4.60  3rd Qu.:1.500
##   Max.    :7.000  Max.    :3.400  Max.    :5.10  Max.    :1.800
## -----
## : virginica
##   Sepal.Length   Sepal.Width   Petal.Length   Petal.Width
##   Min.  :4.900   Min.  :2.200   Min.  :4.500   Min.  :1.400
##   1st Qu.:6.225  1st Qu.:2.800  1st Qu.:5.100  1st Qu.:1.800
##   Median :6.500  Median :3.000  Median :5.550  Median :2.000
##   Mean    :6.588  Mean    :2.974  Mean    :5.552  Mean    :2.026
##   3rd Qu.:6.900  3rd Qu.:3.175  3rd Qu.:5.875  3rd Qu.:2.300
##   Max.    :7.900  Max.    :3.800  Max.    :6.900  Max.    :2.500
##   Species
##   setosa    : 0
##   versicolor: 0
##   virginica:50
## 
## 
## 

```

Group Averages Over Level Combinations of Factors 分组平均

```
str(warpbreaks)
```

```

## 'data.frame':   54 obs. of  3 variables:
## $ breaks : num  26 30 54 25 70 52 51 26 67 18 ...
## $ wool   : Factor w/ 2 levels "A","B": 1 1 1 1 1 1 1 1 1 1 ...
## $ tension: Factor w/ 3 levels "L","M","H": 1 1 1 1 1 1 1 1 1 2 ...

```

```
head(warpbreaks)

##   breaks wool tension
## 1     26    A      L
## 2     30    A      L
## 3     54    A      L
## 4     25    A      L
## 5     70    A      L
## 6     52    A      L

ave(warpbreaks$breaks, warpbreaks$wool)

## [1] 31.03704 31.03704 31.03704 31.03704 31.03704 31.03704 31.03704 31.03704
## [9] 31.03704 31.03704 31.03704 31.03704 31.03704 31.03704 31.03704 31.03704 31.03704
## [17] 31.03704 31.03704 31.03704 31.03704 31.03704 31.03704 31.03704 31.03704 31.03704
## [25] 31.03704 31.03704 31.03704 25.25926 25.25926 25.25926 25.25926 25.25926 25.25926
## [33] 25.25926 25.25926 25.25926 25.25926 25.25926 25.25926 25.25926 25.25926 25.25926
## [41] 25.25926 25.25926 25.25926 25.25926 25.25926 25.25926 25.25926 25.25926 25.25926
## [49] 25.25926 25.25926 25.25926 25.25926 25.25926 25.25926 25.25926 25.25926 25.25926

with(warpbreaks, ave(breaks, tension, FUN = function(x) mean(x, trim = 0.1)))

## [1] 35.6875 35.6875 35.6875 35.6875 35.6875 35.6875 35.6875 35.6875 35.6875
## [10] 26.3125 26.3125 26.3125 26.3125 26.3125 26.3125 26.3125 26.3125 26.3125
## [19] 21.0625 21.0625 21.0625 21.0625 21.0625 21.0625 21.0625 21.0625 21.0625
## [28] 35.6875 35.6875 35.6875 35.6875 35.6875 35.6875 35.6875 35.6875 35.6875
## [37] 26.3125 26.3125 26.3125 26.3125 26.3125 26.3125 26.3125 26.3125 26.3125
## [46] 21.0625 21.0625 21.0625 21.0625 21.0625 21.0625 21.0625 21.0625 21.0625

# 分组求和
with(warpbreaks, ave(breaks, tension, FUN = function(x) sum(x)))

## [1] 655 655 655 655 655 655 655 655 475 475 475 475 475 475 475 475 475 475 390
## [20] 390 390 390 390 390 390 390 390 655 655 655 655 655 655 655 655 655 655 475 475
## [39] 475 475 475 475 475 475 390 390 390 390 390 390 390 390 390 390 390 390 390 390

# 分组求和
with(iris, ave(Sepal.Length, Species, FUN = function(x) sum(x)))

## [1] 250.3 250.3 250.3 250.3 250.3 250.3 250.3 250.3 250.3 250.3 250.3 250.3 250.3
## [13] 250.3 250.3 250.3 250.3 250.3 250.3 250.3 250.3 250.3 250.3 250.3 250.3 250.3
```



```
## [25] 250.3 250.3 250.3 250.3 250.3 250.3 250.3 250.3 250.3 250.3 250.3 250.3 250.3 250.3 250.3
## [37] 250.3 250.3 250.3 250.3 250.3 250.3 250.3 250.3 250.3 250.3 250.3 250.3 250.3 250.3 250.3
## [49] 250.3 250.3 296.8 296.8 296.8 296.8 296.8 296.8 296.8 296.8 296.8 296.8 296.8 296.8 296.8
## [61] 296.8 296.8 296.8 296.8 296.8 296.8 296.8 296.8 296.8 296.8 296.8 296.8 296.8 296.8 296.8
## [73] 296.8 296.8 296.8 296.8 296.8 296.8 296.8 296.8 296.8 296.8 296.8 296.8 296.8 296.8 296.8
## [85] 296.8 296.8 296.8 296.8 296.8 296.8 296.8 296.8 296.8 296.8 296.8 296.8 296.8 296.8 296.8
## [97] 296.8 296.8 296.8 296.8 329.4 329.4 329.4 329.4 329.4 329.4 329.4 329.4 329.4 329.4 329.4
## [109] 329.4 329.4 329.4 329.4 329.4 329.4 329.4 329.4 329.4 329.4 329.4 329.4 329.4 329.4 329.4
## [121] 329.4 329.4 329.4 329.4 329.4 329.4 329.4 329.4 329.4 329.4 329.4 329.4 329.4 329.4 329.4
## [133] 329.4 329.4 329.4 329.4 329.4 329.4 329.4 329.4 329.4 329.4 329.4 329.4 329.4 329.4 329.4
## [145] 329.4 329.4 329.4 329.4 329.4 329.4 329.4 329.4 329.4 329.4 329.4 329.4 329.4 329.4 329.4
```

5.11 表格统计

介绍操作表格的 `table`, `addmargins`, `prop.table`, `xtabs`, `margin.table`, `ftable` 等函数

`table` 多个分类变量分组计数统计

- 介绍 `warpbreaks` 和 `airquality` 纽约空气质量监测数据集二维的数据框
- `UCBAdmissions` 1973 年加州大学伯克利分校的院系录取数据集 3 维的列联表
- `Titanic` 4 维的列联表数据泰坦尼克号幸存者数据集

```
with(warpbreaks, table(wool, tension))
```

```
##      tension
## wool L M H
##   A 9 9 9
##   B 9 9 9
```

以 `iris` 数据集为例, `table` 的第一个参数是自己制造的第二个分类变量, 原始分类变量是 `Species`

```
with(iris, table(Sepal.check = Sepal.Length > 7, Species))
```

```
##           Species
## Sepal.check setosa versicolor virginica
##      FALSE      50       50       38
```

```
##      TRUE      0      0     12
with(iris, table(Sepal.check = Sepal.Length > mean(Sepal.Length), Species))

##          Species
## Sepal.check setosa versicolor virginica
##      FALSE    50      24      6
##      TRUE     0      26     44
```

以 airquality 数据集为例，看看月份中臭氧含量比较高的几天

```
aiq.tab <- with(airquality, table(Oz.high = Ozone > 80, Month))
aiq.tab
```

```
##      Month
## Oz.high 5 6 7 8 9
## FALSE 25 9 20 19 27
## TRUE 1 0 6 7 2
```

对表格按行和列求和，即求表格的边际，查看总体情况

```
addmargins(aiq.tab, 1:2)
```

```
##      Month
## Oz.high 5 6 7 8 9 Sum
## FALSE 25 9 20 19 27 100
## TRUE 1 0 6 7 2 16
## Sum 26 9 26 26 29 116
```

臭氧含量超 80 的天数在每个月的占比，`addmargins` 的第二个参数 1 表示对列求和

```
aiq.prop <- prop.table(aiq.tab, 2)
aiq.prop
```

```
##      Month
## Oz.high      5       6       7       8       9
## FALSE 0.96153846 1.00000000 0.76923077 0.73076923 0.93103448
## TRUE 0.03846154 0.00000000 0.23076923 0.26923077 0.06896552
aiq.marprop <- addmargins(aiq.prop, 1)
aiq.marprop
```

```
##      Month
```



```
## Oz.high      5       6       7       8       9
## FALSE 0.96153846 1.00000000 0.76923077 0.73076923 0.93103448
## TRUE  0.03846154 0.00000000 0.23076923 0.26923077 0.06896552
## Sum   1.00000000 1.00000000 1.00000000 1.00000000 1.00000000
```

转换成百分比，将小数四舍五入转化为百分数，保留两位小数点

```
round(100 * aiq.marprop, 2)
```

```
##          Month
## Oz.high      5       6       7       8       9
## FALSE 96.15 100.00 76.92 73.08 93.10
## TRUE  3.85  0.00 23.08 26.92  6.90
## Sum   100.00 100.00 100.00 100.00 100.00

pairs(airquality, panel = panel.smooth, main = "airquality data")
```

以 UCBAdmissions 数据集为例，使用 `xtabs` 函数把数据组织成列联表，先查看数据的内容

```
UCBAdmissions
```

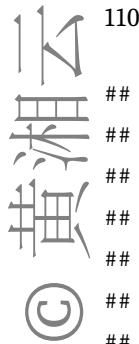
```
## , , Dept = A
##
##          Gender
## Admit     Male Female
## Admitted 512    89
## Rejected 313    19
##
## , , Dept = B
##
##          Gender
## Admit     Male Female
## Admitted 353    17
## Rejected 207     8
##
## , , Dept = C
##
##          Gender
## Admit     Male Female
```

```
##   Admitted 120    202
##   Rejected  205    391
##
## , , Dept = D
##
##          Gender
## Admit      Male Female
##   Admitted 138     131
##   Rejected  279    244
##
## , , Dept = E
##
##          Gender
## Admit      Male Female
##   Admitted 53      94
##   Rejected 138    299
##
## , , Dept = F
##
##          Gender
## Admit      Male Female
##   Admitted 22      24
##   Rejected 351    317

UCBA2DF <- as.data.frame(UCBAdmissions)

UCBA2DF

##      Admit Gender Dept Freq
## 1  Admitted  Male   A  512
## 2  Rejected  Male   A  313
## 3  Admitted Female  A   89
## 4  Rejected Female  A   19
## 5  Admitted  Male   B  353
## 6  Rejected  Male   B  207
## 7  Admitted Female  B   17
## 8  Rejected Female  B    8
## 9  Admitted  Male   C  120
## 10 Rejected  Male   C  205
```



```
## 11 Admitted Female      C  202
## 12 Rejected  Female     C  391
## 13 Admitted   Male      D  138
## 14 Rejected   Male      D  279
## 15 Admitted Female     D  131
## 16 Rejected Female     D  244
## 17 Admitted   Male      E   53
## 18 Rejected   Male      E  138
## 19 Admitted Female     E   94
## 20 Rejected Female     E  299
## 21 Admitted   Male      F   22
## 22 Rejected   Male      F  351
## 23 Admitted Female     F   24
## 24 Rejected Female     F  317
```

接着将 UCBA2DF 数据集转化为表格的形式

```
UCBA2DF.tab <- xtabs(Freq ~ Gender + Admit + Dept, data = UCBA2DF)
ftable(UCBA2DF.tab)
```

```
##                   Dept   A    B    C    D    E    F
## Gender Admit
##   Male   Admitted      512 353 120 138  53  22
##         Rejected       313 207 205 279 138 351
##   Female Admitted      89   17 202 131  94  24
##         Rejected       19    8 391 244 299 317
```

将录取性别和院系进行对比

```
prop.table(margin.table(UCBA2DF.tab, c(1, 3)), 1)
```

```
##                   Dept
##   Gender          A        B        C        D        E        F
##   Male  0.30657748 0.20810108 0.12077295 0.15496098 0.07097733 0.13861018
##   Female 0.05885559 0.01362398 0.32316076 0.20435967 0.21416894 0.18583106
```

男生倾向于申请院系 A 和 B，女生倾向于申请院系 C 到 F，院系 A 和 B 是最容易录取的。

5.12 索引访问

`which` 与引用 `[` 性能比较，在区间 $[0, 1]$ 上生成 10 万个服从均匀分布的随机数，随机抽取其中 $\frac{1}{4}$ 。

```
n <- 100000
x <- runif(n)
i <- logical(n)
i[sample(n, n / 4)] <- TRUE
microbenchmark::microbenchmark(x[i], x[which(i)], times = 1000)
```

TODO: 使用 `subset` 函数与 `[` 比较

5.13 多维数组

多维数组的行列是怎么定义的？`array` 轴的概念，画个图表示数组

```
array(1:27, c(3, 3, 3))
```

```
## , , 1
##
##      [,1] [,2] [,3]
## [1,]    1    4    7
## [2,]    2    5    8
## [3,]    3    6    9
##
## , , 2
##
##      [,1] [,2] [,3]
## [1,]   10   13   16
## [2,]   11   14   17
## [3,]   12   15   18
##
## , , 3
##
##      [,1] [,2] [,3]
## [1,]   19   22   25
## [2,]   20   23   26
```



```
## [3,] 21 24 27
```

垂直于 Z 轴的平面去截三维立方体，3 代表 z 轴，得到三个截面（二维矩阵）

```
asplit(array(1:27, c(3, 3, 3)), 3)
```

```
## [[1]]  
## [,1] [,2] [,3]  
## [1,] 1 4 7  
## [2,] 2 5 8  
## [3,] 3 6 9  
##  
## [[2]]  
## [,1] [,2] [,3]  
## [1,] 10 13 16  
## [2,] 11 14 17  
## [3,] 12 15 18  
##  
## [[3]]  
## [,1] [,2] [,3]  
## [1,] 19 22 25  
## [2,] 20 23 26  
## [3,] 21 24 27
```

对每个二维矩阵按列求和

```
lapply(asplit(array(1:27, c(3, 3, 3)), 3), apply, 2, sum)
```

```
## [[1]]  
## [1] 6 15 24  
##  
## [[2]]  
## [1] 33 42 51  
##  
## [[3]]  
## [1] 60 69 78
```

asplit 和 lapply 组合处理多维数组的计算问题

三维数组的矩阵运算 abind 包提供更多的数组操作，如合并，替换

数组操作 aperm 数组转置 Array Transposition

asplit 数组拆分其后接 lapply 或者 vapply

apply 数组计算

rray 包 <https://github.com/r-lib/rray>

5.14 其它操作

成对的数据操作有 `list` 与 `unlist`、`stack` 与 `unstack`、`class` 与 `unclass`、`attach` 与 `detach` 以及 `with` 和 `within`，它们在数据操作过程中有时会起到一定的补充作用。

5.14.1 列表属性

```
# 创建列表
list(...)
pairlist(...)

# 转化列表
as.list(x, ...)

## S3 method for class 'environment'
as.list(x, all.names = FALSE, sorted = FALSE, ...)
as.pairlist(x)

# 检查列表
is.list(x)
is.pairlist(x)

alist(...)
```

`list` 函数用来构造、转化和检查 R 列表对象。下面创建一个临时列表对象 `tmp`，它包含两个元素 A 和 B，两个元素都是向量，前者是数值型，后者是字符型

```
(tmp <- list(A = c(1, 2, 3), B = c("a", "b")))
```

```
## $A
## [1] 1 2 3
##
## $B
```



```
## [1] "a" "b"
unlist(x, recursive = TRUE, use.names = TRUE)
```

`unlist` 函数将给定的列表对象 `x` 简化为原子向量 (atomic vector)，我们发现简化之后变成一个字符型向量

```
unlist(tmp)
```

```
## A1 A2 A3 B1 B2
## "1" "2" "3" "a" "b"
unlist(tmp, use.names = FALSE)
```

```
## [1] "1" "2" "3" "a" "b"
```

`unlist` 的逆操作是 `relist`

5.14.2 堆叠向量

```
stack(x, ...)
## Default S3 method:
stack(x, drop = FALSE, ...)
## S3 method for class 'data.frame'
stack(x, select, drop = FALSE, ...)

unstack(x, ...)
## Default S3 method:
unstack(x, form, ...)
## S3 method for class 'data.frame'
unstack(x, form, ...)
```

`stack` 与 `unstack` 将多个向量堆在一起组成一个向量

```
# 查看数据集 PlantGrowth
```

```
class(PlantGrowth)
```

```
## [1] "data.frame"
```

```
head(PlantGrowth)
```

```
## weight group
```

```
## 1  4.17  ctrl
## 2  5.58  ctrl
## 3  5.18  ctrl
## 4  6.11  ctrl
## 5  4.50  ctrl
## 6  4.61  ctrl

# 检查默认的公式
formula(PlantGrowth)

## weight ~ group

# 根据公式解除堆叠
# 下面等价于 unstack(PlantGrowth, form = weight ~ group)
(pg <- unstack(PlantGrowth))
```

```
##      ctrl trt1 trt2
## 1  4.17 4.81 6.31
## 2  5.58 4.17 5.12
## 3  5.18 4.41 5.54
## 4  6.11 3.59 5.50
## 5  4.50 5.87 5.37
## 6  4.61 3.83 5.29
## 7  5.17 6.03 4.92
## 8  4.53 4.89 6.15
## 9  5.33 4.32 5.80
## 10 5.14 4.69 5.26
```

现在再将变量 pg 堆叠起来，还可以指定要堆叠的列

```
stack(pg)
```

```
##      values  ind
## 1    4.17  ctrl
## 2    5.58  ctrl
## 3    5.18  ctrl
## 4    6.11  ctrl
## 5    4.50  ctrl
## 6    4.61  ctrl
## 7    5.17  ctrl
```

云
湘
黄
◎

```
## 8    4.53 ctrl
## 9    5.33 ctrl
## 10   5.14 ctrl
## 11   4.81 trt1
## 12   4.17 trt1
## 13   4.41 trt1
## 14   3.59 trt1
## 15   5.87 trt1
## 16   3.83 trt1
## 17   6.03 trt1
## 18   4.89 trt1
## 19   4.32 trt1
## 20   4.69 trt1
## 21   6.31 trt2
## 22   5.12 trt2
## 23   5.54 trt2
## 24   5.50 trt2
## 25   5.37 trt2
## 26   5.29 trt2
## 27   4.92 trt2
## 28   6.15 trt2
## 29   5.80 trt2
## 30   5.26 trt2

stack(pg, select = -ctrl)
```

```
##     values  ind
## 1    4.81 trt1
## 2    4.17 trt1
## 3    4.41 trt1
## 4    3.59 trt1
## 5    5.87 trt1
## 6    3.83 trt1
## 7    6.03 trt1
## 8    4.89 trt1
## 9    4.32 trt1
## 10   4.69 trt1
```

```

## 11   6.31 trt2
## 12   5.12 trt2
## 13   5.54 trt2
## 14   5.50 trt2
## 15   5.37 trt2
## 16   5.29 trt2
## 17   4.92 trt2
## 18   6.15 trt2
## 19   5.80 trt2
## 20   5.26 trt2

```

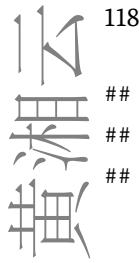
形式上和 `reshape` 有一些相似之处，数据框可以由长变宽或由宽变长。

5.14.3 属性转化

```
class(x)
class(x) <- value
unclass(x)
inherits(x, what, which = FALSE)

oldClass(x)
oldClass(x) <- value
```

`class` 和 `unclass` 函数用来查看、设置类属性和取消类属性，常用于面向对象的编程设计中



```
## [149] 3 3
## attr(,"levels")
## [1] "setosa"      "versicolor"   "virginica"
```

④ 5.14.4 绑定环境

```
attach(what,
  pos = 2L, name = deparse(substitute(what)), backtick = FALSE),
  warn.conflicts = TRUE
)
detach(name,
  pos = 2L, unload = FALSE, character.only = FALSE,
  force = FALSE
)
```

attach 和 detach 是否绑定数据框的列名，不推荐操作，推荐使用 with

```
attach(iris)
head(Species)
```

```
## [1] setosa setosa setosa setosa setosa setosa
## Levels: setosa versicolor virginica
detach(iris)
```

5.14.5 数据环境

```
with(data, expr, ...)
within(data, expr, ...)
## S3 method for class 'list'
within(data, expr, keepAttrs = TRUE, ...)
```

data 参数 data 用来构造表达式计算的环境。其默认值可以是一个环境，列表，数据框。在 within 函数中 data 参数只能是列表或数据框。

expr 参数 expr 包含要计算的表达式。在 within 中通常包含一个复合表达式，比如



```
{  
  a <- somefun()  
  b <- otherfun()  
  ...  
  rm(unused1, temp)  
}
```

`with` 和 `within` 计算一组表达式，计算的环境是由数据构造的，后者可以修改原始的数据

```
with(mtcars, mpg[cyl == 8 & disp > 350])
```

```
## [1] 18.7 14.3 10.4 10.4 14.7 19.2 15.8
```

和下面计算的结果一样，但是更加简洁漂亮

```
mtcars$mpg[mtcars$cyl == 8 & mtcars$disp > 350]
```

```
## [1] 18.7 14.3 10.4 10.4 14.7 19.2 15.8
```

`within` 函数可以修改原数据环境中的多个变量，比如删除、修改和添加等

```
# 原数据集 airquality  
head(airquality)
```

```
##   Ozone Solar.R Wind Temp Month Day  
## 1    41     190  7.4   67      5    1  
## 2    36     118  8.0   72      5    2  
## 3    12     149 12.6   74      5    3  
## 4    18     313 11.5   62      5    4  
## 5    NA      NA 14.3   56      5    5  
## 6    28      NA 14.9   66      5    6
```

```
aq <- within(airquality, {  
  lOzone <- log(Ozone) # 取对数  
  Month <- factor(month.abb[Month]) # 字符串型转因子型  
  cTemp <- round((Temp - 32) * 5 / 9, 1) # 从华氏温度到摄氏温度转化  
  S.cT <- Solar.R / cTemp # 使用新创建的变量  
  rm(Day, Temp)  
})  
# 修改后的数据集
```

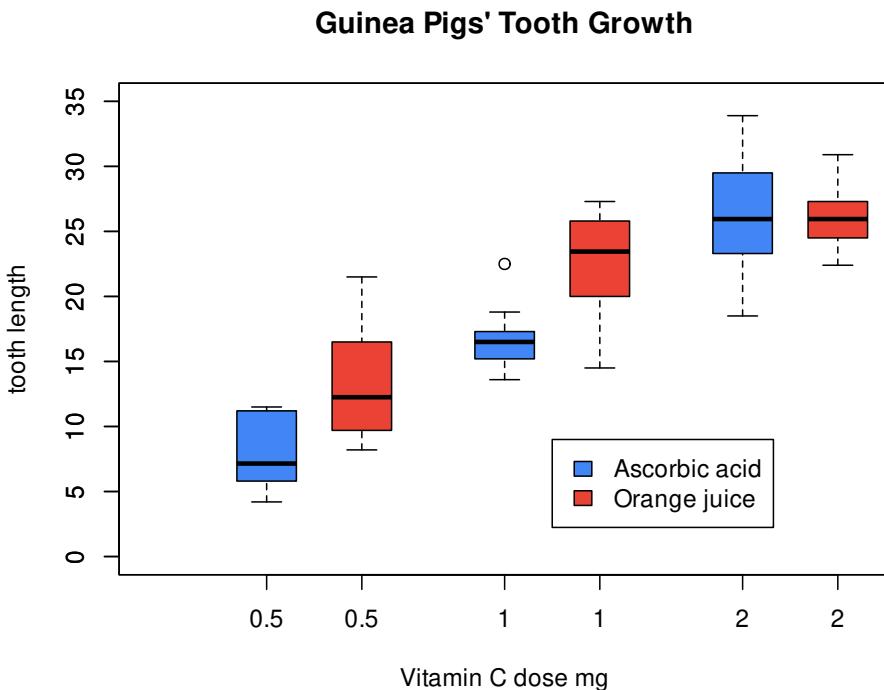


```
head(aq)
```

```
##   Ozone Solar.R Wind Month      S.cT cTemp  lOzone
## 1     41     190  7.4    May  9.793814 19.4 3.713572
## 2     36     118  8.0    May  5.315315 22.2 3.583519
## 3     12     149 12.6    May  6.394850 23.3 2.484907
## 4     18     313 11.5    May 18.742515 16.7 2.890372
## 5     NA      NA 14.3    May       NA 13.3      NA
## 6     28      NA 14.9    May       NA 18.9 3.332205
```

下面再举一个复杂的绘图例子，这个例子来自 `boxplot` 函数

```
with(ToothGrowth, {
  boxplot(len ~ dose,
    boxwex = 0.25, at = 1:3 - 0.2,
    subset = (supp == "VC"), col = "#4285f4",
    main = "Guinea Pigs' Tooth Growth",
    xlab = "Vitamin C dose mg",
    ylab = "tooth length", ylim = c(0, 35)
  )
  boxplot(len ~ dose,
    add = TRUE, boxwex = 0.25, at = 1:3 + 0.2,
    subset = supp == "OJ", col = "#EA4335"
  )
  legend(2, 9, c("Ascorbic acid", "Orange juice"),
    fill = c("#4285f4", "#EA4335")
  )
})
```



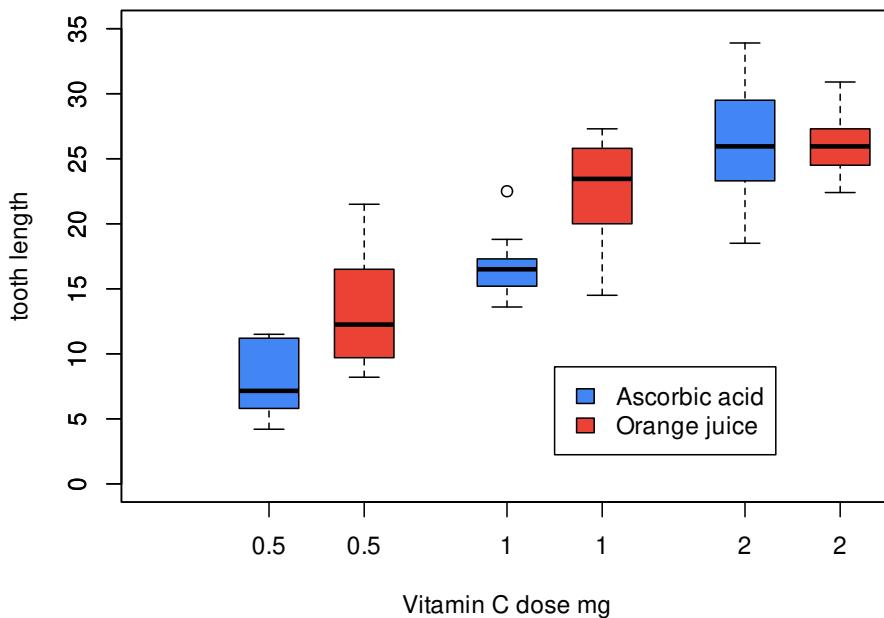
将 `boxplot` 函数的 `subset` 参数单独提出来，调用数据子集选择函数 `subset`，这里 `with` 中只包含一个表达式，所以也可以不用大括号

```
with(  
  subset(ToothGrowth, supp == "VC"),  
  boxplot(len ~ dose,  
    boxwex = 0.25, at = 1:3 - 0.2,  
    col = "#4285f4", main = "Guinea Pigs' Tooth Growth",  
    xlab = "Vitamin C dose mg",  
    ylab = "tooth length", ylim = c(0, 35))  
)  
)  
with(  
  subset(ToothGrowth, supp == "OJ"),  
  boxplot(len ~ dose,  
    add = TRUE, boxwex = 0.25, at = 1:3 + 0.2,  
    col = "#EA4335"))
```

```
)  
legend(2, 9, c("Ascorbic acid", "Orange juice"),  
      fill = c("#4285f4", "#EA4335"))  
)
```

C

Guinea Pigs' Tooth Growth



可以作为数据变换 `transform` 的一种替代，它也比较像 `dplyr` 包的 `mutate` 函数

```
within(mtcars[1:5,1:3],{  
  disp.cc <- disp * 2.54^3  
  disp.l <- disp.cc / 1e3  
})
```

```
##          mpg cyl disp  disp.l  disp.cc  
## Mazda RX4     21.0   6 160 2.621930 2621.930  
## Mazda RX4 Wag 21.0   6 160 2.621930 2621.930  
## Datsun 710    22.8   4 108 1.769803 1769.803  
## Hornet 4 Drive 21.4   6 258 4.227863 4227.863  
## Hornet Sportabout 18.7   8 360 5.899343 5899.343
```



```
# 只能使用已有的列，刚生成的列不能用
# transform(
#   mtcars[1:5, 1:3],
#   disp.cc = disp * 2.54^3,
#   disp.l = disp.cc / 1e3
# )
transform(
  mtcars[1:5, 1:3],
  disp.cc = disp * 2.54^3
)

##          mpg cyl disp  disp.cc
## Mazda RX4     21.0   6 160 2621.930
## Mazda RX4 Wag 21.0   6 160 2621.930
## Datsun 710    22.8   4 108 1769.803
## Hornet 4 Drive 21.4   6 258 4227.863
## Hornet Sportabout 18.7   8 360 5899.343
```

`transform` 只能使用已有的列，变换中间生成的列不能用，所以相比于 `transform` 函数，`within` 显得更为灵活

5.15 apply 族

表 5.1: apply 函数

函数	输入	输出
<code>apply()</code>	矩阵、数据框	向量
<code>lapply()</code>	向量、列表	列表
<code>sapply()</code>	向量、列表	向量、矩阵
<code>mapply()</code>	多个向量	列表
<code>tapply()</code>	数据框、数组	向量
<code>vapply()</code>	列表	矩阵
<code>eapply()</code>	列表	列表
<code>rapply()</code>	嵌套列表	嵌套列表

云
湘
黄
©

除此之外，还有 `dendrapply()` 专门处理层次聚类或分类回归树型结构，而函数 `kernapply()` 用于时间序列的平滑处理

```
# Reproduce example 10.4.3 from Brockwell and Davis (1991) [@Brockwell_1991_Time]
spectrum(sunspot.year, kernel = kernel("daniell", c(11, 7, 3)), log = "no")
```

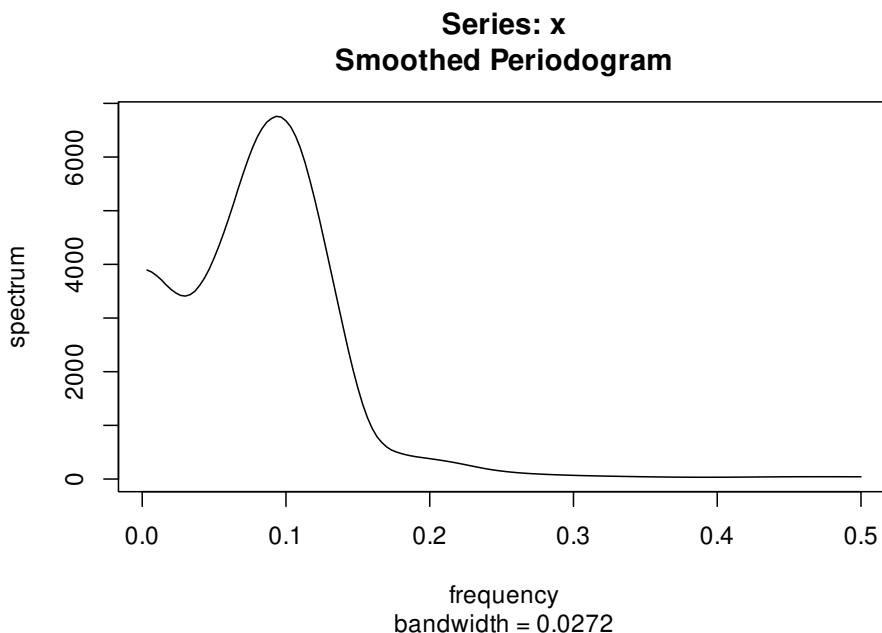


图 5.4: 太阳黑子的频谱

将函数应用到多个向量，返回一个列表，生成四组服从正态分布 $\mathcal{N}(\mu_i, \sigma_i)$ 的随机数，它们的均值和方差依次是 $\mu_i = \sigma_i = 1 \dots 4$

```
means <- 1:4
sds <- 1:4
set.seed(2020)
samples <- mapply(rnorm,
  mean = means, sd = sds,
  MoreArgs = list(n = 10), SIMPLIFY = FALSE
)
samples

## [[1]]
## [1] 1.37697212 1.30154837 -0.09802317 -0.13040590 -1.79653432 1.72057350
```

```
## [7] 1.93912102 0.77062225 2.75913135 1.11736679
##
## [[2]]
## [1] 0.2937544 3.8185184 4.3927459 1.2568322 1.7534795 5.6000862
## [7] 5.4079918 -4.0775292 -2.5779499 2.1166070
##
## [[3]]
## [1] 9.523096 6.294548 3.954661 2.780557 5.502806 3.596252 6.893524 5.810155
## [9] 2.557700 3.331296
##
## [[4]]
## [1] 0.7499813 1.0251913 8.3813803 13.7414948 5.5524739 5.1625107
## [7] 2.8576069 4.3040589 1.7588056 5.7887535
```

我们借用图5.5来看一下 `mapply` 的效果，多组随机数生成非常有助于快速模拟。

```
par(mfrow = c(2, 2), mar = c(2, 2, 2, 2))
invisible(lapply(samples, function(x) {
  plot(x, pch = 16, col = "grey")
  abline(h = mean(x), lwd = 2, col = "darkorange")
}))
```

分别计算每个样本的平均值

```
sapply(samples, mean)
```

```
## [1] 0.8960372 1.7984536 5.0244596 4.9322257
```

分别计算每个样本的 1, 2, 3 分位点

```
lapply(samples, quantile, probs = 1:3 / 4)
```

```
## [[1]]
##      25%      50%      75%
## 0.1191382 1.2094576 1.6346732
##
## [[2]]
##      25%      50%      75%
## 0.5345238 1.9350433 4.2491890
##
## [[3]]
```

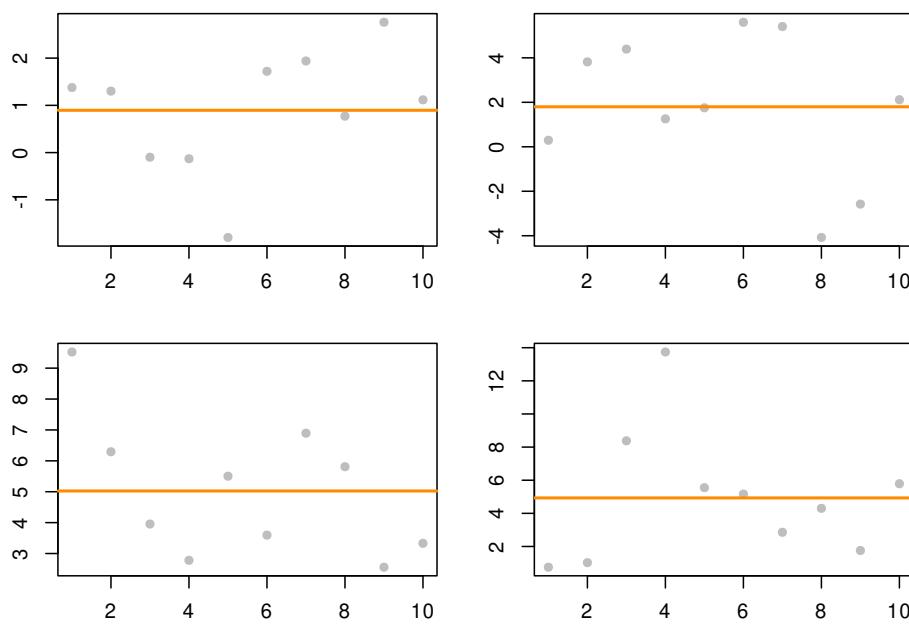


图 5.5: lapply 函数

```

##      25%      50%      75%
## 3.397535 4.728734 6.173450
##
## [[4]]
##      25%      50%      75%
## 2.033506 4.733285 5.729684

```

仅用 `sapply()` 函数替换上面的 `lapply()`，我们可以得到一个矩阵，值得注意的是函数 `quantile()` 和 `fivenum()` 算出来的结果有一些差异

```
sapply(samples, quantile, probs = 1:3 / 4)
```

```

##      [,1]      [,2]      [,3]      [,4]
## 25% 0.1191382 0.5345238 3.397535 2.033506
## 50% 1.2094576 1.9350433 4.728734 4.733285
## 75% 1.6346732 4.2491890 6.173450 5.729684

```

```
vapply(samples, fivenum, c(Min. = 0, "1st Qu." = 0, Median = 0, "3rd Qu." = 0, Max. = 0))
```

```

##      [,1]      [,2]      [,3]      [,4]
##
```

```
## Min. -1.79653432 -4.0775292 2.557700 0.7499813  
## 1st Qu. -0.09802317 0.2937544 3.331296 1.7588056  
## Median 1.20945758 1.9350433 4.728734 4.7332848  
## 3rd Qu. 1.72057350 4.3927459 6.294548 5.7887535  
## Max. 2.75913135 5.6000862 9.523096 13.7414948
```

vapply 和 sapply 类似，但是预先指定返回值类型，这样可以更加安全，有时也更快。

以数据集 presidents 为例，它是一个 ts 对象类型的时间序列数据，记录了 1945 年至 1974 年每个季度美国总统的支持率，这组数据中存在缺失值，以 NA 表示。支持率的变化趋势见图 5.6。

```
plot(presidents)
```

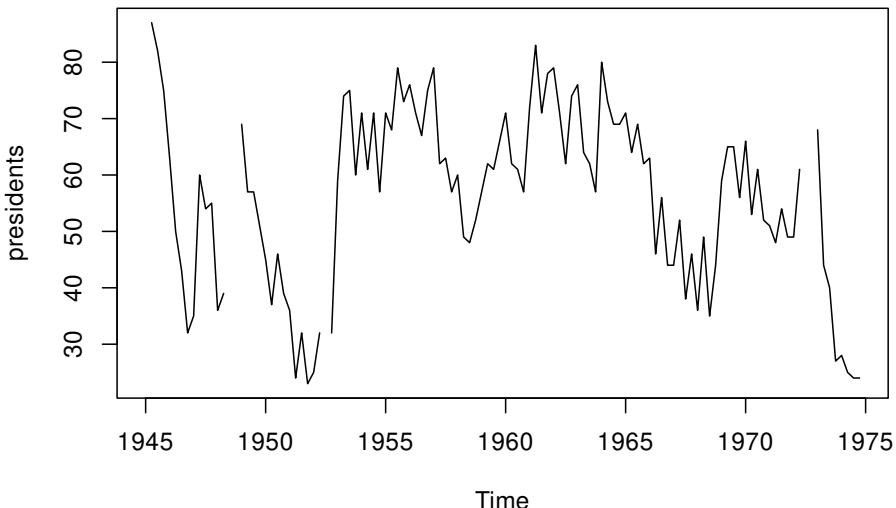


图 5.6: 1945-1974 美国总统的支持率

计算这 30 年每个季度的平均支持率

```
tapply(presidents, cycle(presidents), mean, na.rm = TRUE)
```

```
##      1      2      3      4  
## 58.44828 56.43333 57.22222 53.07143
```



`cycle()` 函数计算序列中每个观察值在周期中的位置, `presidents` 的周期为 4, 根据位置划分组, 然后分组求平均, 也可以化作如下计算步骤, 虽然看起来复杂, 但是数据操作的过程很清晰, 不再看起来像是一个黑箱。

`tapply` 函数来做分组求和

```
# 一个变量分组求和
tapply(warpbreaks$breaks, warpbreaks[, 3, drop = FALSE], sum)

## tension
##   L   M   H
## 655 475 390

# 两个变量分组计数
with(warpbreaks, table(wool, tension))

##      tension
## wool L M H
##   A 9 9 9
##   B 9 9 9

# 两个变量分组求和
dat <- aggregate(breaks ~ wool + tension, data = warpbreaks, sum) |>
  reshape(v.names = "breaks", idvar = "wool", timevar = "tension", direction = "wide",
`colnames<-`(dat, gsub(pattern = "(breaks)", x = colnames(dat), replacement = ""))

##   wool   L   M   H
## 1   A 401 216 221
## 2   B 254 259 169
```

5.16 with 选项

注意 `data.table` 与 Base R 不同的地方

```
# https://github.com/Rdatatable/data.table/issues/4513
# https://d.cosx.org/d/421532-data-table-base-r
library(data.table)
iris <- as.data.table(iris)
```



```
iris[Species == "setosa" & Sepal.Length > 5.5, grep("Sepal", colnames(iris))]

## [1] TRUE TRUE FALSE FALSE FALSE
```

需要使用 `with = FALSE` 选项

```
iris[Species == "setosa" & Sepal.Length > 5.5,
     grep("Sepal", colnames(iris)),
     with = FALSE
]
```

```
##      Sepal.Length Sepal.Width
## 1:          5.8        4.0
## 2:          5.7        4.4
## 3:          5.7        3.8
```

不使用 `with` 选项，用函数 `mget()` 将字符串转变量

```
iris[
  Species == "setosa" & Sepal.Length > 5.5,
  mget(grep("Sepal", colnames(iris), value = TRUE))
]

##      Sepal.Length Sepal.Width
## 1:          5.8        4.0
## 2:          5.7        4.4
## 3:          5.7        3.8
```

更加 `data.table` 风格的方式见

```
iris[Species == "setosa" & Sepal.Length > 5.5, .SD, .SDcols = patterns("Sepal")]

##      Sepal.Length Sepal.Width
## 1:          5.8        4.0
## 2:          5.7        4.4
## 3:          5.7        3.8
```

`with` 还可以这样用，直接修改、添加一列

```
df <- expand.grid(x = 1:10, y = 1:10)
df$z <- with(df, x^2 + y^2)
df <- subset(df, z < 100)
df <- df[sample(nrow(df)), ]
```

```
head(df)
```

④

```
##      x  y  z
## 7    7 1 50
## 8    8 1 65
## 65   5 7 74
## 14   4 2 20
## 37   7 4 65
## 5    5 1 26
```

```
library(ggplot2)
ggplot(df, aes(x, y, z = z)) +
  geom_contour()
```

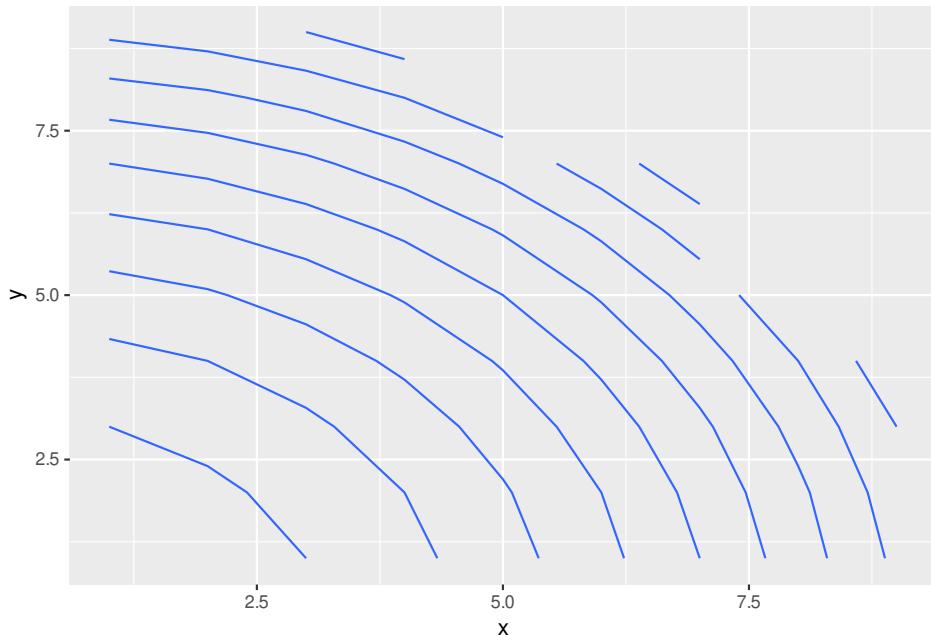


图 5.7: with 操作

5.17 分组聚合

```
methods("aggregate")
## [1] aggregate.data.frame aggregate.default*  aggregate.formula*
## [4] aggregate.ts
## see '?methods' for accessing help and source code
args("aggregate.data.frame")

## function (x, by, FUN, ..., simplify = TRUE, drop = TRUE)
## NULL

args("aggregate.ts")

## function (x, nfrequency = 1, FUN = sum, ndeltat = 1, ts.eps = getOption("ts.eps"),
##        ...)
## NULL

# getAnywhere(aggregate.formula)
```

按 Species 分组，对 Sepal.Length 中大于平均值的数取平均

```
aggregate(Sepal.Length ~ Species, iris, function(x) mean(x[x > mean(x)]))

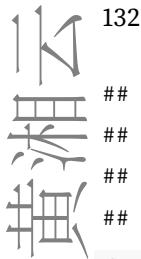
##      Species Sepal.Length
## 1      setosa     5.313636
## 2 versicolor     6.375000
## 3  virginica     7.159091

library(data.table)

dt <- data.table(
  x = rep(1:3, each = 3), y = rep(1:3, 3),
  z = rep(c("A", "B", "C"), 3), w = rep(c("a", "b", "a"), each = 3)
)

dt[, .(x_sum = sum(x), y_sum = sum(y)), by = .(z, w)]

##      z w x_sum y_sum
## 1: A a     4     2
## 2: B a     4     4
```



```
## 3: C a      4      6
## 4: A b      2      1
## 5: B b      2      2
## 6: C b      2      3

dt[, .(x_sum = sum(x), y_sum = sum(y)), by = mget(c("z", "w"))]

##      z w x_sum y_sum
## 1: A a      4      2
## 2: B a      4      4
## 3: C a      4      6
## 4: A b      2      1
## 5: B b      2      2
## 6: C b      2      3
```

shiny 前端传递字符串向量，借助 `mget()` 函数根据选择的变量分组统计计算，只有一个变量可以使用 `get()` 传递变量给 `data.table`

```
library(shiny)

ui <- fluidPage(
  fluidRow(
    column(
      6,
      selectInput("input_vars",
                  label = "变量", # 给筛选框取名
                  choices = c(z = "z", w = "w"), # 待选的值
                  selected = "z", # 指定默认值
                  multiple = TRUE # 允许多选
      ),
      DT::dataTableOutput("output_table")
    )
  )
)

library(data.table)
library(magrittr)

dt <- data.table(
```



```
x = rep(1:3, each = 3), y = rep(1:3, 3),
z = rep(c("A", "B", "C"), 3), w = rep(c("a", "b", "a"), each = 3)
)

server <- function(input, output, session) {
  output$output_table <- DT::renderDataTable(
    {
      dt[, .(x_sum = sum(x), y_sum = sum(y)), by = mget(input$input_vars)] |>
        DT::datatable()
    },
    server = FALSE
  )
}

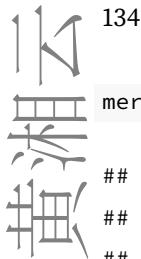
# 执行
shinyApp(ui = ui, server = server)
```

5.18 合并操作

```
dat1 <- data.frame(x = c(0, 0, 10, 10, 20, 20, 30, 30),
dat2 <- data.frame(x = c(0, 10, 20, 30), z = c(3, 4, 5, 6))

data.frame(dat1, z = dat2$z[match(dat1$x, dat2$x)])

##      x y z
## 1  0 1 3
## 2  0 1 3
## 3 10 2 4
## 4 10 2 4
## 5 20 3 5
## 6 20 3 5
## 7 30 4 6
## 8 30 4 6
```



```
merge(dat1, dat2)

##      x y z
## 1  0 1 3
## 2  0 1 3
## 3 10 2 4
## 4 10 2 4
## 5 20 3 5
## 6 20 3 5
## 7 30 4 6
## 8 30 4 6
```

保留两个数据集中的所有行

5.19 长宽转换

```
args("reshape")

## function (data, varying = NULL, v.names = NULL, timevar = "time",
##   idvar = "id", ids = 1L:NROW(data), times = seq_along(varying[[1L]]),
##   drop = NULL, direction, new.row.names = NULL, sep = ".",
##   split = if (sep == "") {
##     list(regexp = "[A-Za-z][0-9]", include = TRUE)
##   } else {
##     list(regexp = sep, include = FALSE, fixed = TRUE)
##   })
## NULL
```

PlantGrowth 数据集的重塑操作也可以使用内置的函数 `reshape()` 实现

```
PlantGrowth$id <- rep(1:10, 3)
dat <- reshape(
  data = PlantGrowth, idvar = "group", v.names = "weight",
  timevar = "id", direction = "wide",
  sep = ""
)
knitr::kable(dat,
```



表 5.2: 不同生长环境下植物的干重

group	1	2	3	4	5	6	7	8	9	10
ctrl	4.17	5.58	5.18	6.11	4.50	4.61	5.17	4.53	5.33	5.14
trt1	4.81	4.17	4.41	3.59	5.87	3.83	6.03	4.89	4.32	4.69
trt2	6.31	5.12	5.54	5.50	5.37	5.29	4.92	6.15	5.80	5.26

```
caption = "不同生长环境下植物的干重", row.names = FALSE,
col.names = gsub("(weight)", "", names(dat)),
align = "c"
)
```

或者，我们也可以使用 **tidyverse** 包提供的 **pivot_wider()** 函数

```
tidyverse::pivot_wider(
  data = PlantGrowth, id_cols = id,
  names_from = group, values_from = weight
)

## # A tibble: 10 x 4
##       id ctrl trt1 trt2
##   <int> <dbl> <dbl> <dbl>
## 1     1  4.17  4.81  6.31
## 2     2  5.58  4.17  5.12
## 3     3  5.18  4.41  5.54
## 4     4  6.11  3.59  5.5 
## 5     5  4.5   5.87  5.37
## 6     6  4.61  3.83  5.29
## 7     7  5.17  6.03  4.92
## 8     8  4.53  4.89  6.15
## 9     9  5.33  4.32  5.8 
## 10    10  5.14  4.69  5.26
```

或者，我们还可以使用 **data.table** 包提供的 **dcast()** 函数，用于将长格式的数据框重塑为宽格式的

```
PlantGrowth_DT <- as.data.table(PlantGrowth)
# 纵
dcast(PlantGrowth_DT, id ~ group, value.var = "weight")
```

黄湘云

```
##      id ctrl trt1 trt2
## 1:  1 4.17 4.81 6.31
## 2:  2 5.58 4.17 5.12
## 3:  3 5.18 4.41 5.54
## 4:  4 6.11 3.59 5.50
## 5:  5 4.50 5.87 5.37
## 6:  6 4.61 3.83 5.29
## 7:  7 5.17 6.03 4.92
## 8:  8 4.53 4.89 6.15
## 9:  9 5.33 4.32 5.80
## 10: 10 5.14 4.69 5.26

# 横
dcast(PlantGrowth_DT, group ~ id, value.var = "weight")
```

```
##      group     1     2     3     4     5     6     7     8     9    10
## 1:  ctrl 4.17 5.58 5.18 6.11 4.50 4.61 5.17 4.53 5.33 5.14
## 2:  trt1 4.81 4.17 4.41 3.59 5.87 3.83 6.03 4.89 4.32 4.69
## 3:  trt2 6.31 5.12 5.54 5.50 5.37 5.29 4.92 6.15 5.80 5.26
```

5.20 对符合条件的列操作

```
# 数值型变量的列的位置
which(sapply(iris, is.numeric))

## Sepal.Length  Sepal.Width Petal.Length  Petal.Width
##                 1             2             3             4

iris[, sapply(iris, is.numeric), with = F][Sepal.Length > 7.5]

##      Sepal.Length Sepal.Width Petal.Length Petal.Width
## 1:        7.6       3.0       6.6       2.1
## 2:        7.7       3.8       6.7       2.2
## 3:        7.7       2.6       6.9       2.3
## 4:        7.7       2.8       6.7       2.0
## 5:        7.9       3.8       6.4       2.0
```

```
## 6:          7.7          3.0          6.1          2.3  
class(iris)  
  
## [1] "data.table" "data.frame"
```

用 Base R 提供的管道符号 |> 将 data.table 数据操作与 ggplot2 数据可视化连接起来

```
library(ggplot2)  
iris |>  
  subset(Species == "setosa" & Sepal.Length > 5.5) |>  
  # 行过滤  
  # subset(select = grep("Sepal", colnames(iris), value = TRUE)) |> # 列过滤  
  subset(select = grepl("Sepal", colnames(iris))) |>  
  ggplot(aes(x = Sepal.Length, y = Sepal.Width)) + # 绘图  
  geom_point()
```

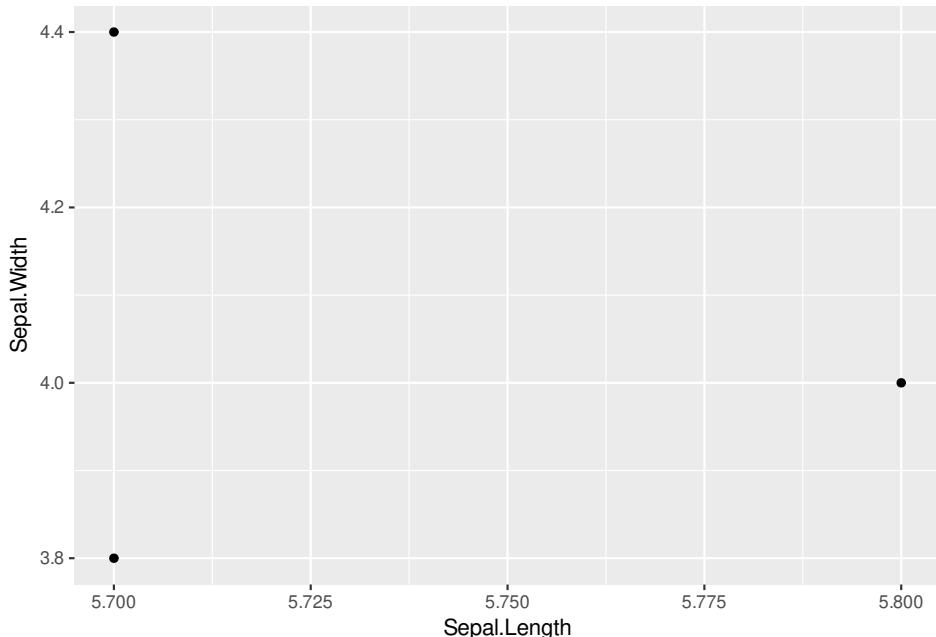


图 5.8: 管道连接数据操作和可视化

5.21 CASE WHEN 和 fcase

CASE WHEN 是 SQL 中的条件判断语句, **data.table** 中的函数 `fcase()` 可与之等价。值得注意的是, `fcase()` 需要 **data.table** 版本 1.13.0 及以上。

```
dat <- data.table(
  weights = c(56.8, 57.2, 46.3, 38.5),
  gender = c("1", "0", "", "0")
)
# 1 表示男, 0表示女, 空表示未知
transform(dat, gender_cn = fcase(
  gender == "1", "男",
  gender == "0", "女",
  gender == "", "未知"
))

##      weights gender gender_cn
## 1:    56.8      1     男
## 2:    57.2      0     女
## 3:    46.3      ""   未知
## 4:    38.5      0     女
```

5.22 数据操作实战

Toby Dylan Hocking 在 useR! 2020 大会上分享的幻灯片 <https://github.com/tdhock/r-devel-emails>

5.23 高频数据操作

以数据集 `dat` 为例介绍常用的数据操作

```
set.seed(2020)
dat <- data.frame(
  num_a = rep(seq(4), each = 4), num_b = rep(seq(4), times = 4),
  group_a = sample(x = letters[1:3], size = 16, replace = T),
  group_b = sample(x = LETTERS[1:3], size = 16, replace = T)
```



```
)  
dat <- as.data.table(dat)  
dat  
  
##      num_a num_b group_a group_b  
## 1:     1     1      c      B  
## 2:     1     2      b      B  
## 3:     1     3      a      B  
## 4:     1     4      a      C  
## 5:     2     1      b      B  
## 6:     2     2      b      C  
## 7:     2     3      a      B  
## 8:     2     4      a      A  
## 9:     3     1      b      C  
## 10:    3     2      b      B  
## 11:    3     3      b      B  
## 12:    3     4      a      B  
## 13:    4     1      b      C  
## 14:    4     2      c      B  
## 15:    4     3      b      C  
## 16:    4     4      a      C
```

5.23.1 循环合并

- 问题来源 [Faster version of Reduce\(merge, list\(DT1,DT2,DT3,...\)\) called mergelist \(a la rbindlist\)](#)

5.23.2 分组计数

```
dat[, .(length(num_a)), by = .(group_a)] # dat[, .N , by = .(group_a)]  
  
##      group_a V1  
## 1:      c   2  
## 2:      b   8  
## 3:      a   6
```

云
湘
黄
◎

```
dat[, .(length(num_a)), by = .(group_b)]
##      group_b V1
## 1:      B  9
## 2:      C  6
## 3:      A  1

dat[, .(length(num_a)), by = .(group_a, group_b)]
##      group_a group_b V1
## 1:      c      B  2
## 2:      b      B  4
## 3:      a      B  3
## 4:      a      C  2
## 5:      b      C  4
## 6:      a      A  1
```

5.23.3 分组抽样

以 group_a 为组别，a、b、c 分别有 6、8、2 条记录

```
# 无放回的抽样
dt_sample_1 <- dat[, .SD[sample(x = .N, size = 2, replace = FALSE)], by = group_a]
# 有放回的随机抽样
dt_sample_2 <- dat[, .SD[sample(x = .N, size = 3, replace = TRUE)], by = group_a]
```

可能存在该组样本不平衡，有的组的样本量不足你想要的样本量。每个组无放回地抽取 4 个样本，如果该组样本量不足 4，则全部抽取全部样本量。

```
dat[, .SD[sample(x = .N, size = min(4, .N))], by = group_a]

##      group_a num_a num_b group_b
## 1:      c     1     1      B
## 2:      c     4     2      B
## 3:      b     3     2      B
## 4:      b     2     2      C
## 5:      b     2     1      B
## 6:      b     3     3      B
## 7:      a     1     3      B
```

```
## 8:      a    2    3    B
## 9:      a    2    4    A
## 10:     a    1    4    C
```

还可以按照指定的比例抽取样本量¹

5.23.4 分组排序

data.table 包的分组排序问题 <https://d.cosx.org/d/421650-datatatable/3>

```
dat[with(dat, order(-ave(num_a, group_a, FUN = max), -num_a)), ]
```

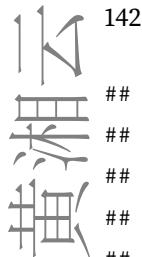
```
##      num_a num_b group_a group_b
## 1:     4     1      b      C
## 2:     4     2      c      B
## 3:     4     3      b      C
## 4:     4     4      a      C
## 5:     3     1      b      C
## 6:     3     2      b      B
## 7:     3     3      b      B
## 8:     3     4      a      B
## 9:     2     1      b      B
## 10:    2     2      b      C
## 11:    2     3      a      B
## 12:    2     4      a      A
## 13:    1     1      c      B
## 14:    1     2      b      B
## 15:    1     3      a      B
## 16:    1     4      a      C
```

num_a 降序排列, 然后对 *group_a* 升序排列

```
dat[with(dat, order(-num_a, group_a)), ]
```

```
##      num_a num_b group_a group_b
## 1:     4     4      a      C
## 2:     4     1      b      C
## 3:     4     3      b      C
```

¹<https://stackoverflow.com/questions/18258690/take-randomly-sample-based-on-groups>



```
## 4:    4    2    c    B
## 5:    3    4    a    B
## 6:    3    1    b    C
## 7:    3    2    b    B
## 8:    3    3    b    B
## 9:    2    3    a    B
## 10:   2    4    a    A
## 11:   2    1    b    B
## 12:   2    2    b    C
## 13:   1    3    a    B
## 14:   1    4    a    C
## 15:   1    2    b    B
## 16:   1    1    c    B
```

简写

```
dat[order(-num_a, group_a)]
```

```
##      num_a num_b group_a group_b
## 1:    4    4    a    C
## 2:    4    1    b    C
## 3:    4    3    b    C
## 4:    4    2    c    B
## 5:    3    4    a    B
## 6:    3    1    b    C
## 7:    3    2    b    B
## 8:    3    3    b    B
## 9:    2    3    a    B
## 10:   2    4    a    A
## 11:   2    1    b    B
## 12:   2    2    b    C
## 13:   1    3    a    B
## 14:   1    4    a    C
## 15:   1    2    b    B
## 16:   1    1    c    B
```

`setorder()` 函数直接修改原始数据记录的排序



```
setorder(dat, -num_a, group_a)
```

参考多个列分组排序²

提示

如果数据集 dat 包含缺失值，考虑去掉缺失值

```
dat[, .(length(!is.na(num_a))), by = .(group_a)]
```

```
##      group_a V1  
## 1:      c   2  
## 2:      b   8  
## 3:      a   6
```

如果数据集 dat 包含重复值，考虑去掉重复值

```
dat[, .(length(unique(num_a))), by = .(group_a)]
```

```
##      group_a V1  
## 1:      c   2  
## 2:      b   4  
## 3:      a   4
```

按 Species 分组，对 Sepal.Length 降序排列，取 Top 3

```
iris <- as.data.table(iris)  
iris[order(-Sepal.Length), .SD[1:3], by = "Species"]
```

```
##      Species Sepal.Length Sepal.Width Petal.Length Petal.Width  
## 1:  virginica     7.9       3.8       6.4       2.0  
## 2:  virginica     7.7       3.8       6.7       2.2  
## 3:  virginica     7.7       2.6       6.9       2.3  
## 4: versicolor     7.0       3.2       4.7       1.4  
## 5: versicolor     6.9       3.1       4.9       1.5  
## 6: versicolor     6.8       2.8       4.8       1.4  
## 7:    setosa      5.8       4.0       1.2       0.2  
## 8:    setosa      5.7       4.4       1.5       0.4  
## 9:    setosa      5.7       3.8       1.7       0.3
```

对 iris 各个列排序

²<https://stackoverflow.com/questions/1296646/how-to-sort-a-dataframe-by-multiple-columns>



```
dat <- head(iris)
ind <- do.call(what = "order", args = dat[, c(5, 1, 2, 3)])
dat[ind, ]

##      Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1:          4.6       3.1        1.5       0.2   setosa
## 2:          4.7       3.2        1.3       0.2   setosa
## 3:          4.9       3.0        1.4       0.2   setosa
## 4:          5.0       3.6        1.4       0.2   setosa
## 5:          5.1       3.5        1.4       0.2   setosa
## 6:          5.4       3.9        1.7       0.4   setosa
```

按 Species 分组，对 Sepal.Length 降序排列，取 Top 3

```
iris = as.data.table(iris)
iris[order(-Sepal.Length), .SD[1:3], by="Species"]
```

```
##      Species Sepal.Length Sepal.Width Petal.Length Petal.Width
## 1: virginica     7.9       3.8        6.4       2.0
## 2: virginica     7.7       3.8        6.7       2.2
## 3: virginica     7.7       2.6        6.9       2.3
## 4: versicolor    7.0       3.2        4.7       1.4
## 5: versicolor    6.9       3.1        4.9       1.5
## 6: versicolor    6.8       2.8        4.8       1.4
## 7: setosa         5.8       4.0        1.2       0.2
## 8: setosa         5.7       4.4        1.5       0.4
## 9: setosa         5.7       3.8        1.7       0.3
```

对 iris 各个列排序，依次对第 5、1、2、3 列升序排列

```
ind <- do.call(what = "order", args = iris[,c(5,1,2,3)])
head(iris[ind, ])
```

```
##      Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1:          4.3       3.0        1.1       0.1   setosa
## 2:          4.4       2.9        1.4       0.2   setosa
## 3:          4.4       3.0        1.3       0.2   setosa
## 4:          4.4       3.2        1.3       0.2   setosa
## 5:          4.5       2.3        1.3       0.3   setosa
## 6:          4.6       3.1        1.5       0.2   setosa
```

表 5.3: iris 数据集原顺序 (左) 和新顺序 (右)

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
4.3	3.0	1.1	0.1	setosa
4.4	2.9	1.4	0.2	setosa
4.4	3.0	1.3	0.2	setosa
4.4	3.2	1.3	0.2	setosa
4.5	2.3	1.3	0.3	setosa
4.6	3.1	1.5	0.2	setosa

第六章 数据搬运

`openxlsx` 可以读写 XLSX 文档

美团使用的大数据工具有很多，最常用的 Hive、Spark、Kylin、Impala、Presto 等，详见 <https://tech.meituan.com/2018/08/02/mt-r-practice.html>。下面主要介绍如何在 R 中连接 MySQL、Presto 和 Spark。

`sparklyr.flint` 支持 Spark 的时间序列库 `flint`, `sparkxgb` 为 Spark 上的 XGBoost 提供 R 接口, `sparkwarc` 支持加载 Web ARCHive 文件到 Spark 里 `sparkavro` 支持从 Apache Avro (<https://avro.apache.org/>) 读取文件到 Spark 里, `sparkbq` 是一个 `sparklyr` 扩展包, 集成 Google BigQuery 服务, `geospark` 提供 GeoSpark 库的 R 接口, 并且以 `sf` 的数据操作方式, `rsparkling` H2O Sparkling Water 机器学习库的 R 接口。

Spark 性能优化，参考三篇博文

- [Spark 在美团的实践](#)
- [Spark 性能优化指南——基础篇](#)
- [Spark 性能优化指南——高级篇](#)

其他材料

- 朱俊晖收集的 Spark 资源列表 <https://github.com/harryprince/awesome-sparklyr>, 推荐使用 `sparklyr` <https://github.com/sparklyr/sparklyr> 连接 Spark
- Spark 与 R 语言 <https://docs.microsoft.com/en-us/azure/databricks/spark/latest/sparkr/>
- Mastering Spark with R <https://therinspark.com/>



6.1 Spark 与 R 语言

6.1.1 sparklyr

警告

Spark 依赖特定版本的 Java、Hadoop，三者之间的版本应该要相融。

在 MacOS 上配置 Java 环境，注意 Spark 仅支持 Java 8 至 11，所以安装指定版本的 Java 开发环境

```
# 安装 openjdk 11
brew install openjdk@11
# 全局设置 JDK 11
sudo ln -sfn /usr/local/opt/openjdk@11/libexec/openjdk.jdk /Library/Java/JavaVirtualMachines/jdk-11.0.1.jdk/Contents/Home
# Java 11 JDK 添加到 .zshrc
export CPPFLAGS="-I/usr/local/opt/openjdk@11/include"
export PATH="/usr/local/opt/openjdk@11/bin:$PATH"
```

配置 R 环境，让 R 能够识别 Java 环境，再安装 **rJava** 包

```
# 配置
sudo R CMD javareconf
# 系统软件依赖
brew install pcre2
# 安装 rJava
Rscript -e 'install.packages("rJava", type="source")'
```

最后安装 **sparklyr** 包，以及 Spark 环境，可以借助 `spark_install()` 安装 Spark，比如下面 Spark 3.0 连同 hadoop 2.7 一起安装。

```
install.packages('sparklyr')
sparklyr::spark_install(version = '3.0', hadoop_version = '2.7')
```

也可以先手动下载 Spark 软件环境，建议选择就近镜像站点下载，比如在北京选择清华站点 <https://mirrors.tuna.tsinghua.edu.cn/apache/spark/spark-3.0.1/spark-3.0.1-bin-hadoop2.7.tgz>，此环境自带 R 和 Python 接口。为了供 `sparklyr` 调用，先设置 `SPARK_HOME` 环境变量指向 Spark 安装位置，再连接单机版 Spark。

```
# 排错 https://github.com/sparklyr/sparklyr/issues/2827
options(sparklyr.log.console = FALSE)
```



```
# 连接 Spark
library(sparklyr)
library(ggplot2)
sc <- spark_connect(
  master = "local",
  # config = list(sparklyr.gateway.address = "127.0.0.1"),
  spark_home = Sys.getenv("SPARK_HOME")
)
# diamonds 数据集导入 Spark
diamonds_tbl <- copy_to(sc, ggplot2::diamonds, "diamonds")
```

做数据的聚合统计，有两种方式。一种是使用用 R 包 `dplyr` 提供的数据操作语法，下面以按 `cut` 分组统计钻石的数量为例，说明 `dplyr` 提供的数据操作方式。

```
library(dplyr)
# 列出数据源下所有的表 tbds
src_tbds(sc)

diamonds_tbl <- diamonds_tbl %>%
  group_by(cut) %>%
  summarise(cnt = n()) %>%
  collect
```

另一种是使用结构化查询语言 SQL，这自不必说，大多数情况下，使用和一般的 SQL 没什么两样。

```
library(DBI)
diamonds_preview <- dbGetQuery(sc, "SELECT count(*) as cnt, cut FROM diamonds GROUP BY
diamonds_preview
```

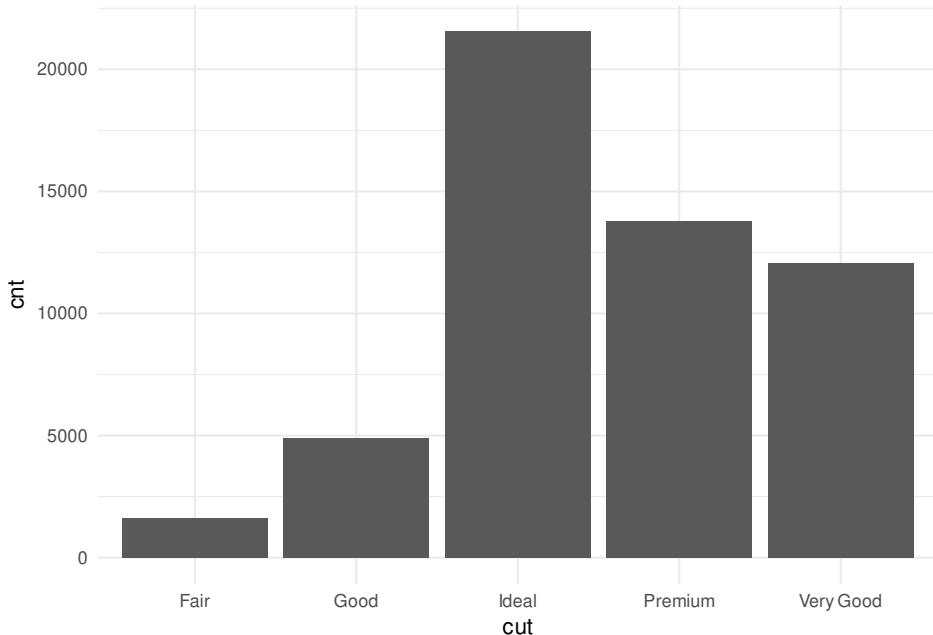
```
##      cnt      cut
## 1 21551     Ideal
## 2 13791    Premium
## 3  4906      Good
## 4  1610      Fair
## 5 12082 Very Good
```



```
# SQL 中的 AVG 和 R 中的 mean 函数是 library(ggplot2)
library(data.table)
diamonds_price <- dbGetQuery(sc, "SELECT * FROM diamonds")
diamonds <- as.data.table(diamonds)
diamonds[,.(mean_price = mean(price)), by = .(cut)]
##   mean_price      cut
## 1    3457.542   Ideal
## 2    4584.258 Premium
## 3    3928.864    Good
## 4    4358.758     Fair
## 5    3981.760 Very Good
```

将结果数据用 ggplot2 呈现出来

```
ggplot(diamonds_preview, aes(cut, cnt)) +
  geom_col() +
  theme_minimal()
```



diamonds 数据集总共 53940 条数据，下面用 BUCKET 分桶抽样，将原数据随机分成 1000 个桶，取其中的一个桶，由于是随机分桶，所以每次的结果都不一样，解释详见<https://spark.apache.org/docs/latest/sql-ref-syntax-qry-select-sampling.html>

```

diamonds_sample <- dbGetQuery(sc, "SELECT * FROM diamonds TABLESAMPLE (BUCKET 1 OUT OF
diamonds_sample

##   carat      cut color clarity depth table price     x     y     z
## 1  0.72  Very Good      F    VS1  62.2     58  2804 5.75 5.70 3.56
## 2  0.71      Ideal      D    SI2  62.0     54  2934 5.77 5.74 3.57
## 3  0.76      Ideal      H    VS2  61.9     55  3016 5.85 5.88 3.64
## 4  0.78      Ideal      G    VS2  62.0     57  3590 5.90 5.86 3.65
## 5  0.31  Premium      D    SI1  60.9     60   571 4.38 4.39 2.67
## 6  0.91  Very Good      G    VS2  62.7     63  3776 6.05 6.00 3.78

```

将抽样的结果用窗口函数 RANK() 排序, 详见 <https://spark.apache.org/docs/latest/sql-ref-syntax-qry-select-window.html>

窗口函数 <https://www.cnblogs.com/ZackSun/p/9713435.html>

```

diamonds_rank <- dbGetQuery(sc, "
  SELECT cut, price, RANK() OVER (PARTITION BY cut ORDER BY price) AS rank
  FROM diamonds TABLESAMPLE (BUCKET 1 OUT OF 1000)
  LIMIT 6
")
diamonds_rank

```

```

##      cut price rank
## 1  Fair   6596    1
## 2  Good    830    1
## 3  Good   3391    2
## 4  Good   4257    3
## 5  Good   4320    4
## 6 Ideal    596    1

```

LATERAL VIEW 把一列拆成多行

<https://liam.page/2020/03/09/LATERAL-VIEW-in-Hive-SQL/> <https://spark.apache.org/docs/latest/sql-ref-syntax-qry-select-lateral-view.html>

创建数据集

```

# 先删除存在的表 person
dbGetQuery(sc, "DROP TABLE IF EXISTS person")
# 创建表 person

```



```
dbGetQuery(sc, "CREATE TABLE IF NOT EXISTS person (id INT, name STRING, age INT, class INT, address STRING)
# 插入数据到表 person
dbGetQuery(sc, "
INSERT INTO person VALUES
    (100, 'John', 30, 1, 'Street 1'),
    (200, 'Mary', NULL, 1, 'Street 2'),
    (300, 'Mike', 80, 3, 'Street 3'),
    (400, 'Dan', 50, 4, 'Street 4')
")
```

查看数据集

```
dbGetQuery(sc, "SELECT * FROM person")
```

```
##   id name age class  address
## 1 100 John  30     1 Street 1
## 2 200 Mary  NA     1 Street 2
## 3 300 Mike  80     3 Street 3
## 4 400 Dan   50     4 Street 4
```

行列转换 <https://www.cnblogs.com/kimbo/p/6208973.html>, LATERAL VIEW 展开

```
dbGetQuery(sc, "
SELECT * FROM person
    LATERAL VIEW EXPLODE(ARRAY(30, 60)) tableName AS c_age
    LATERAL VIEW EXPLODE(ARRAY(40, 80)) AS d_age
LIMIT 6
")
```

```
##   id name age class  address c_age d_age
## 1 100 John  30     1 Street 1     30    40
## 2 100 John  30     1 Street 1     30    80
## 3 100 John  30     1 Street 1     60    40
## 4 100 John  30     1 Street 1     60    80
## 5 200 Mary  NA     1 Street 2     30    40
## 6 200 Mary  NA     1 Street 2     30    80
```

日期相关的函数 <https://spark.apache.org/docs/latest/sql-ref-functions-built-in.html#date-and-timestamp-functions>



```
# 今天
dbGetQuery(sc, "select current_date")

##   current_date()
## 1      2021-08-07

# 昨天
dbGetQuery(sc, "select date_sub(current_date, 1)")

##   date_sub(current_date(), 1)
## 1            2021-08-06

# 本月最后一天 current_date 所属月份的最后一天
dbGetQuery(sc, "select last_day(current_date)")

##   last_day(current_date())
## 1            2021-08-31

# 星期几
dbGetQuery(sc, "select dayofweek(current_date)")

##   dayofweek(current_date())
## 1               7
```

最后，使用完记得关闭 Spark 连接

```
spark_disconnect(sc)
```

6.1.2 SparkR

注意

考虑到和第6.1.1节的重合性，以及 sparklyr 的优势，本节代码都不会执行，仅作为补充信息予以描述。完整的介绍见 [SparkR 包](#)

```
if (nchar(Sys.getenv("SPARK_HOME")) < 1) {
  Sys.setenv(SPARK_HOME = "/opt/spark/spark-3.0.1-bin-hadoop2.7")
}

library(SparkR, lib.loc = c(file.path(Sys.getenv("SPARK_HOME"), "R", "lib")))
sparkR.session(master = "local[*]", sparkConfig = list(spark.driver.memory = "2g"))
```



警告

SparkR 要求 Java 版本满足：大于等于 8，而小于 12，本地 MacOS 安装高版本，比如 oracle-jdk 16.0.1 会报不兼容的错误。

```
Spark package found in SPARK_HOME: /opt/spark/spark-3.1.1-bin-hadoop3.2
```

```
Error in checkJavaVersion() :
```

```
Java version, greater than or equal to 8 and less than 12, is required for this pac
```

sparkConfig 有哪些参数可以传递

Property Name	Property group	spark-submit equivalent
spark.master	Application Properties	--master
spark.kerberos.keytab	Application Properties	--keytab
spark.kerberos.principal	Application Properties	--principal
spark.driver.memory	Application Properties	--driver-memory
spark.driver.extraClassPath	Runtime Environment	--driver-class-path
spark.driver.extraJavaOptions	Runtime Environment	--driver-java-options
spark.driver.extraLibraryPath	Runtime Environment	--driver-library-path

将 `data.frame` 转化为 `SparkDataFrame`

```
faithful_sdf <- as.DataFrame(faithful)
```

`SparkDataFrame`

```
head(faithful_sdf)
```

查看结构

```
str(faithful_sdf)
```

6.2 数据库与 R 语言

Presto 的 R 接口 <https://github.com/prestodb/RPresto> 和文档 <https://prestodb.io/docs/current/index.html>，Presto 数据库

```
install.packages('RPresto')
```

MySQL 为例介绍 `odbc` 的连接和使用，详见 [从 R 连接 MySQL](#)

```
-- !preview conn=DBI::dbConnect(odbc::odbc(), driver = "MariaDB", database = "demo",
--                               uid = "root", pwd = "cloud", host = "localhost", port =
--                               3306)
SELECT * FROM mtcars
LIMIT 10
```

我的系统已经安装了多款数据库的 ODBC 驱动

```
dnf install -y unixODBC unixODBC-devel mariadb mariadb-server mariadb-devel mariadb-connectors
odbc::odbcListDrivers()
```

```
# Driver from the mariadb-connector-odbc package
# Setup from the unixODBC package
[MariaDB]
Description      = ODBC for MariaDB
Driver           = /usr/lib/libmaodbc.so
Driver64         = /usr/lib64/libmaodbc.so
FileUsage        = 1
```

6.3 批量读取 csv 文件

iris 数据转化为 data.table 类型，按照 Species 分组拆成单独的 csv 文件，各个文件的文件名用弯尾花的类别名表示

```
# 批量分组导出
library(data.table)
as.data.table(iris)[, fwrite(.SD, paste0("data/user_", unique(Species), ".csv")), by =
```

读取文件夹 data/ 所有 csv 数据文件

```
library(data.table)
merged_df <- do.call('rbind', lapply(list.files(pattern = "*.csv", path = "data/"), fread))
# 或者
merged_df <- rbindlist(lapply(list.files(pattern = "*.csv", path = "data/"), fread))

xdf$date <- as.Date(xdf$date)
xdf$ts <- as.POSIXct(as.numeric(xdf$ts), origin = "1978-01-01")
split(xdf[order(xdf$ts), ], interaction(xdf$study, xdf$port)) %>%
```

```
lapply(function(.x) {
  .x[nrow(.x), ]
}) %>%
unnname() %>%
Filter(function(.x) {
  nrow(.x) > 0
}, .) %>%
do.call(rbind.data.frame, .)

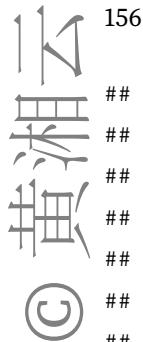
library(dplyr)
xdf %>%
  mutate(
    date = as.Date(date),
    ts = anytime::anytime(as.numeric(ts))
  ) %>%
  arrange(ts) %>%
  group_by(study, port) %>%
  slice(n()) %>%
  ungroup()

library(tibble)
library(magrittr)

mtcars <- tibble(mtcars)

mtcars %>%
  print(n = 16, width = 69)

## # A tibble: 32 x 11
##       mpg     cyl   disp     hp   drat     wt   qsec     vs     am   gear   carb
##       <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1     21      6   160    110    3.9    2.62   16.5     0      1      4      4
## 2     21      6   160    110    3.9    2.88   17.0     0      1      4      4
## 3    22.8     4   108     93    3.85   2.32   18.6     1      1      4      1
## 4    21.4     6   258    110    3.08   3.22   19.4     1      0      3      1
## 5    18.7     8   360    175    3.15   3.44   17.0     0      0      3      2
## 6    18.1     6   225    105    2.76   3.46   20.2     1      0      3      1
```



```
## 7 14.3     8 360    245 3.21 3.57 15.8   0   0   3   4
## 8 24.4     4 147.   62  3.69 3.19  20    1   0   4   2
## 9 22.8     4 141.   95  3.92 3.15 22.9   1   0   4   2
## 10 19.2    6 168.   123 3.92 3.44 18.3   1   0   4   4
## 11 17.8    6 168.   123 3.92 3.44 18.9   1   0   4   4
## 12 16.4    8 276.   180 3.07 4.07 17.4   0   0   3   3
## 13 17.3    8 276.   180 3.07 3.73 17.6   0   0   3   3
## 14 15.2    8 276.   180 3.07 3.78 18     0   0   3   3
## 15 10.4    8 472    205 2.93 5.25 18.0   0   0   3   4
## 16 10.4    8 460    215 3     5.42 17.8   0   0   3   4
## # ... with 16 more rows

mtcars %>%
  print(., n = nrow(.) / 4)

## # A tibble: 32 x 11
##       mpg   cyl  disp    hp  drat    wt  qsec    vs    am  gear  carb
##       <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1     21      6 160    110 3.9   2.62 16.5   0     1     4     4
## 2     21      6 160    110 3.9   2.88 17.0   0     1     4     4
## 3    22.8     4 108    93  3.85  2.32 18.6   1     1     4     1
## 4    21.4     6 258    110 3.08  3.22 19.4   1     0     3     1
## 5    18.7     8 360    175 3.15  3.44 17.0   0     0     3     2
## 6    18.1     6 225    105 2.76  3.46 20.2   1     0     3     1
## 7    14.3     8 360    245 3.21 3.57 15.8   0     0     3     4
## 8    24.4     4 147.   62  3.69 3.19  20    1     0     4     2
## # ... with 24 more rows
```

6.4 批量导出 xlsx 文件

将 R 环境中的数据集导出为 xlsx 表格

```
## 加载 openxlsx 包
library(openxlsx)
## 创建空白的工作薄，标题为鸢尾花数据集
wb <- createWorkbook(title = "鸢尾花数据集")
## 添加 sheet 页
```



```
addWorksheet(wb, sheetName = "iris")
# 将数据写进 sheet 页
writeData(wb, sheet = "iris", x = iris, colNames = TRUE)
# 导出数据到本地
saveWorkbook(wb, file = "iris.xlsx", overwrite = TRUE)

library(openxlsx)
xlsxFile <- system.file("extdata", "readTest.xlsx", package = "openxlsx")
# 导入
dat = read.xlsx(xlsxFile = xlsxFile)
# 导出
write.xlsx(dat, xlsxfile)
```



第七章 数据可视化

```
library(ggplot2)
library(patchwork)          # 图形布局
library(magrittr)           # 管道操作
library(ggrepel)            # 文本注释
library(extrafont)          # 加载外部字体 TTF
library(hrbrthemes)         # 主题
library(maps)               # 地图数据
library(mapdata)             # 地图数据
library(xkcd)                # 漫画字体
library(RgoogleMaps)        # 静态地图
library(data.table)          # 数据操作
library(KernSmooth)          # 核平滑
library(ggnormalviolin)      # 提琴图
library(ggbeeswarm)          # 蜂群图
library(gert)                 # Git 数据操作
library(ggridges)            # 岭线图
library(ggpubr)              # 组合图
library(treemap)             # 树状图
library(treemapify)          # 树状图
library(ggalluvial)          # 桑基图
library(ggmosaic)            # 马赛克图
library(ggbump)               # 凹凸图
library(ggstream)             # 水流图
library(timelineS)           # 时间线
library(ggdendro)             # 聚类图
library(ggfortify)            # 统计分析结果可视化：主成分图
```



```
library(gganimate)      # 动态图
```

David Robinson 给出为何使用 `ggplot2`¹ 当然也有 Jeff Leek 指出在某些重要场合不适合 `ggplot2`² 并且给出强有力的 [证据](#)，其实不管怎么样，适合自己的才是好的。也不枉费 Garrick Aden-Buie 花费 160 页幻灯片逐步分解介绍 [优雅的 ggplot2](#)，Malcolm Barrett 也介绍了 [ggplot2 基础用法](#)，还有 Selva Prabhakaran 精心总结给出了 50 个 `ggplot2` 数据可视化的 [例子](#) 以及 Victor Perrier 为小白用 `ggplot2` 操碎了心地开发 RStudio 插件 `esquisse` 包，Claus O. Wilke 教你一步步创建出版级的图形 https://github.com/clauswilke/practical_ggplot2。

`ggplot2` 是十分方便的统计作图工具，相比 `Base R`，为了一张出版级的图形，不需要去调整每个参数，实现快速出图。集成了很多其它统计计算的 R 包，支持丰富的统计分析和计算功能，如回归、平滑等，实现了作图和模型的无缝连接。比如图7.1，使用 `loess` 局部多项式平滑得到数据的趋势，不仅仅是散点图，代码量也非常少。

```
ggplot(mpg, aes(displ, hwy)) +
  geom_point(aes(color = class)) +
  geom_smooth(se = TRUE, method = "loess") +
  labs(
    title = "Fuel efficiency generally decreases with engine size",
    subtitle = "Two seaters (sports cars) are an exception because of their light
    caption = "Data from fueleconomy.gov"
  )
```

故事源于一幅图片，我不记得第一次见到这幅图是什么时候了，只因多次在多个场合中见过，所以留下了深刻的印象，后来才知道它出自于一篇博文 – [Using R packages and education to scale Data Science at Airbnb](#)，作者 Ricardo Bion 还在其 [Github](#) 上传了相关代码³。除此之外还有几篇重要的参考资料：

1. Pablo Barberá 的 [Data Visualization with R and ggplot2](#)
2. Kieran Healy 的新书 [Data Visualization: A Practical Introduction](#)
3. Matt Leonawicz 的新作 [mapmate](#), 可以去其主页欣赏系列作品⁴
4. [tidytuesday 可视化挑战官方项目](#) 还有 [tidytuesday](#)
5. [ggstatsplot](#) 可视化统计检验、模型的结果

¹<http://varianceexplained.org/r/why-I-use-ggplot2/>

²<https://simplystatistics.org/2016/02/11/why-i-dont-use-ggplot2/>

³https://github.com/ricardo-bion/medium_visualization

⁴<https://leonawicz.github.io/>



Fuel efficiency generally decreases with engine size
Two seaters (sports cars) are an exception because of their light weight

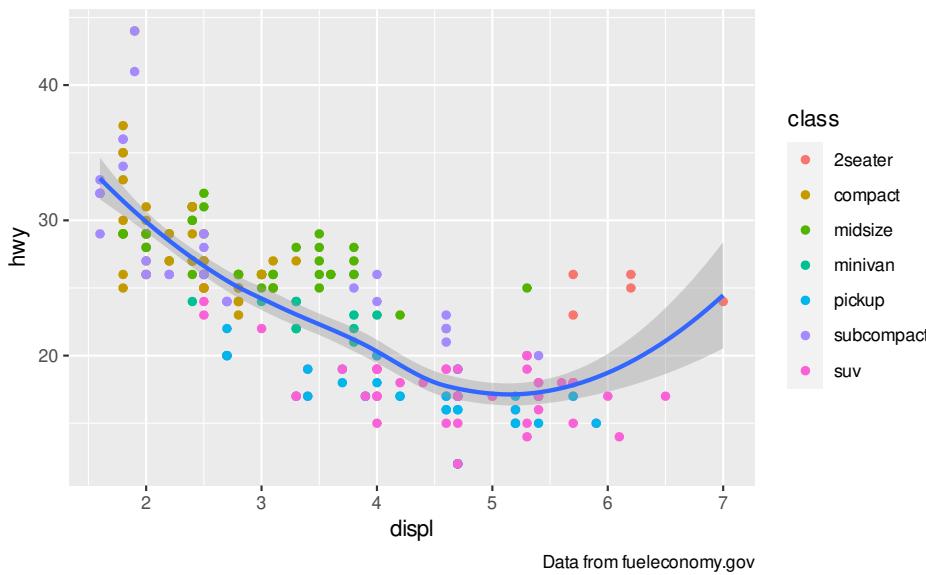


图 7.1: 简洁美观

6. [ggpubr](#) 制作出出版级统计图形
7. Thomas Lin Pedersen [Drawing Anything with ggplot2](#)
8. [Designing ggplots: making clear figures that communicate](#)
9. [ggh4x](#) 提供 ggplot2 的额外定制功能
10. [ggdist](#) Visualizations of distributions and uncertainty
11. [gghighlight](#)
12. [ggnetwork](#)
13. [ggPMX](#) ‘ggplot2’ Based Tool to Facilitate Diagnostic Plots for NLME Models
14. [ggpp](#) ggpp: Grammar Extensions to ‘ggplot2’

如 Berton Gunter 所说，数据可视化只是一种手段，根据数据实际情况作展示才是重要的，并不是要追求酷炫。

3-D bar plots are an abomination. Just because Excel can do them
doesn't mean you should. (Dismount pulpit).

— Berton Gunter⁵

grid 是 **lattice** 和 **ggplot2** 的基础，**gganimate** 是 **ggplot2** 一个扩展，它将静态图

⁵<https://stat.ethz.ch/pipermail/r-help/2007-October/142420.html>



形视为帧，调用第三方工具合成 GIF 动图或 MP4 视频等，要想深入了解 ggplot2，可以去看 Hadley Wickham, Danielle Navarro, and Thomas Lin Pedersen 合著的《ggplot2: elegant graphics for data analysis》第三版 <https://ggplot2-book.org/>。

7.1 元素

7.1.1 标签

图形的标签分为横纵轴标签、刻度标签、主标题、副标题等

```
data.frame(
  dates = seq.Date(
    from = as.Date("1945-01-01"),
    to = as.Date("1974-12-31"),
    by = "quarter"
  ),
  presidents = as.vector(presidents)
) |>
  ggplot(aes(x = dates, y = presidents)) +
  geom_line(color = "slategray", na.rm = TRUE) +
  geom_point(size = 1.5, color = "darkslategray", na.rm = TRUE) +
  scale_x_date(date_breaks = "4 year", date_labels = "%Y") +
  labs(
    title = "1945年至1974年美国总统每季度支持率",
    x = "年份", y = "支持率 (%)",
    caption = "数据源: R 包 datasets"
) +
  theme_minimal(base_size = 10.54, base_family = "source-han-sans-cn")
```

7.1.2 注释

图中注释的作用在于高亮指出关键点，提请读者注意。文本注释可由 **ggrepel** 包提供的标签图层 `geom_label_repel()` 添加，标签数据可独立于之前的数据层，标签所在的位置可以通过参数 `direction` 和 `nudge_y` 精调，图 7.3 模拟了一组数据。

云
相
互
连
网
C

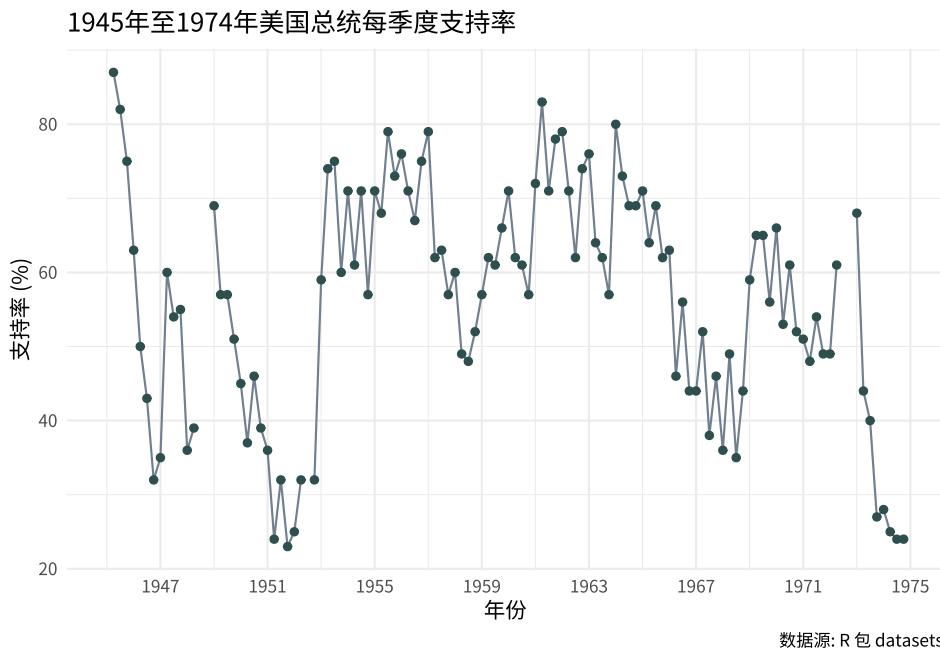


图 7.2: 美国总统支持率: 自 1945 年第一季度至 1974 年第四季度

```
set.seed(2020)
library(ggrepel)
dat <- data.frame(
  x = seq(100),
  y = cumsum(rnorm(100))
)
anno_data <- dat |>
  subset(x %% 25 == 10) |>
  transform(text = "text")

ggplot(data = dat, aes(x, y)) +
  geom_line() +
  geom_label_repel(aes(label = text),
    data = anno_data,
    direction = "y",
    nudge_y = c(-5, 5, 5, 5)
) +
```

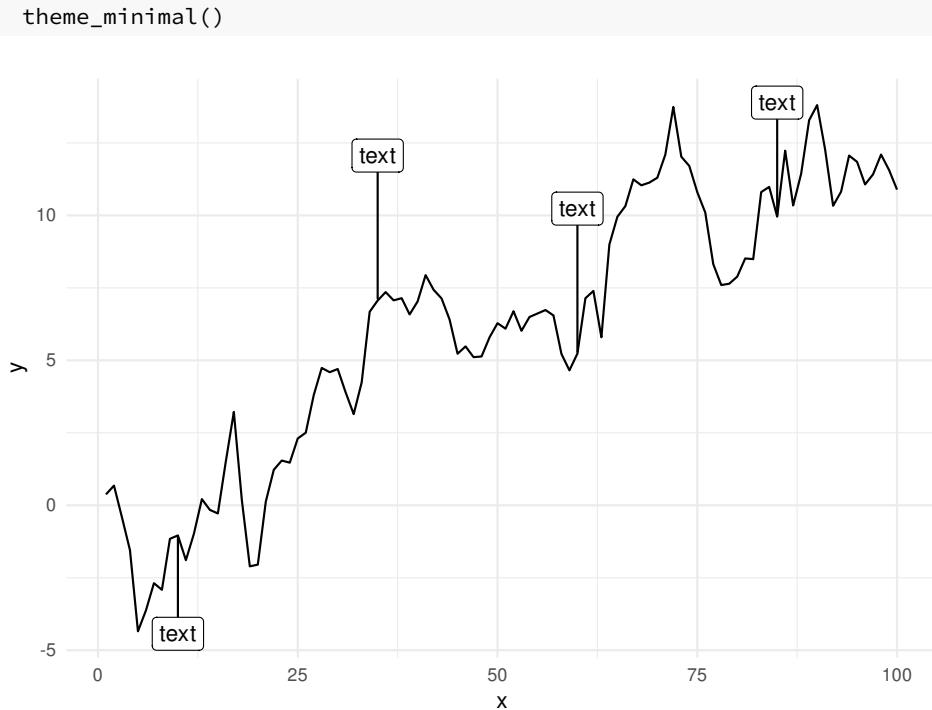


图 7.3: 文本注释

ggrepel 包的图层 `geom_text_repel()` 支持所有数据点的注释，并且自动调整文本的位置，防止重叠，增加辨识度，如图 7.4。当然，数据点如果过于密集也不适合全部注释，高亮其中的关键点即可。

```
mtcars |>  
  transform(cyl = as.factor(cyl)) |>  
  ggplot(aes(wt, mpg, label = rownames(mtcars), color = cyl)) +  
    geom_point() +  
    geom_text_repel(max.overlaps = 12) +  
    theme_minimal()
```

Claus Wilke 开发的 `ggttext` 包支持更加丰富的注释样式，详见网站 <https://wilkelab.org/ggttext/>

7.1.3 主题

`ggcharts` 和 `bbplot prettyB` 美化 Base R 图形 `ggprism`

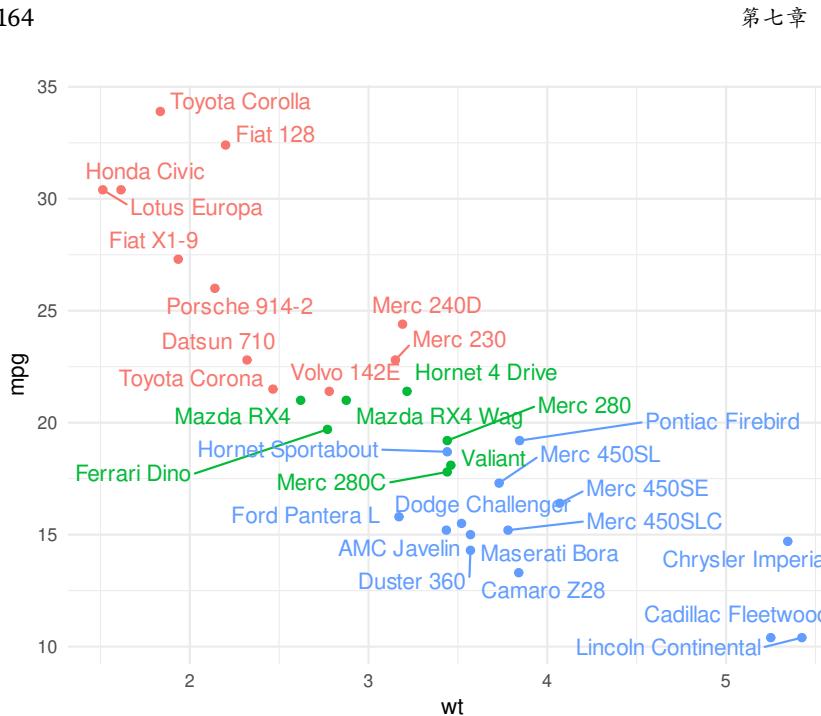


图 7.4: 少量点的情况下可以全部注释，且可以解决注释重叠的问题

7.2 字体

`firatheme` 包提供基于 fira sans 字体的 `ggplot2` 主题，类似的字体主题包还有 `trekfont`、`fontHind`、`fontquiver` 包与 `fontBitstreamVera` (Bitstream Vera 字体)、`fontLiberation` (Liberation 字体) 包和 `fontDejaVu` (DejaVu 字体) 包一道提供了一些可允许使用的字体文件，这样，我们可以不依赖系统制作可重复的图形。Thomas Lin Pedersen 开发的 `systemfonts` 可直接使用系统自带的字体。

7.2.1 系统字体

以 CentOS 系统为例，软件仓库中包含 `Noto`、`DejaVu`、`liberation` 等字体。可以安装自己喜欢的字体类型，比如：

```
sudo dnf install -y \
    google-noto-mono-fonts \
    google-noto-sans-fonts \
    google-noto-serif-fonts \
    dejavu-sans-mono-fonts \
```



```
dejavu-sans-fonts \
dejavu-serif-fonts
# 或者
sudo dnf install -y dejavu-fonts liberation-fonts
```

liberation 系列的四款字体可以用来替换 Windows 系统上对应的四款字体，对应关系见表 7.1

表 7.1: Windows 系统上四款字体的替代品

	CentOS 系统	Windows 系统
衬线体/宋体	liberation-serif-fonts	Times New Roman
无衬线体/黑体	liberation-sans-fonts	Arial
Arial 的细瘦版	liberation-narrow-fonts	Arial Narrow
等宽体/微软雅黑	liberation-mono-fonts	Courier New

此外，我们还可以从网上获取各种个样的字体，特别地，Boryslav Larin 收录的 [awesome-fonts](#) 列表是一个不错的开始，比如图标字体 [Font-Awesome](#)，

```
sudo dnf install -y fontawesome-fonts
```

再安装宏包 [fontawesome](#) 后，即可在 LaTeX 文档中使用，下面这个示例推荐用 XeLaTeX 引擎编译。

```
\documentclass[border=10pt]{standalone}
\usepackage{fontawesome}
\begin{document}
Hello, \faGithub
\end{document}
```

而在 R 绘制的图形中，通过指定 `par()`、`plot()`、`title()` 等函数的 `family` 参数值，比如 `family = "Liberation Sans"` 来调用系统无衬线 Liberation 字体，效果见图 7.5。

```
library(extrafont)
plot(data = pressure, pressure ~ temperature,
      xlab = "Temperature (deg C)", ylab = "Pressure (mm of Hg)",
      col.lab = "red", col.axis = "blue",
      font.lab = 3, font.axis = 2, family = "Liberation Sans")
```



```
title(main = "Vapor Pressure of Mercury as a Function of Temperature",
      family = "Liberation Serif", font.main = 3)
title(sub = "Data Source: Weast, R. C",
      family = "Liberation Mono", font.sub = 1)
```

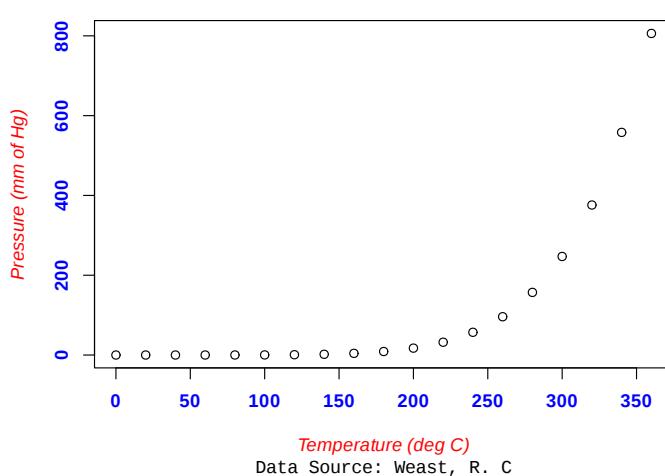


图 7.5: 调用系统字体绘图

为了符合出版的要求，需要在 7.5 中嵌入字体，

```
# embed fonts to pdf
embed_fonts <- function(fig_path) {
  if(knitr:::is_latex_output()){
    embedFonts(
      file = fig_path, outfile = fig_path,
      fontpaths = "~/Library/Fonts"
    )
  }
  return(fig_path)
}
```

设置代码块选项 `fig.process=embed_fonts`，这样生成 PDF 格式图形的时候，会调用此函数处理 PDF 图形。在 `ggplot2` 绘图中的调用方式是类似的，便不再赘述了。值得注意的是，`extrafont` 和 `showtext` 有些不一样，前者只能处理系统字体，后者还能获取网络字体和使用 OTF 字体，下面从 Google 开源的字体库获取



Noto 系列的四款字体，如图 7.6。

```
sysfonts::font_add_google(name = "Noto Sans", family = "Noto Sans")
sysfonts::font_add_google(name = "Noto Serif", family = "Noto Serif")
sysfonts::font_add_google(name = "Noto Serif SC", family = "Noto Serif SC")
sysfonts::font_add_google(name = "Noto Sans SC", family = "Noto Sans SC")
```

警告

在本书中，不要全局加载 `showtext` 包或调用 `showtext::showtext_auto()`，会和 `extrafont` 冲突，使得绘图时默认就只能使用 `showtext` 提供的字体。

```
p1 <- ggplot(pressure, aes(x = temperature, y = pressure)) +
  geom_point() +
  ggtitle(label = "默认字体设置")

p2 <- p1 + theme(
  axis.title = element_text(family = "Noto Sans"),
  axis.text = element_text(family = "Noto Serif"))
) +
  theme(
  title = element_text(family = "Noto Serif SC"))
) +
  ggtitle(label = "英文字体设置")

p3 <- p1 + labs(x = "温度", y = "压力") +
  theme(
  axis.title = element_text(family = "Noto Serif SC"),
  axis.text = element_text(family = "Noto Serif"))
) +
  ggtitle(label = "中文字体设置")

p4 <- p1 + labs(
  x = "温度", y = "压力", title = "散点图",
  subtitle = "Vapor Pressure of Mercury as a Function of Temperature",
  caption = paste("Data on the relation",
                 "between temperature in degrees Celsius and",
                 "vapor pressure of mercury in millimeters (of mercury).",
```

```

    sep = "\n"
)
) +
theme(
  axis.title = element_text(family = "Noto Serif SC"),
  axis.text.x = element_text(family = "Noto Serif"),
  axis.text.y = element_text(family = "Noto Sans"),
  title = element_text(family = "Noto Serif SC"),
  plot.subtitle = element_text(family = "Noto Sans", size = rel(0.7)),
  plot.caption = element_text(family = "Noto Sans", size = rel(0.6))
) +
ggtitle(label = "任意字体设置")

(p1 + p2) / (p3 + p4)

```

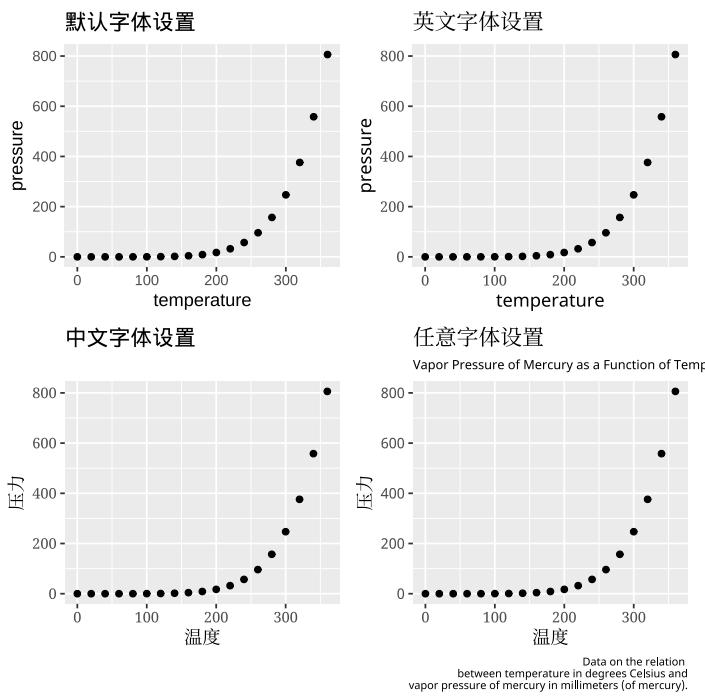


图 7.6: 在 ggplot2 绘图系统中设置中英文字体

另外值得一提的是 `hrbrthemes` 包，除了定制了很多 `ggplot2` 主题，它还打包了很多的字体主题。比如默认主题 `theme_ipsum()` 使用 Arial Narrow 字体，如果没有



该字体就自动寻找系统中的替代品，如图 7.7 实际使用的是 Nimbus Sans Narrow 字体，因为在 GitHub Action 中，我实际使用的测试环境是 Ubuntu 20.04，该系统自带 Nimbus Sans Narrow 字体，Arial Narrow 毕竟是 Windows 上的闭源字体。

```
# brew install font-roboto
# 导入字体
# hrbrthemes::import_roboto_condensed()
sysfonts::font_add_google(name = "Roboto Condensed", family = "Roboto Condensed")

library(hrbrthemes)
ggplot(mtcars, aes(mpg, wt)) +
  geom_point() +
  labs(
    x = "Fuel efficiency (mpg)", y = "Weight (tons)",
    title = "Seminal ggplot2 scatterplot example",
    subtitle = "A plot that is only useful for demonstration purposes",
    caption = "Brought to you by the letter 'g'"
  ) +
  theme_ipsum(base_family = "Roboto Condensed")
```

Seminal ggplot2 scatterplot example

A plot that is only useful for demonstration purposes

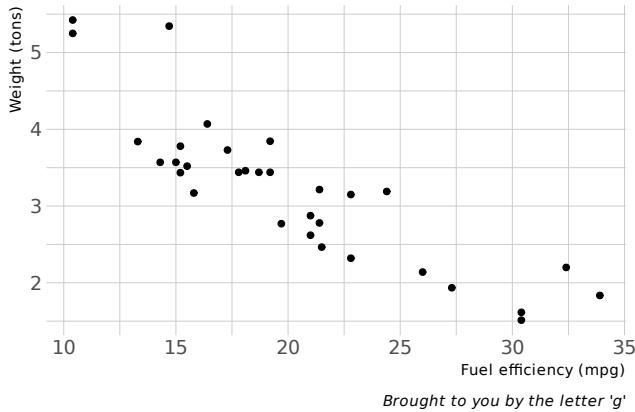


图 7.7: 调用 hrbrthemes 包设置字体主题

如果系统没有安装 Arial Narrow 字体，可以导入 hrbrthemes 包自带的一些



字体，比如 `hrbrthemes::import_roboto_condensed()`，然后调用字体主题 `theme_ipsum_rc()`。如果不使用这个包自带的字体，可以用系统中安装的字体去修改主题 `theme_ipsum()` 和 `theme_ipsum_rc()` 中的字体设置。如图 7.8 使用了 `theme_ipsum()` 中的 Arial Narrow 字体。

```
ggplot(mtcars, aes(mpg, wt)) +
  geom_point() +
  labs(
    x = "Fuel efficiency (mpg)", y = "Weight (tons)",
    title = "Seminal ggplot2 scatterplot example",
    subtitle = "A plot that is only useful for demonstration purposes",
    caption = "Brought to you by the letter 'g'"
  ) +
  theme_ipsum()
```

Seminal ggplot2 scatterplot example

A plot that is only useful for demonstration purposes

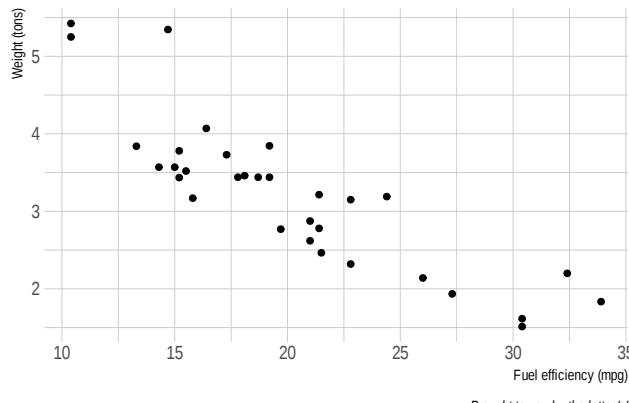


图 7.8: 默认字体 Arial Narrow

提示

hrbrthemes 包提供了一个全局字体加载选项 `hrbrthemes.loadfonts`，如果设置为 TRUE，即 `options(hrbrthemes.loadfonts = TRUE)` 会先调用函数 `extrafont::loadfonts()` 预加载系统字体，就不用一次次手动加载字体了。后续在第 7.2.3 节还会提及 `extrafont` 包的其它功能。



7.2.2 思源字体

邱怡轩开发的 `showtext` 包支持丰富的外部字体，支持 Base R 和 ggplot2 图形，图 7.9 嵌入了 5 号思源宋体，图例和坐标轴文本使用 serif 字体，更多详细的使用文档见 [Qiu, 2015]。

```
# 安装 showtext 包
install.packages('showtext')
# 思源宋体
showtextdb::font_install(showtextdb::source_han_serif())
# 思源黑体
showtextdb::font_install(showtextdb::source_han_sans())

ggplot(iris, aes(Sepal.Length, Sepal.Width)) +
  geom_point(aes(colour = Species)) +
  scale_colour_brewer(palette = "Set1") +
  labs(
    title = "鸢尾花数据的散点图",
    x = "萼片长度", y = "萼片宽度", colour = "鸢尾花类别",
    caption = "鸢尾花数据集最早见于 Edgar Anderson (1935) "
  ) +
  theme(
    title = element_text(family = "source-han-sans-cn"),
    axis.title = element_text(family = "source-han-serif-cn"),
    legend.title = element_text(family = "source-han-serif-cn")
  )
```

斐济是太平洋上的一个岛国，受地壳板块运动的影响，地震活动频繁，图 7.10 清晰展示了它的地震带。

```
library(maps)
library(mapdata)
FijiMap <- map_data("worldHires", region = "Fiji")
ggplot(FijiMap, aes(x = long, y = lat)) +
  geom_map(map = FijiMap, aes(map_id = region), size = .2) +
  geom_point(data = quakes, aes(x = long, y = lat, colour = mag)) +
  xlim(160, 195) +
  scale_colour_distiller(palette = "Spectral") +
  scale_y_continuous(breaks = (-18:18) * 5) +
```

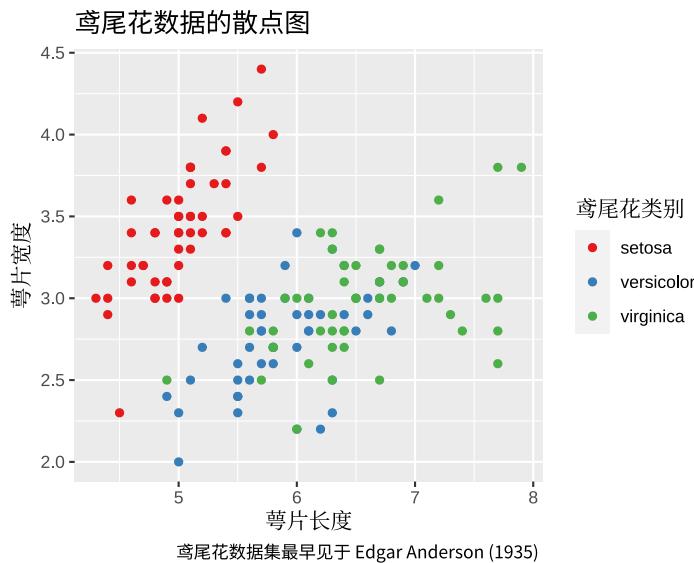


图 7.9: showtext 包处理图里的中文

```
coord_map("ortho", orientation = c(-10, 180, 0)) +
  labs(colour = "震级", x = "经度", y = "纬度", title = "斐济地震带") +
  theme_minimal() +
  theme(
    title = element_text(family = "source-han-sans-cn"),
    axis.title = element_text(family = "source-han-serif-cn"),
    legend.title = element_text(family = "source-han-sans-cn"),
    legend.position = c(1, 0), legend.justification = c(1, 0)
  )
```

7.2.3 数学字体

Winston Chang 将 Paul Murrell 的 Computer Modern 字体文件打包成 **fontcm** 包 [Chang et al., 2014]，**fontcm** 包可以在 Base R 图形中嵌入数学字体⁶，图形中嵌入重音字符⁷。下面先下载、安装、加载字体，

⁶<https://www.stat.auckland.ac.nz/~paul/R/CM/CMR.html>

⁷<https://www.stat.auckland.ac.nz/~paul/Reports/maori/maori.html>

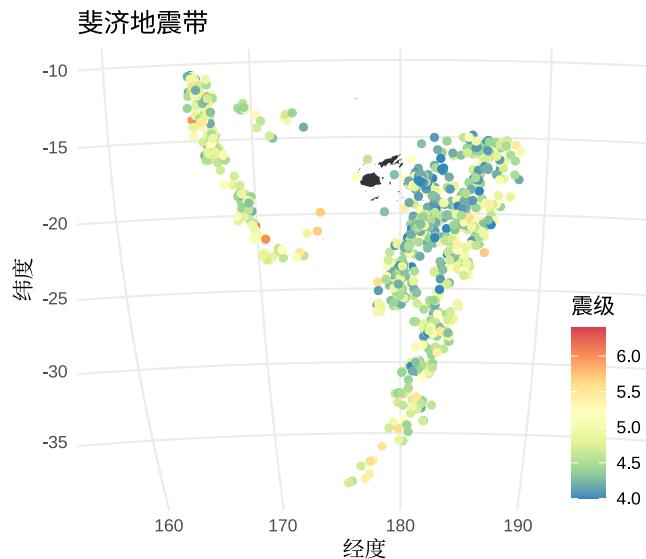


图 7.10: 斐济地震带

```
library(extrafont)
font_addpackage(pkg = "fontcm")
```

查看可被 `pdf()` 图形设备使用的字体列表

```
# 可用的字体
```

```
fonts()
```

```
## [1] "Roboto Condensed"      "xkcd"                  "CM Roman"
## [4] "CM Roman Asian"        "CM Roman CE"          "CM Roman Cyrillic"
## [7] "CM Roman Greek"        "CM Sans"                "CM Sans Asian"
## [10] "CM Sans CE"            "CM Sans Cyrillic"     "CM Sans Greek"
## [13] "CM Symbol"              "CM Typewriter"         "CM Typewriter Asian"
## [16] "CM Typewriter CE"       "CM Typewriter Cyrillic" "CM Typewriter Greek"
```

`fontcm` 包提供数学字体，`grDevices::embedFonts()` 函数调用 `Ghostscript` 软件将数学字体嵌入 `ggplot2` 图形中，达到正确显示数学公式的目的，此方法适用于 `pdf` 设备保存的图形，对 `cairo_pdf()` 保存的 PDF 格式图形无效。

```
library(fontcm)
library(ggplot2)
library(extrafont)
```



```

library(patchwork)
p <- ggplot(
  data = data.frame(x = c(1, 5), y = c(1, 5)),
  aes(x = x, y = y)
) +
  geom_point() +
  labs(
    x = "Made with CM fonts", y = "Made with CM fonts",
    title = "Made with CM fonts"
  )
# 公式
eq <- "italic(sum(frac(1, n*'!'), n==0, infinity) ==
         lim(bgroup('(', 1 + frac(1, n), ')')^n, n %-%>% infinity))"
# 默认字体
p1 <- p + annotate("text",
  x = 3, y = 3,
  parse = TRUE, label = eq
)
# 使用 CM Roman 字体
p2 <- p + annotate("text",
  x = 3, y = 3,
  parse = TRUE, label = eq, family = "CM Roman"
) +
  theme(
    text = element_text(size = 10, family = "CM Roman"),
    axis.title.x = element_text(face = "italic"),
    axis.title.y = element_text(face = "bold")
  )
p1 + p2

```

为实现图 7.11 的最终效果，需要启用一个有超级牛力的 `fig.process` 选项，主要是传递一个函数给它，对用 R 语言生成的图形再操作。

```

# embed math fonts to pdf
embed_math_fonts <- function(fig_path) {
  if(knitr:::is_latex_output()){
    embedFonts(

```

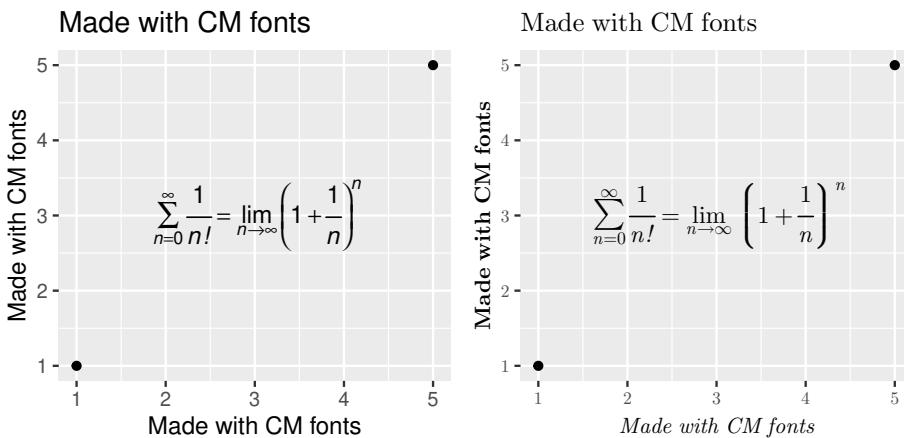


图 7.11: fontcm 处理数学公式

```

    file = fig_path, outfile = fig_path,
    fontpaths = system.file("fonts", package = "fontcm")
)
}

return(fig_path)
}

```

代码块选项中设置 `fig.process=embed_math_fonts` 可在绘图后，立即插入字体，此操作仅限于以 `pdf` 格式保存的图形设备，也适用于 Base R 绘制的图形，见图 7.12。

```

par(mar = c(4.1, 4.1, 1.5, 0.5), family = "CM Roman")
x <- seq(-4, 4, len = 101)
y <- cbind(sin(x), cos(x))
matplot(x, y,
        type = "l", xaxt = "n",
        main = expression(paste(
            plain(sin) * phi, " and ",
            plain(cos) * phi
        )),
        ylab = expression("sin" * phi, "cos" * phi),
        xlab = expression(paste("Phase Angle ", phi)),
        col.main = "blue"
)

```

© 黄湘云

```
)  
axis(1,  
  at = c(-pi, -pi / 2, 0, pi / 2, pi),  
  labels = expression(-pi, -pi / 2, 0, pi / 2, pi))
```

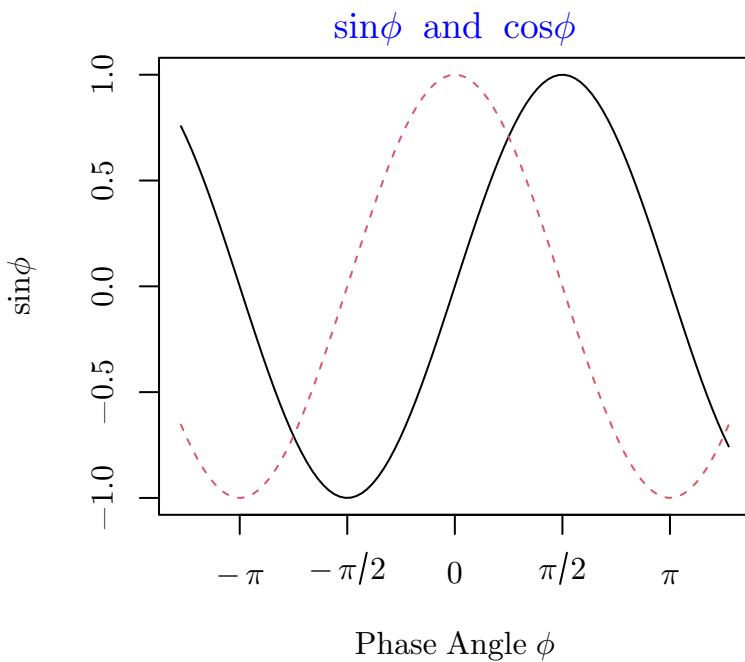


图 7.12: 嵌入数学字体

7.2.4 TikZ 设备

与 7.2.3 小节不同, Ralf Stubner 维护的 **tikzDevice** 包提供了另一种嵌入数学字体的方式, 其提供的 `tikzDevice::tikz()` 绘图设备将图形对象转化为 TikZ 代码, 调用 LaTeX 引擎编译成 PDF 文档。安装后, 先测试一下 LaTeX 编译环境是否正常。

```
tikzDevice::tikzTest()  
  
##  
## Active compiler:  
## /home/runner/.TinyTeX/bin/x86_64-linux/xelatex
```



```
##  XeTeX 3.141592653-2.6-0.999993 (TeX Live 2021)
##  kpathsea version 6.3.3
## [1] 7.90259
```

确认没有问题后，下面图 7.13 的坐标轴标签，标题，图例等位置都支持数学公式，使用 **tikzDevice** 打造出版级的效果图。更多功能的介绍见 <https://www.daqana.org/tikzDevice/>。

```
x <- rnorm(10)
y <- x + rnorm(5, sd = 0.25)
model <- lm(y ~ x)
rsq <- summary(model)$r.squared
rsq <- signif(rsq, 4)
plot(x, y,
      main = "Hello \\LaTeX!",
      xlab = "$x$",
      ylab = "$y$",
      sub = "$\\mathcal{N}(x; \\mu, \\Sigma)$")
abline(model, col = "red")
mtext(paste0("Linear model: $R^2=", rsq, "$"),
      line = 0.5)
legend("bottomright",
      legend = paste0(
        "$y = ",
        round(coef(model)[2], 3),
        "x +",
        round(coef(model)[1], 3),
        "$"),
      bty = "n")
```

推荐的全局 LaTeX 环境配置如下：

```
options(
  tinytex.engine = "xelatex",
  tikzDefaultEngine = "xetex",
  tikzDocumentDeclaration = "\\documentclass[tikz]{standalone}\n",
  tikzXelatexPackages = c(
    "\\usepackage[fontset=adobe]{ctex}",
```

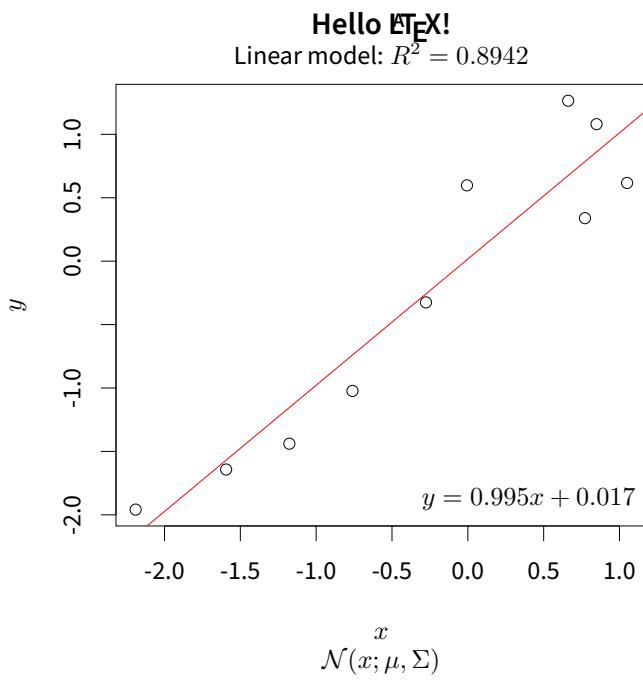


图 7.13: 线性回归模型



```
"\\usepackage[default,semibold]{sourcesanspro}",
"\\usepackage{amsfonts,mathrsfs,amssymb}\\n"
)
)
```

设置默认的 LaTeX 编译引擎为 XeLaTeX，相比于 PDFLaTeX，它对中文的兼容性更好，支持多平台下的中文环境，中文字体这里采用了 Adobe 的字体，默认加载了 `mathrsfs` 宏包支持 `\mathcal`、`\mathscr` 等命令，此外，LaTeX 发行版采用谢益辉自定义的 [TinyTeX](#)。绘制独立的 PDF 图形的过程如下：

```
library(tikzDevice)
tf <- file.path(getwd(), "tikz-regression.tex")
tikz(tf, width = 6, height = 5.5, pointsize = 30, standAlone = TRUE)
# 绘图代码
dev.off()
# 编译成 PDF 图形
tinytex::latexmk(file = "tikz-regression.tex")
```

7.2.5 漫画字体

下载 XKCD 字体，并刷新系统字体缓存

```
mkdir -p ~/.fonts
curl -fLo ~/.fonts/xkcd.ttf http://simonsoftware.se/other/xkcd.ttf
fc-cache -fsv
```

将 XKCD 字体导入到 R 环境，以便后续被 `ggplot2` 图形设备调用。

```
R -e 'library(extrafont);font_import(pattern="[X/x]kcd.ttf", prompt = FALSE)'
```

图 7.14 是一个使用 `xkcd` 字体的简单例子，更多高级特性请看 `xkcd` 包文档 [[Torres-Manzanera, 2018](#)]

```
library(extrafont)
library(xkcd)
ggplot(aes(mpg, wt), data = mtcars) +
  geom_point() +
  theme_xkcd()
```

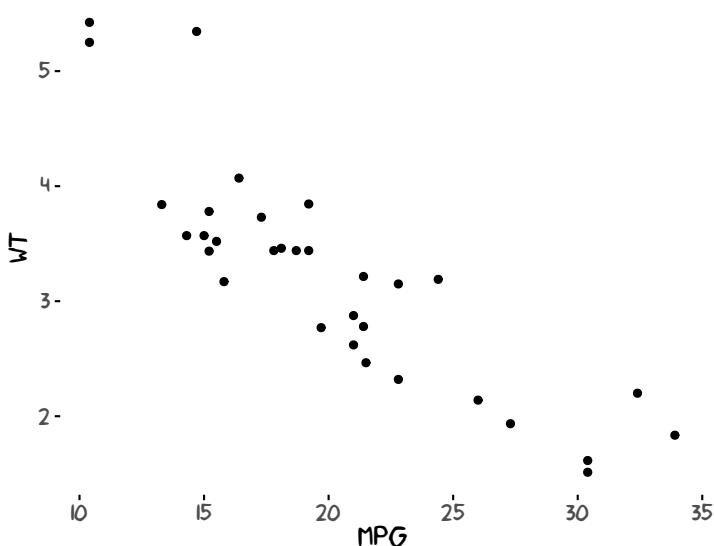


图 7.14: 漫画风格的字体方案

7.2.6 表情字体

余光创开发的 `emojifont` 包和 Hadley 开发的 `emo` 包，下面使用 Noto Emoji 字体，支持的表情图见 <https://www.google.com/get/noto/help/emoji/food-drink/>，下面给出一个示例。先从 GitHub 安装 `emo` 包，目前它还未正式发布到 CRAN 上。

```
remotes::install_github("hadley/emo")
```

除了安装 `emo` 包，系统需要先安装好 `emoji` 字体，图形才会正确地渲染出来，想调用更多 `emoji` 图标请参考 [Emoji 速查手册](#)，给出 `emoji` 对应的名字。

```
# CentOS
sudo dnf install -y google-noto-emoji-color-fonts \
google-noto-emoji-fonts
# MacOS
brew cask install font-noto-color-emoji font-noto-emoji
```

```
data.frame(
  category = c("pineapple", "apple", "watermelon", "mango", "pear"),
  value = c(5, 4, 3, 6, 2)
) |>
  transform(category = sapply(category, emo::ji)) |>
```

```
ggplot(aes(x = category, y = value)) +  
  scale_y_continuous(limits = c(2, 7)) +  
  geom_text(aes(label = category), size = 12, vjust = -0.5) +  
  theme_minimal()
```

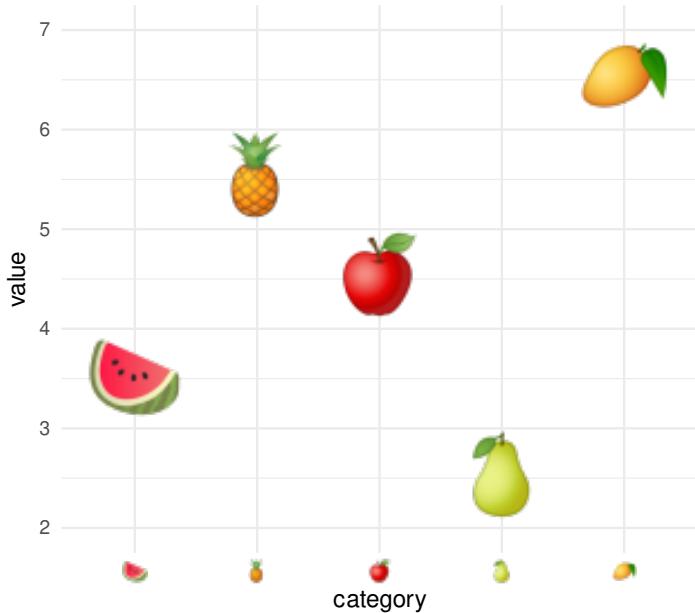


图 7.15: 表情字体

7.3 配色

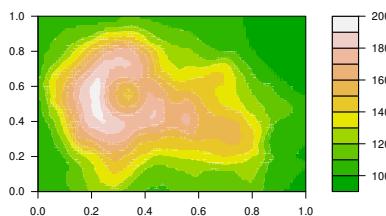
配色真的是一门学问，有的人功力非常深厚，仅用黑白灰就可以创造出一个世界，如中国的水墨画，科波拉执导的《教父》，沃卓斯基姐妹执导的《黑客帝国》等。黑西装、白衬衫和黑领带是《黑客帝国》的经典元素，《教父》开场的黑西装、黑领结和白衬衫，尤其胸前的红玫瑰更是点睛之笔。导演将黑白灰和光影混合形成了层次丰富立体的画面，打造了一场视觉盛宴，无论是呈现在纸上还是银幕上都可以给人留下深刻的印象。正所谓食色性也，花花世界，岂能都是法印眼中的白骨！再说《红楼梦》里，芍药丛中，桃花树下，滴翠亭边，栊翠庵里，处处都是湘云、黛玉、宝钗、妙玉留下的四季诗歌。

为什么需要这么多颜色模式呢？主要取决于颜色输出的通道，比如印刷机，照相机，自然界，网页，人眼等，显示器因屏幕和分辨率的不同呈现的色彩数量是不

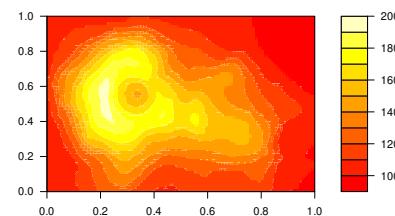


一样的。读者大概都听说过 RGB、CMYK、AdobeRGB、sRGB、P3 广色域等名词，我想这主要归功于各大电子设备厂商的宣传。普清、高清、超高清、全高清、2K、4K、5K、视网膜屏，而 HSV、HCL 估计听说的人就少很多了。本节的目的是简单阐述背后的色彩原理，颜色模式及其之间的转化，在应对天花乱坠的销售时少交一些智商税，同时，告诉读者如何在 R 环境中使用色彩。早些时候我在统计之都论坛上发帖 - R 语言绘图用调色板大全 <https://d.cosx.org/d/419378>，如果读者希望拿来即用，不妨去看看。

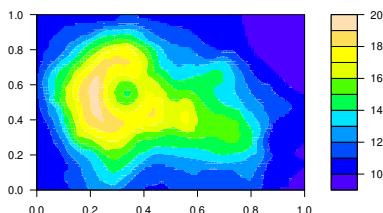
```
filled.contour(volcano, nlevels = 10, color.palette = terrain.colors)
filled.contour(volcano, nlevels = 10, color.palette = heat.colors)
filled.contour(volcano, nlevels = 10, color.palette = topo.colors)
filled.contour(volcano, nlevels = 10, color.palette = cm.colors)
```



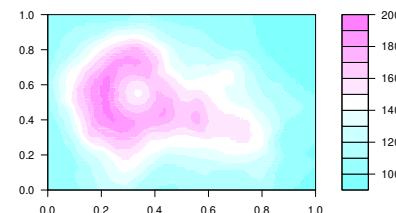
(a) terrain.colors 调色板



(b) heat.colors 调色板



(c) topo.colors 调色板



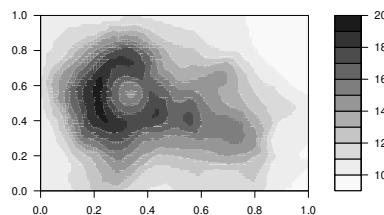
(d) cm.colors 调色板

图 7.16: R 3.6.0 以前的调色板

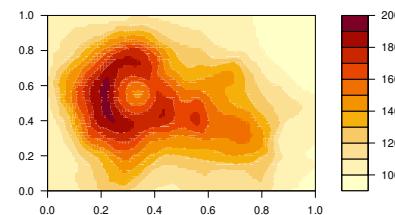
```
filled.contour(volcano,
  nlevels = 10,
  color.palette = function(n, ...) hcl.colors(n, "Grays", rev = TRUE, ...)
)
```



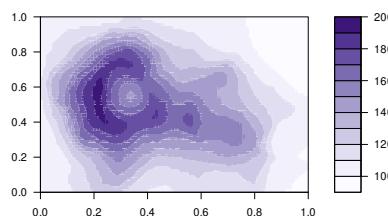
```
filled.contour(volcano,
  nlevels = 10,
  color.palette = function(n, ...) hcl.colors(n, "YlOrRd", rev = TRUE, ...)
)
filled.contour(volcano,
  nlevels = 10,
  color.palette = function(n, ...) hcl.colors(n, "purples", rev = TRUE, ...)
)
filled.contour(volcano,
  nlevels = 10,
  color.palette = function(n, ...) hcl.colors(n, "viridis", rev = FALSE, ...)
```



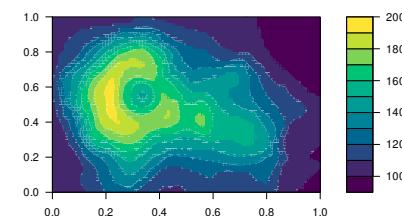
(a) Grays 调色板



(b) YlOrRd 调色板



(c) Purples 3 调色板



(d) Viridis 调色板

图 7.17: R 3.6.0 以后的调色板



注意

`hcl.colors()` 函数是在 R 3.6.0 引入的，之前的 R 软件版本中没有，同时内置了 110 个调色板，详见 `hcl.pals()`。

7.3.1 调色板

R 预置的灰色有 224 种，挑出其中的调色板

```
grep("^(gr(a|e)y)", grep("gr(a|e)y", colors(), value = TRUE),
     value = TRUE, invert = TRUE)

## [1] "darkgray"      "darkgrey"       "darkslategray"  "darkslategray1"
## [5] "darkslategray2" "darkslategray3" "darkslategray4"  "darkslategray"
## [9] "dimgray"        "dimgrey"        "lightgray"      "lightgrey"
## [13] "lightslategray" "lightslategrey" "slategray"      "slategray1"
## [17] "slategray2"     "slategray3"     "slategray4"     "slategrey"

gray_colors <- paste0(rep(c("slategray", "darkslategray"), each = 4), seq(4))
barplot(1:8, col = gray_colors, border = NA)
```

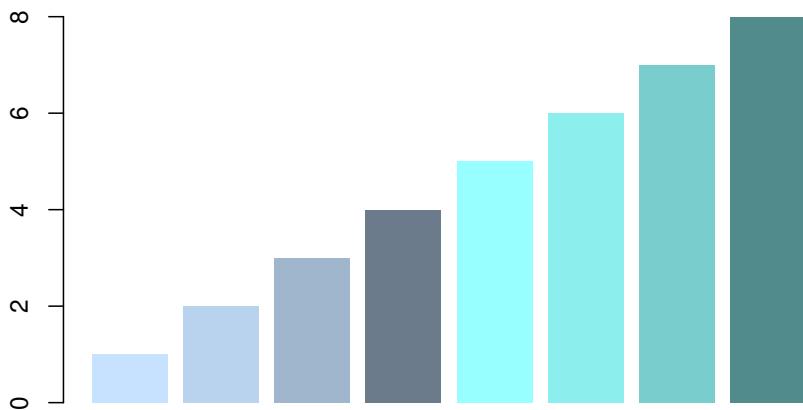


图 7.18: 灰度调色板



gray 与 grey 是一样的，类似 color 和 colour 的关系，可能是美式和英式英语的差别，且看

```
all.equal(col2rgb(paste0("gray", seq(100))), col2rgb(paste0("grey", seq(100))))  
## [1] TRUE  
  
gray100 代表白色，gray0 代表黑色，提取灰色调色板，去掉首尾部分是必要的  
barplot(1:8, col = gray.colors(8, start = .3, end = .9),  
        main = "gray.colors function", border = NA)
```

gray.colors function

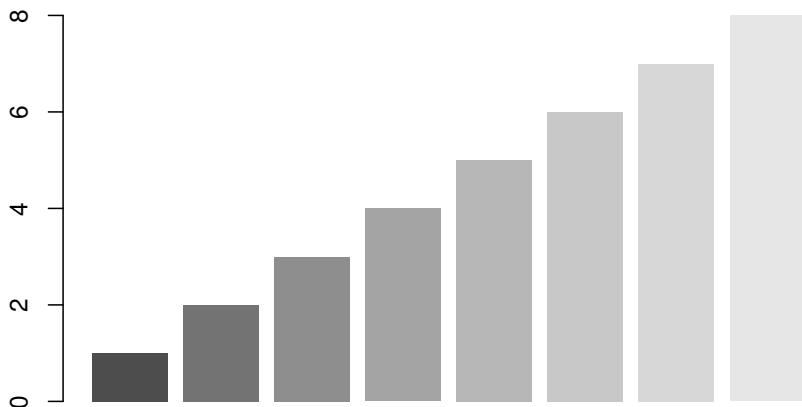


图 7.19: 提取 10 种灰色做调色板

首先选择一组合适的颜色，比如从桃色到梨色，选择 6 种颜色，以此为基础，可以借助 `grDevices::colorRampPalette()` 函数扩充至想要的数目，用 `graphics::rect()` 函数预览这组颜色配制的调色板

```
# Colors from https://github.com/johannesbjork/LaCroixColor  
colors_vec <- c("#FF3200", "#E9A17C", "#E9E4A6",  
                 "#1BB6AF", "#0076BB", "#172869")  
  
# 代码来自 ?colorspace::rainbow_hcl  
pal <- function(n = 20, colors = colors, border = "light gray", ...) {
```

```

colorname <- (grDevices::colorRampPalette(colors))(n)
plot(0, 0,
  type = "n", xlim = c(0, 1), ylim = c(0, 1),
  axes = FALSE, ...
)
rect(0:(n - 1) / n, 0, 1:n / n, 1, col = colorname, border = border)
}
par(mar = rep(0, 4))
pal(n = 20, colors = colors_vec, xlab = "Colors from Peach to Pear", ylab = "")

```

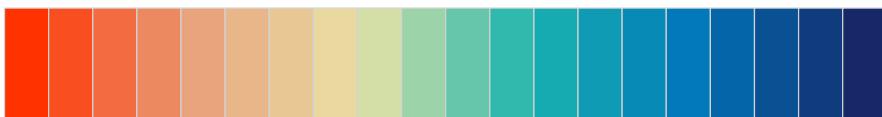


图 7.20: 桃色至梨色的渐变

colorRampPalette() 自制调色板

```

create_palette <- function(n = 1000, colors = c("blue", "orangeRed")) {
  color_palette <- colorRampPalette(colors)(n)
  barplot(rep(1, times = n), col = color_palette,
         border = color_palette, axes = FALSE)
}

par(mfrow = c(3, 1), mar = c(0.1, 0.1, 0.5, 0.1), xaxs = "i", yaxs = "i")
create_palette(n = 1000, colors = c("blue", "orangeRed"))
create_palette(n = 1000, colors = c("darkgreen", "yellow", "orangered"))
create_palette(n = 1000, colors = c("blue", "white", "orangered"))

par(mar = c(0, 4, 0, 0))
RColorBrewer::display.brewer.all()

```

```

# 代码来自 ?palettes
demo.pal <- function(n, border = if (n < 32) "light gray" else NA,
                      main = paste("color palettes: alpha = 1, n=", n),
                      ch.col = c(
                        "rainbow(n, start=.7, end=.1)", "heat.colors(n)",
                        "terrain.colors(n)", "topo.colors(n)",
                        "cm.colors(n)", "gray.colors(n, start = 0.3, end = 0.9)")

```

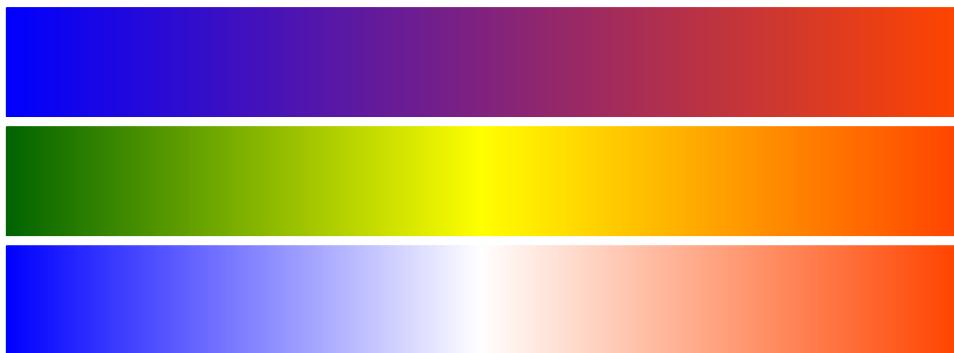


图 7.21: colorRampPalette 自制调色板

```
    )) {
nt <- length(ch.col)
i <- 1:n
j <- n / nt
d <- j / 6
dy <- 2 * d
plot(i, i + d, type = "n", axes = FALSE, ylab = "", xlab = "", main = main)
for (k in 1:nt) {
  rect(i - .5, (k - 1) * j + dy, i + .4, k * j,
        col = eval(parse(text = ch.col[k])), border = border
  )
  text(2 * j, k * j + dy / 4, ch.col[k])
}
}

n <- if (.Device == "postscript") 64 else 16
# Since for screen, larger n may give color allocation problem
par(mar = c(0, 0, 2, 0))
demo.pal(n)

par(mfrow = c(33, 1), mar = c(0, 0, .8, 0))
for (i in seq(32)) {
  pal(
    n = length((1 + 20 * (i - 1)):(20 * i)),
    colors()[(1 + 20 * (i - 1)):(20 * i)],
    main = paste(1 + 20 * (i - 1), "to", 20 * i)
```

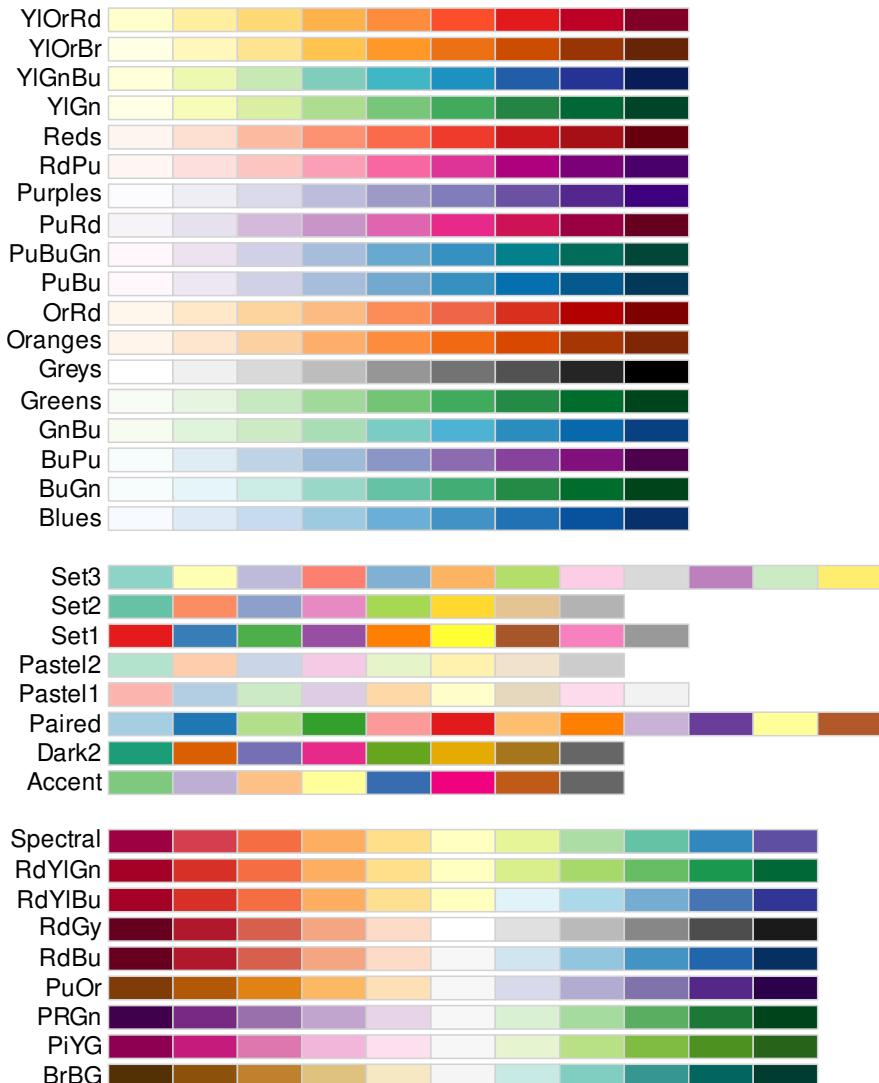


图 7.22: RColorBrewer 调色板

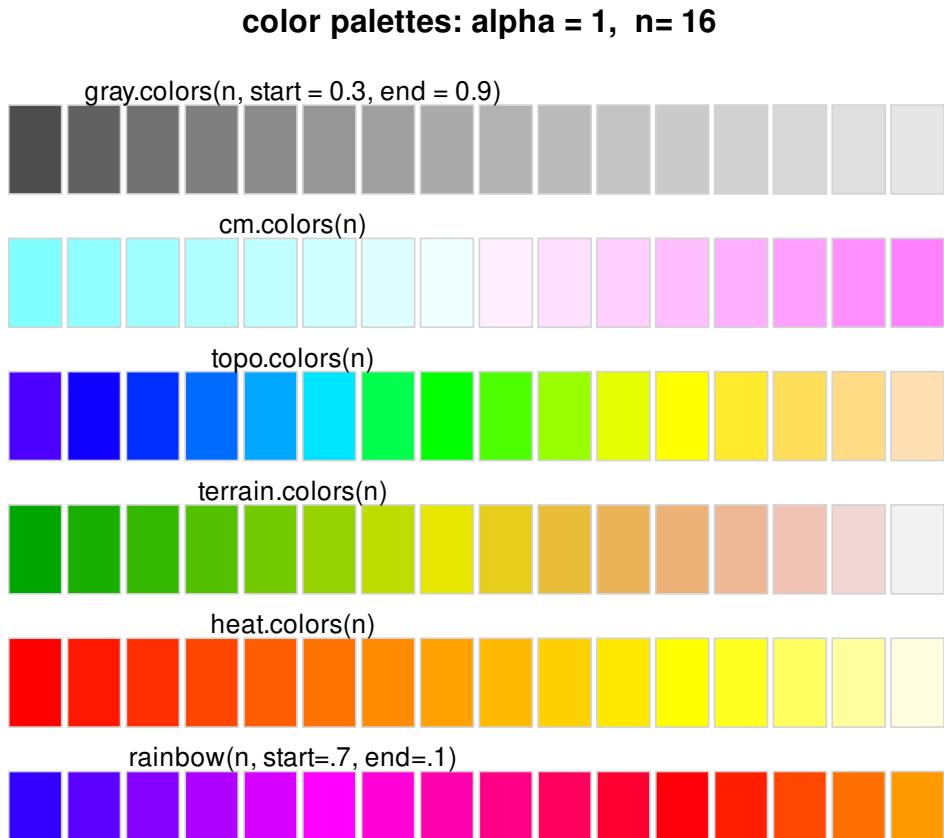


图 7.23: grDevices 调色板



```

)
}

pal(n = 17, colors()[641:657], main = "641 to 657")

library(colorspace)

## a few useful diverging HCL palettes
par(mar = c(0,0,2,0), mfrw = c(16, 2))

pal(n = 16, diverge_hcl(16), main = "diverging HCL palettes")
pal(n = 16, diverge_hcl(16, h = c(246, 40), c = 96, l = c(65, 90)))
pal(n = 16, diverge_hcl(16, h = c(130, 43), c = 100, l = c(70, 90)))
pal(n = 16, diverge_hcl(16, h = c(180, 70), c = 70, l = c(90, 95)))

pal(n = 16, diverge_hcl(16, h = c(180, 330), c = 59, l = c(75, 95)))
pal(n = 16, diverge_hcl(16, h = c(128, 330), c = 98, l = c(65, 90)))
pal(n = 16, diverge_hcl(16, h = c(255, 330), l = c(40, 90)))
pal(n = 16, diverge_hcl(16, c = 100, l = c(50, 90), power = 1))

## sequential palettes
pal(n = 16, sequential_hcl(16), main= "sequential palettes")
pal(n = 16, heat_hcl(16, h = c(0, -100),
                     l = c(75, 40), c = c(40, 80), power = 1))
pal(n = 16, terrain_hcl(16, c = c(65, 0), l = c(45, 95), power = c(1/3, 1.5)))
pal(n = 16, heat_hcl(16, c = c(80, 30), l = c(30, 90), power = c(1/5, 1.5)))

## compare base and colorspace palettes
## (in color and desaturated)
## diverging red-blue colors
pal(n = 16, diverge_hsv(16), main = "diverging red-blue colors")
pal(n = 16, diverge_hcl(16, c = 100, l = c(50, 90)))
pal(n = 16, desaturate(diverge_hsv(16)))
pal(n = 16, desaturate(diverge_hcl(16, c = 100, l = c(50, 90)))))

## diverging cyan-magenta colors
pal(n = 16, cm.colors(16), main = "diverging cyan-magenta colors")
pal(n = 16, diverge_hcl(16, h = c(180, 330), c = 59, l = c(75, 95)))

```

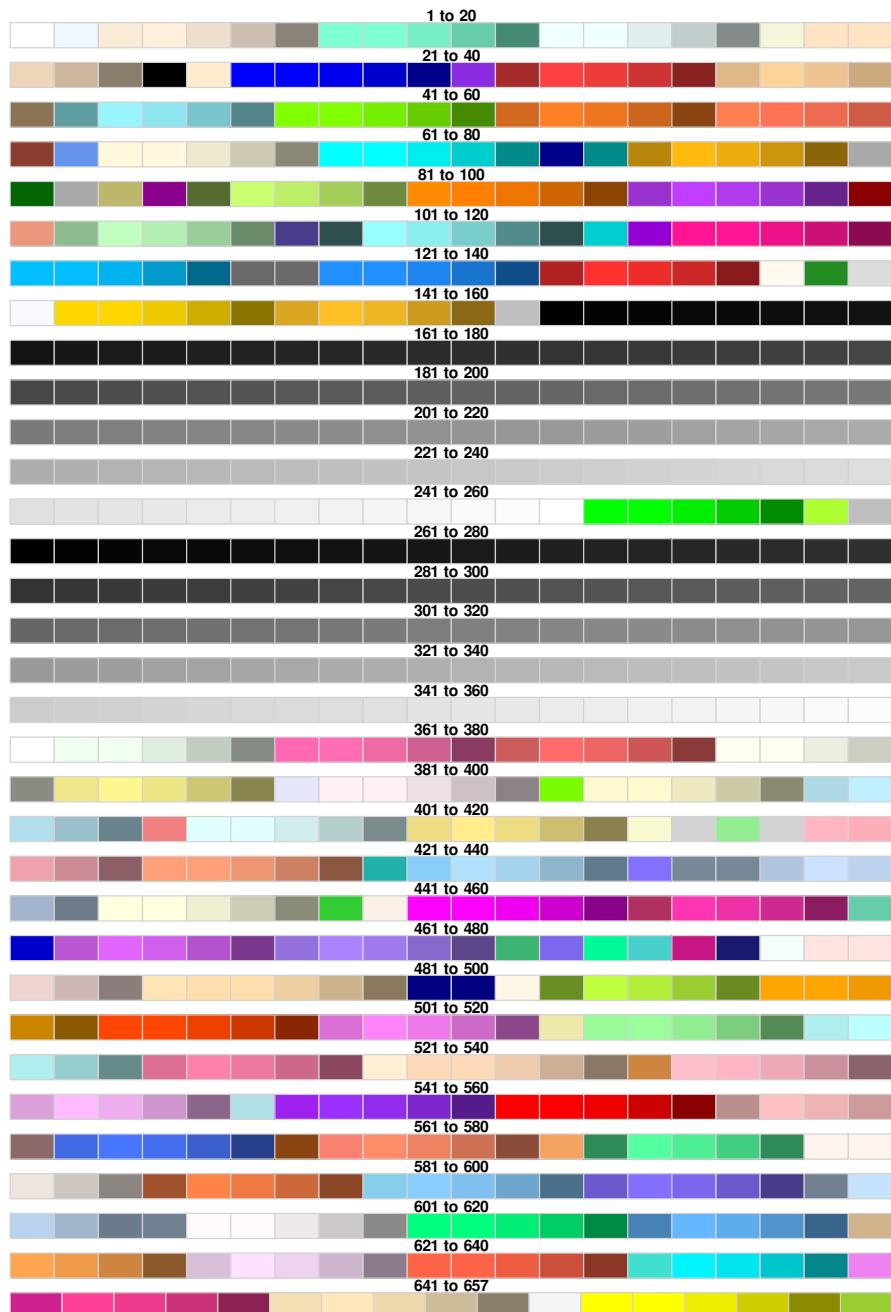


图 7.24: grDevices 调色板



```

pal(n = 16, desaturate(cm.colors(16)))
pal(n = 16, desaturate(diverge_hcl(16, h = c(180, 330), c = 59, l = c(75, 95)))

## heat colors
pal(n = 16, heat.colors(16), main = "heat colors")
pal(n = 16, heat_hcl(16))
pal(n = 16, desaturate(heat.colors(16)))
pal(n = 16, desaturate(heat_hcl(16)))

## terrain colors
pal(n = 16, terrain.colors(16), main = "terrain colors")
pal(n = 16, terrain_hcl(16))
pal(n = 16, desaturate(terrain.colors(16)))
pal(n = 16, desaturate(terrain_hcl(16)))

pal(n = 16, rainbow_hcl(16, start = 30, end = 300), main = "dynamic")
pal(n = 16, rainbow_hcl(16, start = 60, end = 240), main = "harmonic")
pal(n = 16, rainbow_hcl(16, start = 270, end = 150), main = "cold")
pal(n = 16, rainbow_hcl(16, start = 90, end = -30), main = "warm")

```

除之前提到的 **grDevices** 包, **colorspace** (<https://hclwizard.org/>) 包 [Stauffer et al., 2009, Zeileis et al., 2009, 2019], RColorBrewer 包 [Neuwirth, 2014] <https://colorbrewer2.org/>, viridis 包、colourvalues、wesanderson、dichromat 包、pals 包, palr 包, colorRamps 包、ColorPalette 包、colortools 包就不一一详细介绍。

colormap 包基于 node.js 的 colormap 模块提供 44 个预定义的调色板 paletteer 包收集了很多 R 包提供的调色板, 同时也引入了很多依赖。yarrr 包主要是为书籍 《YaRrr! The Pirate's Guide to R》 <https://github.com/ndphillips/ThePiratesGuideToR> 提供配套资源, 兼顾收集了一组调色板。



图 7.25: colorspace 调色板



注意

RColorBrewer 调色板数量必须至少 3 个，这是上游 colorbrewer 的问题，具体体现在调用 RColorBrewer::brewer.pal(n = 2, name = "Set2") 时会有警告。plotly 调用

```
[1] "#66C2A5" "#FC8D62" "#8DA0CB"
```

Warning message:

```
In RColorBrewer::brewer.pal(n = 2, name = "Set2") :
```

```
minimal value for n is 3, returning requested palette with 3 different levels
```

```
par(mar = c(1, 2, 1, 0), mfrow = c(3, 2))
set.seed(1234)
x <- sample(seq(8), 8, replace = FALSE)
barplot(x, col = palette(), border = "white")
barplot(x, col = heat.colors(8), border = "white")
barplot(x, col = gray.colors(8), border = "white")
barplot(x, col = "lightblue", border = "white")
barplot(x, col = colorspace::sequential_hcl(8), border = "white")
barplot(x, col = colorspace::diverge_hcl(8,
  h = c(130, 43),
  c = 100, l = c(70, 90
), border = "white")
```

与图 7.41 对比，图 7.27 的层次更加丰富，识别性更高

```
expand.grid(months = month.abb, years = 1949:1960) |>
  transform(num = as.vector(AirPassengers)) |>
  ggplot(aes(x = years, y = months, fill = num)) +
  scale_fill_distiller(palette = "Spectral") +
  geom_tile(color = "white", size = 0.4) +
  scale_x_continuous(
    expand = c(0.01, 0.01),
    breaks = seq(1949, 1960, by = 1),
    labels = 1949:1960
  ) +
  theme_minimal(
    base_size = 10.54,
    base_family = "source-han-serif-cn"
```

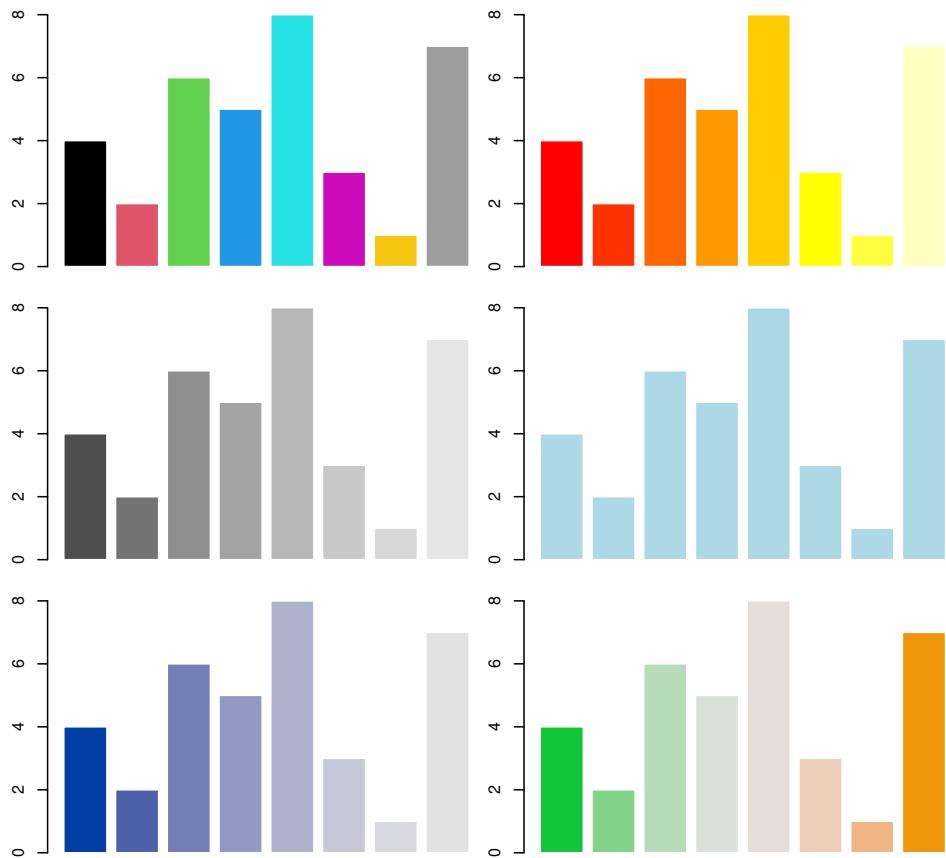


图 7.26: 源起

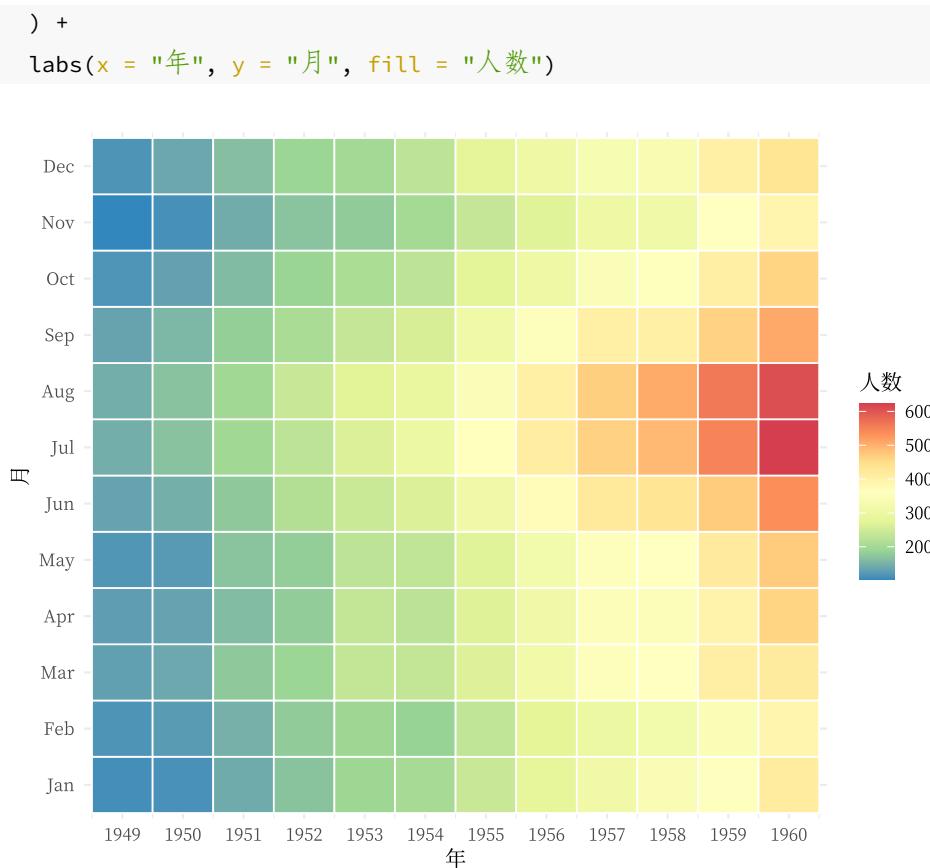


图 7.27: Spectral 调色板

再举栗子，图 7.28 是正负例对比，其中好在哪里呢？这张图要表达美国黄石国家公园的老忠实泉间歇喷发的时间规律，那么好的标准就是层次分明，以突出不同颜色之间的时间差异。这个差异，还要看起来不那么费眼睛，一目了然最好。

```
erupt <- ggplot(faithful, aes(waiting, eruptions, fill = density)) +
  geom_raster() +
  scale_x_continuous(NULL, expand = c(0, 0)) +
  scale_y_continuous(NULL, expand = c(0, 0)) +
  theme(legend.position = "none")
p1 <- erupt + scale_fill_gradientn(colours = gray.colors(7))
p2 <- erupt + scale_fill_distiller(palette = "Spectral")
p3 <- erupt + scale_fill_gradientn(colours = terrain.colors(7))
p4 <- erupt + scale_fill_continuous(type = 'viridis')
```

```
(p1 + p2) / (p3 + p4)
```

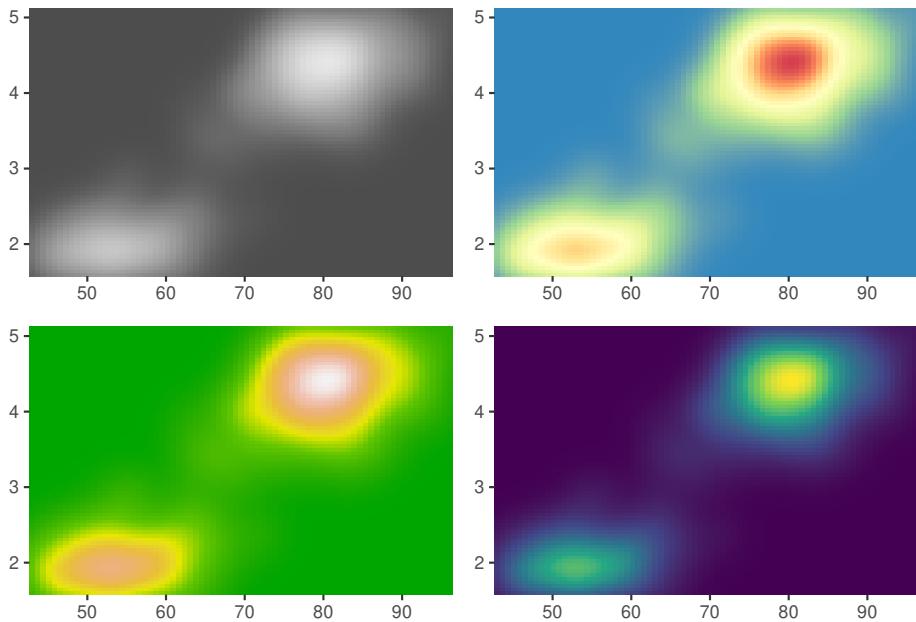


图 7.28: 美国黄石国家公园的老忠实泉

RColorBrewer 包提供了有序 (Sequential)、定性 (Qualitative) 和发散 (Diverging) 三类调色板，一般来讲，分别适用于连续或有序分类变量、无序分类变量、两类分层对比变量的绘图。再加上强大的 ggplot2 包内置的对颜色处理的函数，如 `scale_alpha_*`、`scale_colour_*` 和 `scale_fill_*` 等，详见：

```
ls("package:ggplot2", pattern = "scale_col(ou|o)r_")
```

```
## [1] "scale_color_binned"      "scale_color_brewer"  
## [3] "scale_color_continuous"  "scale_color_date"  
## [5] "scale_color_datetime"    "scale_color_discrete"  
## [7] "scale_color_distiller"   "scale_color_fermenter"  
## [9] "scale_color_gradient"    "scale_color_gradient2"  
## [11] "scale_color_gradientn"   "scale_color_grey"  
## [13] "scale_color_hue"        "scale_color_identity"  
## [15] "scale_color_manual"     "scale_color_ordinal"  
## [17] "scale_color_steps"      "scale_color_steps2"  
## [19] "scale_color_stepsn"     "scale_color_viridis_b"
```



```

## [21] "scale_color_viridis_c"    "scale_color_viridis_d"
## [23] "scale_colour_binned"     "scale_colour_brewer"
## [25] "scale_colour_continuous" "scale_colour_date"
## [27] "scale_colour_datetime"   "scale_colour_discrete"
## [29] "scale_colour_distiller"   "scale_colour_fermenter"
## [31] "scale_colour_gradient"   "scale_colour_gradient2"
## [33] "scale_colour_gradientn"  "scale_colour_grey"
## [35] "scale_colour_hue"        "scale_colour_identity"
## [37] "scale_colour_manual"     "scale_colour_ordinal"
## [39] "scale_colour_steps"      "scale_colour_steps2"
## [41] "scale_colour_stepsn"     "scale_colour_viridis_b"
## [43] "scale_colour_viridis_c"  "scale_colour_viridis_d"

ls("package:ggplot2", pattern = "scale_fill_")

## [1] "scale_fill_binned"    "scale_fill_brewer"    "scale_fill_continuous"
## [4] "scale_fill_date"       "scale_fill_datetime" "scale_fill_discrete"
## [7] "scale_fill_distiller"  "scale_fill_fermenter" "scale_fill_gradient"
## [10] "scale_fill_gradient2"  "scale_fill_gradientn" "scale_fill_grey"
## [13] "scale_fill_hue"        "scale_fill_identity" "scale_fill_manual"
## [16] "scale_fill_ordinal"    "scale_fill_steps"    "scale_fill_steps2"
## [19] "scale_fill_stepsn"    "scale_fill_viridis_b" "scale_fill_viridis_c"
## [22] "scale_fill_viridis_d"

```

7.3.2 颜色模式

7.3.2.1 RGB

红 (red)、绿 (green)、蓝 (blue) 是三原色

```
rgb(red, green, blue, alpha, names = NULL, maxValue = 1)
```

函数参数说明：

- red, blue, green, alpha 取值范围 $[0, M]$, M 是 maxValue
- names 字符向量，给这组颜色值取名
- maxValue 红, 绿, 蓝三色范围的最大值

The colour specification refers to the standard sRGB colorspace (IEC standard 61966).



rgb 产生一种颜色, 如 `rgb(255, 0, 0, maxColorValue = 255)` 的颜色是 "#FF0000" , 这是一串 16 进制数, 每两个一组, 那么一组有 $16^2 = 256$ 种组合, 整个一串有 $256^3 = 16777216$ 种组合, 这就是 RGB 表达的所有颜色。

7.3.2.2 HSL

色相饱和度亮度 hue-saturation-luminance (HSL)

7.3.2.3 HSV

Create a vector of colors from vectors specifying hue, saturation and value. 色相饱和度值

```
hsv(h = 1, s = 1, v = 1, alpha)
```

This function creates a vector of colors corresponding to the given values in HSV space. `rgb` and `rgb2hsv` for RGB to HSV conversion;

`hsv` 函数通过设置色调、饱和度和亮度获得颜色, 三个值都是 0-1 的相对量

RGB HSV HSL 都是不连续的颜色空间, 缺点

7.3.2.4 HCL

基于感知的颜色空间替代 RGB 颜色空间

通过指定色相 (hue), 色度 (chroma) 和亮度 (luminance/lightness), 创建一组 (种) 颜色

```
hcl(h = 0, c = 35, l = 85, alpha, fixup = TRUE)
```

函数参数说明:

- **h** 颜色的色调, 取值范围为 [0,360], 0、120、240 分别对应红色、绿色、蓝色
- **c** 颜色的色度, 其上界取决于色调和亮度
- **l** 颜色的亮度, 取值范围 [0,100], 给定色调和色度, 只有一部分子集可用
- **alpha** 透明度, 取值范围 [0,1], 0 和 1 分别表示透明和不透明

This function corresponds to polar coordinates in the CIE-LUV color space

选色为什么这么难



色相与阴影相比是无关紧要的，色相对于标记和分类很有用，但表示（精细的）空间数据或形状的效果较差。颜色是改善图形的好工具，但糟糕的配色方案 (color schemes) 可能会导致比灰度调色板更差的效果。[\[Stauffer et al., 2009\]](#)

黑、白、灰，看似有三种颜色，其实只有一种颜色，黑和白只是灰色的两极，那么如何设置灰色梯度，使得人眼比较好区分它们呢？这样获得的调色板适用于什么样的绘图环境呢？

7.3.2.5 CMYK

印刷三原色：青 (cyan)、品红 (magenta)、黄 (yellow)

- 颜色模式转化

`col2rgb()`、`rgb2hsv()` 和 `rgb()` 函数 `hex2RGB()` 函数 `colorspace col2hcl()` 函数 `scales col2HSV()` `colortools col2hex()`

```
col2rgb("lightblue") # color to RGB

##      [,1]
## red    173
## green  216
## blue   230

scales::col2hcl("lightblue") # color to HCL

## [1] "#ADD8E6"

# palr::col2hex("lightblue") # color to HEX
# colortools::col2HSV("lightblue") # color to HSV

rgb(173, 216, 230, maxValue = 255) # RGB to HEX

## [1] "#ADD8E6"

colorspace::hex2RGB("#ADD8E6") # HEX to RGB

##          R          G          B
## [1,] 0.6784314 0.8470588 0.9019608

rgb(.678, .847, .902, maxValue = 1) # RGB to HEX

## [1] "#ADD8E6"
```



```
rgb2hsv(173, 216, 230, maxValue = 255) # RGB to HSV  
## [1] 0.5409357  
## s 0.2478261  
## v 0.9019608
```

7.3.3 LaTeX 配色

LaTeX 宏包 `xcolor` 中定义颜色的常用方式有两种, 其一, `\textcolor{green!40!yellow}` 表示 40% 的绿色和 60% 的黄色混合色彩, 其二, `\textcolor[HTML]{34A853}` HEX 表示的色彩直接在 LaTeX 文档中使用的方式, 类似地 `\textcolor[RGB]{52,168,83}` 也表示 Google 图标中的绿色。

```
\documentclass[tikz, border=10pt]{standalone}  
\begin{document}  
\begin{tikzpicture}  
\draw (0,0) rectangle (2,1) node [midway] {\textcolor[RGB]{52,168,83}{Hello}} \textcolor{red}{World}  
\end{tikzpicture}  
\end{document}
```

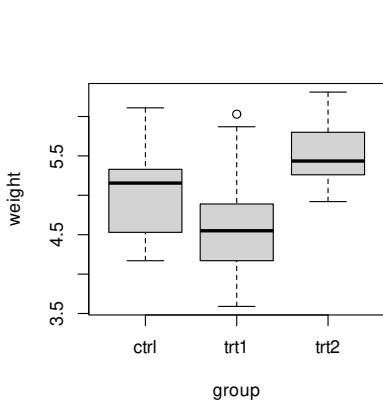
对应于 R 中的调用方式为:

```
rgb(52, 168, 83, maxValue = 255)  
## [1] "#34A853"
```

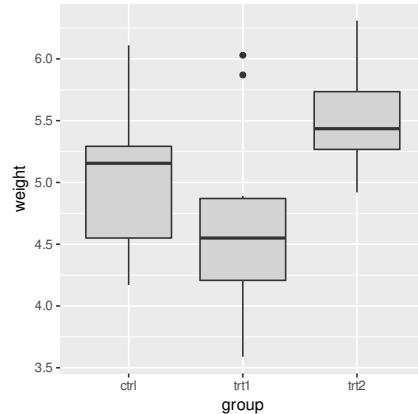
7.3.4 ggplot2 配色

```
boxplot(weight ~ group,  
       data = PlantGrowth, col = "lightgray",  
       notch = FALSE, varwidth = TRUE  
)  
# 类似 boxplot  
ggplot(data = PlantGrowth, aes(x = group, y = weight)) +  
  geom_boxplot(notch = FALSE, varwidth = TRUE, fill = "lightgray")  
  
# 默认调色板
```

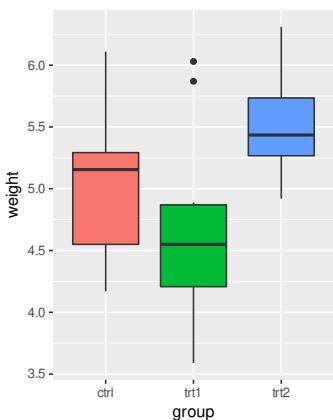
```
ggplot(data = PlantGrowth, aes(x = group, y = weight, fill = group)) +  
  geom_boxplot(notch = FALSE, varwidth = TRUE)  
  
# Google 调色板  
ggplot(data = PlantGrowth, aes(x = group, y = weight, fill = group)) +  
  geom_boxplot(notch = FALSE, varwidth = TRUE) +  
  scale_fill_manual(values = c("#4285f4", "#34A853", "#FBBC05", "#EA4335"))
```



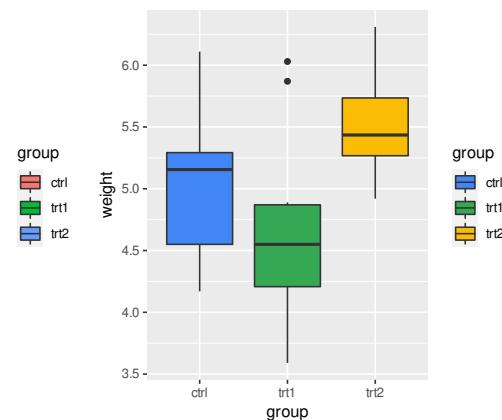
(a) 简单箱线图



(b) ggplot2 绘制的箱线图



(c) ggplot2 调用默认调色板



(d) ggplot2 调用 Google 调色板

图 7.29: 几种不同的箱线图



7.4 图库

7.4.1 饼图

我对饼图是又爱又恨，爱的是它表示百分比的时候，往往让读者联想到蛋糕，份额这类根深蒂固的情景，从而让数字通俗易懂、深入人心，是一种很好的表达方式，恨的也是这一点，我用柱状图表达不香吗？人眼对角度的区分度远不如柱状图呢，特别是当两个类所占的份额比较接近的时候，所以很多时候，除了用饼图表达份额，还会在旁边标上百分比，从数据可视化的角度来说，如图 7.30 所示，这是信息冗余！

```
BOD %>% transform(., ratio = demand / sum(demand)) %>%
  ggplot(., aes(x = "", y = demand, fill = reorder(Time, demand))) +
  geom_bar(stat = "identity", show.legend = FALSE, color = "white") +
  coord_polar(theta = "y") +
  geom_text(aes(x = 1.6, label = paste0(round(ratio, digits = 4) * 100, "%")),
            position = position_stack(vjust = 0.5), color = "black")
  ) +
  geom_text(aes(x = 1.2, label = Time),
            position = position_stack(vjust = 0.5), color = "black")
  ) +
  theme_void(base_size = 14)
```

plot_ly(type = "pie", ...) 和添加图层 add_pie() 的效果是一样的

```
dat = aggregate(formula = carat ~ cut, data = diamonds, FUN = length)
plotly::plot_ly() %>%
  plotly::add_pie(
    data = dat, labels = ~cut, values = ~carat,
    name = "简单饼图1", domain = list(row = 0, column = 0)
  ) %>%
  plotly::add_pie(
    data = dat, labels = ~cut, values = ~carat, hole = 0.6,
    textposition = "inside", textinfo = "label+percent",
    name = "简单饼图2", domain = list(row = 0, column = 1)
  ) %>%
  plotly::layout(
    title = "多图布局", showlegend = F,
```

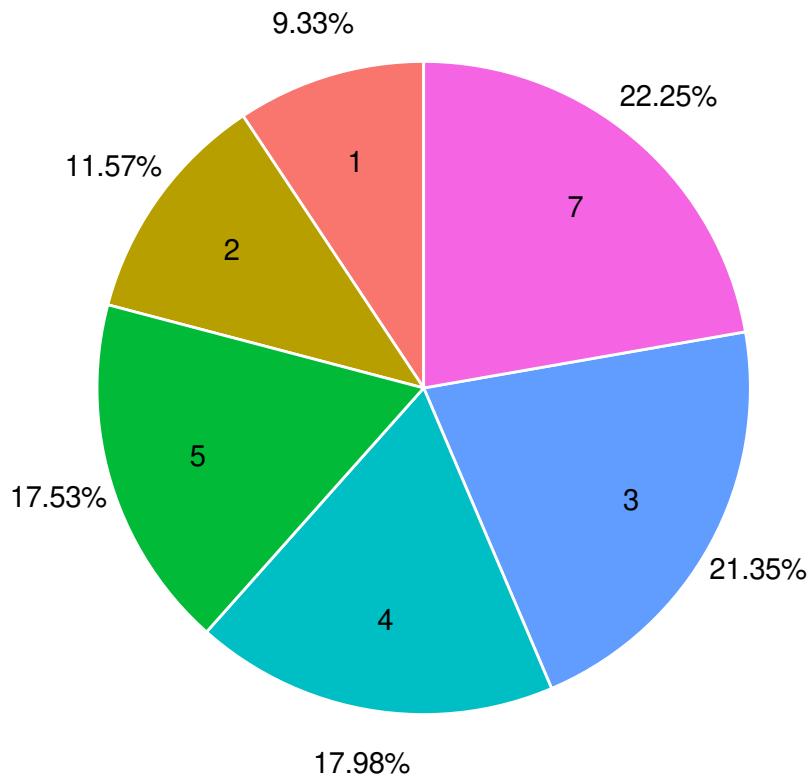


图 7.30: 饼图



```
grid = list(rows = 1, columns = 2),
  xaxis = list(showgrid = FALSE, zeroline = FALSE, showticklabels = FALSE),
  yaxis = list(showgrid = FALSE, zeroline = FALSE, showticklabels = FALSE)
) %>%
plotly::config(displayModeBar = FALSE)
```

设置参数 `hole` 可以绘制环形饼图，比如 `hole = 0.6`

7.4.2 地图

USArrests 数据集描述了 1973 年美国 50 个州每 10 万居民中因袭击、抢劫和强奸而逮捕的人，以及城市人口占比。这里的地图是指按照行政区划为边界的示意图，比如图 7.31

```
library(maps)
crimes <- data.frame(state = tolower(rownames(USArrests)), USArrests)
# 等价于 crimes %>% tidyrr::pivot_longer(Murder:Rape)
vars <- lapply(names(crimes)[-1], function(j) {
  data.frame(state = crimes$state, variable = j, value = crimes[[j]])
})
crimes_long <- do.call("rbind", vars)
states_map <- map_data("state")
ggplot(crimes, aes(map_id = state)) +
  geom_map(aes(fill = Murder), map = states_map) +
  expand_limits(x = states_map$long, y = states_map$lat) +
  scale_fill_binned(type = "viridis") +
  coord_map() +
  theme_minimal()
```

先来看看中国及其周边，见图 7.32，这个地图的缺陷就是中国南海及九段线没有标记，台湾和中国大陆不是一种颜色标记，这里的地图数据来自 R 包 **maps** 和 **mapdata**，像这样的地图就不宜在国内正式刊物上出现。

```
library(maps)
library(mapdata)
east_asia <- map_data("worldHires",
region = c(
  "Japan", "Taiwan", "China",
```

云湘黄
④

206

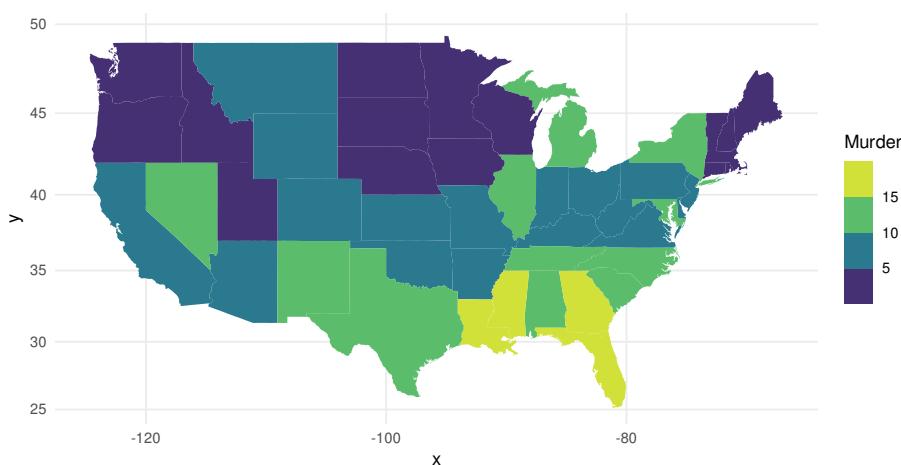


图 7.31: 1975 年美国各州犯罪事件

```
"North Korea", "South Korea"
)
)
ggplot(east_asia, aes(x = long, y = lat, group = group, fill = region)) +
  geom_polygon(colour = "black") +
  scale_fill_brewer(palette = "Set2") +
  theme_minimal()
```

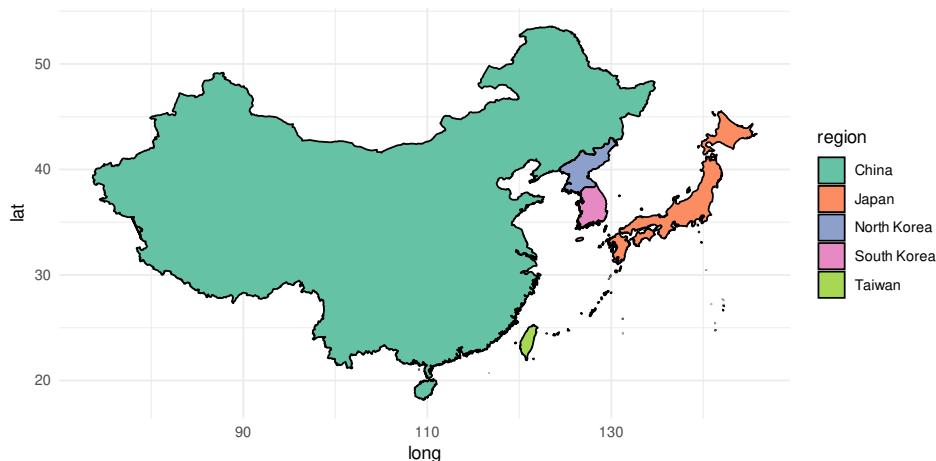


图 7.32: 中国及其周边

绘制真正的地图需要考虑投影坐标系、观察角度、分辨率、政策法规等一系列因素，它是一种复杂的图形，如图 7.33 所示。

```
worldmap <- map_data("world")

# 默认 mercator 投影下的默认视角 c(90, 0, mean(range(x)))
ggplot(worldmap, aes(long, lat, group = group)) +
  geom_polygon(aes(fill = region), show.legend = FALSE) +
  coord_map(
    xlim = c(-120, 40), ylim = c(30, 90)
  )

# 换观察角度
ggplot(worldmap, aes(long, lat, group = group)) +
  geom_polygon(aes(fill = region), show.legend = FALSE) +
  coord_map(
    xlim = c(-120, 40), ylim = c(30, 90),
    orientation = c(90, 0, 0)
  )

# 换投影坐标系
ggplot(worldmap, aes(long, lat, group = group)) +
  geom_polygon(aes(fill = region), show.legend = FALSE) +
  coord_map("ortho",
    xlim = c(-120, 40), ylim = c(30, 90)
  )

# 二者皆换
ggplot(worldmap, aes(long, lat, group = group)) +
  geom_polygon(aes(fill = region), show.legend = FALSE) +
  coord_map("ortho",
    xlim = c(-120, 40), ylim = c(30, 90),
    orientation = c(90, 0, 0)
  )
```

Google 地图

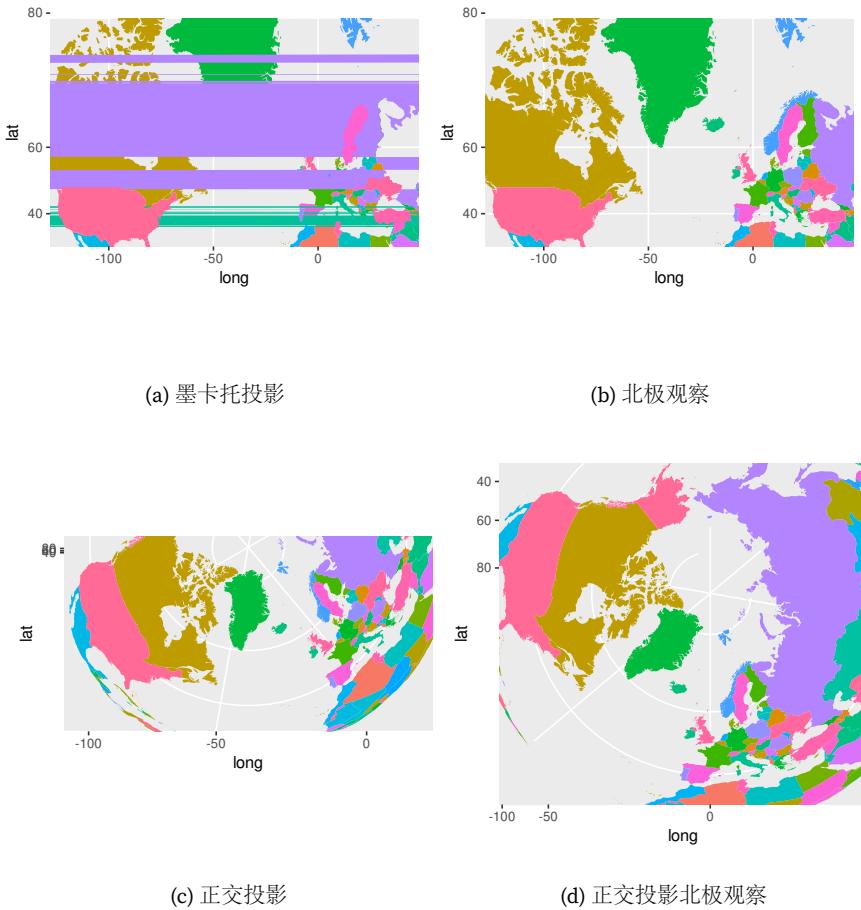


图 7.33: 画地图的正确姿势

```
library(RgoogleMaps)
# 一组坐标的中心位置
lat <- c(40.702147, 40.718217, 40.711614)
lon <- c(-74.012318, -74.015794, -73.998284)
center <- c(mean(lat), mean(lon))
zoom <- min(MaxZoom(range(lat), range(lon)))
# 矩形对角线的两个顶点
bb <- qbbox(lat, lon)
# 获取地图数据
myMap <- GetMap(center, size = c(640, 640), zoom = zoom, type = "osm")
# 在地图上添加红、蓝、绿三个点
PlotOnStaticMap(myMap,
  lat = lat, lon = lon, pch = 20, cex = 10,
  col = c("red", "blue", "green"))
)
```

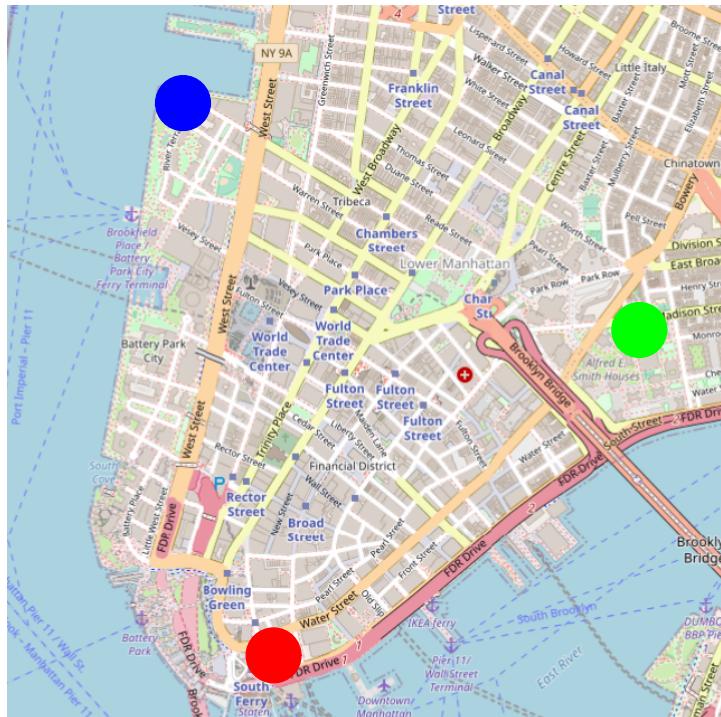


图 7.34: Google 地图示例



7.4.3 热图

Zuguang Gu 开发的 `ComplexHeatmap` 包实现复杂数据的可视化，用以发现关联数据集之间的模式。特别地，比如基因数据、生存数据等，更多应用见开发者的书籍 [ComplexHeatmap 完全手册](#)。R 包发布在 Bioconductor 上 <https://www.bioconductor.org/packages/ComplexHeatmap>。使用之前我要确保已经安装 `BiocManager` 包，这个包负责管理 Bioconductor 上所有的包，需要先安装它，然后安装 `ComplexHeatmap` 包 [Gu et al., 2016]。

```
if (!requireNamespace("BiocManager", quietly = TRUE))
  install.packages("BiocManager")
BiocManager::install("ComplexHeatmap")
```

7.4.4 条形图

```
# 漫谈条形图 https://cosx.org/2017/10/discussion-about-bar-graph
set.seed(2020)
dat <- data.frame(
  age = rep(1:30, 2),
  gender = rep(c("man", "woman"), each = 30),
  num = sample(x = 1:100, size = 60, replace = T)
)
# 重叠
p1 <- ggplot(data = dat, aes(x = age, y = num, fill = gender)) +
  geom_col(position = "identity", alpha = 0.5)
# 堆积
p2 <- ggplot(data = dat, aes(x = age, y = num, fill = gender)) +
  geom_col(position = "stack")
# 双柱
p3 <- ggplot(data = dat, aes(x = age, y = num, fill = gender)) +
  geom_col(position = "dodge")
# 百分比
p4 <- ggplot(data = dat, aes(x = age, y = num, fill = gender)) +
  geom_col(position = "fill") +
  scale_y_continuous(labels = scales::percent_format()) +
  labs(y = "%")
```

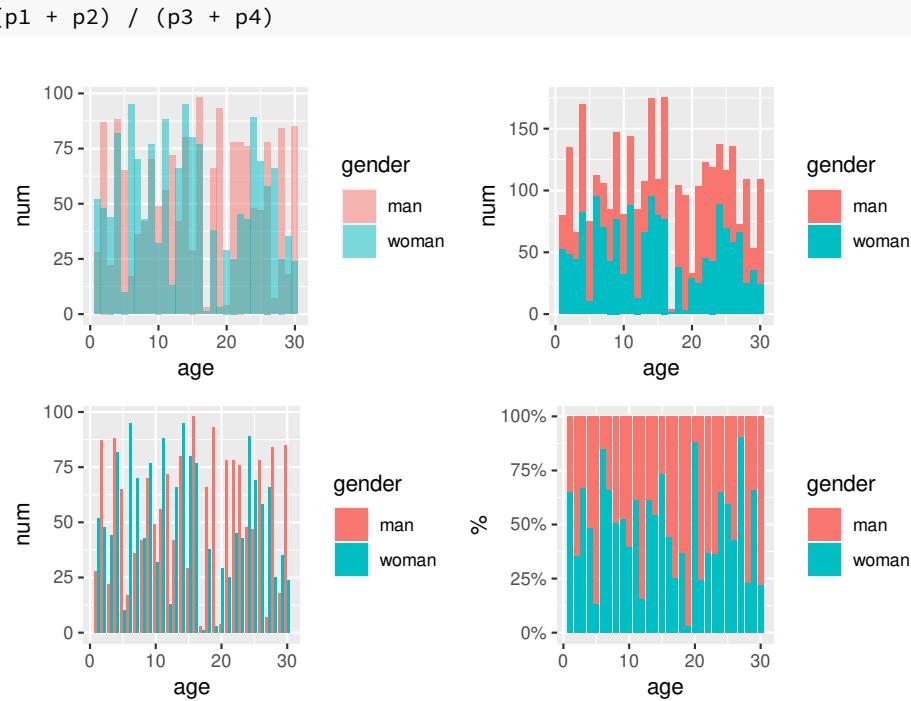


图 7.35: 条形图的四种常见形态

以数据集 diamonds 为例，按照纯净度 clarity 和切工 cut 分组统计钻石的数量，再按切工分组统计不同纯净度的钻石数量占比，如表 7.2 所示

```
library(data.table)
diamonds <- as.data.table(diamonds)
dat <- diamonds[, .(cnt = .N), by = .(cut, clarity)] %>%
  .[, pct := cnt / sum(cnt), by = .(cut)] %>%
  .[, pct_pp := paste0(cnt, " (", scales::percent(pct, accuracy = 0.01), ")")]
# 分组计数 with(diamonds, table(clarity, cut))
dcast(dat, formula = clarity ~ cut, value.var = "pct_pp") %>%
  knitr::kable(alignment = "c", caption = "数值和比例组合呈现")
```

分别以堆积条形图和百分比堆积条形图展示，添加注释到条形图上，见 7.36

```
p1 = ggplot(data = dat, aes(x = cut, y = cnt, fill = clarity)) +
  geom_col(position = "dodge") +
  geom_text(aes(label = cnt), position = position_dodge(1), vjust = -0.5) +
  geom_text(aes(label = scales::percent(pct, accuracy = 0.1)),
```



表 7.2: 数值和比例组合呈现

clarity	Fair	Good	Very Good	Premium	Ideal
I1	210 (13.04%)	96 (1.96%)	84 (0.70%)	205 (1.49%)	146 (0.68%)
SI2	466 (28.94%)	1081 (22.03%)	2100 (17.38%)	2949 (21.38%)	2598 (12.06%)
SI1	408 (25.34%)	1560 (31.80%)	3240 (26.82%)	3575 (25.92%)	4282 (19.87%)
VS2	261 (16.21%)	978 (19.93%)	2591 (21.45%)	3357 (24.34%)	5071 (23.53%)
VS1	170 (10.56%)	648 (13.21%)	1775 (14.69%)	1989 (14.42%)	3589 (16.65%)
VVS2	69 (4.29%)	286 (5.83%)	1235 (10.22%)	870 (6.31%)	2606 (12.09%)
VVS1	17 (1.06%)	186 (3.79%)	789 (6.53%)	616 (4.47%)	2047 (9.50%)
IF	9 (0.56%)	71 (1.45%)	268 (2.22%)	230 (1.67%)	1212 (5.62%)

```
position = position_dodge(1), vjust = 1, hjust = 0.5
) +
scale_fill_brewer(palette = "Spectral") +
labs(fill = "clarity", y = "", x = "cut") +
theme_minimal() +
theme(legend.position = "top")

p2 = ggplot(data = dat, aes(y = cut, x = cnt, fill = clarity)) +
geom_col(position = "fill") +
geom_text(aes(label = cnt), position = position_fill(1), vjust = -0.5) +
geom_text(aes(label = scales::percent(pct, accuracy = 0.1)),
position = position_fill(1), vjust = 1, hjust = 0.5
) +
scale_fill_brewer(palette = "Spectral") +
scale_x_continuous(labels = scales::percent) +
labs(fill = "clarity", y = "", x = "cut") +
theme_minimal() +
theme(legend.position = "top")

p1 / p2
```

借助 `plotly` 制作相应的动态百分比堆积条形图

```
ggplot(data = diamonds, aes(x = cut, fill = clarity)) +
geom_bar(position = "dodge2") +
```

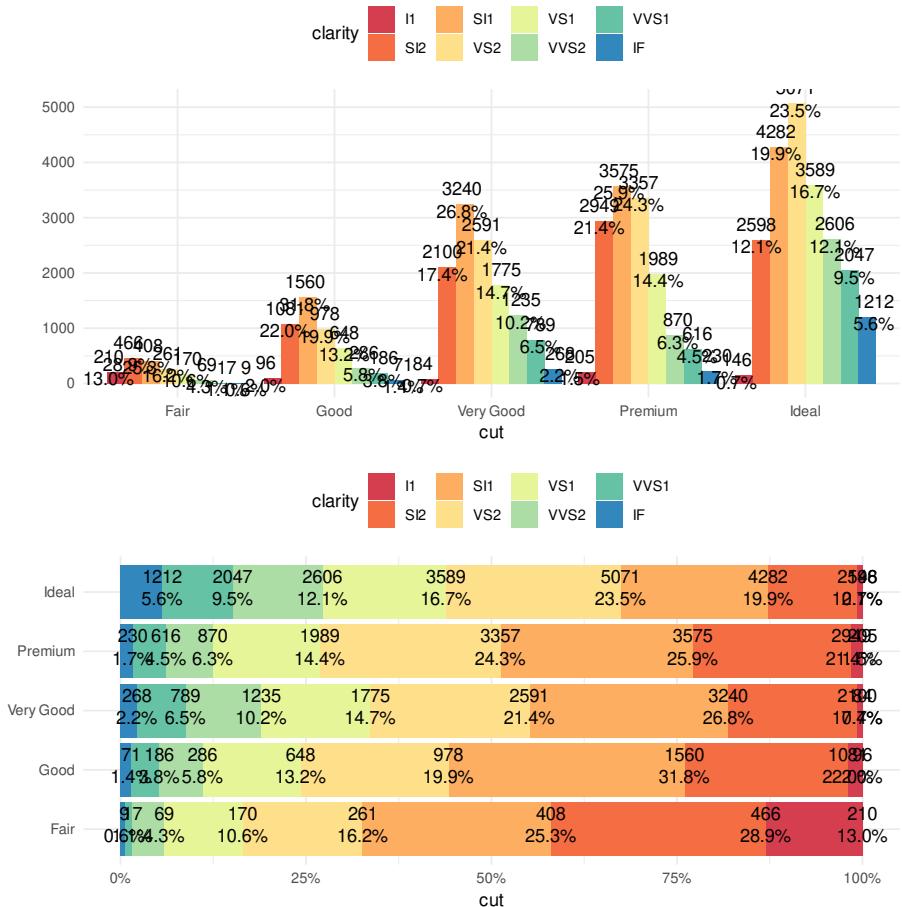


图 7.36: 添加注释到条形图



```
scale_fill_brewer(palette = "Spectral")

plotly::plot_ly(dat,
  y = ~cut, color = ~clarity, x = ~cnt,
  colors = "Spectral", type = "bar",
  text = ~ paste0(
    cnt, "颗 <br>",
    "占比: ", scales::percent(pct, accuracy = 0.1), "<br>"
  ),
  hoverinfo = "text"
) %>%
  plotly::layout(barmode = "stack", barnorm = "percent") %>%
  plotly::config(displayModeBar = FALSE)

# `type = "histogram"` 以 cut 和 clarity 分组计数
plotly::plot_ly(diamonds,
  x = ~cut, color = ~clarity,
  colors = "Spectral", type = "histogram"
) %>%
  plotly::config(displayModeBar = FALSE)

# 堆积图
plotly::plot_ly(diamonds,
  x = ~cut, color = ~clarity,
  colors = "Spectral", type = "histogram"
) %>%
  plotly::layout(
    barmode = "stack",
    yaxis = list(title = "cnt"),
    legend = list(title = list(text = "clarity")))
) %>%
  plotly::config(displayModeBar = FALSE)

# 百分比堆积图
plotly::plot_ly(diamonds,
  x = ~cut, color = ~clarity,
```



```
  colors = "Spectral", type = "histogram"
) %>%
  plotly::layout(
    barmode = "stack", barnorm = "percent",
    yaxis = list(title = "percent"),
    legend = list(title = list(text = "clarity"))
) %>%
  plotly::config(displayModeBar = FALSE)
```

7.4.5 函数图

蝴蝶图的参数方程如下

$$x = \sin t \left(e^{\cos t} - 2 \cos 4t + \sin^5 \left(\frac{t}{12} \right) \right) \quad (7.1)$$

$$y = \cos t \left(e^{\cos t} - 2 \cos 4t + \sin^5 \left(\frac{t}{12} \right) \right), t \in [-\pi, \pi] \quad (7.2)$$

7.4.6 密度图

heatmaps in ggplot2 二维密度图

```
ggplot(faithful, aes(x = eruptions, y = waiting)) +
  stat_density_2d(aes(fill = ..level..), geom = "polygon") +
  xlim(1, 6) +
  ylim(40, 100)

ggplot(faithful, aes(x = eruptions, y = waiting)) +
  stat_density2d(aes(fill = stat(level)), geom = "polygon") +
  scale_fill_viridis_c(option = "viridis") +
  xlim(1, 6) +
  ylim(40, 100)
```

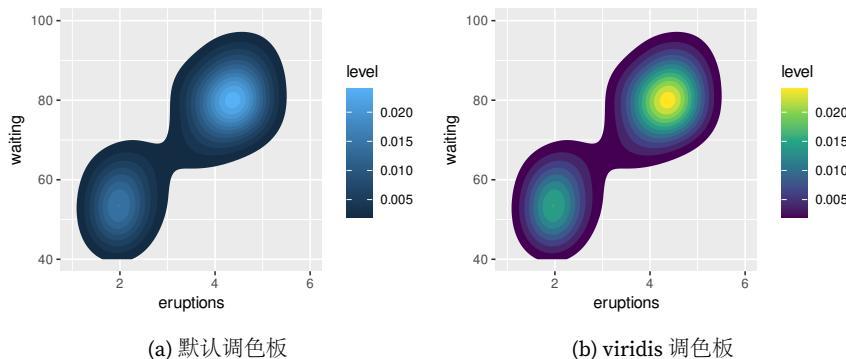


图 7.37: 二维密度图

提示

`MASS::kde2d()` 实现二维核密度估计, `ggplot2` 包提供了两种等价的绘图方式

1. stat_density_2d() 和 ..
 2. stat_density2d() 和 stat()

```
plotly::plot_ly(
  data = faithful, x = ~eruptions,
  y = ~waiting, type = "histogram2dcontour"
) %>%
  plotly::config(displayModeBar = FALSE)

# plot_ly(faithful, x = ~waiting, y = ~eruptions) %>%
#   add_histogram2d() %>%
#   add_histogram2dcontour()
```

延伸一下，热力图

```
library(KernSmooth)
den <- bkde2D(x = faithful, bandwidth = c(0.7, 7))
# 热力图
p1 <- plotly::plot_ly(x = den$x1, y = den$x2, z = den$fhat) %>%
  plotly::config(displayModeBar = FALSE) %>%
  plotly::add_heatmap()
```



```
# 等高线图
p2 <- plotly::plot_ly(x = den$x1, y = den$x2, z = den$fhat) %>%
  plotly::config(displayModeBar = FALSE) %>%
  plotly::add_contour()

htmltools::tagList(p1, p2)
```

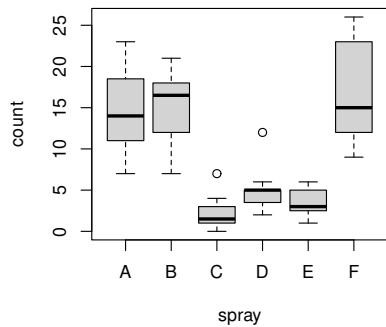
7.4.7 提琴图

2004 年 Daniel Adler 开发 `vioplot` 包实现提琴图的绘制，它可能是最早实现此功能的 R 包，随后 10 余年没有更新却一直坚挺在 CRAN 上，非常难得，好在 Thomas Kelly 已经接手维护。另一款绘制提琴图的 R 包是 Peter Kampstra 开发的 `beanplot` [Kampstra, 2008]，也存在很多年了，不过随着时间的变迁，比较现代的方式是 `ggplot2` 带来的 `geom_violin()` 扔掉了很多依赖，也是各种图形的汇集地，可以看作是最佳实践。提琴图比起箱线图优势在于呈现更多的分布信息，其次在于更加美观，但是就目前来说箱线图的受众比提琴图要多很多，毕竟前者是包含更多统计信息，如图 7.38 所示。

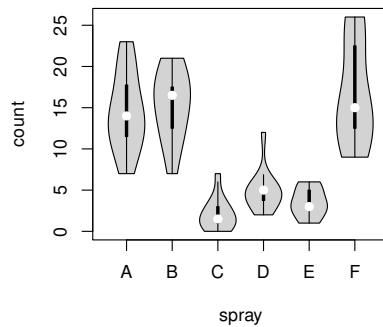
```
boxplot(count ~ spray, data = InsectSprays)
vioplot::vioplot(count ~ spray, data = InsectSprays, col = "lightgray")
ggplot(InsectSprays, aes(x = spray, y = count)) +
  geom_violin(fill = "lightgray") +
  theme_minimal()
beanplot::beanplot(count ~ spray, data = InsectSprays, col = "lightgray")
```

`ggnormalviolin` 包在给定均值和标准差的情况下，绘制正态分布的概率密度曲线，如图 7.39 所示。

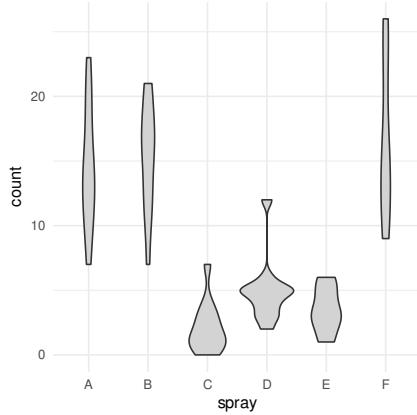
```
library(ggnormalviolin)
with(
  aggregate(
    data = iris, Sepal.Length ~ Species,
    FUN = function(x) c(dist_mean = mean(x), dist_sd = sd(x))
  ),
  cbind.data.frame(Sepal.Length, Species)
) %>%
  ggplot(aes(x = Species, mu = dist_mean, sigma = dist_sd, fill = Species)) +
```



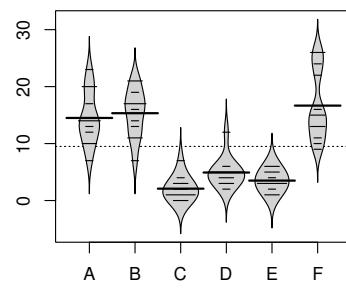
(a) 简单箱线图



(b) vioplot 绘制的提琴图



(c) ggplot2 绘制的提琴图



(d) beanplot 绘制的提琴图

图 7.38: 几种不同的提琴图

```
geom_normalviolin() +  
theme_minimal()
```

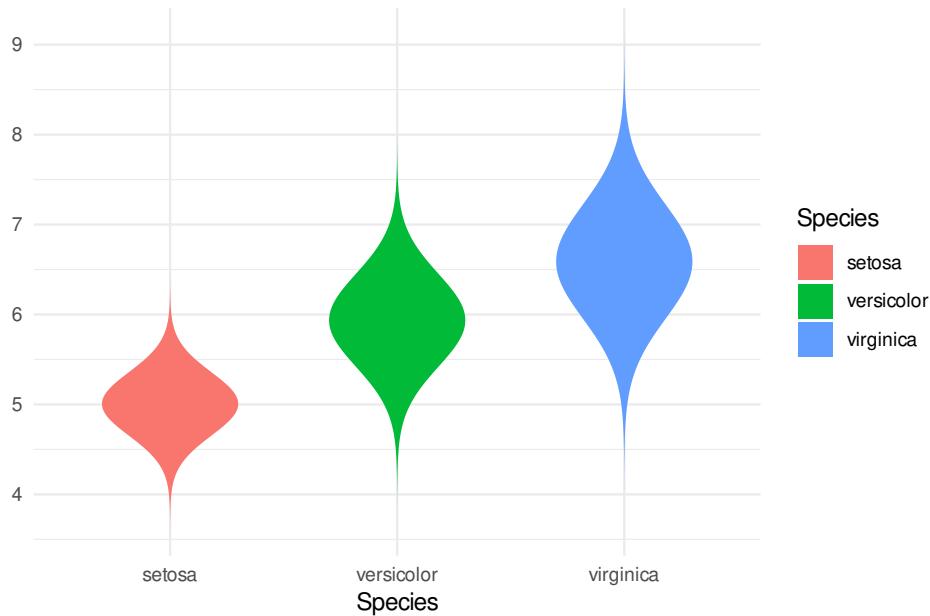


图 7.39: 正态分布的概率密度曲线

7.4.8 蜂群图

在样本点有限的情况下，用蜜蜂图代替普通的抖动图，可视化效果会好很多，如图 7.40 所示。Erik Clarke 开发的 `ggbeeswarm` 包可以将随机抖动的散点图朝着比较规律的方向聚合，又不丢失数据本身的准确性。

```
library(ggbeeswarm)  
p1 <- ggplot(iris, aes(Species, Sepal.Length)) +  
  geom_jitter() +  
  theme_minimal()  
p2 <- ggplot(iris, aes(Species, Sepal.Length)) +  
  geom_quasirandom() +  
  theme_minimal()  
p1 + p2
```

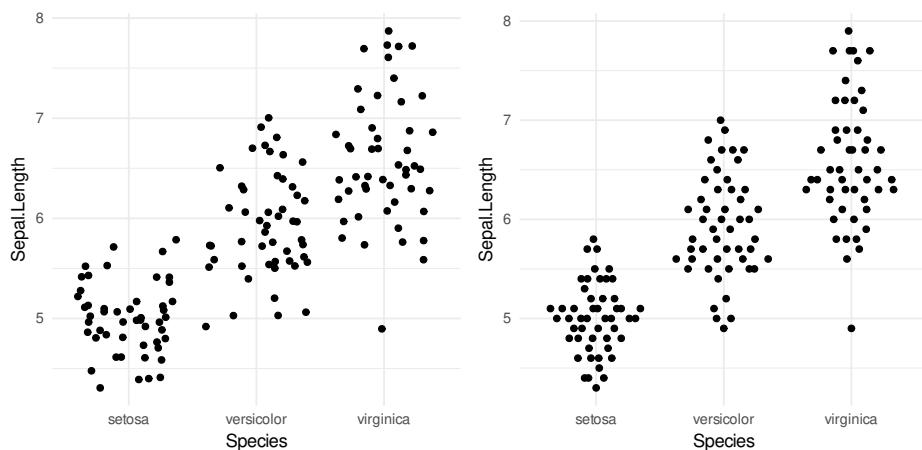


图 7.40: 蜜蜂图可视化效果比抖动图好

7.4.9 瓦片图

```
p1 <- expand.grid(months = month.abb, years = 1949:1960) %>%
  transform(num = as.vector(AirPassengers)) %>%
  ggplot(aes(x = years, y = months, fill = num)) +
  scale_fill_continuous(type = "viridis") +
  geom_tile(color = "white", size = 0.4) +
  scale_x_continuous(
    expand = c(0.01, 0.01),
    breaks = seq(1949, 1960, by = 1), labels = 1949:1960
  ) +
  theme_minimal(base_size = 10.54, base_family = "source-han-serif-cn") +
  theme(legend.position = "top") +
  labs(x = "年", y = "月", fill = "人数")

p2 <- expand.grid(months = month.abb, years = 1949:1960) %>%
  transform(num = as.vector(AirPassengers)) %>%
  ggplot(aes(x = years, y = months, color = num)) +
  geom_point(pch = 15, size = 8) +
  scale_color_distiller(palette = "Spectral") +
  scale_x_continuous(
    expand = c(0.01, 0.01),
```



```
breaks = seq(1949, 1960, by = 1), labels = 1949:1960
) +
theme_minimal(base_size = 10.54, base_family = "source-han-sans-cn") +
theme(legend.position = "top") +
labs(x = "年", y = "月", color = "人数")
p1 + p2
```

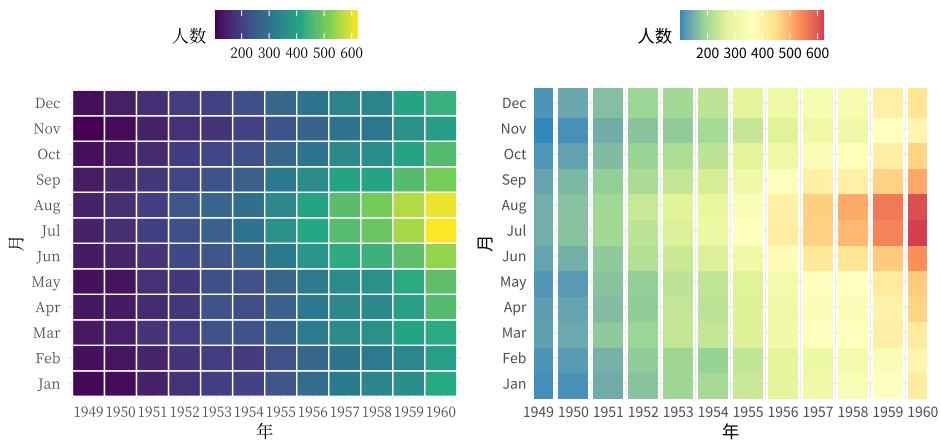


图 7.41: 1949-1960 年国际航线乘客数量的月度趋势

7.4.10 日历图

airquality 数据集记录了 1973 年 5 月至 9 月纽约的空气质量，包括气温（华氏度）、风速（米/小时）、紫外线强度、臭氧含量四个指标，图 7.42 展示了每日的气温变化。

```
airquality %>%
  transform(Date = seq.Date(
    from = as.Date("1973-05-01"),
    to = as.Date("1973-09-30"), by = "day"
  )) %>%
  transform(
    Week = as.integer(format(Date, "%w")),
    Year = as.integer(format(Date, "%Y")),
    Weekdays = factor(weekdays(Date, abbreviate = T),
```

```
levels = c("Mon", "Tue", "Wed", "Thu", "Fri", "Sat", "Sun")
)
) %>%
ggplot(aes(x = Week, y = Weekdays, fill = Temp)) +
scale_fill_distiller(name = "Temp (F)", palette = "Spectral") +
geom_tile(color = "white", size = 0.4) +
facet_wrap(~Year, ncol = 1) +
scale_x_continuous(
  expand = c(0, 0),
  breaks = seq(1, 52, length = 12),
  labels = month.abb
)
```

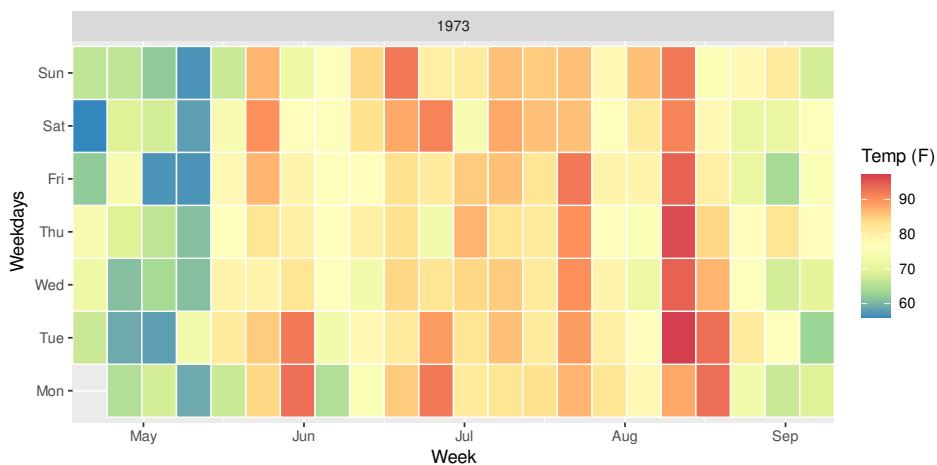


图 7.42: 1973 年 5 月至 9 月纽约的气温变化



注意

图 7.42 横轴的刻度标签换成了月份，一个月为四周，一年 52~53 周，每周的第一天约定为星期一，1973 年 05 月 01 日为星期二。代码中颇为技巧的在于 `format()` 函数从 Date 日期类型的数据提取第几周，用 `weekdays()` 函数提取星期几，而 `month.abb` 则是一个内置常量，12 个月份的英文缩写。在调用其它 R 包处理日期数据时要特别小心，要留意一周的第一天是星期几，有的是星期一，有的是星期日，这往往和宗教信仰相关，星期日在西方也叫礼拜天。上面 Base R 提供的日期函数认为一周的第一天是星期一，而调用 `data.table` 的话，默认一周是从星期日（礼拜天）开始的。

```
# https://d.cosx.org/d/421230
weekdays(Sys.Date(), abbreviate = TRUE)
## [1] "Sat"
data.table::wday(Sys.Date())
## [1] 7
```

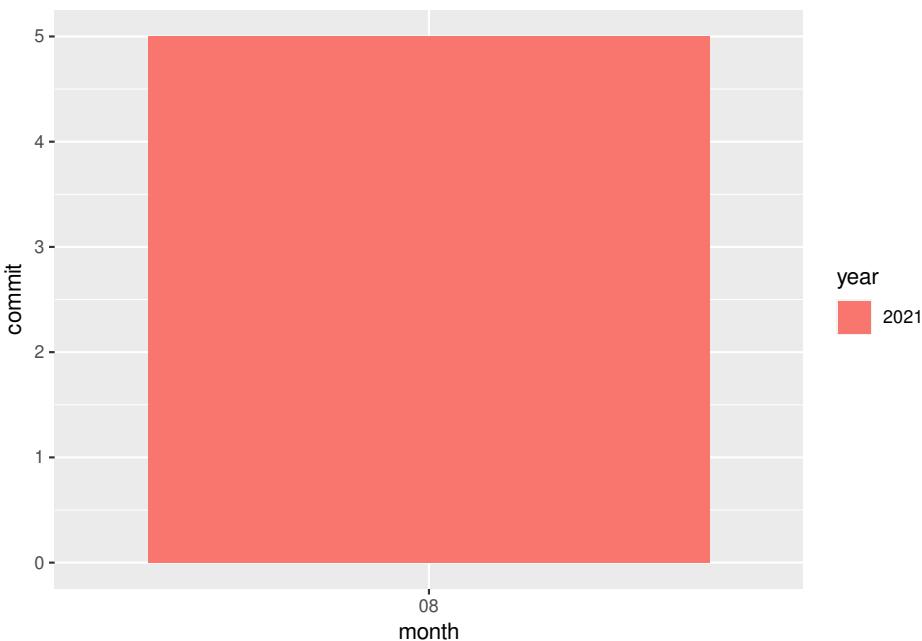
```
library(gert)
library(ggplot2)

dat <- git_log(max = 1000)

dat <- transform(dat,
  date = format(time, "%Y-%m-%d"),
  year = format(time, "%Y") ,
  month = format(time, "%m"),
  weekday = factor(format(time, "%a")),
  levels = c("Sun", "Mon", "Tue", "Wed", "Thu", "Fri", "Sat"),
),
  week = as.integer(format(time, "%W"))
)
```

本书的活跃情况

```
dat1 <- aggregate(formula = commit ~ year + month, data = dat, FUN = length)
# 条形图
ggplot(data = dat1, aes(x = month, y = commit, fill = year)) +
  geom_bar(stat = "identity", position = "identity")
```



日历图

```
dat2 <- aggregate(formula = commit ~ year + week + weekday, data = dat, FUN = length)

dat2 <- transform(dat2, colorBin = cut(commit, breaks = c(0, 5, 10, 15, 20, 25)))

ggplot(data = dat2, aes(x = week, y = weekday, fill = colorBin)) +
  scale_fill_brewer(name = "commit", palette = "Greens") +
  geom_tile(color = "white", size = 0.4) +
  facet_wrap("year", ncol = 1) +
  scale_x_continuous(
    expand = c(0, 0),
    breaks = seq(1, 52, length = 12),
    labels = month.abb
  ) +
  labs(x = "", y = "")
```



图 7.43: 《现代统计图形》的活跃情况

7.4.11 岭线图

ggridges 包, [于森](#) 对此图形的来龙去脉做了比较系统的阐述, 详见统计之都[主站文章叠嶂图的前世今生](#)

```
library(ggridges)
ggplot(lincoln_weather, aes(x = `Mean Temperature [F]`, y = Month, fill = stat(x)))
  geom_density_ridges_gradient(scale = 3, rel_min_height = 0.01, gradient_lwd = 1,
    scale_x_continuous(expand = c(0, 0)) +
    scale_y_discrete(expand = expansion(mult = c(0.01, 0.25))) +
    scale_fill_viridis_c(name = "Temp. [F]", option = "C") +
  labs(
    title = 'Temperatures in Lincoln NE',
    subtitle = 'Mean temperatures (Fahrenheit) by month for 2016'
  ) +
  theme_ridges(font_size = 13, grid = TRUE) +
  theme(axis.title.y = element_blank())
```

通过数据可视化的手段帮助肉眼检查两组数据的分布

```
p1 <- ggplot(sleep, aes(x = extra, y = group, fill = group)) +
  geom_density_ridges() +
  theme_ridges()
```

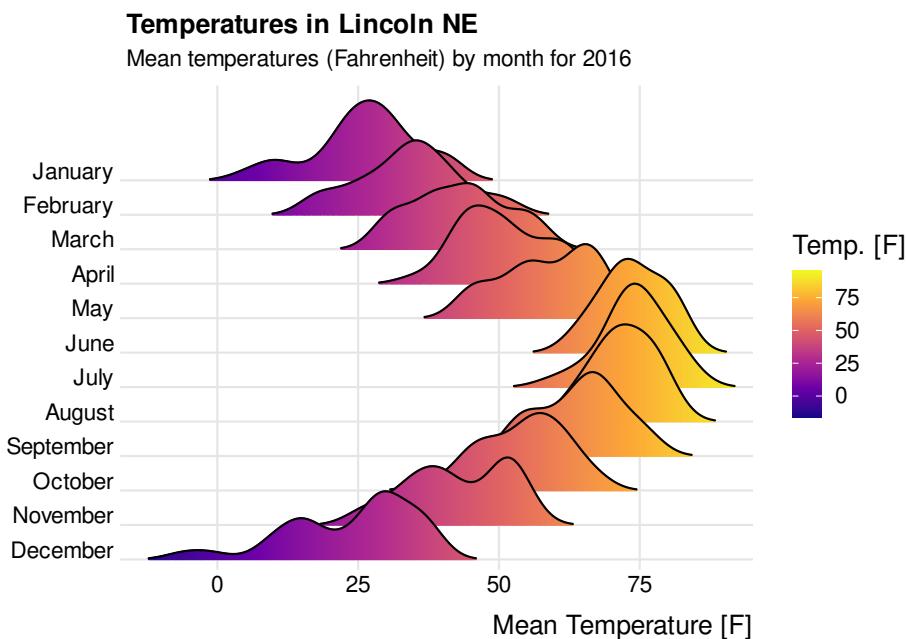


图 7.44: 2016 年在内布拉斯加州林肯市的天气变化

```
p2 <- ggplot(diamonds, aes(x = price, y = color, fill = color)) +
  geom_density_ridges() +
  theme_ridges()

p1 / p2
```

[ridgeline](#) 提供 Base R 绘图方案

7.4.12 椭圆图

type 指定多元分布的类型，type = "t" 和 type = "norm" 分别表示 t 分布和正态分布，geom = "polygon"，以 eruptions > 3 分为两组

```
ggplot(faithful, aes(x = waiting, y = eruptions)) +
  geom_point() +
  stat_ellipse()
```

```
ggplot(faithful, aes(waiting, eruptions, color = eruptions > 3)) +
```

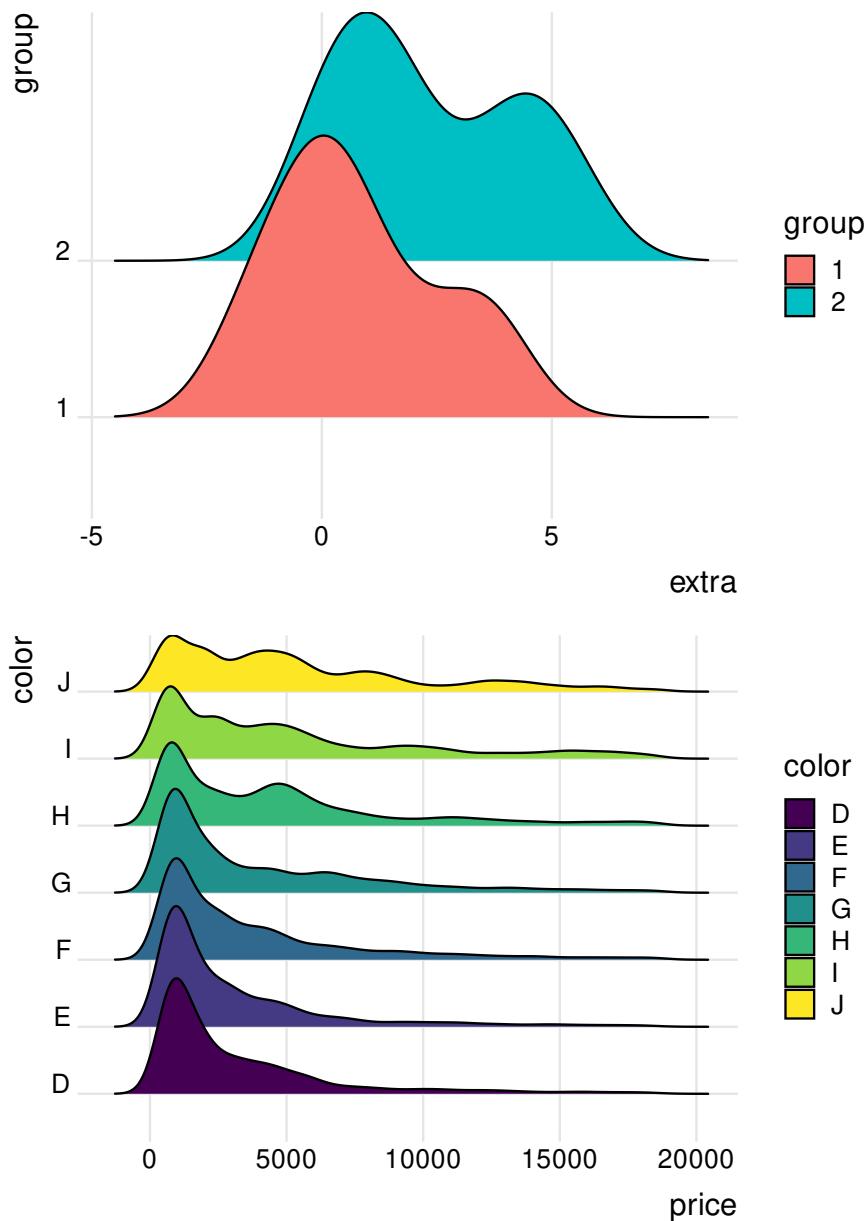
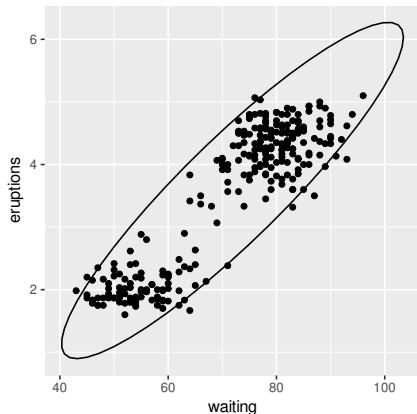
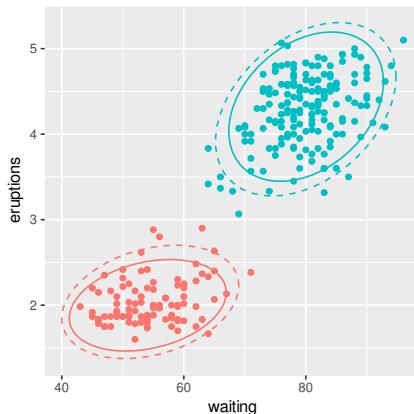


图 7.45: 比较数据的分布

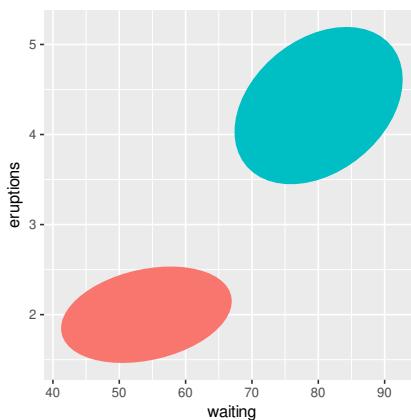
```
geom_point() +  
stat_ellipse(type = "norm", linetype = 2) +  
stat_ellipse(type = "t") +  
theme(legend.position = "none")  
  
④ ggplot(faithful, aes(waiting, eruptions, fill = eruptions > 3)) +  
stat_ellipse(geom = "polygon") +  
theme(legend.position = "none")
```



(a) 简单椭圆图



(b) 正态和 t 分布



(c) 填充几何图形

图 7.46: 几种不同的椭圆图

7.4.13 包络图

ggpubr 包提供了 stat_chull() 图层

```
library(ggpubr)
ggscatter(mpg, x = "displ", y = "hwy", color = "drv") +
  stat_chull(aes(color = drv, fill = drv), alpha = 0.1, geom = "polygon")
```

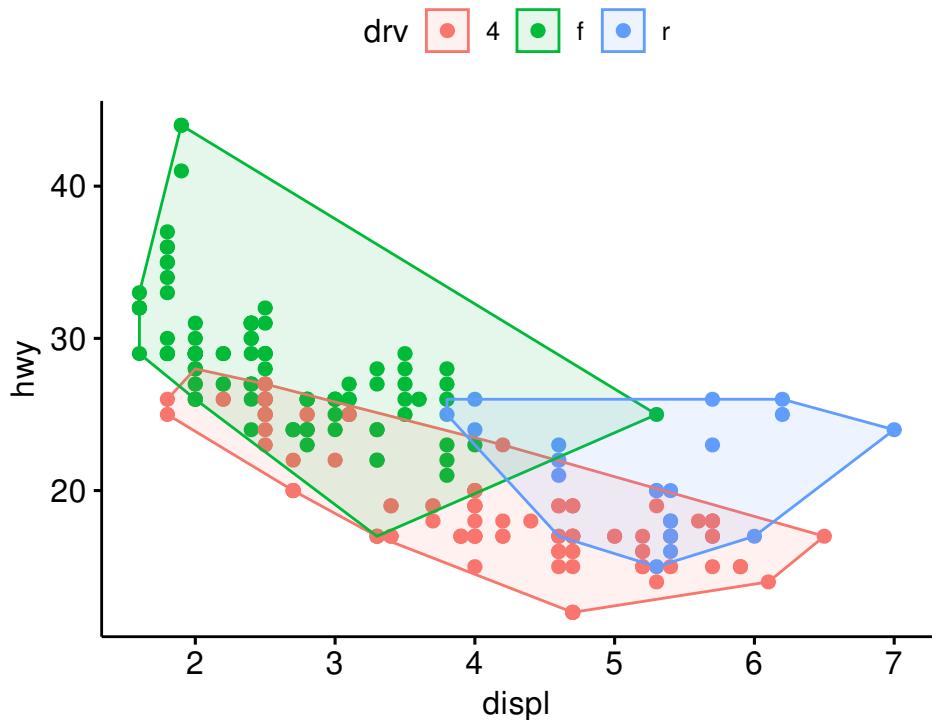


图 7.47: 包络图

其背后的原理如下

```
stat_chull

## function (mapping = NULL, data = NULL, geom = "path", position = "identity",
##           na.rm = FALSE, show.legend = NA, inherit.aes = TRUE, ...)
## {
##   layer(stat = StatChull, data = data, mapping = mapping, geom = geom,
##         position = position, show.legend = show.legend, inherit.aes = inherit.aes,
##         params = list(na.rm = na.rm, ...))
```

```

## }
## <bytecode: 0x564c5d3b5330>
## <environment: namespace:ggpubr>

StatChull <- ggproto("StatChull", Stat,
  compute_group = function(data, scales) {
    data[chull(data$x, data$y), , drop = FALSE]
  },
  required_aes = c("x", "y")
)

stat_chull <- function(mapping = NULL, data = NULL, geom = "polygon",
  position = "identity", na.rm = FALSE, show.legend = NA,
  inherit.aes = TRUE, ...) {
  layer(
    stat = StatChull, data = data, mapping = mapping, geom = geom,
    position = position, show.legend = show.legend, inherit.aes = inherit.aes,
    params = list(na.rm = na.rm, ...))
}

ggplot(mpg, aes(displ, hwy)) +
  geom_point() +
  stat_chull(fill = NA, colour = "black")

ggplot(mpg, aes(displ, hwy, colour = drv)) +
  geom_point() +
  stat_chull(fill = NA)

```

7.4.14 拟合图

```

xx <- -9:9
yy <- sqrt(abs(xx))
plot(xx, yy,
  col = "red",
  xlab = expression(x),

```

```
ylab = expression(sqrt(abs(x)))
)
lines(spline(xx, yy, n = 101, method = "fmm", ties = mean), col = "pink")

myspline <- function(formula, data, ...) {
  dat <- model.frame(formula, data)
  res <- splinefun(dat[[2]], dat[[1]])
  class(res) <- "myspline"
  res
}

predict.myspline <- function(object, newdata, ...) {
  object(newdata[[1]])
}

data.frame(x = -9:9) %>%
  transform(y = sqrt(abs(x))) %>%
  ggplot(aes(x = x, y = y)) +
  geom_point(color = "red", pch = 1, size = 2) +
  stat_smooth(method = myspline, formula = y~x, se = F, color = "pink") +
  labs(x = expression(x), y = expression(sqrt(abs(x)))) +
  theme_minimal()
```

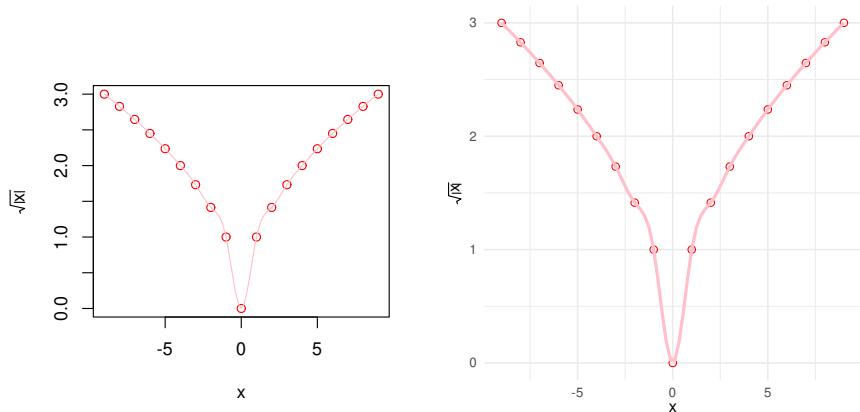


图 7.48: 自定义样条函数

下面以真实数据集 `trees` 为例，介绍 `geom_smooth()` 支持的拟合方法，比如 "lm"



线性回归和 "nls" 非线性回归

```
ggplot(trees, aes(x = log(Girth), y = log(Volume))) +
  geom_point() +
  geom_smooth(method = "lm", formula = y ~ x, se = FALSE)

ggplot(trees, aes(x = Girth, y = Volume)) +
  geom_point() +
  geom_smooth(
    method = "nls", formula = y ~ a * x^2 + b, se = F,
    method.args = list(start = list(a = 5, b = -36))
  )
```

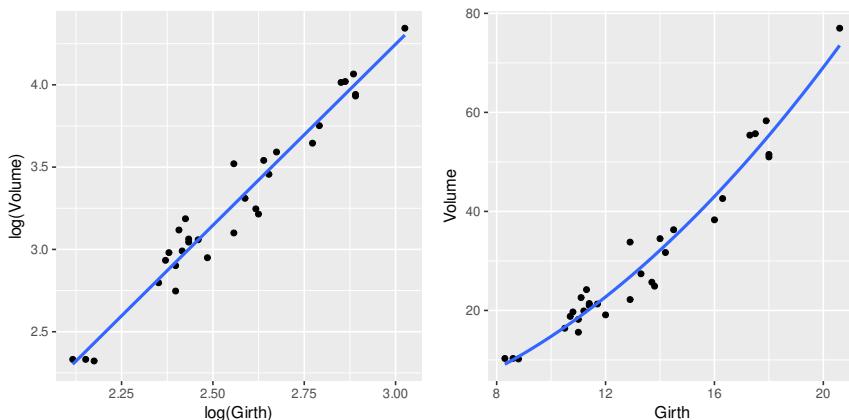


图 7.49: 平滑方法

7.4.15 地形图

区域之间以轮廓分割，轮廓之间以相同的颜色填充，Cleveland 把这个叫做 level plot，**lattice** 包中 `levelplot()` 函数正来源于此。

[Auckland's Maunga Whau Volcano](#) 是火山喷发后留下的渣堆，位于新西兰奥克兰伊甸山郊区。Ross Ihaka 收集了它的地形数据，命名为 `volcano`，打包在 R 软件环境中，见图 7.50

```
filled.contour(volcano,
  color.palette = terrain.colors,
  plot.title = title(
```

```
main = "The Topography of Maunga Whau",
      xlab = "Meters North", ylab = "Meters West"
    ),
  plot.axes = {
    axis(1, seq(100, 800, by = 100))
    axis(2, seq(100, 600, by = 100))
  },
  key.title = title(main = "Height\n(meters)" ),
  key.axes = axis(4, seq(90, 190, by = 10))
}
```

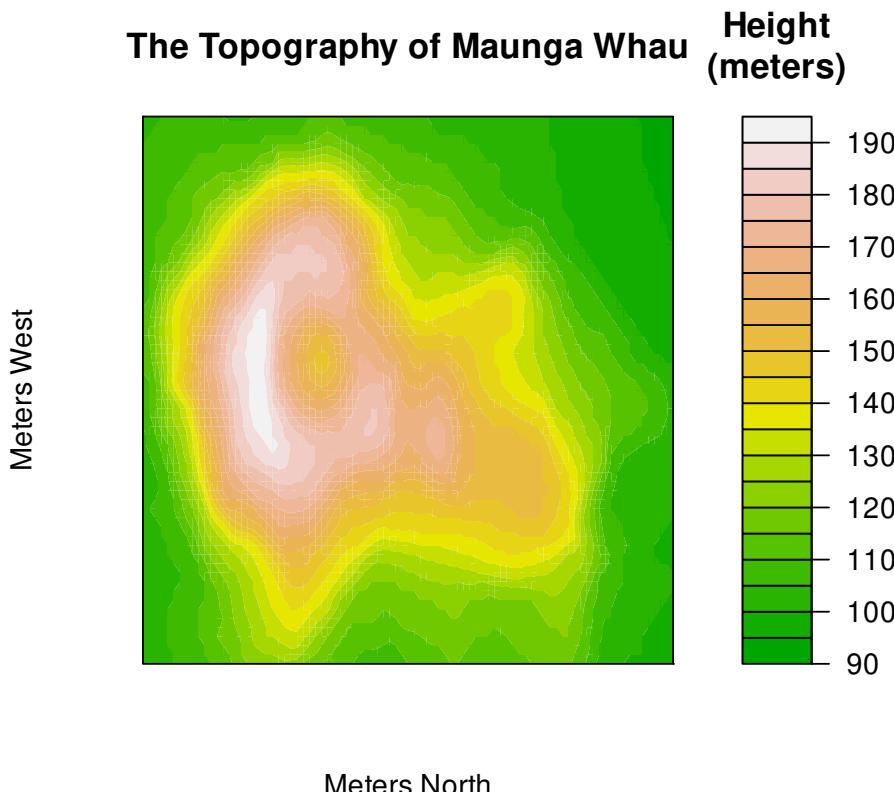


图 7.50: image 图形

美国西南部犹他州锡安国家公园的高程栅格数据 `elevation`, 该数据集由 Jakub Nowosad 收集于 `spDataLarge` 包内, 由于该 R 包收集的地理信息数据很多又很

大，超出了 CRAN 对 R 包的大小限制，需要从作者制作的 drat 站点下载。

```
install.packages("spDataLarge", repos = "https://nowosad.github.io/drat/")
```

④ elevation 数据集通过雷达地形测绘 SRTM (Shuttle Radar Topography Mission) 获得，其分辨率为 $90m \times 90m$ ，属于高精度地形网格数据，更多细节描述见 <http://srtm.csi.cgiar.org/>，图 7.51 将公园的地形清晰地展示出来了，读者不妨再借助维基百科词条 (https://en.wikipedia.org/wiki/Zion_National_Park) 从整体上了解该公园的情况，结合丰富的实景图可以获得更加直观的感受。

```
data("elevation", package = "spDataLarge")
raster::plot(elevation, asp = NA)
```

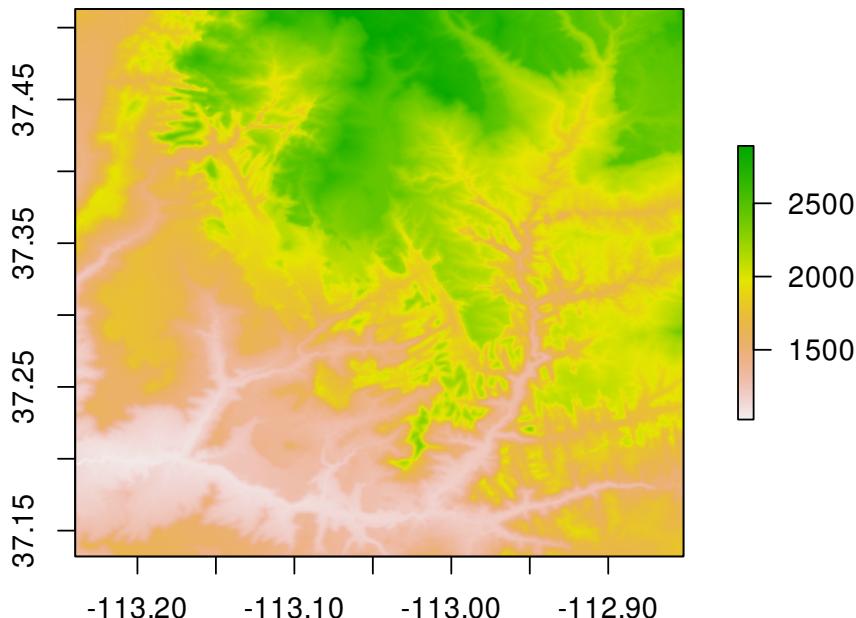


图 7.51：锡安国家公园的高程栅格数据

7.4.16 树状图

数据集 GNI2014 来自 **treemap** 包，是一个 `data.frame` 类型的数据对象，记录了 2014 年每个国家的人口总数 `population` 和国民人均收入 `GNI`，数据样例见下方：

```
library(treemap)
data(GNI2014, package = "treemap")
subset(GNI2014, subset = grepl(x = country, pattern = 'China'))
```



```
##      iso3          country continent population    GNI
##  7   MAC     Macao SAR, China      Asia  559846 76270
## 33  HKG Hong Kong SAR, China      Asia 7061200 40320
## 87  CHN           China      Asia 1338612970  7400
```

数据呈现明显的层级结构，从大洲到国家记录人口数量和人均收入，矩阵树图以方块大小表示人口数量，以颜色深浅表示人均收入，见图7.52

```
treemap(GNI2014,
        index = c("continent", "iso3"),
        vSize = "population",
        vColor = "GNI",
        type = "value",
        format.legend = list(scientific = FALSE, big.mark = " ")
)
```

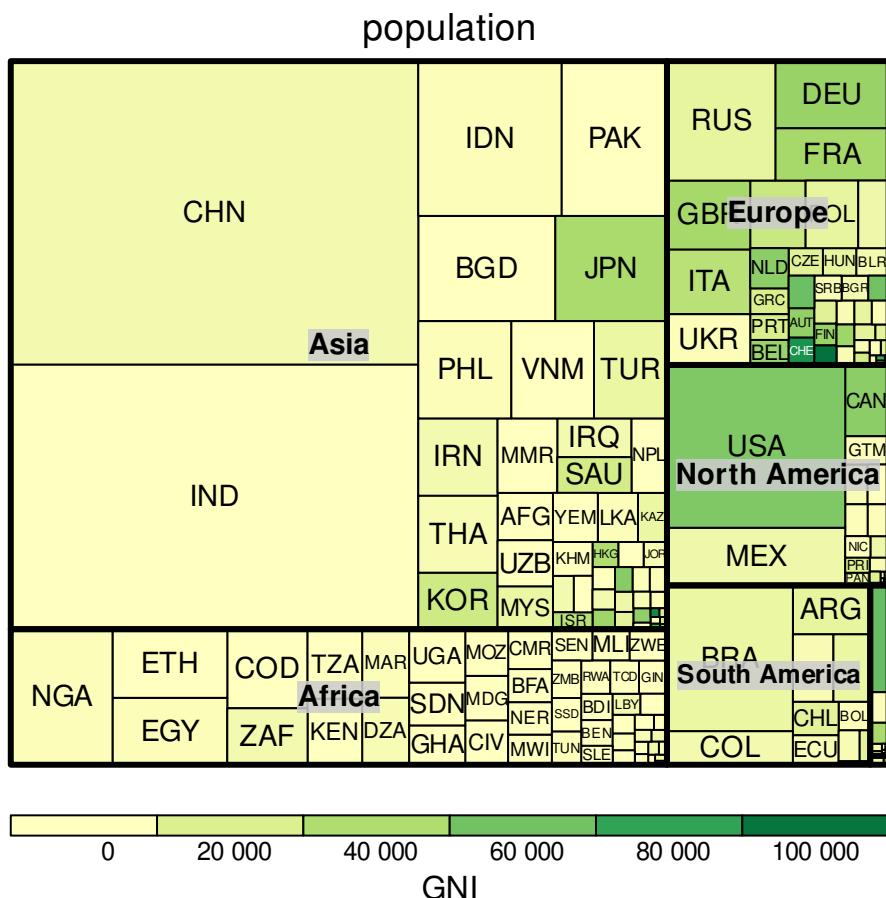


图 7.52: 矩阵树图



提示

数据集 GNI2014 的另一种呈现方式是将数据铺到地图上，可以借助 **highcharter** 包的 `hcmap()` 函数来实现。

```
# 代码块不要启用缓存
data(GNI2014, package = "treemap")
library(highcharter)
hcmap(
  "custom/world-robinson-lowres",
  data = GNI2014,
  name = "Gross national income per capita",
  value = "GNI",
  borderWidth = 0,
  nullColor = "gray",
  joinBy = c("iso-a3", "iso3")
) %>%
  hc_colorAxis(
    stops = color_stops(
      colors = terrain.colors(n = 10)
    ),
    type = "logarithmic"
)
```

treemapify 包基于 `ggplot2` 制作树状图，类似地，该 R 包内置了数据集 `G20`，记录了世界主要经济体 `G20` (<https://en.wikipedia.org/wiki/G20>) 的经济和人口信息，国家 `GDP`（单位：百万美元）`gdp_mil_usd` 和人类发展指数 `hdi`。相比于 `GNI2014`，它还包含了两列标签信息：经济发展阶段和所处的半球。图 @`(fig:treemap-ggplot2)` 以南北半球 `hemisphere` 分面，以色彩填充区域 `region`，以 `gdp_mil_usd` 表示区域大小

```
library(treemapify)
ggplot(G20, aes(
  area = gdp_mil_usd, fill = region,
  label = country, subgroup = region
)) +
  geom_treemap() +
  geom_treemap_text(grow = T, reflow = T, colour = "black") +
```



```

facet_wrap(~hemisphere) +
scale_fill_brewer(palette = "Set1") +
theme(legend.position = "bottom") +
labs(
  title = "The G-20 major economies by hemisphere",
  caption = "The area of each tile represents the country's GDP as a
            proportion of all countries in that hemisphere",
  fill = "Region"
)

```

The G-20 major economies by hemisphere

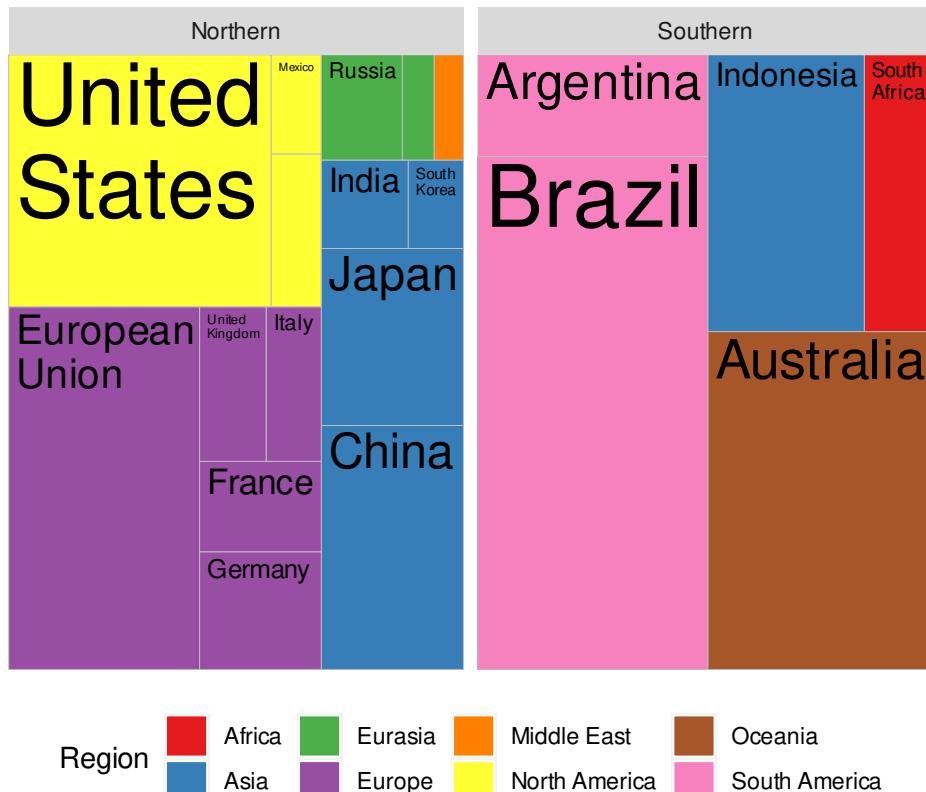
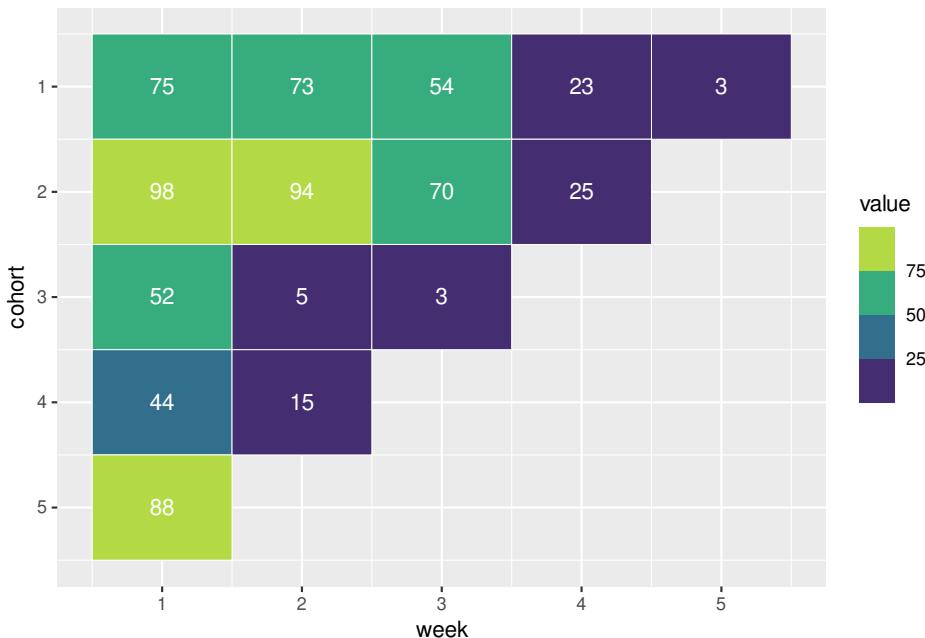


图 7.53: 世界主要经济体 G20 的人口和经济信息

7.4.17 留存图

```
cohort <- data.frame(  
  cohort = rep(1:5, times = 5:1),  
  week = c(1:5, 1:4, 1:3, 1:2, 1),  
  value = c(  
    75, 73, 54, 23, 3,  
    98, 94, 70, 25,  
    52, 5, 3,  
    44, 15,  
    88  
)  
)  
  
ggplot(cohort, aes(x = week, y = cohort, fill = value)) +  
  geom_tile(color = "white") +  
  geom_text(aes(label = value), color = "white") +  
  scale_y_reverse() +  
  scale_fill_binned(type = "viridis")
```





7.4.18 瀑布图

瀑布图 waterfall 与上月相比，谁增谁减，用瀑布图分别表示占比和绝对数值。[瀑布图 waterfall](#)

```
balance <- data.frame(
  event = c(
    "Starting\nCash", "Sales", "Refunds",
    "Payouts", "Court\nLosses", "Court\nWins", "Contracts", "End\nCash"
  ),
  change = c(2000, 3400, -1100, -100, -6600, 3800, 1400, -2800)
)

balance$balance <- cumsum(c(0, balance$change[-nrow(balance)])) # 累计值
balance$time <- 1:nrow(balance)
balance$flow <- factor(sign(balance$change)) # 变化为正还是为负

ggplot(balance) +
  geom_hline(yintercept = 0, colour = "white", size = 2) +
  geom_rect(aes(
    xmin = time - 0.45, xmax = time + 0.45,
    ymin = balance, ymax = balance + change, fill = flow
  )) +
  geom_text(aes(
    x = time,
    y = pmin(balance, balance + change) - 50,
    label = scales::dollar(change)
  ),
  hjust = 0.5, vjust = 1, size = 3
) +
  scale_x_continuous(
    name = "",
    breaks = balance$time,
    labels = balance$event
  ) +
  scale_y_continuous(
    name = "Balance",
```

```
  labels = scales::dollar
) +
scale_fill_brewer(palette = "Spectral") +
theme_minimal()
```



图 7.54: 瀑布图

7.4.19 桑基图

[ggalluvial](#)

```
titanic_wide <- data.frame(Titanic)
```

```
head(titanic_wide)
```

```
##   Class     Sex   Age Survived Freq
## 1   1st     Male Child      No     0
## 2   2nd     Male Child      No     0
## 3   3rd     Male Child      No    35
## 4 Crew     Male Child      No     0
## 5   1st Female Child      No     0
```

```

## 6 2nd Female Child      No      0

library(ggalluvial)
ggplot(data = titanic_wide,
       aes(axis1 = Class, axis2 = Sex, axis3 = Age,
            y = Freq)) +
  scale_x_discrete(limits = c("Class", "Sex", "Age"), expand = c(.2, .05)) +
  xlab("Demographic") +
  geom_alluvium(aes(fill = Survived)) +
  geom_stratum() +
  geom_text(stat = "stratum", aes(label = after_stat(stratum))) +
  theme_minimal() +
  ggtitle("passengers on the maiden voyage of the Titanic",
          "stratified by demographics and survival")

```

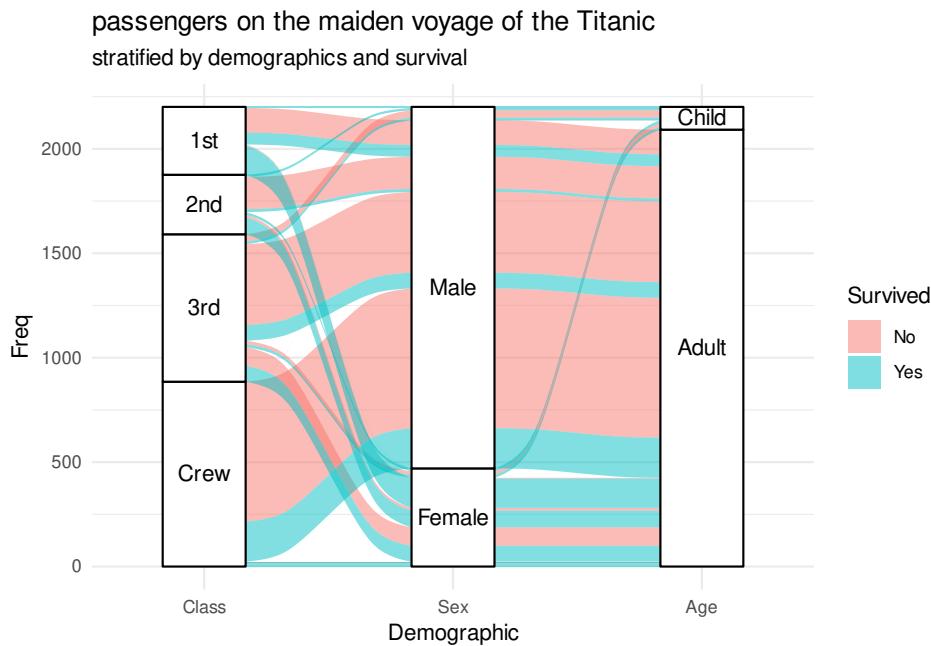


图 7.55: 桑基图

7.4.20 马赛克图

```
library(ggmosaic)
ggplot(data = as.data.frame(UCBAdmissions)) +
  geom_mosaic(aes(weight = Freq, x = product(Gender, Admit), fill = Dept)) +
  coord_flip() +
  theme_minimal() +
  labs(x = "Admit", y = "Gender")
```

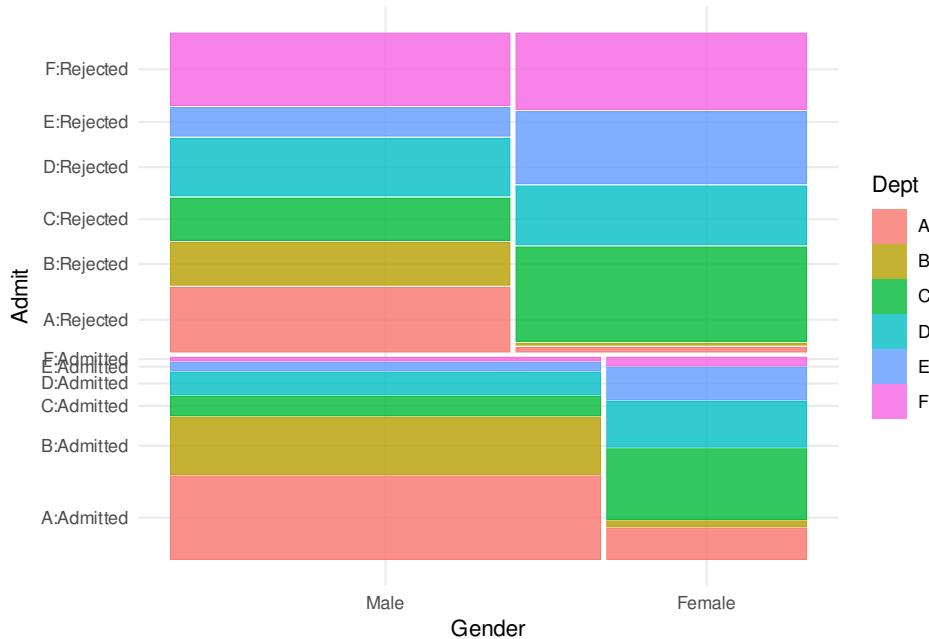


图 7.56: UCBAdmissions 马赛克图

7.4.21 凹凸图

ggbump 排序随位置的变化

```
# remotes::install_github("davidsjoberg/ggbump")
library(ggbump)
# 代码修改自 https://github.com/davidsjoberg/ggbump
df <- data.frame(
  season = c(
```



```
"Spring", "Pre-season", "Summer", "Season finale", "Autumn", "Winter",
"Spring", "Pre-season", "Summer", "Season finale", "Autumn", "Winter",
"Spring", "Pre-season", "Summer", "Season finale", "Autumn", "Winter",
"Spring", "Pre-season", "Summer", "Season finale", "Autumn", "Winter"
),
rank = c(
  1, 3, 4, 2, 1, 4,
  2, 4, 1, 3, 2, 3,
  4, 1, 2, 4, 4, 1,
  3, 2, 3, 1, 3, 2
),
player = c(
  rep("David", 6),
  rep("Anna", 6),
  rep("Franz", 6),
  rep("Ika", 6)
)
)

# Create factors and order factor
df <- transform(df, season = factor(season, levels = unique(season)))

# Add manual axis labels to plot
ggplot(df, aes(season, rank, color = player)) +
  geom_bump(size = 2, smooth = 20, show.legend = F) +
  geom_point(size = 5, aes(shape = player)) +
  theme_minimal(base_size = 10, base_line_size = 0) +
  theme(panel.grid.major = element_blank(),
        axis.ticks = element_blank()) +
  scale_color_manual(values = RColorBrewer::brewer.pal(name = "Set2", n = 4))
```

7.4.22 水流图

常用于时间序列数据展示的堆积区域图，`ggstream` 和 `streamgraph`

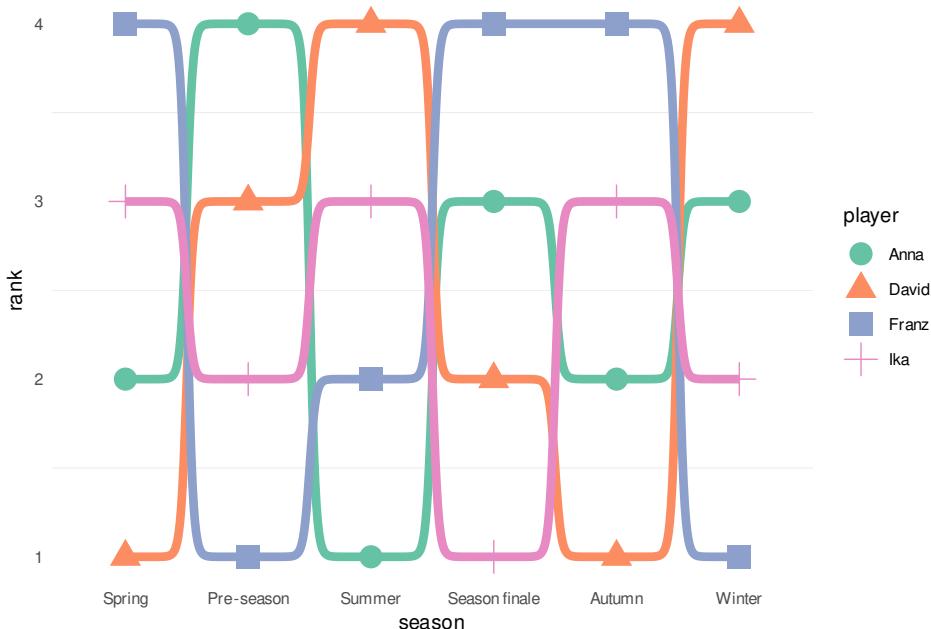


图 7.57: 凹凸图

```
library(ggstream)

ggplot(blockbusters, aes(year, box_office, fill = genre)) +
  geom_stream() +
  theme_minimal()
```

7.4.23 时间线

```
# 交互动态图 https://github.com/shosaco/vistime
# 刘思喆 2018 数据科学的时间轴 https://bjt.name/2018/11/18/timeline.html
x <- read.table(
  textConnection("
```

The Future of Data Analysis,1962

Relational Database,1970

Data science(Peter Naur),1974

Two-Way Communication,1975

```
)
```

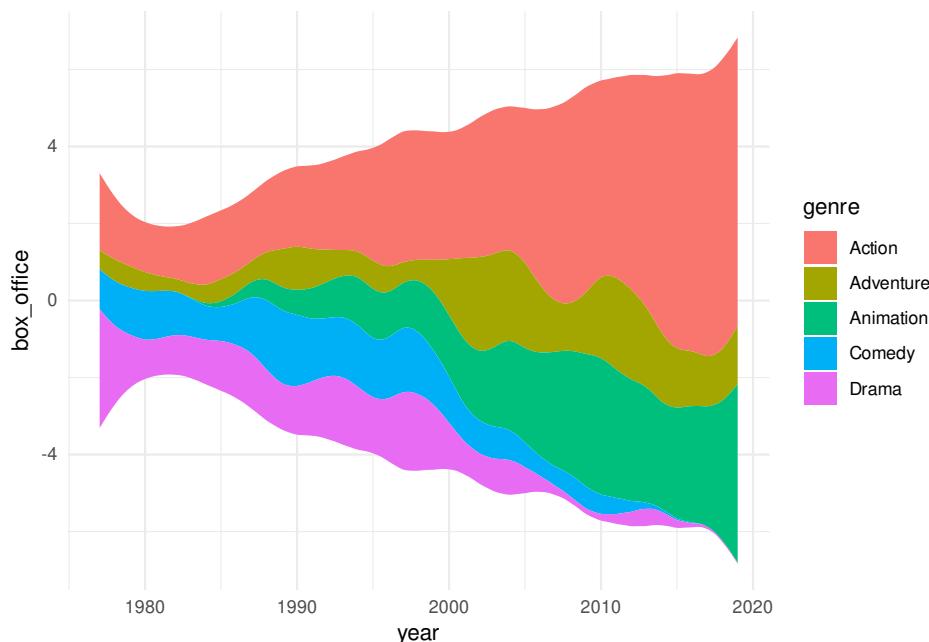


图 7.58: 堆积区域图

```
Exploratory Data Analysis,1977
Business Intelligence,1989
The First Database Report,1992
The World Wide Web Explodes,1995
Data Mining and Knowledge Discovery,1997
S(ACM Software System Award),1998
Statistical Modeling: The Two Cultures,2001
Hadoop,2006
Data scientist,2008
NOSQL,2009
Deep Learning,2015
"),
  sep = ","
)
names(x) <- c("Event", "EventDate")
x$EventDate <- as.Date(paste(x$EventDate, "/01/01", sep = ""))
```



```
library(timelines)
timelines(x,
  labels = paste(x[[1]], format(x[[2]], "%Y")),
  line.color = "blue", label.angle = 15
)
```

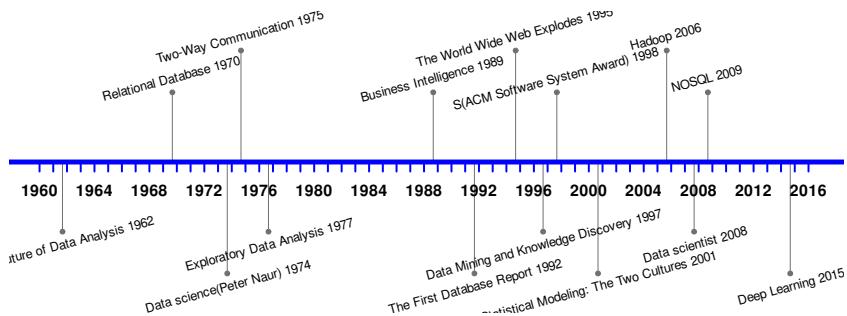


图 7.59: 数据科学的时间轴

```
library(timeline)
data(ww2, package = 'timeline')
timeline(ww2, ww2.events, event.spots=2, event.label='', event.above=FALSE)

# 适合放在动态幻灯片
# 美团风格的写轮眼
# 时间线

library(vistime)
# presidents and vice presidents
pres <- data.frame(
  Position = rep(c("President", "Vice"), each = 3),
  Name = c("Washington", rep(c("Adams", "Jefferson"), 2), "Burr"),
  start = c("1789-03-29", "1797-02-03", "1801-02-03"),
  end = c("1797-02-03", "1801-02-03", "1809-02-03"),
  color = c("#cbb69d", "#603913", "#c69c6e")
)
```

```
hc_vistime(pres, col.event = "Position", col.group = "Name",
           title = "Presidents of the USA")
```

7.4.24 三元图

 Ternary 使用基础图形库，而 ggtern 使用 ggplot2 绘制

```
library(ggtern)
library(ggalt)
data("Fragments")
ggtern(Fragments, aes(
  x = Qm, y = Qp, z = Rf + M,
  fill = GrainSize, shape = GrainSize
)) +
  geom_encircle(alpha = 0.5, size = 1) +
  geom_point() +
  labs(
    title = "Example Plot",
    subtitle = "using geom_encircle"
  ) +
  theme_bw() +
  theme_legend_position("tr")
```

7.4.25 四象限图

```
dat <- data.frame(
  perc = c(54, 18, 5, 15),
  wall_policy = c("oppose", "favor", "oppose", "favor"),
  dreamer_policy = c("favor", "favor", "oppose", "oppose"),
  stringsAsFactors = FALSE
) %>%
  transform(
    xmin = ifelse(wall_policy == "oppose", -sqrt(perc), 0),
    xmax = ifelse(wall_policy == "favor", sqrt(perc), 0),
    ymin = ifelse(dreamer_policy == "oppose", -sqrt(perc), 0),
```



```
ymax = ifelse(dreamer_policy == "favor", sqrt(perc), 0)
)

ggplot(data = dat) +
  geom_rect(aes(
    xmin = xmin, xmax = xmax,
    ymin = ymin, ymax = ymax
  ), fill = "grey") +
  geom_text(aes(
    x = xmin + 0.5 * sqrt(perc),
    y = ymin + 0.5 * sqrt(perc),
    label = perc
  ),
  color = "white", size = 10
) +
  coord_equal() +
  geom_hline(yintercept = 0) +
  geom_vline(xintercept = 0) +
  theme_minimal() +
  labs(x = "", y = "", title = "")
```

7.4.26 韦恩图

`ggVennDiagram`

7.4.27 龙卷风图

```
dat <- data.frame(
  variable = c("A", "B", "A", "B"),
  Level = c("Top-2", "Top-2", "Bottom-2", "Bottom-2"),
  value = c(.8, .7, -.2, -.3)
)
ggplot(dat, aes(x = variable, y = value, fill = Level)) +
  geom_bar(position = "identity", stat = "identity") +
  scale_y_continuous(labels = abs)
```

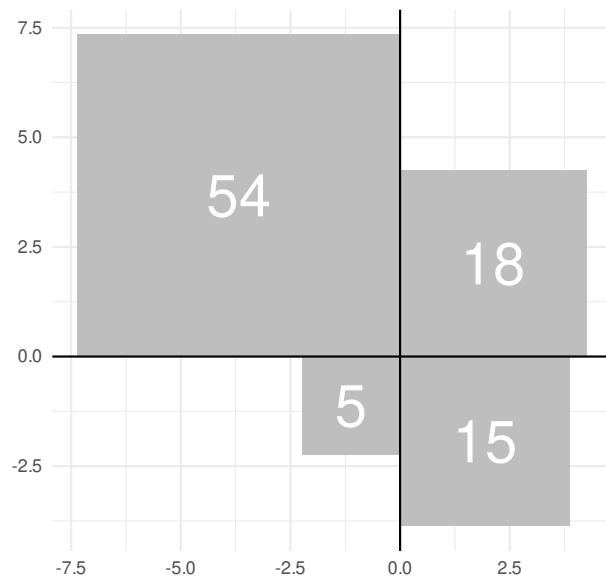


图 7.60: 四象限图

```
coord_flip() +  
theme_minimal()
```

Tornado diagram 主要用于敏感性分析，比较不同变量的重要性程度。条形图 `geom_bar()` 图层的变体，模型权重可视化的手段，仅限于广义线性模型。

7.4.28 聚类图

`ggdendro` 的 `dendro_data()` 函数支持 `tree`、`hclust`、`dendrogram` 和 `rpart` 结果的整理，进而绘图

```
library(ggdendro)  
hc <- hclust(dist(USArrests), "ave")  
hcdata <- dendro_data(hc, type = "rectangle")  
ggplot() +  
  geom_segment(data = segment(hcdata),  
               aes(x = x, y = y, xend = xend, yend = yend))  
  ) +
```

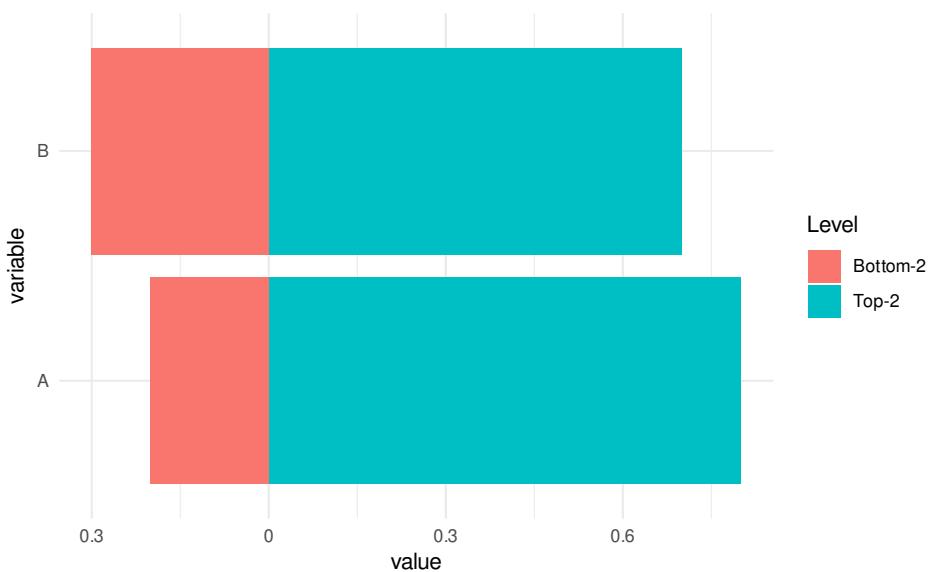
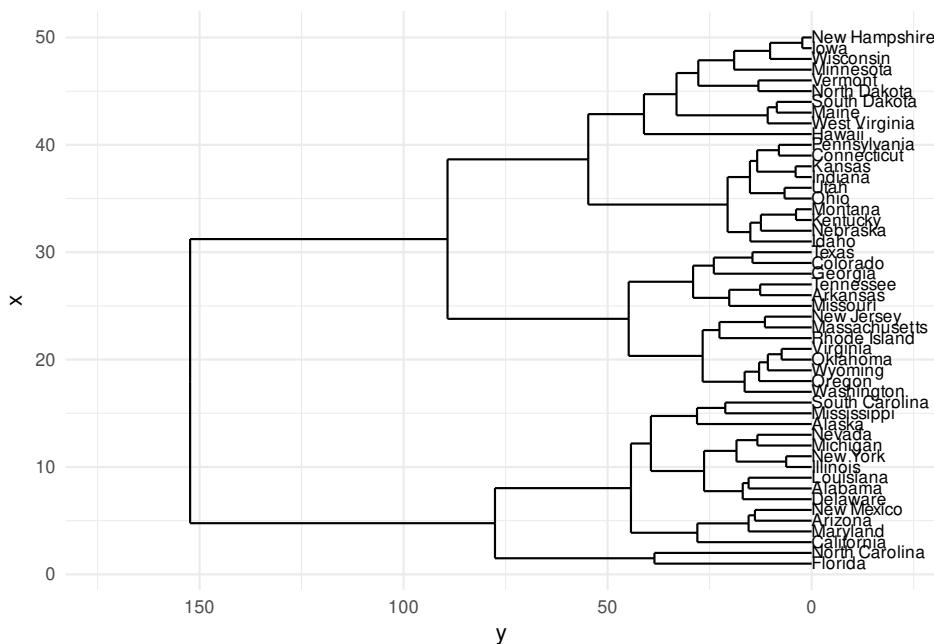


图 7.61: 龙卷风图展示变量重要性

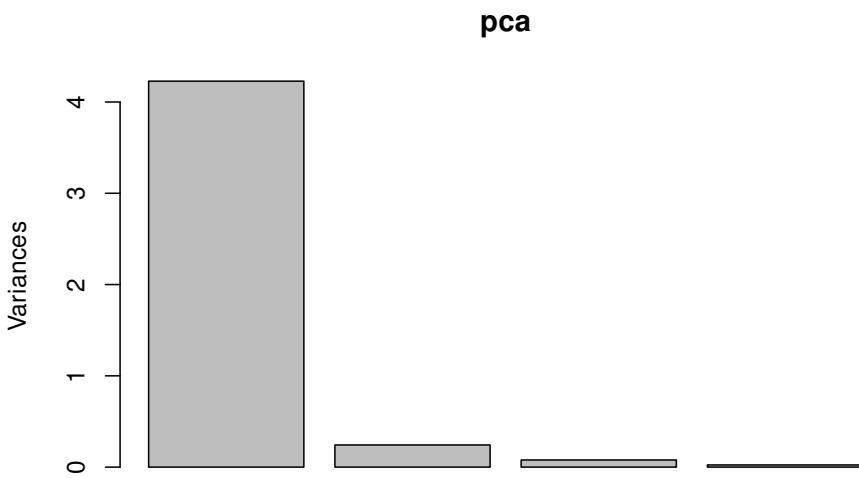
```
geom_text(data = label(hcdata),
          aes(x = x, y = y, label = label, hjust = 0),
          size = 3
) +
coord_flip() +
scale_y_reverse(expand = c(0.2, 0)) +
theme_minimal()
```



7.4.29 主成分图

借助 **autoplotly** 包 [Tang, 2018] 可将函数 `stats:::prcomp` 生成的结果转化为交互图形

```
pca <- prcomp(iris[c(1, 2, 3, 4)])
plot(pca)
```



```
library(autoplotly)
autoplotly(pca,
  data = iris, colour = "Species",
  label = TRUE, label.size = 3, frame = TRUE
)
```

ggfortify [Tang et al., 2016] 包将主成分分析图转化为静态图形

```
library(ggfortify)
autoplot(pca, data = iris, colour = 'Species')
```

7.4.30 组合图

组合的意思是将不同种类的图形绘制在一个区域中，比如密度曲线和地毯图⁸组合。**GGally**、**ggupset**、**ggcharts** 和 **ggbpbr** 高度定制了一些组合统计图形，以 **ggbpbr** 为例，见图 7.63。

```
library(ggbpbr)
ggdensity(sleep,
```

⁸其实是轴须图 rug plot，只因样子看起来像铺在地上的毛毯，故而称之为地毯图，对应于 R 内置的 rug() 函数或 ggplot2 提供的图层 geom_rug()，更多解释详见 https://en.wikipedia.org/wiki/Rug_plot。

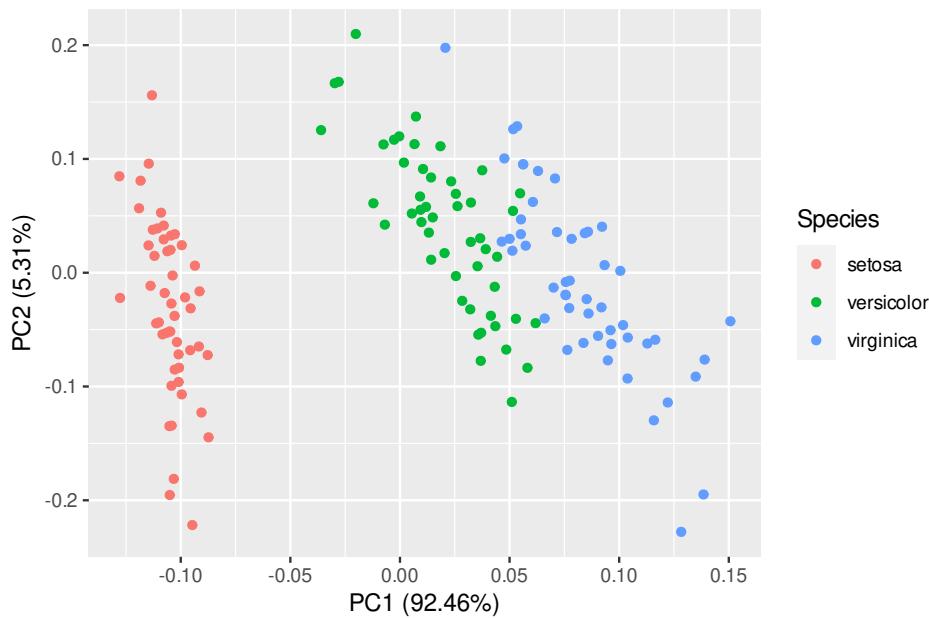


图 7.62: 主成分分析

```
x = "extra", add = "mean", rug = TRUE, color = "group",
fill = "group", palette = c("#00AFBB", "#E7B800")
)
```

上面介绍的都是已经固化的组合方式，一般地，将多个图形组合到一个图中，可以有很多办法，比如 Claus Wilke 开发的 `cowplot`，在他的书里 `Fundamentals of Data Visualization` 大量使用，后起之秀 `patchwork` 则提供更加简洁的组合语法，非常受欢迎，更加底层的拼接方法可以去看 [一页多图](#) 和 R 内置的 `grid` 系统。

7.4.31 动态图

`av` 包基于 `FFmpeg` 将静态图片合成视频，而 `gifski` 包基于 `gifski` 将静态图片合成 GIF 动画，`animation` 包 [Xie, 2013] 将 Base R 绘制的图形转化为动画或视频，`mapmate` 制作地图相关的三维可视化图形，`ganimate` 包支持将 `ggplot2` 生成的图形，`magick` 可以将一系列静态图形合成动态图形，借助 `gifski` 包转化为动态图片或视频。推荐读者从 [ganimate 案例合集](#) 开始制作动态图形。`rgl` 可以制作真三维动态图形，支持缩放、拖拽、旋转等操作，`rayshader` 还支持转化 `ggplot2` 对象为 3D 图形。

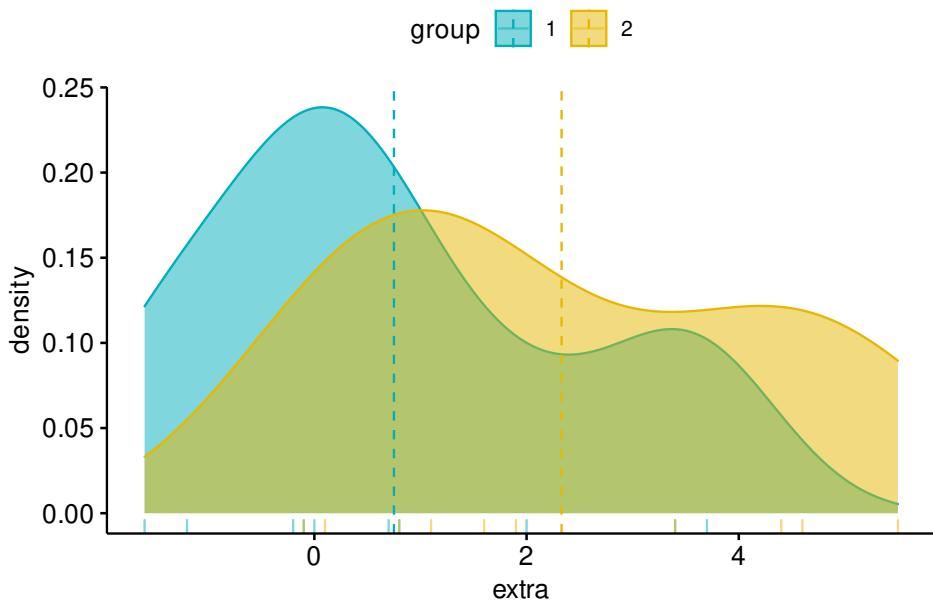


图 7.63: 组合图形

数据集 Indometh 记录了药物在人体中的代谢情况，给 6 个人分别静脉注射了吲哚美辛，每隔一段时间抽血检查药物在血浆中的浓度，收集的数据见表 7.3

```
reshape(Indometh, v.names = "conc", idvar = "Subject",
       timevar = "time", direction = "wide", sep = "") %>%
knitr::kable(),
caption = "吲哚美辛在人体中的代谢情况",
row.names = FALSE, col.names = gsub("(conc)", "", names(.)),
align = "c"
)
```

如图 7.64 所示，药物在人体中浓度变化情况

```
p <- ggplot(
  data = Indometh,
  aes(x = time, y = conc, color = Subject)
) +
  geom_point() +
  geom_line() +
  theme_minimal() +
```

表 7.3: 吲哚美辛在人体中的代谢情况

Subject	0.25	0.5	0.75	1	1.25	2	3	4	5	6	8
1	1.50	0.94	0.78	0.48	0.37	0.19	0.12	0.11	0.08	0.07	0.05
2	2.03	1.63	0.71	0.70	0.64	0.36	0.32	0.20	0.25	0.12	0.08
3	2.72	1.49	1.16	0.80	0.80	0.39	0.22	0.12	0.11	0.08	0.08
4	1.85	1.39	1.02	0.89	0.59	0.40	0.16	0.11	0.10	0.07	0.07
5	2.05	1.04	0.81	0.39	0.30	0.23	0.13	0.11	0.08	0.10	0.06
6	2.31	1.44	1.03	0.84	0.64	0.42	0.24	0.17	0.13	0.10	0.09

```
  labs(  
    x = "time (hr)",  
    y = "plasma concentrations of indometacin (mcg/ml)"  
  )  
p
```

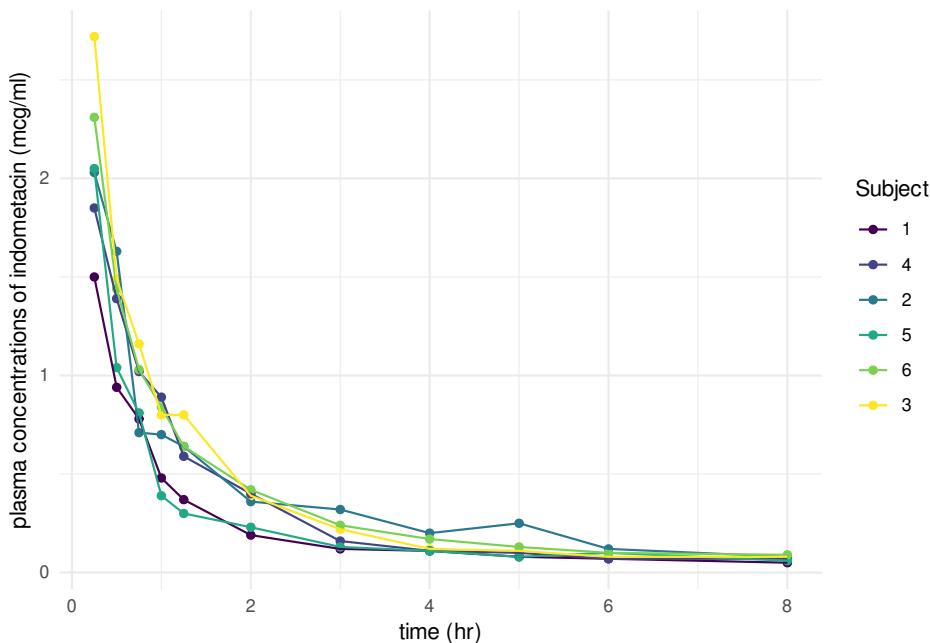


图 7.64: 药物在人体中的代谢情况

```
library(gganimate)
p + transition_reveal(time)
```

提示

书籍目标输出格式是 PDF，则在代码块选项设置里必须指定参数 `fig.show='animate'` 否则插入的只是图片而不是动画，目标格式是 HTML 网页，就不必指定参数，默认会将图片合成 GIF 动态图，嵌入 PDF 里面的动画需要 Acrobat Reader 阅读器才能正确地显示。

动态图形制作的原理，简单来说，就是将一帧帧静态图形以较快的速度播放，人眼形成视觉残留，以为是连续的画面，相比于 `animation`, **gganimate** 借助 **tweenr** 包添加了过渡效果，动态图形显得非常自然。下面以 `cup` 函数⁹为例

$$f(x; \theta, \phi) = \theta x \log(x) - \frac{1}{\phi} e^{-\phi^4(x - \frac{1}{e})^4}, \quad \theta \in (2, 3), \phi \in (30, 50), x \in (0, 1)$$

函数图像随着 θ 和 ϕ 的变化情况见图 7.65。

⁹函数来自余光创的博客 – [3D 版邪恶的曲线](#)，此处借用 `gganimate` 将其动态化，前方高能，少儿不宜，R 还能这么不正经的玩。



```
library(tweenr)
cup_curve <- function(n = 100, theta = 3, phi = 30, cup = "A") {
  data.frame(x = seq(0.00001, 1, length.out = n), cup = cup) %>%
    transform(y = theta * x * log(x, base = 10)
              - 1 / phi * exp(-(phi * x - phi / exp(1))^4))
}
mapply(
  FUN = cup_curve, theta = c(E = 3, D = 2.8, C = 2.5, B = 2.2, A = 2),
  phi = c(30, 33, 36, 40, 50), cup = c("E", "D", "C", "B", "A"),
  MoreArgs = list(n = 50), SIMPLIFY = FALSE, USE.NAMES = TRUE
) %>%
tween_states(
  data = .,
  tweenlength = 2, statelength = 1,
  ease = rep("cubic-in-out", 4), nframes = 100
) %>%
ggplot(data = ., aes(x, y, color = cup, frame = .frame)) +
  geom_path() +
  coord_flip() +
  theme_void()
```

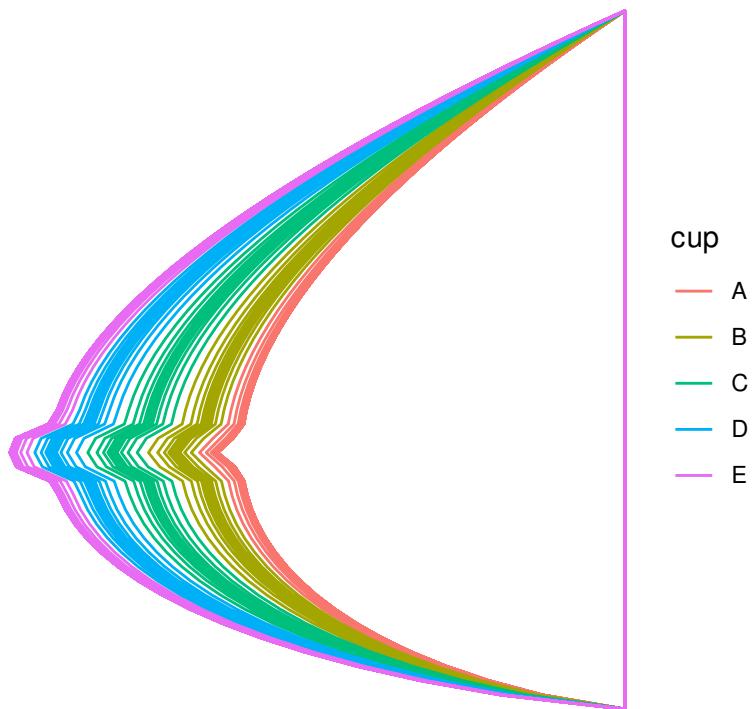


图 7.65: 添加过渡效果

第八章 动态文档



图 8.1: R Markdown 极其周边生态

`WrapRmd` 将 R Markdown 里很长的文本自动断行，但不产生空行。`regeexplain` 帮助检查正则表达式，`rdoc` 支持 R 帮助文档的语法高亮。`shinyComponents` 实现在 R Markdown 中写 shiny。`wordcountaddin` 统计 R Markdown 文档中的单词数量。`styler` 格式化 R Markdown 文档中的代码块。`reprex` 添加代码执行的软件环境，提供可重复的例子，方便在论坛/Github 上发问。`carbonate` 将源代码截图。`downloadthis` 在 R Markdown 文档中添加下载按钮。`icon` 添加各种各样的图标，`thematic` 定制 R Markdown 主题。`datadrivencv`、`vitae` 制作基于 R Markdown 文档的简历。`addinslist` 收集了一系列 RStudio 插件，提高写作和编码的效率。`posterdown` 写宣传海报，`redoc` 实现 R Markdown 和 Microsoft Word 两种文档格式之间互相转化，`rrtools` 写可重复性的研究论文和报告，提供一套自动化的软件环境的配置，节省科研人员的时间。`butteRfly` 快速获取 Github 等社交网络上活动记录，以日历图的形式展现出来。`flow` 可以非常方便地制作函数内部调用执行的流程图。



`minidown` 提供轻量级的 CSS 框架打磨的网页模版, `rmdformats` 和 `prettydoc` 提供不同主题样式的网页输出, `govdown` 提供 GOV.UK 风格的网页模版。

`uiucthemes` 伊利诺伊大学主题的 R Markdown 模版, `rmdshower` 提供 `shower` 引擎打造的幻灯片, 而 `xaringan` 是基于 `remark.js`。`xaringanthemer` 和 `xaringanExtra` 包含丰富的 `xaringan` 的主题。

`slidex` 可以将 PowerPoint 幻灯片转化为粗燥的 `xaringan` 幻灯片。

`gluedown` 用 R 代码写格式化的 Markdown 文本,

- Reproducible Research Data and Project Management in R <https://annakrystalli.me/rrresearchACCE20/>
- Higher, further, faster with Marvelous R Markdown <https://bit.ly/marvelRMD>
- R Markdown for Scientists <https://rmd4sci.njtierney.com/>
- Getting Used to R, RStudio, and R Markdown <https://rbasics.netlify.app/>
- R Markdown 指南手册 <https://www.dataquest.io/blog/r-markdown-guide-cheatsheet/>
- Statistical Inference via Data Science: A Modern Dive into R and the tidyverse <https://moderndive.com/>
- 参数化报告 <https://github.com/jenniferthompson/ParamRmdExample> 和 <https://elastic-lovelace-155848.netlify.app/gallery/themes/flatly.html>
- Sharing analyses with R Markdown https://andrewbtran.github.io/NICAR/2018/workflow/docs/02_rmarkdown.html
- Introduction to the Normal Distribution https://tinystats.github.io/teacups-giraffes-and-statistics/02_bellCurve.html
- 混合效应模型的 workshop https://github.com/singmann/mixed_model_workshop
- 基于 `thematic` 和 `bslib` 包美化 Rmd 文档 <https://www.tillac-data.com/2020-fast-rmd-theming-with-thematic-and-bootstraplib/>
- 借助 `flipbookr` 在 `xaringan` 制作的幻灯片里逐行展示代码执行的效果, 特别适合用于 `ggplot2` 的教学 https://evamaerey.github.io/little_flipbooks_library/flipbookr/skeleton
- 制作 note/tips 等自定义块 <https://desiree.rbind.io/post/2019/making-tip-boxes-with-bookdown-and-rmarkdown/>

8.1 文档元素

knitr 将 R Markdown 文件转化为 Markdown 文件，Pandoc 可以将 Markdown 文件转化为 HTML5、Word、PowerPoint 和 PDF 等文档格式。



图 8.2: rmarkdown 支持的输出格式

rmarkdown 自 2014 年 09 月 17 日在 CRAN 上发布第一个正式版本以来，逐渐形成了一个强大的生态系统，世界各地的开发者贡献各种各样的扩展功能，见图 8.3

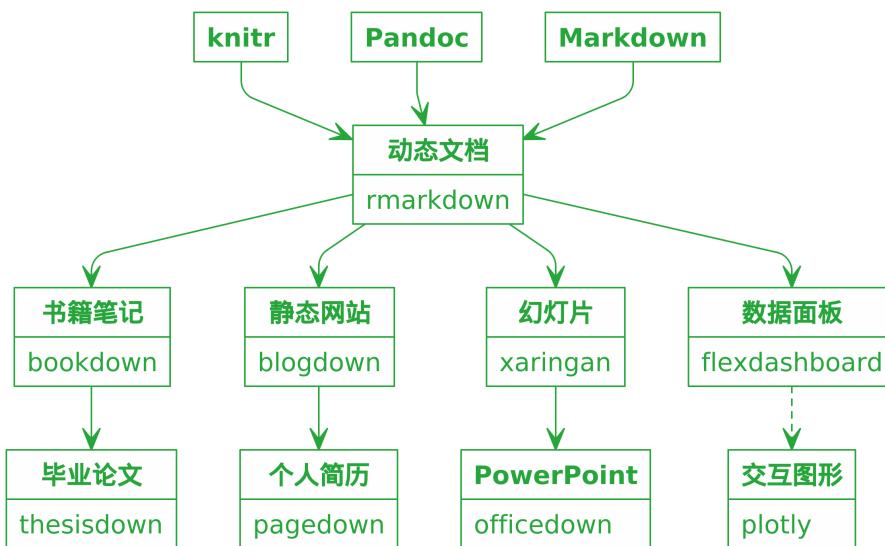


图 8.3: rmarkdown 生态系统

8.1.1 控制选项

[Using SQL in RStudio](#)

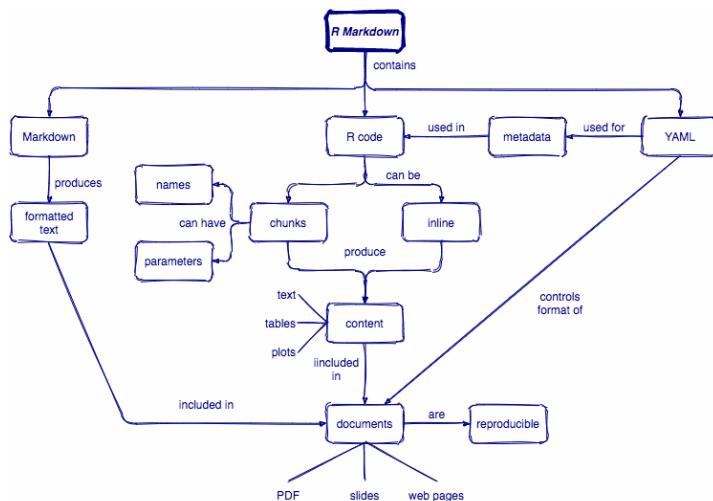


图 8.4: R Markdown 概念图

```

library(DBI)
conn <- DBI::dbConnect(RSQLite::SQLite(),
  dbname = system.file("db", "datasets.sqlite", package = "RSQLite")
)
  
```

Base R 内置的数据集都整合进 RSQLite 的样例数据库里了，

```
dbListTables(conn)
```

```

## [1] "BOD"          "CO2"          "ChickWeight"   "DNase"
## [5] "Formaldehyde" "Indometh"     "InsectSprays"  "LifeCycleSavings"
## [9] "Loblolly"      "Orange"        "OrchardSprays" "PlantGrowth"
## [13] "Puromycin"    "Theoph"        "ToothGrowth"   "USArrests"
## [17] "USJudgeRatings" "airquality"   "anscombe"     "attenu"
## [21] "attitude"      "cars"         "chickwts"     "esoph"
## [25] "faithful"       "freeny"        "infert"       "iris"
## [29] "longley"        "morley"        "mtcars"       "npk"
## [33] "pressure"       "quakes"        "randu"        "rock"
## [37] "sleep"          "stackloss"     "swiss"        "trees"
## [41] "warpbreaks"    "women"
  
```

随意选择 5 行数据记录，将结果保存到变量 iris_preview



```
SELECT * FROM iris LIMIT 5;
```

查看变量 `iris_preview` 的内容

```
iris_preview
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1         5.1       3.5        1.4       0.2   setosa
## 2         4.9       3.0        1.4       0.2   setosa
## 3         4.7       3.2        1.3       0.2   setosa
## 4         4.6       3.1        1.5       0.2   setosa
## 5         5.0       3.6        1.4       0.2   setosa
```

结束后关闭连接

```
dbDisconnect(conn = conn)
```

8.1.2 表格

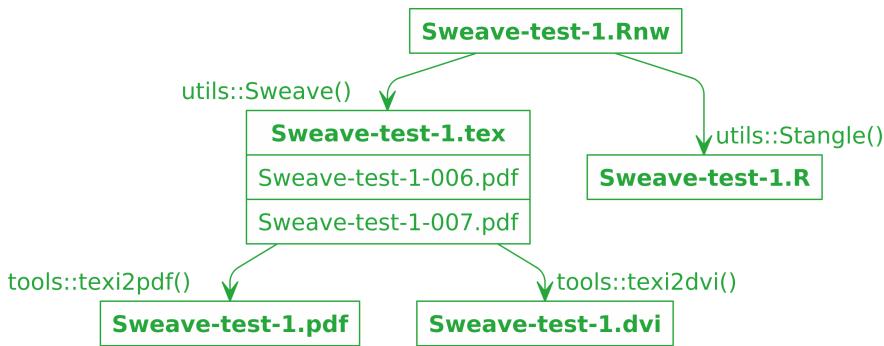
knitr 的 `kable()` 函数提供了制作表格的基本功能 <https://bookdown.org/yihui/rmarkdown-cookbook/tables.html>, **flextable** 支持更加细粒度的表格定制功能。**beautifyR** 整理 Markdown 表格非常方便, **datapasta** 快速复制粘贴 `data.frame` 和 `tibble` 类型的数据表格。**rpivotTable** 不更新了, **pivottabler** 在更新, 内容似乎更好。**remedy** 提供了更加通用的 Markdown 写作功能, 简化创作的技术难度。

8.1.3 流程图

nomnoml 流程图、思维导图

```
nomnoml::nomnoml("
#stroke: #26A63A
#.box: fill=#8f8 dashed visual=note
#direction: down

[Sweave-test-1.Rnw] -> utils:::Sweave() [Sweave-test-1.tex|Sweave-test-1-006.pdf|Sweave-
[Sweave-test-1.Rnw] -> utils:::Stangle() [Sweave-test-1.R]
[Sweave-test-1.tex] -> tools:::texi2pdf() [Sweave-test-1.pdf]
[Sweave-test-1.tex] -> tools:::texi2dvi() [Sweave-test-1.dvi]
")
```



8.2 便携式文档

8.2.1 文档汉化

从 R Markdown 到 beamer 幻灯片，如何迁移 LaTeX 模版

默认的 PDF 文档 [PDF 文档案例](#)

详见[PDF 文档案例](#)

8.2.2 添加水印

[draftwatermark](#)

8.2.3 双栏排版

普通单栏排版改为双栏排版，只需添加文档类选项 "twocolumn"，将 YAML 元数据中的

```
classoption: "UTF8,a4paper,fontset=adobe,zihao=false"
```

变为

```
classoption: "UTF8,a4paper,fontset=adobe,zihao=false,twocolumn"
```

其中，参数 `UTF8` 设定文档编码类型，`a4paper` 设置版面为 A4 纸大小，`fontset=adobe` 指定中文字体为 Adobe 字体，`zihao=false` 不指定字体大



小，使用文档类 `ctexart` 默认的字号，

8.2.4 参数化报告

④ 参数化文档案例

进一步将文档类型做成参数化，实现在运行时自由选择，只需将如下两行替换掉上述一行

```
params:  
  classoption: twocolumn  
  classoption: `r params$classoption`"
```

如果想要双栏的排版风格，编译时传递 `documentclass` 参数值，覆盖掉默认的参数值即可

```
rmarkdown::render(  
  input = "examples/pdf-document.Rmd",  
  params = list(classoption = c("twocolumn"))  
)
```

8.2.5 学术幻灯片

`beamer` 幻灯片也是一种 PDF 文档 [PDF 文档案例](#)

Dirk Eddelbuettel 将几个大学的 `beamer` 幻灯片转化成 R Markdown 模板，收录在 `binb` 包里，方便调用。伊利诺伊大学的 James J Balamuta 在 R Markdown 基础上专门为自己的学校开发了一套的幻灯片模版，全部打包在 `uiucthemes` 包里。

[komaletter](#) 用 Markdown 写信件

```
memor memor::pdf_memo()
```

`hrbrthemes` 提供两个文档模版 `hrbrthemes::ipsum_pdf()` 和 `hrbrthemes::ipsum()`

此汉风主题由 [林莲枝](#) 开发，LaTeX 宏包已发布在 [CTAN](#) 上，使用此幻灯片主题需要将相关的 LaTeX 宏包一块安装。

```
tlmgr install pgfornament pgfornament-han needspace xpatch
```



8.2.6 文档模版

字体设置

```
--  
output:  
  pdf_document:  
    extra_dependencies:  
      DejaVuSansMono:  
        - scaled=0.9  
      DejaVuSerif:  
        - scaled=0.9  
      DejaVuSans:  
        - scaled=0.9  
--
```

```
--  
output:  
  pdf_document:  
    extra_dependencies:  
      sourcecodepro:  
        - scale=0.85  
      sourceserifpro:  
        - rmdefault  
      sourcesanspro:  
        - sfdefault  
--
```

8.2.7 引用文献

Getting started with Zotero, Better BibTeX, and RMarkdown

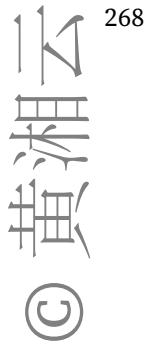
`knitcitations` 包可以根据文献数字对象标识符（英文 Digital Object Identifier，简称 DOI）生成引用，以文章《A Probabilistic Grammar of Graphics》[Pu and Kay, 2020] 为例，其 DOI 为 `10.1145/3313831.3376466`，总之，DOI 就像是文章的身份证，是一一对应的关系¹。

```
library(knitcitations)  
citet(x = '10.1145/3313831.3376466')  
  
[1] "(Pu and Kay, 2020)"
```

在表格的格子中引用参考文献

```
data.frame(  
  author = c("Yihui Xie", "Yihui Xie", "Yihui Xie"),  
  citation = c("[@xie2019]", "[@xie2015]", "[@xie2016]")  
) |>  
knitr::kable(format = "pandoc")
```

¹<https://zh.wikipedia.org/wiki/DOI>



author	citation
Yihui Xie	[Xie, 2019]
Yihui Xie	[Xie, 2015]
Yihui Xie	[Xie, 2016]

citr 包提供了快速查找参考文献的 RStudio 插件，不用去原始文献库 *.bib 搜索查找，也会自动生成引用，非常方便，极大地提高了工作效率。**citr** 还支持集成 **Zotero** 文献管理软件，可以直接从 **Zotero** 中导入参考文献数据库。**rbbt** 包也提供了类似的功能，只要系统安装 **Zotero** 软件及其插件 **Better Bibtex for Zotero connector**。

8.2.8 自定义块

```
tinytex::tlmgr_install(c('awesomibox', 'fontawesome5'))
```

安装 **awesomibox** 包，开发仓库在 <https://github.com/milouse/latex-awesomibox>，这个 LaTeX 宏包的作用是提供几类常用的块，比如提示、注意、警告等



注意这是注意



提示这是提示信息



警告这是警告信息



重要这是重要信息



8.3 网页文档

丘怡轩开发的 [prettydoc](#) 包提供了一系列模版，方便快速提高网页逼格。另有 Atsushi Yasumoto 开发的 [minidown](#) 包非常轻量，但是常用功能都覆盖了。

谢益辉开发的 [xaringan](#) 用于制作网页幻灯片，[xaringanthemer](#) 为 xaringan 提供主题定制，[xaringanExtra](#) 在 xaringan 之上提供各种功能扩展，[xaringanBuilder](#) 为 xaringan 提供多种输出格式。

8.4 编写书籍

此外，[ElegantTufteBookdown](#) 项目提供了 tufte 风格的书籍模板，本书配套的仓库目录 examples/ 下准备了一系列常用模板。

8.5 个人网站

8.6 微软文档

[docxtools](#)、[officer](#) 和 [officedown](#) 大大扩展了 rmarkdown 在制作 Word/PPT 方面的功能。

本节探索 Markdown + Pandoc 以 Word 格式作为最终交付的可能性。R Markdown 借助 Pandoc 将 Markdown 转化为 Word 文档，继承自 Pandoc 的扩展性，R Markdown 也支持自定义 Word 模版，那如何自定义呢？首先，我们需要知道 Pandoc 内建的 Word 模版长什么样子，然后我们依样画葫芦，制作适合实际需要的模版。获取 Pandoc 2.10.1 自带的 Word 和 PPT 模版，只需在命令行中执行

```
# DOCX 模版
pandoc -o custom-reference.docx --print-default-data-file reference.docx
# PPTX 模版
pandoc -o custom-reference.pptx --print-default-data-file reference.pptx
```

这里其实是将 Pandoc 自带的 docx 文档 reference.docx 拷贝一份到 custom-reference.docx，而后将 custom-reference.docx 文档自定义一番，但仅限于借助 MS Word 去自定义样式。Word 文档的 YAML 元数据定义详情见 <https://pandoc.org/MANUAL.html#option--reference-doc>，如何深度自定义



文档模版见 <https://bookdown.org/yihui/rmarkdown/word-document.html>，其它模版见 GitHub 仓库 [pandoc-templates](#)。这里提供一个Word 文档案例供读者参考。**bookdown** 提供的函数 `word_document2()` 相比于 **rmarkdown** 提供的 `word_document()` 支持图表的交叉引用，更多细节详见帮助 `?bookdown::word_document2`。

注意

R Markdown 文档支持带编号的 Word 文档格式输出要求 Pandoc 版本 2.10.1 及以上，**rmarkdown** 版本 2.4 及以上。

8.7 发送邮件

emayili 是非常轻量的实现邮件发送的 R 包，其它功能类似的 R 包有 **blastula**、**mailR**。Rahul Premraj 基于 rJava 开发的 **mailR** 虽然还未在 CRAN 上正式发布，但是已得到很多人的关注，也被广泛的使用，目前作者已经不维护了，继续使用有一定风险。RStudio 公司 Richard Iannone 新开发的 **blastula** 扔掉了 Java 的重依赖，更加轻量化、现代化，支持发送群组邮件²。**curl** 包提供的函数 `send_mail()` 本质上是在利用 **curl** 软件发送邮件，举个例子，邮件内容如下：

```
From: "黄湘云" <邮箱地址>
To: "黄湘云" <邮箱地址>
Subject: 测试邮件
```

你好：

这是一封测试邮件！

将邮件内容保存为 `mail.txt` 文件，然后使用 `curl` 命令行工具将邮件内容发出去。

```
curl --url 'smtp://公司邮件服务器地址:开放的端口号' \
--ssl-reqd --mail-from '发件人邮箱地址' \
--mail-rcpt '收件人邮箱地址' \
--upload-file data/mail.txt \
--user '发件人邮箱地址:邮箱登陆密码'
```

²<https://thecoatlessprofessor.com/programming/r/sending-an-email-from-r-with-blastula-to-groups-of-students/>



注意

Gmail 出于安全性考虑，不支持这种发送邮件的方式，会将邮件内容阻挡，进而接收不到邮件。

下面以 `blastula` 包为例怎么支持 Gmail/Outlook/QQ 等邮件发送，先安装系统软件依赖，CentOS 8 上安装依赖

```
sudo dnf install -y libsecret-devel libsodium-devel
```

然后安装 `keyring` 和 `blastula`

```
install.packages(c("keyring", "blastula"))
```

接着配置邮件帐户，这一步需要邮件账户名和登陆密码，配置一次就够了，不需要每次发送邮件的时候都配置一次

```
library(blastula)
create_smtp_creds_key(
  id = "outlook",
  user = "xiangyunfaith@outlook.com",
  provider = "outlook"
)
```

第二步，准备邮件内容，包括邮件主题、发件人、收件人、抄送人、密送人、邮件主体和附件等。

```
library(blastula)

attachment <- "data/mail.txt" # 如果没有附件，引号内留空即可。
# 这个Rmd文件渲染后就是邮件的正文，交互图形和交互表格不适用
body <- "examples/html-document.Rmd"
# 渲染邮件内容，生成预览
email <- render_email(body) |>
  add_attachment(file = attachment)
email
```

最后，发送邮件

```
smtp_send(
  from = c("张三" = "xxx@outlook.com"), # 发件人
  to = c("李四" = "xxx@foxmail.com",
```



```

    "王五" = "xxx@gmail.com"), # 收件人
cc = c("赵六" = "xxx@outlook.com"), # 抄送人
subject = "这是一封测试邮件",
email = email,
credentials = creds_key(id = "outlook")
)

```

密送人实现群发单显，即一封邮件同时发送给多人，每个收件人只能看到发件人地址而看不到其它收件人地址。

8.8 工作流

`drake` 一站式可重复性研究工作空间打造者，用户手册 <https://books.ropensci.org/drake/> 和学习材料 <https://github.com/wlandau/learndrake>

8.9 运行环境

```

sessionInfo()

## R version 4.1.0 (2021-05-18)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 20.04.2 LTS
##
## Matrix products: default
## BLAS:    /usr/lib/x86_64-linux-gnublas/libblas.so.3.9.0
## LAPACK:  /usr/lib/x86_64-linux-gnulapack/liblapack.so.3.9.0
##
## locale:
## [1] LC_CTYPE=en_US.UTF-8          LC_NUMERIC=C
## [3] LC_TIME=en_US.UTF-8          LC_COLLATE=en_US.UTF-8
## [5] LC_MONETARY=en_US.UTF-8       LC_MESSAGES=en_US.UTF-8
## [7] LC_PAPER=en_US.UTF-8          LC_NAME=C
## [9] LC_ADDRESS=C                  LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8   LC_IDENTIFICATION=C

```



```
##  
## attached base packages:  
## [1] stats      graphics   grDevices utils     datasets  methods   base  
##  
## other attached packages:  
## [1] DBI_1.1.1  
##  
## loaded via a namespace (and not attached):  
## [1] Rcpp_1.0.7       knitr_1.33       magrittr_2.0.1    bit_4.0.4  
## [5] R6_2.5.0        rlang_0.4.11     fastmap_1.1.0    highr_0.9  
## [9] blob_1.2.2      stringr_1.4.0    tools_4.1.0      webshot_0.5.2  
## [13] nomnoml_0.2.3   xfun_0.24       htmltools_0.5.1.1 yaml_2.2.1  
## [17] bit64_4.0.5    digest_0.6.27    lifecycle_1.0.0   bookdown_0.22  
## [21] processx_3.5.2 callr_3.7.0     vctrs_0.3.8     htmlwidgets_1.5.3  
## [25] ps_1.6.0       cachem_1.0.5    memoise_2.0.0    evaluate_0.14  
## [29] RSQLite_2.2.7   rmarkdown_2.9    stringi_1.7.3    compiler_4.1.0  
## [33] jsonlite_1.7.2  pkgconfig_2.0.3
```

第九章 交互图形

提示

plotly 包的函数使用起来还是比较复杂的，特别是需要打磨细节以打造数据产品时，此外，其依赖相当重，仅数据处理就包含两套方法 – `dplyr` 和 `data.table`，引起很多函数冲突，可谓「苦其久矣」！因此，准备另起炉灶，开发一个新的 R 包 `qplotly`，取意 quick plotly，以 `qplot_ly()` 替代 `plot_ly()`。类似简化 API 的工作有 `simplevis`、`autoplotly`、`ggfortify` 和 `plotme`。
plotly 团队开发了 `plotly.js` 库，且维护了 R 接口文档 (<https://plotly.com/r/>)，Carson Sievert 开发了 `plotly` 包，配套书 *Interactive web-based data visualization with R, plotly, and shiny*。Paul C. Bauer 的书 *Applied Data Visualization* 介绍 plotly <https://bookdown.org/paul/applied-data-visualization/what-is-plotly.html>

`echarts4r` 包基于 [Apache ECharts \(incubating\)](#)，ECharts 的 Python 接口 `pyecharts` 也非常受欢迎，基于 `apexcharts.js` 的 `apexcharter`。`ECharts2Shiny` 包将 ECharts 嵌入 `shiny` 框架中。

`timevis` 创建交互式的时间线的时序可视化，它基于 `Vis` 的 `vis-timeline` 模块，支持 `shiny` 集成。`dygraphs` 包基于 `dygraphs` 可视化库，将时序数据可视化，更多情况见 <https://dygraphs.com/>。`leaflet` 提供 `leaflet` 的 R 接口。`rAmCharts4` 基于 `amCharts 4` 库，`apexcharter` 提供 `apexcharts.js` 的 R 接口。还有 `billboarder` 等。更完整地，请看 Etienne Bacher 维护的 R 包列表 [r.js-adaptation](#)。

对于想了解 `htmlwidgets` 框架，JavaScript 响应式编程的读者，推荐 John Coene 新书 [JavaScript for R](#)



提示

学习 `plotly` 和 `highcharter` 为代表的基于 JavaScript 的 R 包，共有四重境界：第一重是照着帮助文档的示例，示例有啥我们做啥；第二重是明白帮助文档中 R 函数和 JavaScript 函数的对应关系，能力达到 JS 库的功能边界；第三重是深度自定义一些扩展性的 JS 功能，放飞自我；第四重是重新造轮子，为所欲为。下面的介绍希望能帮助读者到达第二重境界。

`plotly` 是一个功能非常强大的绘制交互式图形的 R 包，支持图片下载、背景图片¹、工具栏²和注释³等一系列细节的自定义控制。下面结合 JavaScript 库 `plotly.js` 一起介绍，帮助文档 `?config` 没有太详细地介绍，所以我们看看 `config()` 函数中参数 ... 和 JS 库 `plot_config.js` 中的功能函数是怎么对应的。图中图片下载按钮对应 `toImageButtonOptions` 参数，看 `toImageButtonOptions` 源代码，可知它接受任意数据类型，对应到 R 里面就是列表。`watermark` 和 `displaylogo` 都是传递布尔值 (TRUE/FALSE)，具体根据 JS 代码中的 `valType` (参数值类型) 决定，其它参数类似。另一个函数 `layout` 和函数 `config()` 是类似的，怎么传递参数值是根据 JS 代码来的。

```
toImageButtonOptions: {
  valType: 'any',
  dflt: {},
  description: [
    'Statically override options for toImage modebar button',
    'allowed keys are format, filename, width, height, scale',
    'see ../components/modebar/buttons.js'
  ].join(' ')
},
displaylogo: {
  valType: 'boolean',
  dflt: true,
  description: [
    'Determines whether or not the plotly logo is displayed',
    'on the end of the mode bar.'
  ].join(' ')
},
```

¹<https://plotly.com/r/logos/>

²<https://plotly-r.com/control-modebar.html>

³<https://plotly.com/r/reference/#layout-scene-annotations-items-annotation-font>

```
watermark: {  
    valType: 'boolean',  
    dflt: false,  
    description: 'watermark the images with the company\'s logo'  
},  
  
library(plotly, warn.conflicts = FALSE)  
plot_ly(diamonds,  
        x = ~clarity, y = ~price,  
        color = ~clarity, colors = "Set1", type = "box"  
) %>%  
config(  
    toImageButtonOptions = list(  
        format = "svg", filename = paste("plot", Sys.Date(), sep = "_"),  
        width = 450, height = 300  
        # 设置下载图片的尺寸 https://github.com/ropensci/plotly/issues/1556#issuecomment-  
    ), # 还可设置为 PNG 格式, 可用 rsvg 的 rsvg_pdf 函数转化为 PDF  
    modeBarButtons = list(list("toImage")), # 保留下载按钮  
    # 完整的列表见 https://github.com/plotly/plotly.js/blob/master/src/components/modebar.js  
    watermark = F,  
    displaylogo = FALSE, # 移除 Plotly 的 logo  
    locale = "zh-CN", # 汉化  
    # staticPlot = TRUE, # 静态图形而不是交互图形  
    # modeBarButtonsToRemove = c(  
    #     "zoom2d", "zoomIn2d", "zoomOut2d", "autoScale2d", "resetScale2d", "pan2d",  
    #     "hoverClosestCartesian", "hoverCompareCartesian", "toggleSpikelines"  
    # ), # 去掉任意一个按钮  
    # displayModeBar = FALSE, # 去掉整个顶部工具栏  
    showLink = FALSE  
) %>%  
layout(  
    images = list(  
        source = "https://images.plot.ly/language-icons/api-home/r-logo.png",  
        xref = "paper",  
        yref = "paper",  
        x = 1.0,
```

```
y = 0.25,  
sizex = 0.2,  
sizey = 0.2,  
opacity = 0.5  
),  
annotations = list(  
    text = "watermark", # 文本注释  
    font = list(  
        size = 40, # 字号  

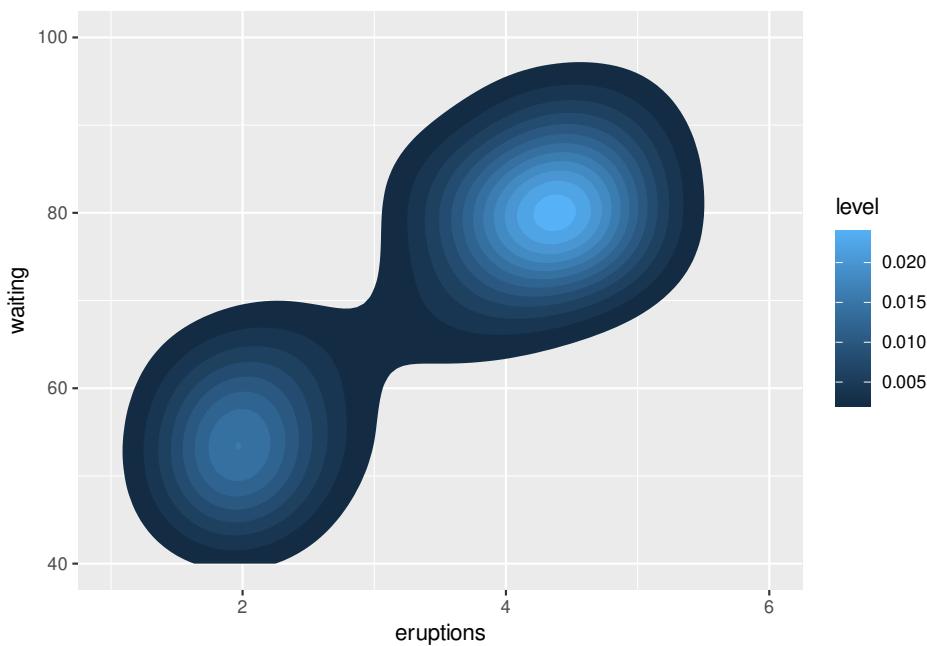
```

函数 `ggplotly()` 将 `ggplot` 对象转化为交互式 `plotly` 对象

```
gg <- ggplot(faithful, aes(x = eruptions, y = waiting)) +  
    stat_density_2d(aes(fill = ..level..), geom = "polygon") +  
    xlim(1, 6) +  
    ylim(40, 100)
```

静态图形

```
gg
```



转化为 `plotly` 对象

```
ggplotly(gg)
```

添加动态点的注释，比如点横纵坐标、坐标文本，整个注释标签的样式（如背景色）

```
ggplotly(gg, dynamicTicks = "y") %>%
  style(., hoveron = "points", hoverinfo = "x+y+text",
        hoverlabel = list(bgcolor = "white"))
```

9.1 散点图

表 9.1: 散点图类型

类型	名称
<code>scattercarpet</code>	地毯图
<code>scatterternary</code>	三元图
<code>scatter3d</code>	三维散点图
<code>scattergeo</code>	地图散点图



类型	名称
scattermapbox	地图散点图 Mapbox
scatter	散点图
scattergl	散点图 GL
scatterpolar	极坐标散点图
scatterpolargl	极坐标散点图 GL

plotly.js 提供很多图层用于绘制各类图形 <https://github.com/plotly/plotly.js/tree/master/src/traces>

9.2 条形图

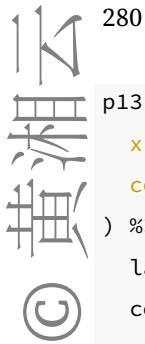
日常使用最多的图形无外乎散点图、柱形图（分组、堆积、百分比堆积等）

```
# 简单条形图
library(data.table)
diamonds <- as.data.table(diamonds)

p11 <- diamonds[, .(cnt = .N), by = .(cut)] %>%
  plot_ly(x = ~cut, y = ~cnt, type = "bar") %>%
  add_text(
    text = ~ scales::comma(cnt), y = ~cnt,
    textposition = "top middle",
    cliponaxis = FALSE, showlegend = FALSE
  ) %>%
  config(displayModeBar = F)

# 分组条形图
p12 <- plot_ly(diamonds,
  x = ~cut, color = ~clarity,
  colors = "Accent", type = "histogram"
) %>%
  config(displayModeBar = F)

# 堆积条形图
```



```
p13 <- plot_ly(diamonds,
  x = ~cut, color = ~clarity,
  colors = "Accent", type = "histogram"
) %>%
  layout(barmode = "stack") %>%
  config(displayModeBar = F)

# 百分比堆积条形图

p14 <- plot_ly(diamonds,
  x = ~cut, color = ~clarity,
  colors = "Accent", type = "histogram"
) %>%
  layout(barmode = "stack", barnorm = "percent") %>%
  config(displayModeBar = F)

htmltools::tagList(p11, p12, p13, p14)
```

9.3 折线图

其它常见的图形还要折线图、直方图、箱线图和提琴图

```
# 折线图

p21 <- plot_ly(Orange,
  x = ~age, y = ~circumference, color = ~Tree,
  type = "scatter", mode = "markers+lines"
)
```

9.4 双轴图

双轴图

模拟一组数据

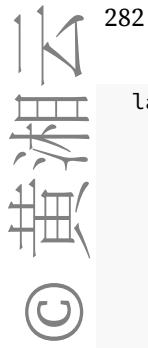
```
set.seed(2020)
dat <- data.frame(
```



```
dt = seq(from = as.Date("2020-01-01"), to = as.Date("2020-01-31"), by = "day"),
search_qv = sample(100000:1000000, size = 31, replace = T)
) %>%
  transform(valid_click_qv = sapply(search_qv, rbinom, n = 1, prob = 0.5)) %>%
  transform(qv_ctr = valid_click_qv / search_qv)
```

hoverinfo = "text" 表示 tooltips 使用指定的 text 映射，而 visible = "legendonly" 表示图层默认隐藏不展示，只在图例里显示，有时候很多条线，默认只是展示几条而已。举例如下

```
plot_ly(data = dat) %>%
  add_bars(
    x = ~dt, y = ~search_qv, color = I("#4285f4"), name = "搜索 QV",
    text = ~ paste0(
      "日期: ", dt, "<br>",
      "点击 QV: ", format(valid_click_qv, big.mark = ","), "<br>",
      "搜索 QV: ", format(search_qv, big.mark = ","), "<br>",
      "QV_CTR: ", scales::percent(qv_ctr, accuracy = 0.01), "<br>"
    ),
    hoverinfo = "text"
  ) %>
  add_bars(
    x = ~dt, y = ~valid_click_qv, color = I("#FBBC05"), name = "点击 QV",
    text = ~ paste0(
      "日期: ", dt, "<br>",
      "点击 QV: ", format(valid_click_qv, big.mark = ","), "<br>",
      "搜索 QV: ", format(search_qv, big.mark = ","), "<br>",
      "QV_CTR: ", scales::percent(qv_ctr, accuracy = 0.01), "<br>"
    ), visible = "legendonly",
    hoverinfo = "text"
  ) %>
  add_lines(
    x = ~dt, y = ~qv_ctr, name = "QV_CTR", yaxis = "y2", color = I("#34A853"),
    text = ~ paste("QV_CTR: ", scales::percent(qv_ctr, accuracy = 0.01), "<br>"),
    hoverinfo = "text",
    line = list(shape = "spline", color = "Set1", width = 3, dash = "line")
  ) %>%
```



```
layout(
  title = "",
  yaxis2 = list(
    tickfont = list(color = "black"),
    overlaying = "y",
    side = "right",
    title = "QV_CTR (%)",
    # ticksuffix = "%", # 设置坐标轴单位
    tickformat = '.1%', # 设置坐标轴刻度
    showgrid = F, automargin = TRUE
  ),
  xaxis = list(title = "日期", showgrid = F, showline = F),
  yaxis = list(title = " ", showgrid = F, showline = F),
  margin = list(r = 20, autoexpand = T),
  legend = list(
    x = 0, y = 1, orientation = "h",
    title = list(text = " ")
  )
) %>%
config(displayModeBar = F)
```

9.5 直方图

```
# 分组直方图
p22 <- plot_ly(iris,
  x = ~Sepal.Length,
  color = ~Species, type = "histogram"
)
```

9.6 箱线图

```
# 箱线图
p23 <- plot_ly(diamonds,
```



```
x = ~clarity, y = ~price,
color = ~clarity, type = "box"
)

# 箱线图
plot_ly(diamonds, x = ~cut, y = ~price) %>%
add_boxplot()

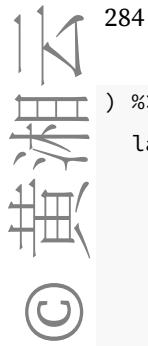
# 不同的类别使用不同的颜色上色
plot_ly(diamonds, x = ~clarity, y = ~price, color = ~clarity) %>%
add_boxplot()

# 使用 colors 参数设置调色板
plot_ly(diamonds,
x = ~clarity, y = ~price,
color = ~clarity, colors = "Set1"
) %>%
add_boxplot()

# 或者使用 qplot 式绘图风格
plot_ly(diamonds,
x = ~clarity, y = ~price,
color = ~clarity, colors = "Set1", type = "box"
)

# 分组箱线图 https://github.com/ropensci/plotly/issues/994
plot_ly(diamonds,
x = ~cut, y = ~price,
color = ~clarity, type = "box"
) %>%
layout(boxmode = "group")

# 修改图例的标题, R 的嵌套 list 对象对应于 JS 的 JSON 数据对象
plot_ly(diamonds,
x = ~cut, y = ~price,
color = ~clarity, colors = "Set1", type = "box"
```



```
) %>%
  layout(
    boxmode = "group",
    legend = list(
      bgcolor = "white",
      title = list(text = "clarity")
    )
  )

# 提琴图
plot_ly(diamonds, x = ~cut, y = ~price) %>%
  add_trace(type = "violin")

plot_ly(diamonds,
        x = ~cut, y = ~price, split = ~cut, type = "violin",
        box = list(visible = T),
        meanline = list(visible = T)
) %>%
  layout(
    xaxis = list(title = "Cut"),
    yaxis = list(title = "Price", zeroline = F)
  )
```

9.7 提琴图

```
# 提琴图
p24 <- plot_ly(sleep,
  x = ~group, y = ~extra, split = ~group, type = "violin",
  box = list(visible = T),
  meanline = list(visible = T)
)

htmltools::tagList(p21, p22, p23, p24)
```

plotly 包含图层 27 种，见表 9.2



表 9.2: 图层

A	B	C
add_annotations	add_histogram	add_polygons
add_area	add_histogram2d	add_ribbons
add_bars	add_histogram2dcontour	add_scattergeo
add_boxplot	add_image	add_segments
add_choropleth	add_lines	add_sf
add_contour	add_markers	add_surface
add_data	add_mesh	add_table
add_fun	add_paths	add_text
add_heatmap	add_pie	add_trace

9.8 气泡图

简单图形 scatter, 分布图几类, 其中 scatter、heatmap、scatterpolar 支持 WebGL 绘图引擎

```
# https://plotly.com/r/bubble-charts/
dat <- diamonds[, .(
  carat = mean(carat),
  price = sum(price),
  cnt = .N
), by = .(cut)]  
  
plot_ly(
  data = dat,
  x = ~carat, y = ~price, color = ~cut, size = ~cnt,
  type = "scatter", mode = "markers",
  marker = list(
    symbol = "circle", sizemode = "diameter",
    line = list(width = 2, color = "#FFFFFF"), opacity = 0.4
  ),
  text = ~ paste(
    sep = " ", "重量: ", round(carat, 2), "克拉",
    "<br>价格:", round(price / 10^6, 2), "百万"
  ),
)
```

```

  hoverinfo = 'text'
) %>%
add_annotations(
  x = ~carat, y = ~price, text = ~cnt,
  showarrow = F, font = list(family = "sans")
) %>%
layout(
  xaxis = list(hoverformat = ".2f"),
  yaxis = list(hoverformat = ".0f")
) %>%
config(displayModeBar = F)

```

9.9 曲线图

```

plot_ly(
  x = c(1, 2.2, 3), y = c(5.3, 6, 7), type = "scatter",
  mode = "markers+lines", line = list(shape = "spline"), color = I("#EA4335")
) %>%
add_annotations(
  x = 2, y = 6, size = I(100),
  text = TeX("x_i \sim N(\mu, \sigma)")
) %>%
layout(
  xaxis = list(showgrid = F, title = TeX("\mu")),
  yaxis = list(showgrid = F, title = TeX("\alpha"))
) %>%
config(displayModeBar = FALSE, mathjax = 'cdn')

```

9.10 堆积图

```

plot_ly(
  data = PlantGrowth, y = ~weight,
  color = ~group,

```



```
    type = "scatter", line = list(shape = "spline"),
    mode = "lines", fill = "tozeroY"
)
```

9.11 热力图

其他基础图形

```
# Heatmaps
plot_ly(z = volcano, type = 'heatmap')
```

9.12 地图 I

`plot_mapbox()` 使用 Mapbox 提供的地图服务，因此，需要注册一个账户，获取 MAPBOX_TOKEN

```
data("quakes")
plot_mapbox(
  data = quakes,
  lon = ~long, lat = ~lat,
  color = ~mag, size = 2,
  type = "scattermapbox",
  mode = "markers",
  marker = list(opacity = 0.5)
) %>%
  layout(
    title = "Fiji Earthquake",
    mapbox = list(
      zoom = 3,
      center = list(
        lat = ~ median(lat - 5),
        lon = ~ median(long)
      )
    )
) %>%
```



```
config(
  mapboxAccessToken = Sys.getenv("MAPBOX_TOKEN"),
  displayModeBar = FALSE
)

plot_ly(
  data = quakes,
  lon = ~long, lat = ~lat,
  type = "scattergeo", mode = "markers",
  text = ~ paste0(
    "站点: ", stations, "<br>",
    "震级: ", mag
  ),
  marker = list(
    color = ~mag,
    size = 10, opacity = 0.8,
    line = list(color = "white", width = 1)
  )
) %>%
  layout(geo = list(
    showland = TRUE,
    landcolor = toRGB("gray95"),
    subunitcolor = toRGB("gray85"),
    countrycolor = toRGB("gray85"),
    countrywidth = 0.5,
    subunitwidth = 0.5,
    lonaxis = list(
      showgrid = TRUE,
      gridwidth = 0.5,
      range = c(160, 190),
      dtick = 5
    ),
    lataxis = list(
      showgrid = TRUE,
      gridwidth = 0.5,
      range = c(-40, -10),
      dtick = 5
    )
  ))
```



```
    dtick = 5
)
)) %>%
config(
  displayModeBar = FALSE
)

dat = data.frame(state.x77, stats = rownames(state.x77), stats_abbr = state.abb)
plot_ly(data = dat,
  type = "choropleth",
  locations = ~stats_abbr,
  locationmode = "USA-states",
  colorscale = "Viridis",
  z = ~Income
) %>%
  layout(geo = list(scope = "usa"))
```

9.13 拟合图

```
plot_ly(economics,
  type = "scatter",
  x = ~date,
  y = ~uempmed,
  name = "observed unemployment",
  mode = "markers+lines",
  marker = list(
    color = "red"
  ),
  line = list(
    color = "red",
    dash = "dashed"
  )
) %>%
  add_trace(
    x = ~date,
```



```
y = ~fitted(loess(uempmed ~ as.numeric(date))),  
name = "fitted unemployment",  
mode = "markers+lines",  
marker = list(  
  color = "orange"  
,  
line = list(  
  color = "orange"  
)  
) %>%  
layout(  
  title = "失业时间",  
  xaxis = list(  
    title = "日期",  
    showgrid = F  
,  
  yaxis = list(  
    title = "失业时间 (周) "  
,  
  legend = list(  
    x = 0, y = 1, orientation = "v",  
    title = list(text = "")  
)  
)  
)
```

9.14 轨迹图

rasterly 百万量级的散点图

```
library(rasterly)  
plot_ly(quakes, x = ~long, y = ~lat) %>%  
  add_rasterly_heatmap()  
  
quakes %>%  
  rasterly(mapping = aes(x = long, y = lat)) %>%
```

```
rasterly_points()

library(plotly)
# 读取数据
# uber 轨迹数据来自 https://github.com/plotly/rasterly
ridesDf <- readRDS(file = 'data/uber.rds')

ridesDf %>%
  rasterly(mapping = aes(x = Lat, y = Lon)) %>%
  rasterly_points()
```

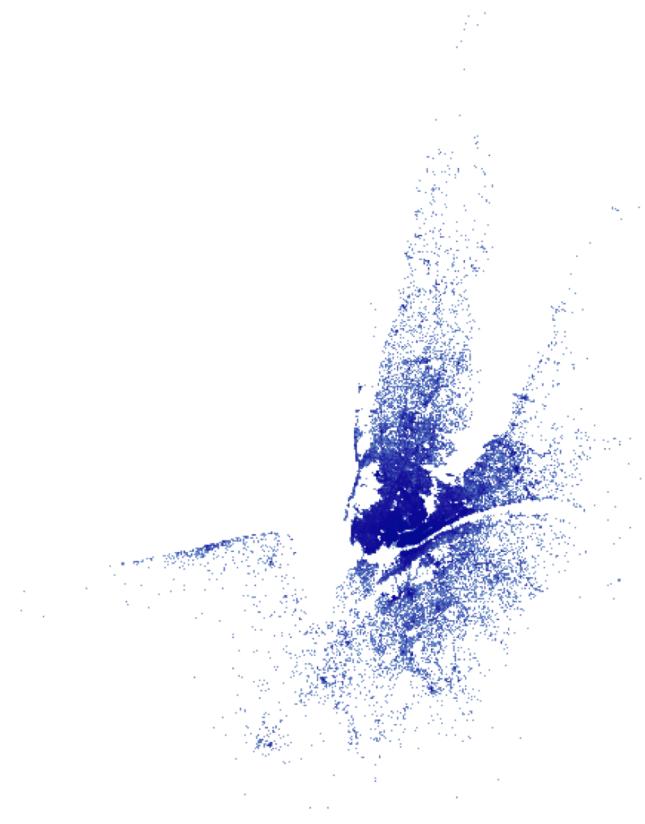


图 9.1: 轨迹数据



9.15 三维图

```
plot_ly(z = ~volcano) %>%  
  add_surface()  
  
⑤ plot_ly(x = c(0, 0, 1), y = c(0, 1, 0), z = c(0, 0, 0)) %>%  
  add_mesh()  
  
# https://plot.ly/r/reference/#scatter3d  
transform(mtcars, am = ifelse(am == 0, "Automatic", "Manual")) %>%  
  plot_ly(x = ~wt, y = ~hp, z = ~qsec,  
          color = ~am, colors = c("#BF382A", "#0C4B8E")) %>%  
  add_markers() %>%  
  layout(scene = list(  
    xaxis = list(title = "Weight"),  
    yaxis = list(title = "Gross horsepower"),  
    zaxis = list(title = "1/4 mile time"))  
)
```

9.16 甘特图

项目管理必备，如图所示，本项目拆分成 7 个任务，一共使用 3 种项目资源

```
# https://plotly.com/r/gantt/  
# 项目拆解为一系列任务，每个任务的开始时间，持续时间和资源类型  
df <- data.frame(  
  task = paste("Task", 1:8),  
  start = as.Date(c(  
    "2016-01-01", "2016-02-20", "2016-01-01",  
    "2016-04-10", "2016-06-09", "2016-04-10",  
    "2016-09-07", "2016-11-26"  
  )),  
  duration = c(50, 25, 100, 60, 30, 150, 80, 10),  
  resource = c("A", "B", "C", "C", "C", "A", "B", "B")  
) %>%
```



```
transform(end = start + duration) %>%
  transform(y = 1:nrow(.))

plot_ly(data = df) %>%
  add_segments(
    x = ~start, xend = ~end,
    y = ~y, yend = ~y,
    color = ~resource,
    mode = "lines",
    colors = "Set2",
    line = list(width = 20),
    showlegend = F,
    hoverinfo = "text",
    text = ~ paste(
      "任务: ", task, "<br>",
      "启动时间: ", start, "<br>",
      "周期: ", duration, "天<br>",
      "资源: ", resource
    )
  ) %>%
  layout(
    xaxis = list(
      showgrid = F,
      title = list(text = ""))
  ),
  yaxis = list(
    showgrid = F,
    title = list(text = ""),
    tickmode = "array",
    tickvals = 1:nrow(df),
    ticktext = unique(df$task),
    domain = c(0, 0.9)
  ),
  annotations = list(
    list(
      xref = "paper", yref = "paper",
```



```
x = 0.80, y = 0.1,
text = paste0(
  "项目周期: ", sum(df$duration), " 天<br>",
  "资源类型: ", length(unique(df$resource)), " 个<br>"
),
font = list(size = 12),
ax = 0, ay = 0,
align = "left"
),
list(
  xref = "paper", yref = "paper",
  x = 0.1, y = 1,
  xanchor = "left",
  text = "项目资源管理",
  font = list(size = 20),
  ax = 0, ay = 0,
  align = "left",
  showarrow = FALSE
)
)
)
) %>%
config(
  displayModeBar = FALSE
)
```

9.17 帕雷托图

帕雷托图 20/80 法则

```
# 数据来自 https://github.com/plotly/datasets
dat <- data.frame(
  complaint = c(
    "Small portions", "Overpriced",
    "Wait time", "Food is tasteless", "No atmosphere", "Not clean",
    "Too noisy", "Food is too salty", "Unfriendly staff", "Food not fresh"
```



```
),
count = c( 621L, 789L, 109L, 65L, 45L, 30L, 27L, 15L, 12L, 9L)
)

dat <- dat[order(-dat$count), ] %>%
  transform(cumulative = round(100 * cumsum(count) / sum(count), digits = 2))

# complaint 按 count 降序排列
dat$complaint <- reorder(x = dat$complaint, X = dat$count, FUN = function(x) 1/(1

plot_ly(data = dat) %>%
  add_bars(x = ~complaint, y = ~count, showLegend = F, color = I("#4285f4")) %>%
  add_lines(x = ~complaint, y = ~cumulative, yaxis = "y2", showlegend = F) %>%
  layout(
    yaxis2 = list(
      tickfont = list(color = "black"),
      overlaying = "y",
      side = "right",
      title = "累积百分比 (%) ",
      showgrid = F
    ),
    xaxis = list(title = "投诉类型", showgrid = F, showline = F),
    yaxis = list(title = "数量", showgrid = F, showline = F)
  ) %>%
  config(
    displayModeBar = FALSE
  )
)
```

提示

`reorder()` 对 `complaint` 按照降序还是升序由 `FUN` 函数的单调性决定，单调增对应升序，单调减对应降序



9.18 时间线

```
library(vistime)

pres <- data.frame(
  Position = rep(c("President", "Vice"), each = 3),
  Name = c("Washington", rep(c("Adams", "Jefferson"), 2), "Burr"),
  start = c("1789-03-29", "1797-02-03", "1801-02-03"),
  end = c("1797-02-03", "1801-02-03", "1809-02-03"),
  color = c("#cbb69d", "#603913", "#c69c6e"),
  fontcolor = c("black", "white", "black")
)

vistime(pres, col.event = "Position", col.group = "Name") %>%
  config(
    displayModeBar = FALSE
  )
```

9.19 漏斗图

```
dat <- data.frame(
  category = c("访问", "下载", "潜客", "报价", "下单"),
  value = c(39, 27.4, 20.6, 11, 2)
) %>%
  transform(percent = value / cumsum(value))
plot_ly(data = dat) %>%
  add_trace(
    type = "funnel",
    y = ~category,
    x = ~value,
    color = ~category,
    text = ~ paste0(value, "<br>", sprintf("%.2f%%", 100*percent)) ,
    hoverinfo = "text",
    showlegend = FALSE
```



```
) %>%  
  layout(yaxis = list(  
    categoryarray = ~category,  
    title = "")  
) %>%  
  config(  
    displayModeBar = FALSE  
)  
  
plotly::plot_ly(data = dat) %>%  
  plotly::add_trace(  
    type = "funnel",  
    y = ~category,  
    x = ~value,  
    marker = list(color = RColorBrewer::brewer.pal(n = 5, name = "Set2")),  
    textposition = "auto",  
    textinfo = "value+percent previous",  
    hoverinfo = "none"  
) %>%  
  plotly::layout(yaxis = list(categoryarray = ~category, title = "")) %>%  
  plotly::config(displayModeBar = FALSE)
```

9.20 雷达图

```
plot_ly(  
  type = "scatterpolar", mode = "markers", fill = "toself"  
) %>%  
  add_trace(  
    r = c(39, 28, 8, 7, 28, 39),  
    theta = c("数学", "物理", "化学", "英语", "生物", "数学"),  
    name = "学生 A"  
) %>%  
  add_trace(  
    r = c(1.5, 10, 39, 31, 15, 1.5),  
    theta = c("数学", "物理", "化学", "英语", "生物", "数学"),
```



```

name = "学生 B"
) %>%
layout(
  polar = list(
    radialaxis = list(
      visible = T,
      range = c(0, 50)
    )
  )
)

```

9.21 瀑布图

盈亏图

```

library(plotly)
library(dplyr)

dat <- data.frame(
  x = c(
    "销售", "咨询", "净收入",
    "购买", "其他费用", "税前利润"
  ),
  y = c(60, 80, 10, -40, -20, 0),
  measure = c(
    "relative", "relative", "relative",
    "relative", "relative", "total"
  )
) %>%
  mutate(text = case_when(
    y > 0 ~ paste0("+", y),
    y == 0 ~ "",
    y < 0 ~ as.character(y)
  )) %>%
  mutate(x = factor(x, levels = c(

```



```
"销售", "咨询", "净收入",
"购买", "其他费用", "税前利润"
)))

n_rows <- nrow(dat)
dat[nrow(dat), "text"] <- "累计"

# measure 取值为 'relative'/'total'/'absolute'
plotly::plot_ly(dat,
  x = ~x, y = ~y, measure = ~measure, type = "waterfall",
  text = ~text, textposition = "outside",
  name = "收支", hoverinfo = "final",
  connector = list(line = list(color = "gray")),
  increasing = list(marker = list(color = "#66C2A5")),
  decreasing = list(marker = list(color = "#FC8D62")),
  totals = list(marker = list(color = "#8DA0CB"))

) %>%
  plotly::layout(
    title = "2018 年收支状态",
    xaxis = list(title = "业务"),
    yaxis = list(title = "金额"),
    showlegend = FALSE
) %>%
  plotly::config(displayModeBar = FALSE)
```

9.22 树状图

plotly 绘制 treemap 和 sunburst 图比较复杂，接口不友好，[plotme](#) 正好弥补不足。

9.23 旭日图

[plotme](#)



9.24 调色板

```
plot_ly(iris,
        x = ~Petal.Length, y = ~Petal.Width,
        mode = "markers", type = "scatter",
        color = ~ Sepal.Length > 6, colors = c("#132B43", "#56B1F7")
)
plot_ly(iris, x = ~Petal.Length, y = ~Petal.Width, color = ~Sepal.Length>6,
        mode = "markers", type = "scatter")

plot_ly(iris, x = ~Petal.Length, y = ~Petal.Width, color = ~Sepal.Length>6,
        mode = "markers", type = "scatter", colors = "Set2")

plot_ly(iris, x = ~Petal.Length, y = ~Petal.Width, color = ~Sepal.Length>6,
        mode = "markers", type = "scatter", colors = "Set1")
```

构造 20 个类别超出 Set1 调色板的范围，会触发警告说 Set1 没有那么多色块，但还是返回足够多的色块，也可以使用 viridis、plasma、magma 或 inferno 调色板

```
dat <- data.frame(
  dt = rep(seq(
    from = as.Date("2021-01-01"),
    to = as.Date("2021-01-31"), by = "day"
  ), each = 20),
  bu = rep(LETTERS[1:20], 31),
  qv = rbinom(n = 20 * 31, size = 10000, prob = runif(20 * 31))
)
# viridis
plot_ly(dat,
        x = ~dt, y = ~qv, color = ~bu,
        mode = "markers", type = "scatter", colors = "viridis"
)
```



9.25 面积图

Joshua Kunst 在他的博客里 <https://jkunst.com/> 补充了很多数据可视化案例，另一个关键的参考资料是 [highcharts API 文档](#)，文档主要分两部分 全局选项 `Highcharts.setOptions` 和绘图函数 `Highcharts.chart`。下面以 `data_to_boxplot()` 为例解析 R 中的数据结构是如何和 highcharts 的 JSON 以及 绘图函数对应的。

```
library(highcharter)
highchart() %>%
  hc_xAxis(type = "category") %>%
  hc_add_series_list(x = data_to_boxplot(
    data = iris,
    variable = Sepal.Length,
    group_var = Species,
    add_outliers = TRUE,
    name = "iris"
))
```

除了箱线图 `boxplot` 还有折线图、条形图、密度图等一系列常用图形，共计 50 余种，详见表9.3，各类图形示例见 <https://www.highcharts.com/demo>。

表 9.3: 图形种类

A	B	C	D	E
area	columnrange	item	pyramid3d	treemap
arearange	cylinder	line	sankey	variablepie
areaspline	dependencywheel	lollipop	scatter	variwide
areasplinerange	dumbbell	networkgraph	scatter3d	vector
bar	errorbar	organization	solidgauge	venn
bellcurve	funnel	packedbubble	spline	waterfall
boxplot	funnel3d	pareto	streamgraph	windbarb
bubble	gauge	pie	sunburst	wordcound
column	heatmap	polygon	tilemap	xrange
columnpyramid	histogram	pyramid	timeline	NA

```
library(highcharter)
hchart(iris, "scatter",
       hcaes(x = Sepal.Length, y = Sepal.Width, group = Species))
```

有的图形种类包含多个变体，如 area 面积图，还有 arearange、areaspline 和 areasplinerange，而 area 图其实是折线图，只是线与坐标轴围成的区域用颜色填充了。一个基本示例见[基础面积图](#)，数据结构如下：

```
Highcharts.chart('container', {
    chart: {
        type: 'area'
    },
    accessibility: {
        description: 'Image description: An area chart compares the nuclear stockpiles
    },
    title: {
        text: 'US and USSR nuclear stockpiles'
    },
    subtitle: {
        text: 'Sources: <a href="https://thebulletin.org/2006/july/global-nuclear-stock
            'thebulletin.org</a> &amp; <a href="https://www.armscontrol.org/factsheets/
            'armscontrol.org</a>'
    },
    xAxis: {
        allowDecimals: false,
        labels: {
            formatter: function () {
                return this.value; // clean, unformatted number for year
            }
        },
        accessibility: {
            rangeDescription: 'Range: 1940 to 2017.'
        }
    },
    yAxis: {
        title: {
            text: 'Nuclear weapon states'
        }
    }
})
```



```
        },
        labels: {
            formatter: function () {
                return this.value / 1000 + 'k';
            }
        }
    },
    tooltip: {
        pointFormat: '{series.name} had stockpiled <b>{point.y:,.0f}</b><br/>warhe
    },
    plotOptions: {
        area: {
            pointStart: 1940,
            marker: {
                enabled: false,
                symbol: 'circle',
                radius: 2,
                states: {
                    hover: {
                        enabled: true
                    }
                }
            }
        }
    }
},
series: [
    {
        name: 'USA',
        data: [
            null, null, null, null, null, 6, 11, 32, 110, 235,
            369, 640, 1005, 1436, 2063, 3057, 4618, 6444, 9822, 15468,
            20434, 24126, 27387, 29459, 31056, 31982, 32040, 31233, 29224, 27342,
            26662, 26956, 27912, 28999, 28965, 27826, 25579, 25722, 24826, 24605,
            24304, 23464, 23708, 24099, 24357, 24237, 24401, 24344, 23586, 22380,
            21004, 17287, 14747, 13076, 12555, 12144, 11009, 10950, 10871, 10824,
            10577, 10527, 10475, 10421, 10358, 10295, 10104, 9914, 9620, 9326,
            5113, 5113, 4954, 4804, 4761, 4717, 4368, 4018
        ]
    }
]
```

```

        ]
    }, {
        name: 'USSR/Russia',
        data: [null, null, null, null, null, null, null, null, null,
            5, 25, 50, 120, 150, 200, 426, 660, 869, 1060,
            1605, 2471, 3322, 4238, 5221, 6129, 7089, 8339, 9399, 10538,
            11643, 13092, 14478, 15915, 17385, 19055, 21205, 23044, 25393, 27935,
            30062, 32049, 33952, 35804, 37431, 39197, 45000, 43000, 41000, 39000,
            37000, 35000, 33000, 31000, 29000, 27000, 25000, 24000, 23000, 22000,
            21000, 20000, 19000, 18000, 17000, 16000, 15537, 14162, 12787,
            12600, 11400, 5500, 4512, 4502, 4502, 4500, 4500
        ]
    }]
});
```

对应到 R 包 **highcharter** 中，绘图代码如下：

```

library(highcharter)
options(highcharter.theme = hc_theme_hcrt(tooltip = list(valueDecimals = 2)))

usa <- ts(
    data = c(
        NA, NA, NA, NA, NA, 6, 11, 32, 110, 235,
        369, 640, 1005, 1436, 2063, 3057, 4618, 6444, 9822, 15468,
        20434, 24126, 27387, 29459, 31056, 31982, 32040, 31233, 29224, 27342,
        26662, 26956, 27912, 28999, 28965, 27826, 25579, 25722, 24826, 24605,
        24304, 23464, 23708, 24099, 24357, 24237, 24401, 24344, 23586, 22380,
        21004, 17287, 14747, 13076, 12555, 12144, 11009, 10950, 10871, 10824,
        10577, 10527, 10475, 10421, 10358, 10295, 10104, 9914, 9620, 9326,
        5113, 5113, 4954, 4804, 4761, 4717, 4368, 4018
    ),
    start = 1940, end = 2017
)

russia <- ts(
    data = c(
        NA, NA, NA, NA, NA, NA, NA, NA, NA,
```



```
5, 25, 50, 120, 150, 200, 426, 660, 869, 1060,
1605, 2471, 3322, 4238, 5221, 6129, 7089, 8339, 9399, 10538,
11643, 13092, 14478, 15915, 17385, 19055, 21205, 23044, 25393, 27935,
30062, 32049, 33952, 35804, 37431, 39197, 45000, 43000, 41000, 39000,
37000, 35000, 33000, 31000, 29000, 27000, 25000, 24000, 23000, 22000,
21000, 20000, 19000, 18000, 18000, 17000, 16000, 15537, 14162, 12787,
12600, 11400, 5500, 4512, 4502, 4502, 4500, 4500
),
start = 1940, end = 2017
)

unit_format <- JS("function(){
  return this.value / 10000 + 'M';
}")

highchart() %>%
  hc_xAxis(type = "datetime") %>%
  hc_yAxis(
    title = list(text = "Nuclear weapon states"),
    labels = list(formatter = unit_format)
  ) %>%
  hc_title(text = "US and USSR nuclear stockpiles") %>%
  hc_subtitle(text = paste(
    'Sources: <a href="https://thebulletin.org/2006/july/global-nuclear-stockpiles">
      thebulletin.org</a> &amp; <a href="https://www.armscontrol.org/factsheets/Nuclear
      armscontrol.org</a>"'
  )) %>%
  hc_add_series(data = russia, type = "area", name = "USSR/Russia") %>%
  hc_add_series(data = usa, type = "area", name = "USA") %>%
  hc_exporting(
    enabled = TRUE,
    filename = paste(Sys.Date(), "nuclear", sep = "-")
)
```

可以看出来，JS API 文档里 `chart -> plotOptions` 对应于 R 包 API 的 `hc_plotOptions()` 函数，`hchart()` 函数对应于 <https://api.highcharts.com/highcharts/series>，为了绘图方便起见，作者还直接支持 R 中一些数据对象，比

如数据框 `data.frame` 和时间序列 `ts` 等，完整的支持列表见：

```
library(highcharter)
methods(hchart)

## [1] hchart.acf*      hchart.character* hchart.data.frame* hchart.default*
## [5] hchart.density*   hchart.dist*     hchart.ets*       hchart.factor*
## [9] hchart.forecast*  hchart.histogram* hchart.igraph*    hchart.list*
## [13] hchart.matrix*   hchart.mforecast* hchart.mts*      hchart.numeric*
## [17] hchart.prcomp*   hchart.princomp* hchart.stl*      hchart.survfit*
## [21] hchart.tibble*   hchart.ts*      hchart.xts*
## see '?methods' for accessing help and source code
```

更多 API 细节描述见 <https://jkunst.com/highcharter/articles/modules.html>。桑基图描述能量的流动⁴

```
library(jsonlite)
# 转化为 JSON 格式的字符串
dat <- toJSON(data.frame(
  from = c("AT", "DE", "CH", "DE"),
  to = c("DE", "CH", "DE", "FI"),
  weight = c(10, 5, 15, 5)
))

highchart() %>%
  hc_chart(type = "sankey") %>%
  hc_add_series(data = dat)
```

此外，`highcharter` 提供 `highchartOutput()` 和 `renderHighchart()` 函数支持在 shiny 中使用 `highcharts` 图形。

```
library(shiny)
library(highcharter)

shinyApp(
  ui = fluidPage(
    highchartOutput("plot_hc")
  ),
  server = function(input, output) {
```

⁴<https://antv-2018.alipay.com/zh-cn/vis/chart/sankey.html>



```
output$plot_hc <- renderHighchart({  
  hchart(PlantGrowth, "area", hcAES(y = weight, group = group))  
})  
})
```

借助 `htmlwidgets` 和 `reactR` 创建新的基于 JS 库的 R 包，这样就快速将可视化图形库赋能 R 环境，关于网页可视化，JS 一定是优于 R 的，毕竟人家是专业前端工具，我们做的就是快速套模板，让 R 数据操作和分析的结果以非常精美的方式展现出来。这里有一篇基于 `reactR` 框架引入 React.js 衍生 JS 库到 R 环境中的资料 <https://github.com/react-R/nivocal>，一读就懂，非常适合上手。

提示

点击图例隐藏某一类别，可以看到图形纵轴会自适应展示区域的大小，这个特性对于所有图形都是支持的。

```
library(highcharter)  
# 折线图  
hchart(sleep, "line", hcAES(ID, extra, group = group))  
# 堆积区域图  
# 堆积折线图
```

9.26 动画 I

动态条形图

```
library(highcharter) # highcharter 的依赖也很重  
library(idbr)  
library(purrr)  
library(dplyr) # 未来替代一下  
  
# the US Census Bureau International Data Base API  
# 美国人口普查局国际数据库 API  
idb_api_key("35f116582d5a89d11a47c7fffc2ba309133f09d")  
yrs <- seq(1980, 2030, by = 5)  
  
df <- map_dfr(c("male", "female"), function(sex) {
```

```
    transform(get_idb("US", yrs, sex = sex), sex_label = sex)
})

df <- df %>%
  transform(population = pop * ifelse(sex_label == "male", -1, 1))

# 数据变换
series <- df %>%
  group_by(sex_label, age) %>%
  do(data = list(sequence = .\$population)) %>%
  ungroup() %>%
  group_by(sex_label) %>%
  do(data = .\$data) %>%
  mutate(name = sex_label) %>%
  list_parse()

maxpop <- max(abs(df\$population))

xaxis <- list(
  categories = sort(unique(df$age)),
  reversed = FALSE, tickInterval = 5,
  labels = list(step = 5)
)

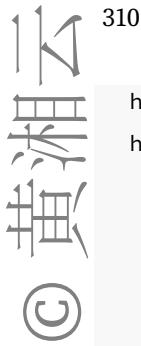
highchart() %>%
  hc_chart(type = "bar") %>%
  hc_motion(
    enabled = TRUE,
    labels = yrs,
    series = c(0, 1),
    autoplay = TRUE,
    updateInterval = 10,
    playIcon = "fa fa-play",
    pauseIcon = "fa fa-pause"
) %>%
  hc_add_series_list(series) %>%
```



```
hc_plotOptions(  
    series = list(stacking = "normal"),  
    bar = list(groupPadding = 0, pointPadding = 0, borderWidth = 0)  
) %>%  
hc_tooltip(  
    shared = FALSE,  
    formatter = JS(  
        function() {  
            return '<b>' + this.series.name +  
                ', age ' + this.point.category +  
                '</b><br/>' + 'Population: ' +  
                Highcharts.numberFormat(Math.abs(this.point.y), 0);  
        }  
    ")  
) %>%  
hc_yAxis(  
    labels = list(  
        formatter = JS(  
            function() {  
                return Math.abs(this.value) / 1000000 + 'M';  
            }  
        ")  
,  
        tickInterval = 0.5e6,  
        min = -maxpop,  
        max = maxpop  
) %>%  
hc_xAxis(  
    xaxis,  
    rlist::list.merge(xaxis, list(opposite = TRUE, linkedTo = 0))  
)
```

动态气泡图

```
highchart() %>%  
    hc_xAxis(min = 0, max = 10) %>%  
    hc_yAxis(min = 0, max = 10) %>%
```



```
hc_motion(enabled = TRUE) %>%
hc_add_series(
  type = "bubble",
  data = list(
    list(
      sequence = list(
        list(x = 1, y = 1, z = 10),
        list(x = 2, y = 3, z = 5),
        list(x = 3, y = 5, z = 8)
      )
    )
  )
)

highchart() %>%
  hc_xAxis(min = 0, max = 10) %>%
  hc_yAxis(min = 0, max = 10) %>%
  hc_add_series(
    type = "bubble",
    name = "气泡图",
    data = list(
      list(x = 1, y = 1, z = 10)
    )
  )
```

动态散点图

```
library(highcharter)

highchart() %>%
  hc_chart(type = "scatter") %>%
  hc_yAxis(max = 6, min = 0) %>%
  hc_xAxis(max = 6, min = 0) %>%
  hc_add_series(
    name = "Australia",
    data = list(
      list(sequence = list(c(1, 1), c(2, 2), c(3, 3), c(4, 4)))
    )
  )
```



```
) %>%  
hc_add_series(  
  name = "United States",  
  data = list(  
    list(sequence = list(c(0, 0), c(3, 2), c(4, 3), c(4, 1)))  
  )  
) %>%  
hc_add_series(  
  name = "China",  
  data = list(  
    list(sequence = list(c(3, 2), c(2, 2), c(1, 1), c(2, 5)))  
  )  
) %>%  
hc_motion(  
  enabled = TRUE,  
  labels = 2000:2003,  
  series = c(0, 1, 2)  
)
```

动态柱状图

```
highchart() %>%  
  hc_chart(type = "column") %>%  
  hc_yAxis(max = 6, min = 0) %>%  
  hc_add_series(name = "A", data = c(2, 3, 4), zIndex = -10) %>%  
  hc_add_series(  
    name = "B",  
    data = list(  
      list(sequence = c(1, 2, 3, 4)),  
      list(sequence = c(3, 2, 1, 3)),  
      list(sequence = c(2, 5, 4, 3))  
    )  
) %>%  
  hc_add_series(  
    name = "C",  
    data = list(  
      list(sequence = c(3, 2, 1, 3)),
```



```
list(sequence = c(2, 5, 4, 3)),
  list(sequence = c(1, 2, 3, 4))
)
) %>%
hc_motion(
  enabled = TRUE,
  labels = 2000:2003,
  series = c(1, 2),
  playIcon = "fa fa-play",
  pauseIcon = "fa fa-pause"
)
```

9.27 时序图

dygraphs 专门用来绘制交互式时间序列图形，下面以美团股价为例，展示时间窗口筛选、坐标轴名称、刻度标签、注释、事件标注、缩放等功能

```
meituan <- quantmod::getSymbols("3690.HK", auto.assign = FALSE, src = "yahoo")
library(dygraphs)

# 缩放
dyUnzoom <- function(dygraph) {
  dyPlugin(
    dygraph = dygraph,
    name = "Unzoom",
    path = system.file("plugins/unzoom.js", package = "dygraphs")
  )
}

# 年月
getYearMonth <- '
function(d) {
  var monthNames = ["01", "02", "03", "04", "05", "06", "07", "08", "09", "10", "11",
    date = new Date(d);
    return date.getFullYear() + "-" + monthNames[date.getMonth()];
}
'
```



```
dygraph(meituan[, "3690.HK.Adjusted"], main = "美团股价走势") |>
  dyRangeSelector(dateWindow = c(format(Sys.Date(), "%Y-01-01"), as.character(Sys.
  dyAxis(name = "x", axisLabelFormatter = getYearMonth) |>
  dyAxis("y", valueRange = c(0, 500), label = "美团股价") |>
  dyEvent("2020-01-23", "武汉封城", labelLoc = "bottom") |>
  dyShading(from = "2020-01-23", to = "2020-04-08", color = "#FFE6E6") |>
  dyAnnotation("2020-01-23", text = "武汉封城", tooltip = "武汉封城", width = 60)
  dyAnnotation("2020-04-08", text = "武汉解封", tooltip = "武汉解封", width = 60)
  dyHighlight(highlightSeriesOpts = list(strokeWidth = 2)) |>
  dySeries(label = "调整股价") |>
  dyLegend(show = "follow", hideOnMouseOut = FALSE) |>
  dyOptions(fillGraph = TRUE, drawGrid = FALSE, gridLineColor = "lightblue") |>
  dyUnzoom()
```

9.28 图形导出

orca (Open-source Report Creator App) 软件针对 plotly.js 库渲染的图形具有很强的导出功能，[安装 orca](#) 后，`plotly::orca()` 函数可以将基于 `htmlwidgets` 的 `plotly` 图形对象导出为 PNG、PDF 和 SVG 等格式的高质量静态图片。

```
p <- plot_ly(x = 1:10, y = 1:10, color = 1:10)
orca(p, "plot.svg")
```

9.29 地图 II

相比于 `plotly`, `echarts4r` 更加轻量，这得益于 JavaScript 库 [Apache ECharts](#)。前者 MIT 协议，后者采用 Apache-2.0 协议，都可以商用。Apache ECharts 是 Apache 旗下顶级开源项目，由百度前端技术团队贡献，中文文档也比较全，学习起来门槛会低一些。

```
library(echarts4r)
quakes |>
  e_charts(long) |>
  e_geo(
```



```

roam = TRUE,
boundingCoords = list(
  c(185, -10),
  c(165, -40)
)
) |>
e_scatter(
  lat, mag,
  coord_system = "geo"
) |>
e_visual_map(mag, scale = e_scale)

```

leaflet 包制作地图，斐济是太平洋上的一个岛国，处于板块交界处，经常发生地震，如下图所示，展示 1964 年来 1000 次震级大于 4 级的地震活动。

```

library(leaflet)
data(quakes)
# Pop 提示
quakes$popup_text <- lapply(paste(
  "编号:", "<strong>", quakes$stations, "</strong>", "<br>",
  "震深:", quakes$depth, "<br>",
  "震级:", quakes$mag
), htmltools::HTML)
# 构造调色板
pal <- colorBin("Spectral", bins = pretty(quakes$mag), reverse = TRUE)
p <- leaflet(quakes) |>
  addProviderTiles(providers$CartoDB.Positron) |>
  addCircles(lng = ~long, lat = ~lat, color = ~pal(mag), label = ~popup_text) |>
  addLegend("bottomright",
    pal = pal, values = ~mag,
    title = "地震震级"
  ) |>
  addScaleBar(position = c("bottomleft"))
p

```

将上面的绘图部分保存为独立的 HTML 网页文件

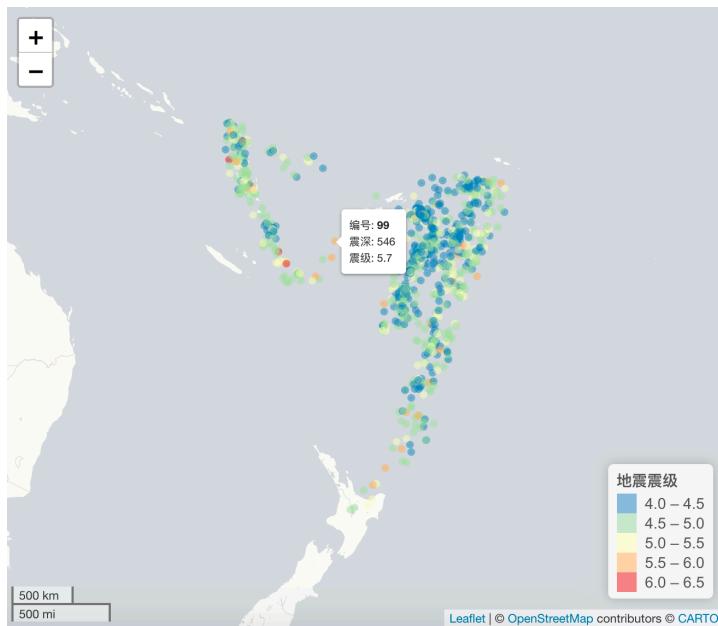


图 9.2: 斐济地震带

```
library(htmlwidgets)
# p 就是绘图部分的数据对象
saveWidget(p, "fiji-map.html", selfcontained = T)

library(leaflet)
library(leaflet.extras)

quakes |>
  leaflet() |>
  addTiles() |>
  addProviderTiles(providers$OpenStreetMap.DE) |>
  addHeatmap(
    lng = ~long, lat = ~lat, intensity = ~mag,
    max = 100, radius = 20, blur = 10
  )
```

leafletCN 提供汉化

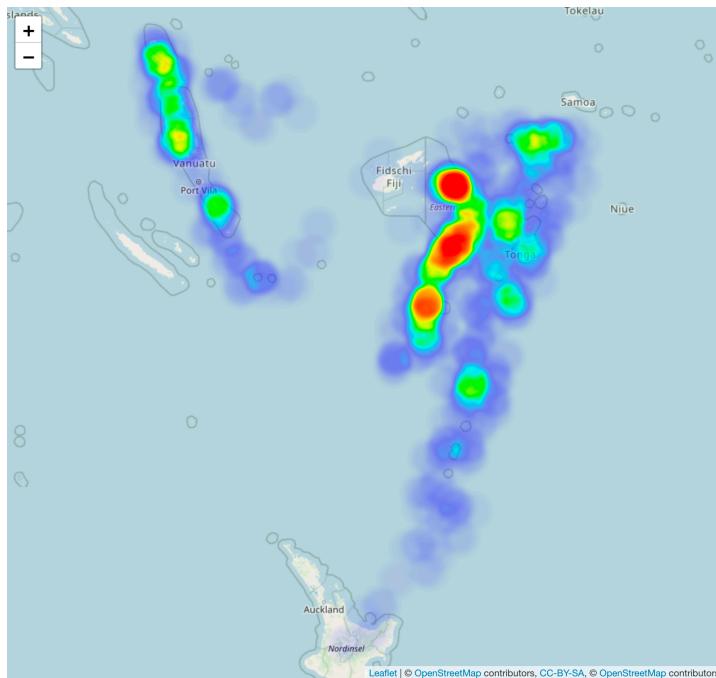


图 9.3: 斐济地震带热力图

```
# 地图默认放大倍数
zoom      <- 4

# 地图可以放大的倍数区间
minZoom   <- 1
maxZoom   <- 18

library(leaflet)
library(leafletCN)
library(maptools)
library(leaflet.extras)

# 热力图 heatmap
leaflet(res, options = leafletOptions(minZoom = minZoom, maxZoom = maxZoom)) |>
  amap() |>
  # setView(lng = mean(data$long), lat = mean(data$lat), zoom = zoom) |>
  setView(lng = 109, lat = 38, zoom = 4) |>
  addHeatmap()
```



```
    lng = ~long2, lat = ~lat2, intensity = ~uv, max = max(res$uv),
    blur = blur, minOpacity = minOpacity, radius = radius
  )

quakes$popup_text <- lapply(paste(
  "编号:", "<strong>", quakes$stations, "</strong>", "<br>",
  "震深:", quakes$depth, "<br>",
  "震级:", quakes$mag
), htmltools::HTML)
# 构造调色板
pal <- colorBin("Spectral", bins = pretty(quakes$mag), reverse = TRUE)

leaflet(quakes) |>
  addProviderTiles(providers$CartoDB.Positron) |>
  addCircles(
    lng = ~long, lat = ~lat,
    color = ~ pal(mag), label = ~popup_text
  ) |>
  setView(178, -20, 5) |>
  addHeatmap(
    lng = ~long, lat = ~lat, intensity = ~mag,
    blur = 20, max = 0.05, radius = 15
  ) |>
  addLegend("bottomright",
    pal = pal, values = ~mag,
    title = "地震震级"
  ) |>
  addScaleBar(position = c("bottomleft"))
```

9.30 动画 II

```
# https://d.cosx.org/d/422311
library(purrr)
library(echarts4r)
```

```
data("gapminder", package = "gapminder")

titles <- map(unique(gapminder$year), function(x) {
  list(
    text = "Gapminder",
    left = "center"
  )
})

years <- map(unique(gapminder$year), function(x) {
  list(
    subtext = x,
    left = "center",
    top = "center",
    z = 0,
    subtextStyle = list(
      fontSize = 100,
      color = "rgb(170, 170, 170, 0.5)",
      fontWeight = "bolder"
    )
  )
})

# 添加一列颜色，各大洲和颜色的对应关系可自定义，调整 levels 或 labels 里面的顺序即可，也
gapminder <- gapminder |>
  transform(
    color = factor(
      continent,
      levels = c("Asia", "Africa", "Americas", "Europe", "Oceania"),
      labels = RColorBrewer::brewer.pal(n = 5, name = "Spectral")
    )
  )

gapminder |>
  group_by(year) |>
```



```
e_charts(x = gdpPercap, timeline = TRUE) |>  
e_scatter(  
  serie = lifeExp, size = pop, bind = country,  
  symbol_size = 5, name = ""  
) |>  
e_add("itemStyle", color) |>  
e_y_axis(  
  min = 20, max = 85, nameGap = 30,  
  name = "Life Exp", nameLocation = "center"  
) |>  
e_x_axis(  
  type = "log", min = 100, max = 100000,  
  nameGap = 30, name = "GDP / Cap", nameLocation = "center"  
) |>  
e_timeline_serie(title = titles) |>  
e_timeline_serie(title = years, index = 2) |>  
e_timeline_opts(playInterval = 1000) |>  
e_grid(bottom = 100) |>  
e_tooltip()
```

```
# params.name 对应 bind  
# params.value[0] 对应 x  
# params.value[1] 对应 serie  
# params.value[2] 对应 size  
# tooltips 自定义  
# https://stackoverflow.com/questions/50554304/displaying-extra-variables-in-tooltips  
# 百分数处理  
# https://stackoverflow.com/questions/11832914/how-to-round-to-at-most-2-decimal-places  
mtcars |>  
  tibble::rownames_to_column("model") |>  
  e_charts(x = wt) |>  
  e_scatter(serie = mpg, size = qsec, bind = model) |>  
  e_tooltip(formatter = htmlwidgets::JS(  
    function(params) {  
      return (  
        '<strong>' + params.name + '</strong>' +
```

黄湘云
©

```
'<br />wt: ' + params.value[0] +
'<br />mpg: ' + params.value[1] +
'<br />qsec- ' + params.value[2]
)
}
"))
})
```

9.31 网络图

[gephi](#) 探索和可视化网络图 GraphViz

```
library(igraph)
```

9.31.1 networkD3

[networkD3](#) D3 非常适合绘制网络图，如网络、树状、桑基图

```
library(networkD3)
data(MisLinks, MisNodes) # 加载数据
head(MisLinks) # 边
```

```
##   source target value
## 1      1      0     1
## 2      2      0     8
## 3      3      0    10
## 4      3      2     6
## 5      4      0     1
## 6      5      0     1
```

```
head(MisNodes) # 节点
```

	name	group	size
## 1	Myriel	1	15
## 2	Napoleon	1	20
## 3	Mlle.Baptistine	1	23
## 4	Mme.Magloire	1	30
## 5	CountessdeLo	1	11



```
## 6 Geborand 1 9
```

构造网络图

```
forceNetwork(  
  Links = MisLinks, Nodes = MisNodes, Source = "source",  
  Target = "target", Value = "value", NodeID = "name",  
  Group = "group", opacity = 0.4  
)
```

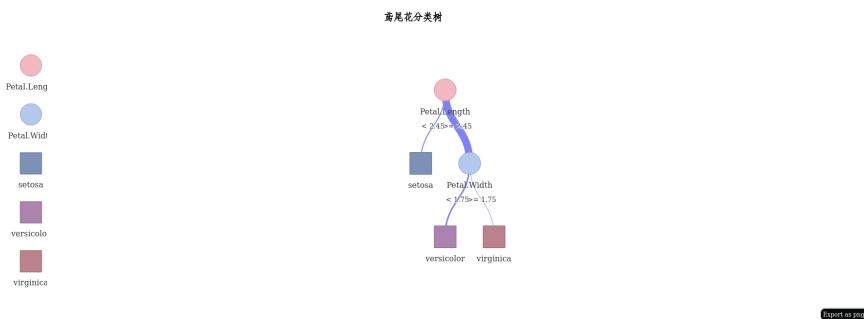
9.31.2 visNetwork

visNetwork 使用 [vis-network.js](https://datastorm-open.github.io/visNetwork) 库绘制网络关系图 <https://datastorm-open.github.io/visNetwork>

```
library(visNetwork)
```

调用函数 `visTree()` 可可视化分类模型结果

```
library(rpart)  
library(sparkline) # 函数 visTree 需要导入 sparkline 包  
res <- rpart(Species~., data=iris)  
visTree(res, main = "鸢尾花分类树", width = "100%")
```



节点、边的属性都可以映射数据指标

9.31.3 r2d3

D3 是非常流行的 JavaScript 库，r2d3 提供了 R 接口

```
library(r2d3)
```

更加具体的使用介绍，一个复杂的案例，如何从简单配置过来，以条形图为例，D3 是一个相当强大且成熟的库，提供的案例功能要覆盖 plotly

r2d3 提供了两个样例 JS 库 baranims.js 和 barchart.js

```
list.files(system.file("examples/", package = "r2d3"))
```

```
## [1] "baranims.js" "barchart.js"
```

```
library(r2d3)
```

```
r2d3(
```

```
  data = c(0.3, 0.6, 0.8, 0.95, 0.40, 0.20),
  script = system.file("examples/barchart.js", package = "r2d3")
)
```

```
r2d3(  
  data = c(0.3, 0.6, 0.8, 0.95, 0.40, 0.20),  
  script = system.file("examples/baranims.js", package = "r2d3")  
)
```

TODO: 提供一个 R 包和 HTML Widgets 小练习：给 roughViz.js 写个 R 包装
<https://d.cosx.org/d/421030-r-html-widgets-roughviz-js-r> <https://github.com/XiangyunHuang/roughviz>

9.32 Python 交互图形

Plotly 的图形库

```
import plotly.express as px  
  
px.scatter(  
  px.data.iris(),  
  x="sepal_width",  
  y="sepal_length",  
  color="species",  
  trendline="ols",  
  template="simple_white",  
  labels={  
    "sepal_length": "Sepal Length (cm)",  
    "sepal_width": "Sepal Width (cm)",  
    "species": "Species of Iris",  
  },  
  title="Edgar Anderson's Iris Data",  
  color_discrete_sequence=px.colors.qualitative.Set2  
)
```

9.33 运行环境

```
sessionInfo()
```



图 9.4: 插入图片

```
## R version 4.1.0 (2021-05-18)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 20.04.2 LTS
##
## Matrix products: default
## BLAS:    /usr/lib/x86_64-linux-gnublas/libblas.so.3.9.0
## LAPACK: /usr/lib/x86_64-linux-gnulapack/liblapack.so.3.9.0
##
## locale:
## [1] LC_CTYPE=en_US.UTF-8          LC_NUMERIC=C
## [3] LC_TIME=en_US.UTF-8          LC_COLLATE=en_US.UTF-8
## [5] LC_MONETARY=en_US.UTF-8       LC_MESSAGES=en_US.UTF-8
## [7] LC_PAPER=en_US.UTF-8          LC_NAME=C
## [9] LC_ADDRESS=C                  LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8   LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats      graphics   grDevices utils      datasets   methods   base
##
## other attached packages:
## [1] sparkline_2.0           rpart_4.1-15        visNetwork_2.0.9
```



```
## [4] networkD3_0.4           igraph_1.2.6          leaflet.extras_1.0.0
## [7] maptools_1.1-1            sp_1.4-5             leafletCN_0.2.1
## [10] leaflet_2.0.4.1          echarts4r_0.4.1    r2d3_0.2.5
## [13] dygraphs_1.1.1.6         highcharter_0.8.2   plotly_4.9.4.1
## [16] ggplot2_3.3.5            reticulate_1.20
##
## loaded via a namespace (and not attached):
##
## [1] xts_0.12.1           lubridate_1.7.10   webshot_0.5.2   httr_1.4.2
## [5] tools_4.1.0            backports_1.2.1   utf8_1.2.2     R6_2.5.0
## [9] DBI_1.1.1              lazyeval_0.2.2    colorspace_2.0-2 withr_2.4.2
## [13] tidyselect_1.1.1       processx_3.5.2   curl_4.3.2      compiler_4.1.0
## [17] isoband_0.2.5          labeling_0.4.2    bookdown_0.22   scales_1.1.1
## [21] callr_3.7.0            stringr_1.4.0    digest_0.6.27   foreign_0.8-81
## [25] rmarkdown_2.9           pkgconfig_2.0.3   htmltools_0.5.1.1 fastmap_1.1.0
## [29] highr_0.9              htmlwidgets_1.5.3  rlang_0.4.11   TTR_0.24.2
## [33] rstudioapi_0.13        quantmod_0.4.18   shiny_1.6.0    farver_2.1.0
## [37] generics_0.1.0          zoo_1.8-9        jsonlite_1.7.2  crosstalk_1.1.1
## [41] dplyr_1.0.7             magrittr_2.0.1    rlist_0.4.6.1   Matrix_1.3-4
## [45] Rcpp_1.0.7              munsell_0.5.0    fansi_0.5.0    lifecycle_1.0.0
## [49] stringi_1.7.3           yaml_2.2.1      MASS_7.3-54    grid_4.1.0
## [53] promises_1.2.0.1        crayon_1.4.1    lattice_0.20-44 knitr_1.33
## [57] ps_1.6.0                pillar_1.6.2    glue_1.4.2     evaluate_0.14
## [61] data.table_1.14.0       png_0.1-7      vctrs_0.3.8    httpuv_1.6.1
## [65] gtable_0.3.0             purrr_0.3.4    tidyverse_1.1.3 assertthat_0.2.1
## [69] xfun_0.24                mime_0.11      xtable_1.8-4   broom_0.7.9
## [73] later_1.2.0              viridisLite_0.4.0 tibble_3.1.3   ellipsis_0.3.2
```

第十章 交互表格

Greg Lin 开发的 **reactable** 包覆盖测试达到惊人的 99%，它基于 JavaScript 库 **react-table**，是 **react** 框架的衍生品，Nick Raienko 整理了一份超棒的 **react** 模块合集也许机智如你，可以引入更多优秀的 **react** 模块到 R 语言社区。[reactablefmtr](#) 提供一些函数简化 **reactable** 定制表格的复杂性

谢益辉开发的 **DT** 包覆盖测试 31%，它基于 **DataTables** 库，是 **jQuery** 框架的衍生品。益辉评价 **reactable** 在多个方面优于 **DT**，比如行分组和聚合，嵌入 **HTML widgets**，甚至说要是 **reactable** 存在于 **DT** 之前，他就不会新开发 **DT** 这个 R 包了，不过这是后话了¹。

Richard Iannone 开发的 **gt** 包覆盖测试 78%，类似 **ggplot2** 的设计哲学，试图打造制作表格的语法，相比于 **reactable** 和 **DT**，它不依赖于 JavaScript 库，更加轻量，一般来讲，持续维护更新重 JS 库依赖的 R 包比较累人，JS 库可能会不断重构，进而变动 API。

朱昊开发的 **kableExtra** 大大扩展了 **knitr** 包的 **kable()** 函数的功能，虽没有覆盖测试，但中英文文档特别详细，见官网 <https://haozhu233.github.io/kableExtra/>。

目前，Greg Lin、谢益辉和 Richard Iannone 都是 RStudio 公司雇员，他们背靠开源组织和大公司，开发的这些 R 包的生命力都比较强。**gt** 和 **kableExtra** 摆脱了 JavaScript 库的依赖，网页形式的表格可以嵌入到邮件内容中，这是一个不太引人注意的优势。**kableExtra** 还支持高度自定义的 LaTeX 输出，详见案例 <https://github.com/XiangyunHuang/bookdown-kableExtra>，**gt** 包据说未来也会支持，拭目以待吧，也许在成书之日能看到！

此外，还有任坤开发的 **formattable** 和 David Gohel 开发的 **flextable** 包等，一份综合介绍见博文 [How to Make Beautiful Tables in R](#)。

rtables 处于原型开发的阶段，针对复杂表格，有比较好的设计。**tablesgg** 使用

¹<https://bookdown.org/yihui/rmarkdown-cookbook/table-other.html>



ggplot2 将表格渲染成图片。

10.1 DT 和 reactable

DT 基于 jQuery 的 JS 库 `DataTables` 提供了一个 R 的封装，封装工具和许多其他基于 JS 库的 R 包一样，比如即将介绍的 `reactable` 包，都依赖于 `htmlwidgets`。

```
library(magrittr)

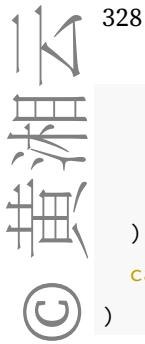
if (!is.na(Sys.getenv('CI', NA))) {
  Sys.setenv(R_CRAN_WEB = "https://cloud.r-project.org/")
} else {
  Sys.setenv(R_CRAN_WEB = "https://mirrors.tuna.tsinghua.edu.cn/CRAN")
}

pdb <- tools::CRAN_package_db()
sub_pdb <- subset(pdb, subset = !duplicated(pdb[, "Package"]) & pdb[, "Package"] %in%
  c("gridExtra", "grid", "gridBase", "gridSVG", "gridUnit", "gridExtra", "gridBase", "grid"))
pkg_pdb <- subset(sub_pdb,
  subset = grepl("Yihui Xie", sub_pdb[, "Maintainer"]) | grepl("Hadley Wickham", sub_pdb[, "Maintainer"]))
select = c("Maintainer", "Package", "Version", "Published", "Title")
)

pkg_pdb <- transform(pkg_pdb, Title = gsub("\\\\n", " ", Title))

library(DT)

datatable(pkg_pdb[order(pkg_pdb$Maintainer, decreasing = T), ],
  rownames = F, # 不显示行名
  extensions = c("Buttons", "RowGroup"),
  options = list(
    pageLength = 10, # 每页显示的行数
    language = list(url = "//cdn.datatables.net/plug-ins/1.10.11/i18n/Chinese.json"),
    dom = "Brtp", # 去掉显示行数 i、过滤 f 的能力，翻页用 p 表示
    ordering = F, # 去掉列排序
    buttons = c("copy", "csv", "excel", "pdf", "print"), # 提供打印按钮
    rowGroup = list(dataSrc = 0), # 按 Maintainer 列分组
    columnDefs = list(
```



```
list(className = "dt-center", targets = 0), # 不显示行名, 则 targets 从 0 开始, 省略
      list(visible = FALSE, targets = 0) # 不显示 Maintainer 列
    )
),
caption = "谢大和哈神维护的 R 包"
)

colorize_num <- function(x) {
  ifelse(x > 0,
    sprintf("<span style='color:%s'>%s</span>", "green", x),
    sprintf("<span style='color:%s'>%s</span>", "red", x)
  )
}

colorize_pct <- function(x) {
  ifelse(x > 0,
    sprintf("<span style='color:%s'>%s</span>", "green", scales::percent(x, accuracy = 0)),
    sprintf("<span style='color:%s'>%s</span>", "red", scales::percent(x, accuracy = 0))
  )
}

colorize_pp <- function(x) {
  ifelse(x > 0,
    sprintf("<span style='color:%s'>%s</span>", "green", paste0(round(100*x, digits = 2))),
    sprintf("<span style='color:%s'>%s</span>", "red", paste0(round(100*x, digits = 2)))
  )
}

colorize_text <- function(x, color = "red") {
  sprintf("<span style='color:%s'>%s</span>", color, x)
}

library(tibble)

dat = tribble(
  ~name1, ~name2,
  as.character(htmltools::tags$b("加粗")), as.character(htmltools::a(href = "https://rstatistician.com")))
```



```
as.character(htmltools::em("强调")), '<a href="#" onclick="alert(\'Hello World\')'
as.character(htmltools::span(style = 'color:red', "正常")), '正常'
)

datatable(
  data = dat,
  escape = F, # 设置 escape = F
  colnames = c(colorize_text("第1列", "red"), as.character(htmltools::em("第2列")))
  caption = htmltools::tags$caption(
    style = "caption-side: top; text-align: center;",
    "表格 2: ", htmltools::em("表格标题")
  ), # 在表格底部显示标题，默认在表格上方显示标题
  # filter = "top", # 过滤框
  options = list(
    pageLength = 5, # 每页显示5行
    dom = "t"
  )
)
```

下面重点介绍 reactable 包，看看 React.js 和 Shiny 是如何集成的，这是比较高级的主题，主要参考 [Alan Dipert](#) 的演讲材料 [Integrating React.js and Shiny](#)。

```
library(reactable)
```

下面这个例子来自 React.js 官网 <https://reactjs.org/>

```
```js
class HelloMessage extends React.Component {
 render() {
 return (
 <div>
 Hello {this.props.name}
 </div>
);
 }
}

ReactDOM.render(
```



```

<HelloMessage name="Taylor" />,
document.getElementById('hello-example')
);
...
```

```

更多细节定制见 Thomas Mock 的博文 [reactable - An Interactive Tables Guide](#)

reactable 制作表格

```

library(shiny)
library(reactable)

ui <- fluidPage(
  reactableOutput("table")
)

server <- function(input, output) {
  output$table <- renderReactable({
    reactable(iris,
      filterable = TRUE, # 过滤
      searchable = TRUE, # 搜索
      showPageSizeOptions = TRUE, # 页面大小
      pageSizeOptions = c(5, 10, 15), # 页面大小可选项
      defaultPageSize = 10, # 默认显示10行
      highlight = TRUE, # 高亮选择
      striped = TRUE, # 隔行高亮
      fullWidth = FALSE, # 默认不要全宽填充, 适应数据框的宽度
      defaultSorted = list(
        Sepal.Length = "asc", # 由小到大排序
        Petal.Length = "desc" # 由大到小
      ),
      columns = list(
        Sepal.Width = colDef(style = function(value) { # Sepal.Width 添加颜色标记
          if (value > 3.5) {
            color <- "#008000"
          } else if (value > 2) {
            color <- "#e00000"
          } else {
            color <- "#cccccc"
          }
        })
      )
    )
  })
}

shinyApp(ui, server)
```

```



```
 color <- "#777"
 }
 list(color = color, fontWeight = "bold")
})
}

)
}
}
}

shinyApp(ui, server)
```

```
修改自 Code: https://gist.github.com/jthomas/mock/f085dce3e70e42ca49b052bbe25de49
library(reactable)
library(htmltools)

barchart function from: https://glin.github.io/reactable/articles/building-twitt
bar_chart <- function(label, width = "100%", height = "14px", fill = "#00bfc4", ba
 bar <- div(style = list(background = fill, width = width, height = height))
 chart <- div(style = list(flexGrow = 1, marginLeft = "6px", background = backgr
 div(style = list(display = "flex", alignItems = "center"), label, chart)
}

data <- mtcars |>
 subset(select = c("cyl", "mpg")) |>
 subset(subset = sample(x = c(TRUE, FALSE), size = 6, replace = T))

reactable(
 data,
 defaultPageSize = 20,
 columns = list(
 cyl = colDef(align = "center"),
 mpg = colDef(
 name = "mpg",
 defaultSortOrder = "desc",
```



```
minWidth = 250,
cell = function(value, index) {
 width <- paste0(value * 100 / max(mtcars$mpg), "%")
 value <- format(value, width = 9, justify = "right", nsmall = 1)

 # output the value of another column
 # that aligns with current value
 cyl_val <- data$cyl[index]

 # Color based on the row's cyl value
 color_fill <- if (cyl_val == 4) {
 "#3686d3" # blue
 } else if (cyl_val == 6) {
 "#88398a" # purple
 } else {
 "#fcab27" # orange
 }
 bar_chart(value, width = width, fill = color_fill, background = "#e1e1e1")
},
align = "left",
style = list(fontFamily = "monospace", whiteSpace = "pre")
)
)
)
```

## 10.2 gt 和 kableExtra

如表 10.1 所示，我们可以自定义表格样式，比如配色，例子修改自 kableExtra 帮助文档 <https://haozhu233.github.io/kableExtra/bookdown/cross-format-tables-in-bookdown.html>，同时支持 HTML 和 LaTeX 输出，但是 LaTeX 输出需要在文档类选项中增加 table 选项，即 classoption: "table"，这样就可以加载 colortbl 宏包，进而提供 \rowcolor 等 LaTeX 命令，在表格中给每个格子定制颜色。我们推荐在 classoption 中添加 table 选项，而不是再次加载 xcolor 包，比



如像这样 `\usepackage[table]{xcolor}`，这会在 R Markdown 中引起冲突<sup>2</sup>。

```
library(kableExtra)

iris[1:10,] %>%
 transform(
 Sepal.Length =
 cell_spec(Sepal.Length,
 bold = T,
 color = spec_color(Sepal.Length, end = 0.9),
 font_size = spec_font_size(Sepal.Length)
)
) %>%
 transform(Species = cell_spec(
 Species,
 color = "white", bold = T,
 background = spec_color(1:10,
 end = 0.9,
 option = "A", direction = -1
)
)) %>%
 kable(
 escape = F, align = "c", booktabs = T,
 caption = "自定义表格样式"
) %>%
 kable_styling(c("striped", "condensed"),
 latex_options = "striped",
 full_width = F
)
```

一个非常基本的 gt 制作的表格

```
library(gt)
iris %>%
 head() %>%
 gt()
```

<sup>2</sup><https://stackoverflow.com/questions/50094698/rmarkdown-beamer-presentation-option-clash-clash-for-xcolor>



表 10.1: 自定义表格样式

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5	3.4	1.5	0.2	setosa
4.4	2.9	1.4	0.2	setosa
4.9	3.1	1.5	0.1	setosa

然后添加表格的标题和副标题，套上 `md()` 函数后，标题和副标题支持 Markdown 语法，告别 HTML 的制表方式吧！其它表格元素，如脚注支持和表格的列指标关联

```
library(data.table)

iris %>%
 as.data.table %>%
 .[, head(.SD, 2), by = .(Species)] %>%
 gt() %>%
 tab_header(
 title = md("★★鸢尾花★★数据集"),
 subtitle = "R 内置数据集"
) %>%
 data_color(
 columns = vars(Sepal.Length),
 colors = scales::col_numeric(palette = terrain.colors(5, rev = T), domain = NULL)
) %>%
 data_color(
 columns = vars(Species),
 colors = scales::col_factor(palette = hcl.colors(3), domain = NULL)
) %>%
```

```
tab_footnote(
 footnote = md("据说数据集最早收集自 Fisher's or Anderson's"),
 locations = cells_column_labels(columns = vars(Sepal.Length))
) %>%
tab_footnote(
 footnote = "鸢尾花的类别",
 locations = cells_column_labels(
 columns = vars(Species)
)
)
```

更多细节的设置见 Thomas Mock 的博文[gt - a \(G\)rammar of \(T\)ables](#)

注意

当前 `gt` 包对 LaTeX 的支持比较弱，上述表格在 HTML 网页环境中可以看到的效果并不能一一对应到 LaTeX 输出中。且 `gt` 包生成 LaTeX 表格会自动加载宏包 `amsmath`、`booktabs`、`caption` 和 `longtable`，`gt_latex_dependencies()` 且不能控制

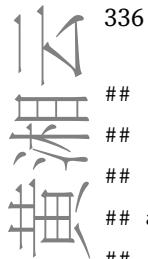
## 10.3 运行环境

```
sessionInfo()

R version 4.1.0 (2021-05-18)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 20.04.2 LTS

Matrix products: default
BLAS: /usr/lib/x86_64-linux-gnublas/libblas.so.3.9.0
LAPACK: /usr/lib/x86_64-linux-gnulapack/liblapack.so.3.9.0

locale:
[1] LC_CTYPE=en_US.UTF-8 LC_NUMERIC=C
[3] LC_TIME=en_US.UTF-8 LC_COLLATE=en_US.UTF-8
[5] LC_MONETARY=en_US.UTF-8 LC_MESSAGES=en_US.UTF-8
[7] LC_PAPER=en_US.UTF-8 LC_NAME=C
```



```
[9] LC_ADDRESS=C LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
attached base packages:
[1] stats graphics grDevices utils datasets methods base
##
other attached packages:
[1] kableExtra_1.3.4 reactable_0.2.3 DT_0.18 magrittr_2.0.1
##
loaded via a namespace (and not attached):
[1] rstudioapi_0.13 knitr_1.33 xml2_1.3.2 rvest_1.0.1
[5] munsell_0.5.0 viridisLite_0.4.0 colorspace_2.0-2 R6_2.5.0
[9] rlang_0.4.11 stringr_1.4.0 httr_1.4.2 tools_4.1.0
[13] webshot_0.5.2 xfun_0.24 systemfonts_1.0.2 htmltools_0.5.1.1
[17] yaml_2.2.1 digest_0.6.27 lifecycle_1.0.0 bookdown_0.22
[21] htmlwidgets_1.5.3 glue_1.4.2 evaluate_0.14 rmarkdown_2.9
[25] stringi_1.7.3 compiler_4.1.0 scales_1.1.1 svglite_2.0.0
```

## 第十一章 交互报表

学习 shiny 应用开发，建议多看看 [Learn Shiny](#)。了解 shiny server，推荐从 [Shiny Server Professional Administrator's Guide](#) 开始。了解 shiny 相关的生态，建议从 shiny 资源列表 <https://github.com/grabear/awesome-rshiny> 和 shiny 扩展合集 <https://github.com/nanxstats/awesome-shiny-extensions> 开始，希望读者能从中打造属于自己的最佳实践。

RStudio 首席技术官 CTO Joe Cheng 在 2019 年 RStudio 大会上介绍 [企业级 shiny 应用原理、实践和工具](#) 可以作为 shiny 从新技术到生产力的蜕变节点。支持高并发的异步编程，比如 Heather Nolis 和 Dr. Jacqueline Nolis 的报告介绍了日百万访问量下的 shiny 应用如何搭建<sup>1</sup>。Colin Fay, Sébastien Rochette, Vincent Guyader, Cervan Girard 的书 [Engineering Production-Grade Shiny Apps](#)、David Granjon 的书 [Outstanding User Interfaces with Shiny](#) 和 Hadley Wickham 的书 [Mastering Shiny](#) 的问世宣告 shiny 的成熟稳定，以及生态的形成，在此之前 shiny 一直不被看好。shiny 生态意味着一个完整的工业级的应用圈，满足安全性、稳定性、高效性、维护性、扩展性的要求。

iSEE is winner of the Most Technically Impressive award of the 2019 Shiny Contest.  
源码地址 <https://github.com/iSEE/isee-shiny-contest>

Six Years of Shiny in Research - Collaborative Development of Web Tools in R  
[Kasprzak et al., 2021]

以 RStudio 为核心，开发 Shiny 应用扩展的社区组织有 [RStudio](#)、[Apppsilon](#)、[RinTeRface](#)、[ThinkR-open](#)、[dreamRs](#) 和 [datastorm-open](#)

---

<sup>1</sup><https://resources.rstudio.com/rstudio-conf-2020/we-re-hitting-r-a-million-times-a-day-so-we-made-a-talk-about-it-heather-nolis-dr-jacqueline-nolis>

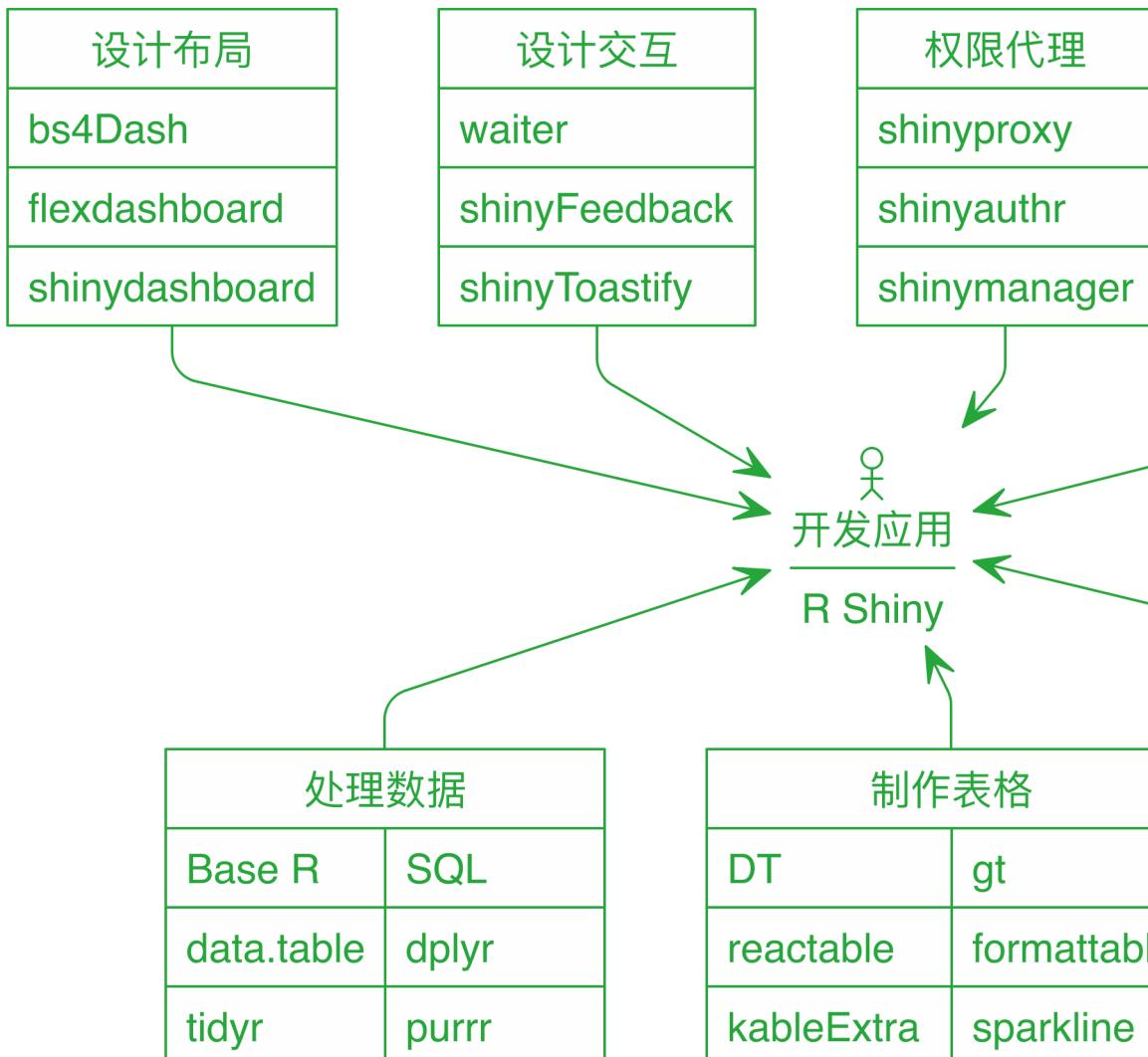


图 11.1: Shiny 生态系统

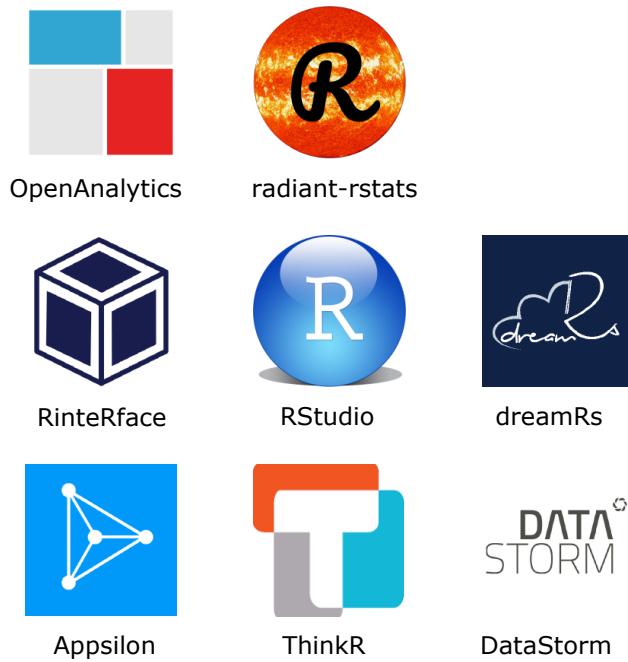


图 11.2: 开发 Shiny 应用扩展的组织



## 11.1 开发流程

报表开发从数据仓库的 DWD 层开始，可能一些业务原因，我们需要从 ODS 层甚至从点击流的日志数据开始，经过数据清洗、提取、聚合成为支撑 BI 报表最底层的基础表，存储在 Hive 中，然后对这一系列的基础表根据 BI 展示的需要进行第二层聚合形成中间表，这两层数据根据业务情况做增量更新或者全量更新，并将中间表同步到 MySQL 仓库中，全量更新的情况，往往更新数据比较大，建议用 sqoop 做数据的同步。创建第二层的中间表稍有些灵活性，原则是在中间表之上对应的数据操作和可视化是容易实现且效率较高的，否则应该构造第三层的中间表，绝不能将大规模的数据集直接导入 R 中进行分析和可视化，拖慢前端展示的速度，占用过多的服务器资源。

## 11.2 开发工具

除了在第 11.1 节介绍的和数据库紧密相关的工具外，我们还需要 Git 做代码管理，Azkaban 做任务调度（或者其它工具做任务调度器），RStudio IDE 做开发工具（或者 VS Code 等），Shiny Server 做报表支撑，做报表管理。具体到 shiny 页面开发，我们需要：

- RMySQL 做数据库连接，推荐 odbc 这个包，它支持连接相当广泛的数据  
库。
- data.table 或者 dplyr 做数据操作，推荐和管道操作 magrittr 一起使用，增  
加代码可读性。
- plotly 或者 highcharter 做数据可视化，reactable 和 DT 做数据呈现，也  
可以 ggplot2 和 plotly 的 ggplotly() 函数共同实现静态图到动态图的交互可  
视化。
- shiny 及其扩展工具做页面设计，比如 shinythemes 可以统一配色，dash-  
boardthemes 提供更加深度的主题，shinytableau 提供仿 Tableau 的 dash-  
board 框架。sass 在 CSS 样式层面重定义网站风格，比如借助 sass 修改  
Bootstrap 4，shiny 的布局其实就是魔改了 Bootstrap 库。
- 针对特定应用场景的其它交互可视化工具包，比如 leaflet 可以将地图嵌入  
Shiny 应用，dygraphs 可以将时间序列塞进去。
- 其它加强 shiny 页面的小功能，比如 shinyFeedback 提供用户输入的反馈，  
miniUI 专为小屏幕设计，shinyMobile 在 IOS 和安卓手机上访问 shiny 应  
用，大大加强 miniUI 的功能，shinyWidgets 提供自定义 widget 的功能，  
shinymanager 支持单个 shiny 应用的权限管理，firebase 提供访问权限设



置 <https://firebase.john-coene.com/>。

- `shiny-server` 以网络服务的方式支持 `shiny` 应用，是企业级 `shiny` 应用的核心，`shinyproxy` 提供企业级部署 `shiny` 应用的开源解决方案，`ShinyStudio` 打造基于容器架构的协作开发环境的开源解决方案，`golem` 构建企业级 `shiny` 应用的框架，`RinteRface` 开发的系列 R 包也试图打造一套完整的解决方案，并配有速查小抄 `cheatsheets`
- `radiant` 探索性分析解决方案

`shinyauthr` 应用授权

```
library(shiny)
```

## 11.3 基础知识

1920s 汽车数据分析和建模

## 11.4 基础组件

### 11.4.1 书签

链接可以指向页面状态

```
library(shiny)

ui <- function(request) {
 fluidPage(
 plotOutput("plot"),
 sliderInput("n", "Number of observations", 1, nrow(faithful), 100),
 bookmarkButton()
)
}

server <- function(input, output, session) {
 output$plot <- renderPlot({
 hist(faithful$eruptions[seq_len(input$n)], breaks = 40)
 })
}
```

```
}

enableBookmarking(store = "url")
shinyApp(ui, server)
```



### 11.4.2 表格

`reactable` 基于 JS 库 `React Table` 提供交互式表格渲染，和 `shiny` 无缝集成，是替代 `DT` 的不二选择，在 `app.R` 用 `reactable` 包的 `reactableOutput()` 和 `renderReactable()` 函数替代 `shiny` 里面的 `dataTableOutput()` 和 `renderDataTable()`。再也不用忍受 `DT` 和 `shiny` 的函数冲突了，且其覆盖测试达到 99%。

```
library(shiny)
library(data.table)
```

`gt` 高度自定义 `gt` 表格样式，支持 `shiny` 集成，`data.table` 提供高效的数据操作，`formattable` 支持自定义格子。

`kableExtra` 包

```
library(shiny)
library(data.table)
library(magrittr)
library(kableExtra)

ui <- fluidPage(
 title = "mtcars datasets",
 titlePanel("mtcars 数据集"),

 sidebarLayout(
 sidebarPanel(
 sliderInput("mpg", "mpg 范围",
 min = 11, max = 33, value = 15
)
),
 mainPanel(
 tableOutput("mtcars_kable")
```



```
)
)
)

设置列序 https://stackoverflow.com/questions/19619666/change-column-position-in-kable
server <- function(input, output) {
 output$mtcars_kable <- function() {
 # 转化数据类型
 mtcars_dt <- as.data.table(mtcars)
 # 添加新的列
 mtcars_dt[, car := rownames(mtcars)][mpg <= input$mpg] %>%
 setcolorder(., c("car", setdiff(names(.), "car")))) %>%
 knitr::kable("html") %>%
 kable_styling("striped", full_width = F) %>%
 add_header_above(c(" ", "Group 1" = 5, "Group 2" = 6))
 }
}

执行程序
shinyApp(ui = ui, server = server)
```

reactable 包

```
library(shiny)
library(reactable)

ui <- fluidPage(
 reactableOutput("table")
)

server <- function(input, output) {
 output$table <- renderReactable({
 reactable(iris,
 filterable = TRUE, # 过滤
 searchable = TRUE, # 搜索
 showPageSizeOptions = TRUE, # 页面大小
 pageSizeOptions = c(5, 10, 15), # 页面大小可选项
```



```
defaultPageSize = 10, # 默认显示10行
highlight = TRUE, # 高亮选择
striped = TRUE, # 隔行高亮
fullWidth = FALSE, # 默认不要全宽填充，适应数据框的宽度
defaultSorted = list(
 Sepal.Length = "asc", # 由小到大排序
 Petal.Length = "desc" # 由大到小
),
columns = list(
 Sepal.Width = colDef(style = function(value) { # Sepal.Width 添加颜色标记
 if (value > 3.5) {
 color <- "#008000"
 } else if (value > 2) {
 color <- "#e00000"
 } else {
 color <- "#777"
 }
 list(color = color, fontWeight = "bold") # 字体加粗
 })
)

)
})
})
}

shinyApp(ui, server)
```

下面介绍 DT

```
library(magrittr)
ui.R 前端
library(shiny)
shinyUI(fluidPage(
 # 应用的标题名称
 titlePanel("鸢尾花数据集"),
 # 边栏
 fluidRow(
```



```
column(
 12,
 DT::dataTableOutput("table")
)
)
)

server.R 服务端
library(shiny)
shinyServer(function(input, output, session) {
 output$table <- iris %>%
 `colnames<-`(., gsub("\\\\.", "_", tolower(colnames(.)))) %>%
 DT::renderDataTable(.,
 options = list(
 pageLength = 5, # 每页显示5行
 initComplete = I("function(settings, json) {alert('Done.')}")
, server = F
)
})
```

**注意**

加载 shiny 包后再加载 DT 包，函数 dataTableOutput() 和 renderDataTable() 显示冲突，因为两个 R 包都有这两个函数。在创建 shiny 应用的过程中，如果我们需要呈现动态表格，就需要使用 DT 包的 DT::dataTableOutput() 和 DT::renderDataTable() 否则会报错，详见 <https://github.com/rstudio/shiny/issues/2653>，DT 包官方文档 <https://rstudio.github.io/DT/>。

**提示**

在 server.R 里我们对数据集 iris 做了重命名的操作，如果不使用管道操作，通常是下面这样操作。

```
colnames(iris) <- gsub("\\\\.", "_", tolower(colnames(iris)))
```

换成管道操作，函数 colnames() 要换成 colnames<-，这其实类似于  $1 + 2$  换成  $+ (1, 2)$ ，保持函数在左边，参数值在右边的一致性。

设置页面默认显示的行数和列的宽度



```
https://stackoverflow.com/questions/45509501/set-names-of-values-in-lengthmenu-page-l
相关例子见 https://github.com/rstudio/shiny-examples/tree/master/018-datatable-option
DT 选项 https://rstudio.github.io/DT/options.html

library(shiny)
library(DT)

ui <- fluidPage(
 DT::dataTableOutput("table")
)

server <- function(input, output) {
 output$table <- DT::renderDataTable({
 DT::datatable(iris, options = list(
 language = list(url = "//cdn.datatables.net/plug-ins/1.10.11/i18n/Chinese.json"),
 pageLength = 24, # 设置页面默认显示的行数
 lengthMenu = list(
 c(24, 48, 72, 96, -1),
 c("24", "48", "72", "96", "All")
),
 paging = T,
 # 设置第一列和第三列的宽度 https://rstudio.github.io/DT/options.html
 autoWidth = TRUE, columnDefs = list(list(width = '400px', targets = c(1, 3)))
)))
 })
}

shinyApp(ui, server)
```

按指定格式显示数据

```
data <- data.frame(x = c(100.0011, 80.0011, -90.0011, -110.0011, -70))
#
library(shiny)
runApp(list(
ui = fluidPage(dataTableOutput("num")),
server = function(input, output) {
```



```
output$num = renderDataTable(format(round(data, 3), nsmall = 3))
}
))
```

```
library(DT)
```

```
dat <- data.frame(x = c(100.0011, 80.0011, -90.0011, -110.0067, -70))
```

```
rowCallback <- c(
 "function(row, data, index){",
 " var N = data.length;",
 " for(var j=1; j<data.length; j++){",
 " $('td:eq('+j+')',row)",
 " .html(parseFloat(data[j]).toFixed(3));", # 四舍五入保留 3 位小数
 " }",
 "}"
)
```

```
https://github.com/rstudio/shiny/issues/2277
datatable(dat,
 options = list(
 rowCallback = JS(rowCallback)
)
)
```

## 11.5 高级主题

异步编程，并发访问

```
shiny 异步编程
```

```
解决问题，多人同时访问 shiny 应用的情况下，必须等另一个人完成访问的情况下才能继
```

```
library(shiny)
library(future)
library(promises)
```



```
plan(multiprocess)

ui <- fluidPage(
 h2("测试异步下载"),
 tags$ol(
 tags$li("Verify that plot appears below"),
 tags$li("Verify that pressing Download results in 5 second delay, then rock.csv bei"),
 tags$li("Check 'Throw on download?' checkbox and verify that pressing Download resu"),
),
 hr(),
 checkboxInput("throw", "Throw on download?"),
 downloadButton("download", "下载 (等待5秒)"),
 plotOutput("plot")
)

server <- function(input, output, session) {
 output$download <- downloadHandler("rock.csv", function(file) {
 future({Sys.sleep(5)}) %...>%
 {
 if (input$throw) {
 stop("boom")
 } else {
 write.csv(rock, file)
 }
 }
 })

 output$plot <- renderPlot({
 plot(cars)
 })
}

shinyApp(ui, server)
```



## 11.6 部署应用

## 11.7 最佳实践

提升 shiny 仪表盘访问性能的 4 个建议

## 11.8 仪表盘

dashboard 翻译过来叫仪表盘，就是驾驶仓的那个玩意，形象地表达作为掌舵者应该关注的对象。R 包 shiny 出现后，仪表盘的制作显得非常容易，也很快形成了一个生态，比如 shinydashboard、flexdashboard 等，此外 bs4Dash 基于 Bootstrap 4 的仪表盘，目前 shiny 和 rmarkdown 都在向 Bootstrap 4 升级，这是未来的方向。shinydashboardPlus 主要目的在于扩展 shinydashboard 包

shinydashboard 包

```
app.R
library(shiny)
library(shinydashboard)

ui <- dashboardPage(
 dashboardHeader(title = "Basic dashboard"),
 ## Sidebar content
 dashboardSidebar(
 sidebarMenu(
 menuItem("Dashboard", tabName = "dashboard", icon = icon("dashboard")),
 menuItem("Widgets", tabName = "widgets", icon = icon("th"))
)
),
 ## Body content
 dashboardBody(
 tabItems(
 # First tab content
 tabItem(tabName = "dashboard",
 fluidRow(
 box(plotOutput("plot1", height = 250)),
 box(plotOutput("plot2", height = 250))
)
)
)
)
)
```



```
box(
 title = "Controls",
 sliderInput("slider", "Number of observations:", 1, 100, 50
)
),
)

Second tab content
tabItem(tabName = "widgets",
 h2("Widgets tab content")
)
)
)

server <- function(input, output) {
 set.seed(122)
 histdata <- rnorm(500)

 output$plot1 <- renderPlot({
 data <- histdata[seq_len(input$slider)]
 hist(data)
 })
}

shinyApp(ui, server)
```

### shinydashboardPlus 包

```
library(shiny)
library(shinydashboard)
library(shinydashboardPlus)

shinyApp(
 ui = dashboardPage(
 dashboardHeader(),
```



```
dashboardSidebar(),
dashboardBody(
 box(
 solidHeader = FALSE,
 title = "Status summary",
 background = NULL,
 width = 4,
 status = "danger",
 footer = fluidRow(
 column(
 width = 6,
 descriptionBlock(
 number = "17%",
 numberColor = "green",
 numberIcon = "fa fa-caret-up",
 header = "$35,210.43",
 text = "TOTAL REVENUE",
 rightBorder = TRUE,
 marginBottom = FALSE
)
),
 column(
 width = 6,
 descriptionBlock(
 number = "18%",
 numberColor = "red",
 numberIcon = "fa fa-caret-down",
 header = "1200",
 text = "GOAL COMPLETION",
 rightBorder = FALSE,
 marginBottom = FALSE
)
)
)
)
),
```



```
 title = "Description Blocks"
),
server = function(input, output) { }
)
```



## shinymaterial 包

```
library(shiny)
library(shinymaterial)

https://ericrayanderson.github.io/shinymaterial/
https://github.com/ericrayanderson/shinymaterial

Wrap shinymaterial apps in material_page
ui <- material_page(
 title = "用户画像",
 nav_bar_fixed = TRUE,
 # 每个 sidebar 内容
 material_side_nav(
 fixed = TRUE,
 # Place side-nav tabs within side-nav
 material_side_nav_tabs(
 side_nav_tabs = c(
 "数据汇总" = "tab_1",
 "趋势信息" = "tab_2"
),
 icons = c("cast", "insert_chart")
)
),
 # 每个 tab 页面的内容
 material_side_nav_tab_content(
 side_nav_tab_id = "tab_1",
 tags$h2("第一个tab页")
),
 material_side_nav_tab_content(
 side_nav_tab_id = "tab_2",
 tags$h2("第二个tab页")
)
)
```



```
)
)

server <- function(input, output) {
}
shinyApp(ui = ui, server = server)
```

### miniUI 包

```
library(shiny)
library(miniUI)
library(leaflet)
library(ggplot2)

ui <- miniPage(
 gadgetTitleBar("Shiny gadget example"),
 miniTabstripPanel(
 miniTabPanel("Parameters", icon = icon("sliders"),
 miniContentPanel(
 sliderInput("year", "Year", 1978, 2010, c(2000, 2010), sep = "")
)
,
 miniTabPanel("Visualize", icon = icon("area-chart"),
 miniContentPanel(
 plotOutput("cars", height = "100%")
)
,
 miniTabPanel("Map", icon = icon("map-o"),
 miniContentPanel(padding = 0,
 leafletOutput("map", height = "100%")
),
 miniButtonBlock(
 actionButton("resetMap", "Reset")
)
,
 miniTabPanel("Data", icon = icon("table"),
```



```
miniContentPanel(
 DT::dataTableOutput("table")
)
),
selected = "Map"
)
)

server <- function(input, output, session) {
 output$cars <- renderPlot({
 require(ggplot2)
 ggplot(cars, aes(speed, dist)) + geom_point()
 })

 output$map <- renderLeaflet({
 force(input$resetMap)

 leaflet(quakes, height = "100%") %>% addTiles() %>%
 addMarkers(lng = ~long, lat = ~lat)
 })

 output$table <- DT::renderDataTable({
 diamonds
 })

 observeEvent(input$done, {
 stopApp(TRUE)
 })
}

shinyApp(ui, server)
```

## 11.9 交互式数据报表 dash

```
library(dash)
library(dashHtmlComponents)
library(dashCoreComponents)
library(dashTable)
```

## 11.10 运行环境

```
sessionInfo()

R version 4.1.0 (2021-05-18)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 20.04.2 LTS
##
Matrix products: default
BLAS: /usr/lib/x86_64-linux-gnublas/libblas.so.3.9.0
LAPACK: /usr/lib/x86_64-linux-gnulapack/liblapack.so.3.9.0
##
locale:
[1] LC_CTYPE=en_US.UTF-8 LC_NUMERIC=C
[3] LC_TIME=en_US.UTF-8 LC_COLLATE=en_US.UTF-8
[5] LC_MONETARY=en_US.UTF-8 LC_MESSAGES=en_US.UTF-8
[7] LC_PAPER=en_US.UTF-8 LC_NAME=C
[9] LC_ADDRESS=C LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
attached base packages:
[1] stats graphics grDevices utils datasets methods base
##
other attached packages:
[1] data.table_1.14.0 shiny_1.6.0
##
loaded via a namespace (and not attached):
[1] Rcpp_1.0.7 bookdown_0.22 png_0.1-7 later_1.2.0
```

```
[5] digest_0.6.27 mime_0.11 R6_2.5.0 lifecycle_1.0.0
[9] xtable_1.8-4 magrittr_2.0.1 evaluate_0.14 rlang_0.4.11
[13] stringi_1.7.3 promises_1.2.0.1 ellipsis_0.3.2 rmarkdown_2.9
[17] tools_4.1.0 stringr_1.4.0 fastmap_1.1.0 httpuv_1.6.1
[21] xfun_0.24 yaml_2.2.1 compiler_4.1.0 htmltools_0.5.1.1
[25] knitr_1.33
```



## 第十二章 字符串操作

```
stringdist stringfish stringb stringi stringr

shopping_list <- c("apples x4", "bag of flour", "bag of sugar", "milk x2")

stringr::str_replace(string = shopping_list, pattern = "\\d", replacement = "aa")

[1] "apples xaa" "bag of flour" "bag of sugar" "milk xaa"
https://github.com/hadley/stringb/issues/5
x is vector
str_replace <- function(x, pattern, fun, ...) {
 loc <- gregexpr(pattern, text = x, perl = TRUE)
 matches <- regmatches(x, loc)
 out <- lapply(matches, fun, ...)

 regmatches(x, loc) <- out
 x
}

loc <- gregexpr(pattern = "\\d", text = shopping_list, perl = TRUE)

matches = regmatches(x = shopping_list, loc)

matches

out <- lapply(matches, transform, "aa")
```



```
regmatches(x = shopping_list, loc) <- out

shopping_list

str_replace(shopping_list, pattern = "\\\d", replace = "aa")
```

## 12.1 字符串加密

字符串编码加密，**openssl** 包提供了 sha1 函数<sup>1</sup>

```
library(openssl)
encode_mobile <- function(phone_number) paste("*", paste(toupper(sh1(sh1(charToRaw(pa
随意模拟两个手机号
mobile_vec <- c("18601013453", "13811674545")
sapply(mobile_vec, encode_mobile)

18601013453
"*B1D46D1D62C7280137F0E14249EE500865247B7B"
13811674545
"*0554DA6E403491F58F1567DF2EDEB19186B77173"
```

<sup>1</sup>参考刘思喆的两篇博文：[利用 R 函数生成差异化密码](#) 和 [在 R 中各种码的转换](#)

## 第十三章 正则表达式

批量转换驼峰式命名

```
old_name <- list.files(".", pattern = "^[A-Z].*.Rmd$")
new_name <- gsub("rmd", "Rmd", tolower(old_name))
file.rename(from = old_name, to = new_name)
```

```
html_lines <- readLines("https://movie.douban.com/top250")
doc <- paste0(html_lines, collapse = "")
```

```
title_lines <- grep('class="title"', html_lines, value = T)
titles <- gsub(".*>(.*)<.*", "\\\1", title_lines, perl = T)
```

```
gsub(".*>(.*)<.*", "\\\1", '肖生克的救赎', perl = T)
```

解析术之 XPath

```
library(xml2)
dom = read_html(doc)
title_nodes = xml_find_all(dom, './/span[@class="title"]')
xml_text(title_nodes)
```

解析术之 CSS Selector

```
library(rvest)
read_html(doc) %>%
html_nodes('.title') %>% # class="title"的标签
html_text()
```



## 第十四章 文本分析

PDFR 和 pdftools 从 PDF 文档抽取文本，tesseract 从扫描件中抽取文本

fastTextR <https://github.com/facebookresearch/fastText>



## 第十五章 抽样分布

分布我们已经听说过很多了，可是它们都是凭空臆测的吗？肯定不是，那它们是怎么产生的呢？谁提出了正态分布，他/她是怎么提出的？一定有故事背景，一定有数据记录，即观察值，我们的样本数据

抽样分布其中抽样二字更加贴近生活，说明它源于实际生产场景，而不是光靠大脑思维理论推导出来的东西，它是最本质的

### 15.1 正态分布

分三块介绍

- 历史背景
- 分布性质
- 应用场景

来源，为啥叫逻辑斯谛？历史故事

逻辑斯谛分布

1. 正态分布
2. t 分布
3. F 分布
4.  $\chi^2$  分布
5. 霍特林  $T^2$  分布 Hoteling's T<sup>2</sup> Distribution
6. 威沙特分布 Wishart Distribution

分一元和多元情况阐述正态分布、t 分布、F 分布、卡方分布及分布拟合

常见分布之间的关系图需要用 TikZ 来绘制



完整的关系图 <http://www.math.wm.edu/~leemis/2008amstat.pdf> 参考自 <https://www.math.wustl.edu/~jmding/math494/dist.pdf>

图来自 [Leemis, 1986]



## 15.2 指数族

谁提出的指数族，有哪些性质，指数族 quasi-poisson 是什么含义，拟族

如何判别一个分布是否属于指数族

常见的高斯、二项、正态分布、伽马分布、泊松分布

指数族

推广到一般情况

三大抽样分布 t 分布， $\chi$  分布和 F 分布，一元和多元情形，一元分布知识范围是本科，多元分布范围是研究生和博士，参考数理统计引论。一元分布多用于本科假设检验，多元分布常用于均值向量和协方差阵以及统计量的极限分布。介绍各个分布的形式、历史来源、各个特征量、密度、分布函数推导，数值计算

三大抽样的发现、历史、多元、非中心形式的推广

多元 t 分布函数 (MVT)

$$T(\mathbf{a}, \mathbf{b}, \Sigma, \nu) = \frac{2^{1-\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})} \int_0^{\infty} s^{\nu-1} e^{-\frac{s^2}{2}} \Phi\left(\frac{s\mathbf{a}}{\sqrt{\nu}}, \frac{s\mathbf{b}}{\sqrt{\nu}}, \Sigma\right) ds$$

多元正态分布函数 (MVN)

$$\Phi(\mathbf{a}, \mathbf{b}, \Sigma) = \frac{1}{\sqrt{|\Sigma|(2\pi)^m}} \int_{a_1}^{b_1} \int_{a_2}^{b_2} \cdots \int_{a_m}^{b_m} e^{-\frac{1}{2}x^\top \Sigma^{-1} x} dx$$

其中  $x = (x_1, x_2, \dots, x_m)^\top$ ,  $\forall i, -\infty \leq a_i \leq b_i \leq \infty$ ,  $\Sigma$  是  $m \times m$  对称非负定的矩阵

多元 t 分布分位数计算

```
library(mvtnorm)
n <- c(26, 24, 20, 33, 32)
```



```
V <- diag(1 / n)
df <- 130
C <- matrix(c(
 1, 1, 1, 0, 0, -1, 0, 0, 1, 0,
 0, -1, 0, 0, 1, 0, 0, 0, -1, -1,
 0, 0, -1, 0, 0
), ncol = 5)
cv <- C %*% V %*% t(C) ## covariance matrix
dv <- t(1 / sqrt(diag(cv)))
cr <- cv * (t(dv) %*% dv) ## correlation matrix
delta <- rep(0, 5)
Tn <- qmvn(0.95,
 df = df, delta = delta, corr = cr,
 abseps = 0.0001, maxpts = 100000, tail = "both"
)
Tn
```

```
$quantile
[1] 2.561154
##
$f.quantile
[1] 1.492994e-07
##
attr(,"message")
[1] "Normal Completion"
```

计算多元正态分布的概率，这个例子来自 <https://stackoverflow.com/questions/36704081>

```
模拟一个协方差矩阵
sigma <- as.matrix(read.csv(file = "data/sigma.csv", header = F, sep = ","))
rownames(sigma) <- colnames(sigma)
matrixcalc::is.symmetric.matrix(sigma) # 判断 sigma 是否为对称的矩阵
matrixcalc::is.positive.definite(sigma) # 判断 sigma 是否为正定的矩阵
isTRUE(all.equal(sigma, t(sigma)))
m <- nrow(sigma)
Fn <- pmvnorm(
 lower = rep(-Inf, m), upper = rep(0, m),
```

```
mean = rep(0, m), sigma = sigma
)
Fn
```

④ **mvrnorm()** 函数来自 **MASS** 包，模拟多元正态分布的样本

```
library(MASS)
n <- 1000 # 样本量
X <- mvrnorm(n, mu = rep(0, 2), Sigma = matrix(c(1, 0.8, 0.8, 1), ncol = 2, byrow = TRUE))
plot(X,
 pch = 20, panel.first = grid(), cex = 1,
 col = densCols(X, colramp = terrain.colors),
 xlab = expression(X[1]), ylab = expression(X[2]))
)
points(x = 0, y = 0, pch = 3, cex = 2)
```

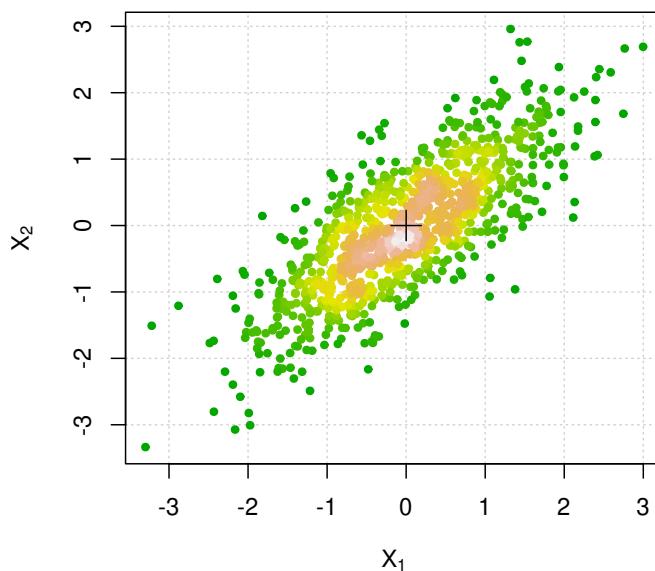
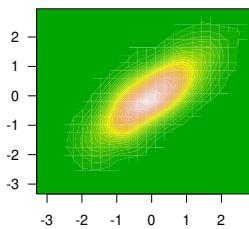


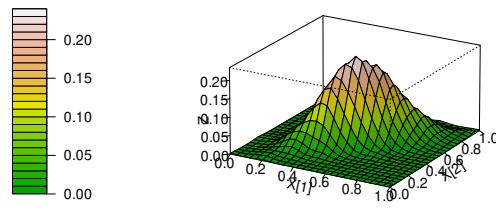
图 15.1: 二维正态分布

```
f1 <- kde2d(X[, 1], X[, 2], n = 25)
filled.contour(f1, color.palette = terrain.colors)

library(shape)
persp(f1$z,
 xlab = expression(X[1]), ylab = expression(X[2]),
 zlab = expression(Z),
 col = drapecol(f1$z, col = terrain.colors(20)),
 theta = 30, phi = 20,
 r = 50, d = 0.1, expand = 0.5, ltheta = 90, lphi = 180,
 shade = 0.1, ticktype = "detailed", nticks = 5, box = TRUE
)
```



(a) 等高线图



(b) 透视图

图 15.2: 二维正态分布

Wishart 分布文献 [Eaton, 2007] 第八章



## 第十六章 参数估计

Jeremy Koster: My students were looking at the estimated varying intercepts for each higher-level group (or the “BLUP’s”, as some people seem to call them).

Douglas Bates: As Alan James once said, “these values are just like the BLUPs - Best Linear Unbiased Predictors - except that they aren’t linear and they aren’t unbiased and there is no clear sense in which they are “best”, but other than that ...”

— Jeremy Koster and Douglas Bates<sup>1</sup>

### 16.1 点估计

- 矩估计
- 极大似然估计
- 最小二乘估计
- 同变估计
- 稳健估计

单参数和多参数模型的参数估计，比如指数分布、泊松分布、二项分布、正态分布，线性模型各个估计的推导过程

---

<sup>1</sup><https://stat.ethz.ch/pipermail/r-sig-mixed-models/2012q3/018817.html>



注意

应当考虑  $(X^\top X)^{-1}$  不存在的情况下，在均方误差最小的意义下，不必要求  $\beta$  的估计  $\hat{\beta}$  满足无偏性的要求，所以介绍岭回归估计  $\hat{\beta}_{ridge}$ 、压缩估计  $\hat{\beta}_{jse}$ 、主成分估计  $\hat{\beta}_{pca}$  和偏最小二乘估计  $\hat{\beta}_{pls}$ 。相比于  $\hat{\beta}_{pca}$ ,  $\hat{\beta}_{pls}$  考虑了响应变量的作用。《数理统计引论》第 5 章第 5 节线性估计类从改进 LS 估计出发，牺牲一部分估计的偏差，即采用有偏的估计，达到总体均方误差更小的效果 [陈希孺, 1981]

James-Stein 估计可不可以看作一种压缩估计？从它牺牲一部分偏差，获取整体方差的降低来看和上面应该有某种联系

- 昔日因，今日意 讲线性混合效应模型和很多模型之间的联系
- 那些年，我们一起追的 EB James-Stein 估计和岭回归估计的联系
- 统计学习那些事 lasso 和 boosting 之间的联系

### 16.1.1 矩估计

### 16.1.2 最小二乘估计

谈非线性最小二乘，这段话的意思是非线性模型不要谈 ANOVA 和 R^2 之类的东西

As one of the developers of the nls function I would like to state that the lack of automatic ANOVA,  $R^2$  and adj. $R^2$  from nls is a feature, not a bug :-)

— Douglas Bates<sup>2</sup>

最小二乘估计是一种非参数估计方法（对数据分布没有假设，只要预测误差达到最小即可），而极大似然估计是一种参数估计方法（观测数据服从带参数的多元分布）

非线性最小二乘估计

```
Nonlinear least-squares using nlm()
demo(nlm)

Helical Valley Function
非线性最小二乘
```

<sup>2</sup><https://stat.ethz.ch/pipermail/r-help/2000-August/007778.html>



```

theta <- function(x1, x2) (atan(x2 / x1) + (if (x1 <= 0) pi else 0)) / (2 * pi)
更加简洁的表达
theta <- function(x1, x2) atan2(x2, x1) / (2 * pi)
目标函数
f <- function(x) {
 f1 <- 10 * (x[3] - 10 * theta(x[1], x[2]))
 f2 <- 10 * (sqrt(x[1]^2 + x[2]^2) - 1)
 f3 <- x[3]
 return(f1^2 + f2^2 + f3^2)
}

explore surface {at x3 = 0}
x <- seq(-1, 2, length.out = 50)
y <- seq(-1, 1, length.out = 50)
z <- apply(as.matrix(expand.grid(x, y)), 1, function(x) f(c(x, 0)))

contour(x, y, matrix(log10(z), 50, 50))

nlm.f <- nlm(f, c(-1, 0, 0), hessian = TRUE)

points(rbind(nlm.f$estim[1:2]), col = "red", pch = 20)

```

### ### the Rosenbrock banana valley function 香蕉谷函数

```

fR <- function(x) {
 x1 <- x[1]
 x2 <- x[2]
 100 * (x2 - x1 * x1)^2 + (1 - x1)^2
}

explore surface
fx <- function(x) { ## `vectorized' version of fR()
 x1 <- x[, 1]
 x2 <- x[, 2]
 100 * (x2 - x1 * x1)^2 + (1 - x1)^2
}

```



```
x <- seq(-2, 2, length.out = 100)
y <- seq(-0.5, 1.5, length.out = 100)
z <- fx(expand.grid(x, y))
op <- par(mfrow = c(2, 1), mar = 0.1 + c(3, 3, 0, 0))
contour(x, y, matrix(log10(z), length(x)))

nlm.f2 <- nlm(fR, c(-1.2, 1), hessian = TRUE)
points(rbind(nlm.f2$estim[1:2]), col = "red", pch = 20)

Zoom in :
rect(0.9, 0.9, 1.1, 1.1, border = "orange", lwd = 2)
x <- y <- seq(0.9, 1.1, length.out = 100)
z <- fx(expand.grid(x, y))
contour(x, y, matrix(log10(z), length(x)))
mtext("zoomed in")
box(col = "orange")
points(rbind(nlm.f2$estim[1:2]), col = "red", pch = 20)
par(op)

with(
 nlm.f2,
 stopifnot(
 all.equal(estimate, c(1, 1), tol = 1e-5),
 minimum < 1e-11, abs(gradient) < 1e-6, code %in% 1:2
)
)

fg <- function(x) {
 gr <- function(x1, x2) {
 c(-400 * x1 * (x2 - x1 * x1) - 2 * (1 - x1), 200 * (x2 - x1 * x1))
 }
 x1 <- x[1]
 x2 <- x[2]
 structure(100 * (x2 - x1 * x1)^2 + (1 - x1)^2,
 gradient = gr(x1, x2)
)
}
```



```
 nfg <- nlm(fg, c(-1.2, 1), hessian = TRUE)
 str(nfg)

 with(
 nfg,
 stopifnot(
 minimum < 1e-17, all.equal(estimate, c(1, 1)),
 abs(gradient) < 1e-7, code %in% 1:2
)
)

or use deriv to find the derivatives

 fd <- deriv(~ 100 * (x2 - x1 * x1)^2 + (1 - x1)^2, c("x1", "x2"))
 fdd <- function(x1, x2) {}
 body(fdd) <- fd

 nlfdf <- nlm(function(x) fdd(x[1], x[2]), c(-1.2, 1), hessian = TRUE)
 str(nlfdf)

 with(
 nlfdf,
 stopifnot(
 minimum < 1e-17, all.equal(estimate, c(1, 1)),
 abs(gradient) < 1e-7, code %in% 1:2
)
)

 fgh <- function(x) {
 gr <- function(x1, x2) {
 c(-400 * x1 * (x2 - x1 * x1) - 2 * (1 - x1), 200 * (x2 - x1 * x1))
 }
 h <- function(x1, x2) {
 a11 <- 2 - 400 * x2 + 1200 * x1 * x1
 a21 <- -400 * x1
 matrix(c(a11, a21, a21, 200), 2, 2)
 }
 }
```



```
x1 <- x[1]
x2 <- x[2]

structure(100 * (x2 - x1 * x1)^2 + (1 - x1)^2,
 gradient = gr(x1, x2),
 hessian = h(x1, x2)
)
}

nlfgh <- nlm(fgh, c(-1.2, 1), hessian = TRUE)

str(nlfgh)

NB: This did _NOT_ converge for R version <= 3.4.0
with(
 nlfgh,
 stopifnot(
 minimum < 1e-15, # see 1.13e-17 .. slightly worse than above
 all.equal(estimate, c(1, 1), tol = 9e-9), # see 1.236e-9
 abs(gradient) < 7e-7, code %in% 1:2
)
) # g[1] = 1.3e-7
```

### 16.1.3 极大似然估计

教材简短一句话，这里面有很多信息值得发散，一个数学家提出了统计学领域极其重要的一个核心思想，他是在研究什么的时候提出了这个想法，为什么后来没有得到重视，虽然这可能有点离题，但是对于读者可能有很多别的启迪。整整 100 年以后，Fisher 又是怎么提出这一思想的呢？他做了什么使得这个思想被广泛接受和应用？

统计决策理论，任何统计推断都应该依赖损失函数，而极大似然估计未曾考虑到，这是它的局限性。Lasso 和贝叶斯先验的关系，和损失函数的关系

是最大似然估计还是极大似然估计？当然是极大似然估计，如果有人告诉你是最大似然估计那一定是假的，这两个概念归根结底是极值和最值得区别

书本定义和性质，在后续章节介绍



介绍线性模型为何引入 REML 减少偏差

极大似然估计是费舍尔提出来的

- 边际似然 Marginal Likelihood
- 条件似然 conditional likelihood
- 完全似然 complete Likelihood
- 层次似然 Hierarchical likelihood
- 部分似然 partial likelihood
- 剖面似然 Profile Likelihood
- 限制似然 Restricted Likelihood
- 惩罚/边际拟似然 (PQL/MQL) Penalized Quasi-Likelihood/Marginal Quasi-Likelihood
- 分布边际分布条件分布
- 似然边际似然条件似然
- 极大似然估计 Maximum likelihood 简称 ML
- 限制极大似然 Restricted Maximum likelihood, 简称 REML
- 惩罚拟似然 Penalized Quasi-Likelihood, 简称 PQL 和边际拟似然 Marginal Quasi-Likelihood, 简称 MQL, Profile Maximal Likelihood, 简称 PML

拟似然估计 极大似然估计 似然函数

Penalized maximum likelihood estimates are calculated using optimization methods such as the limited memory Broyden-Fletcher-Goldfarb-Shanno algorithm (L-BFGS).

BFGS 拟牛顿法和采样器 <https://bookdown.org/rdpeng/advstatcomp>

## 16.2 区间估计

### 16.2.1 正态分布

正态分布  $\mathcal{N}(\mu, \sigma^2)$ ,  $\sigma^2$  未知, 关于参数  $\mu$  的置信水平为  $1 - \alpha$  的区间估计

1. 构造统计量  $t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t(n - 1)$
2. 参数  $\mu$  的  $1 - \alpha$  置信区间为

$$\bar{x} \pm t_{1-\alpha/2}(n - 1)s/\sqrt{n}$$



其中,  $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$  是  $\sigma^2$  的无偏估计。若取  $\alpha = 0.05$ , 则置信水平  $1 - \alpha = 0.95$ 。

```
set.seed(2020) # 为了可重复, 设置随机数种子
mu_ci <- function(alpha = 0.05, n = 100, mu = 4) {
 x <- rnorm(n = n, mean = mu, sd = 1)
 x_bar <- mean(x)
 d <- qt(p = 1 - alpha / 2, df = n - 1, lower.tail = TRUE) * var(x) / sqrt(n)
 c(mu = mu, lower = x_bar - d, upper = x_bar + d)
}
重抽样 100 次, 获得 100 个置信区间
dat <- t(replicate(n = 100, mu_ci(alpha = 0.05, n = 100, mu = 4)))
dat <- transform(dat, idx = 1:100, cover = ifelse(mu >= lower & mu <= upper, TRUE,
```

真实的参数值  $\mu = 4$ , 重抽样 100 次, 覆盖真值的次数为 97 次, 覆盖概率为 0.97

```
覆盖概率
mean(dat$cover)
```

```
[1] 0.97
library(ggplot2)
ggplot() +
 geom_segment(data = dat, aes(
 x = idx, xend = idx,
 y = lower, yend = upper, color = cover
)) +
 geom_hline(yintercept = 4) +
 theme_minimal() +
 labs(x = "", y = "")
```

方差  $\sigma^2$  已知的情况下, 标准正态分布  $N(\mu, \sigma^2), \mu = 0, \sigma^2 = 1$  的参数  $\mu$  的区间估计和覆盖概率 <https://yihui.org/animation/example/conf-int/>

### 16.2.2 0-1 分布

设 0-1 分布  $B(1, p)$  的成功概率  $p = 0.95$ , 假定是抛硬币的场景, 成功概率对应正面朝上的概率为 0.95。一次实验, 重复抛 10 次, 有两次正面朝上。现在要根据这次实验结果估计成功概率  $p$  的值, 及其置信区间

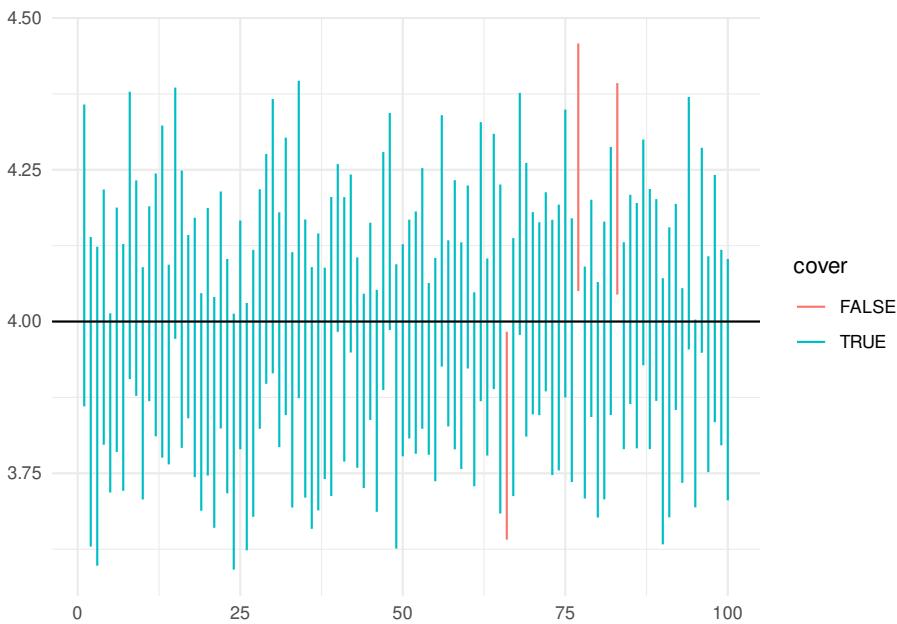


图 16.1:  $\mu$  的置信水平为 0.95 的置信区间

# 卡方近似

```
prop.test(x = 2, n = 10, p = 0.95, conf.level = 0.95, correct = TRUE)
```

```
Warning in prop.test(x = 2, n = 10, p = 0.95, conf.level = 0.95, correct =
TRUE): Chi-squared approximation may be incorrect

##
1-sample proportions test with continuity correction
##
data: 2 out of 10, null probability 0.95
X-squared = 103.16, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.95
95 percent confidence interval:
0.03542694 0.55781858
sample estimates:
p
0.2
```

```
二项精确估计
binom.test(x = 2, n = 10, p = 0.95, conf.level = 0.95)

Exact binomial test

data: 2 and 10
number of successes = 2, number of trials = 10, p-value = 1.605e-09
alternative hypothesis: true probability of success is not equal to 0.95
95 percent confidence interval:
0.02521073 0.55609546
sample estimates:
probability of success
0.2
```

可知，在置信水平都是 0.95 的情况下，带连续矫正的单样本比例检验方法获得的区间估计是 (0.0354, 0.5578)，区间长度 0.5224。精确二项检验方法获得的区间估计是 (0.0252, 0.5560)，区间长度 0.5308。

从二项分布  $B(30, 0.2)$  中随机抽取一个样本，为可重复记，设置随机数种子为 2020

```
set.seed(2020)
rbinom(1, size = 30, prob = 0.2)
```

```
[1] 7
```

得到样本观测值为 7，

```
7 - qnorm(1 - 0.95 / 2) * sqrt(0.2 * 0.8 / 30) # 6.995
```

```
[1] 6.995421
```

```
7 + qnorm(1 - 0.95 / 2) * sqrt(0.2 * 0.8 / 30) # 7.0045
```

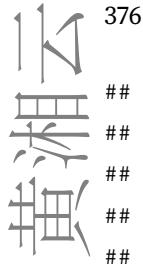
```
[1] 7.004579
```

样本观测值 7 对应的参数  $p$  的区间估计，如下

```
prop.test(x = 7, n = 30, p = 0.2, conf.level = 0.95, correct = TRUE)
```

```

1-sample proportions test with continuity correction
```



```

data: 7 out of 30, null probability 0.2
X-squared = 0.052083, df = 1, p-value = 0.8195
alternative hypothesis: true p is not equal to 0.2
95 percent confidence interval:
0.1063502 0.4270023
sample estimates:

p
0.2333333
```

随机变量  $X$  服从二项分布  $B(30, 0.2)$ , 则概率值  $P(x \leq 7) = 0.7607$

```
pbinom(7, size = 30, prob = 0.2, lower.tail = TRUE)
```

```
[1] 0.7607906
```

已知概率值为 0.95, 即  $P(x \leq m) = 0.95$  且  $X \sim B(30, 0.2)$ , 现在计算  $m$  的值, 即求下分位点, 为 10

```
qbinom(p = 0.95, size = 30, prob = 0.2, lower.tail = TRUE)
```

```
[1] 10
```

提示

二项分布的特点, 主要用于计算期望, 概率  $P\{C_1 + 1 \leq x \leq C_2 - 1\}$

$$\sum_{x=C_1+1}^{C_2-1} x \binom{n}{x} p^x (1-p)^{n-x} = np \sum_{x=C_1+1}^{C_2-1} \binom{n-1}{x-1} p^{x-1} (1-p)^{(n-1)-(x-1)}$$

```
n = 30
c2 = 20
c1 = 10
p = 0.2
n * p * (pbinom(q = c2 - 2, size = n - 1, prob = p) - pbinom(q = c1 - 1, size = n -
```

```
[1] 0.2955803
```

### 16.2.3 置信区间和信仰区间

计算置信区间的覆盖概率 `binom`



二项分布的参数估计，包括点估计和区间估计 [Clopper and Pearson, 1934]

给定样本量  $n = 10$  0-1 分布成功概率  $p$  分别取 0.1, 0.2, ..., 1 置信度为 95% 观测到  $x$  取 1, 2, 3, ..., 10 时估计  $p$  的上下限

```
set.seed(2019)
```

```
x <- rbinom(n = 1, size = 10, prob = 0.1) # 结果解读
```

抛掷硬币 10 次，观测到 2 次正面朝上，估计正面朝上的概率

观测到正面朝上 2 次此时请以 95% 的信心给出  $p$  的区间 ( $p_{\text{low}}$ ,  $p_{\text{up}}$ )

绘制曲线  $p$  关于  $x$  的曲线

```
set.seed(2019)
```

```
p <- seq(from = 0, to = 1, length.out = 11)
```

```
成功概率 总体参数 p 值
```

```
sapply(rep(p, each = 10), rbinom, n = 1, size = 10)
```

```
[1] 0 0 0 0 0 0 0 0 0 0 2 1 0 1 0 0 2 0 0 1 3 2 1 1 3
[26] 2 0 3 1 2 3 3 5 2 1 0 3 5 3 3 5 1 5 3 1 2 3 4 1 5
[51] 5 4 7 6 6 5 7 7 4 4 7 8 9 7 6 7 2 4 8 8 8 8 6 8 6
[76] 4 8 9 6 7 9 9 9 9 8 4 9 8 9 7 10 8 7 10 9 10 9 9 8 10
[101] 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10
```

计算每一次抽样获得的上下限

Clopper-Pearson 方法，即求和搜索，在保持累积概率

$$B(x, n; n, p) = \sum_{r=x}^n \binom{n}{r} p^r (1-p)^{n-r} = \alpha/2$$

其中  $n$  表示试验次数，这里是 10， $p$  是未知待求，已知  $\alpha = 0.05$ ，而  $1 - \alpha$  表示置信水平，意思是说对于我给出的区间估计，长期来看，我有 95% 的信心认为，真实值  $p$  会落在此区间内。

对上尾部从  $x$  到  $n$  求和，计算  $p$ ，对每一个  $x$  都能计算出一个  $p$ ，根据二项分布的对称性，区间  $[0, x]$  和  $[x, n]$  的累积概率是相同的，各占  $\alpha/2$

```
精确计算二项分布检验的 p
```

```
调用符号计算
```

```
x = 7
```

```
fun <- function(p, r = 8, n = 10) {
```



```
choose(n, n-2)*p^r*(1-p)^(n-r) + choose(n, n-1)*p^(n-1)*(1-p) + choose(n, n)*p^n - 0.
}
uniroot(fun, lower = 0, upper = 1)

$root
[1] 0.4439038
##
$f.root
[1] -2.707352e-07
##
$iter
[1] 9
##
$init.it
[1] NA
##
$estim.prec
[1] 6.103516e-05

x = 8
fun <- function(p) {
 45*p^8*(1-p)^2 + 10*p^9*(1-p) + p^10 - 0.025
}
uniroot(fun, lower = 0, upper = 1)

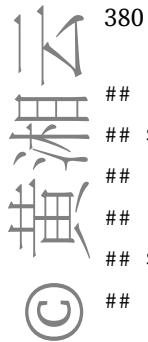
$root
[1] 0.4439038
##
$f.root
[1] -2.707352e-07
##
$iter
[1] 9
##
$init.it
[1] NA
##
$estim.prec
```

```
[1] 6.103516e-05
x = 9
fun <- function(x) {
 9 * x^10 - 10 * x^9 + 0.025
}
0.555 计算下限
uniroot(fun, lower = 0, upper = 1)

$root
[1] 0.5549828
##
$f.root
[1] 3.773379e-07
##
$iter
[1] 10
##
$init.it
[1] NA
##
$estim.prec
[1] 6.462529e-05

x = 10
fun <- function(x) {
 x^10 - 0.025
}
0.691
uniroot(fun, lower = 0, upper = 1)

$root
[1] 0.6914996
##
$f.root
[1] -1.194136e-06
##
$iter
[1] 9
```



```

$init.it
[1] NA

$estim.prec
[1] 6.103516e-05
```

累积二项概率

找到最小的 p 使得其等于 9

```
已知概率求上分位点

等于
qbinom(0.025, size = 10, prob = 0.565, lower.tail = F)

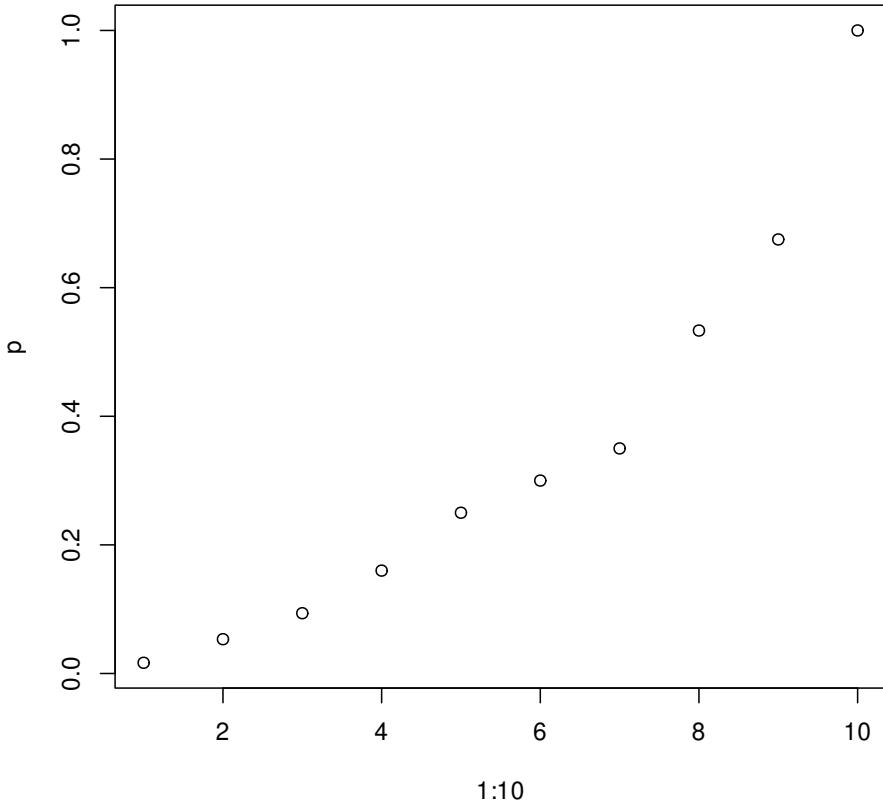
[1] 9
```

找到使得函数为 0 的 p 中最小的那个，找到所有的根，然后取最小的那个

```
fun <- function(p, r = 9) qbinom(0.025, size = 10, prob = p, lower.tail = F) - r
计算每个 x 对应的 p
(p <- sapply(1:10, function(x) uniroot(fun, lower = 0, upper = 1, r = x)$root))

[1] 0.01666667 0.05333333 0.09375000 0.16000000 0.25000000 0.30000000
[7] 0.35000000 0.53333333 0.67500000 1.00000000

plot(x = 1:10, y = p)
```



```
二项检验 菱形置信带
set.seed(2019)

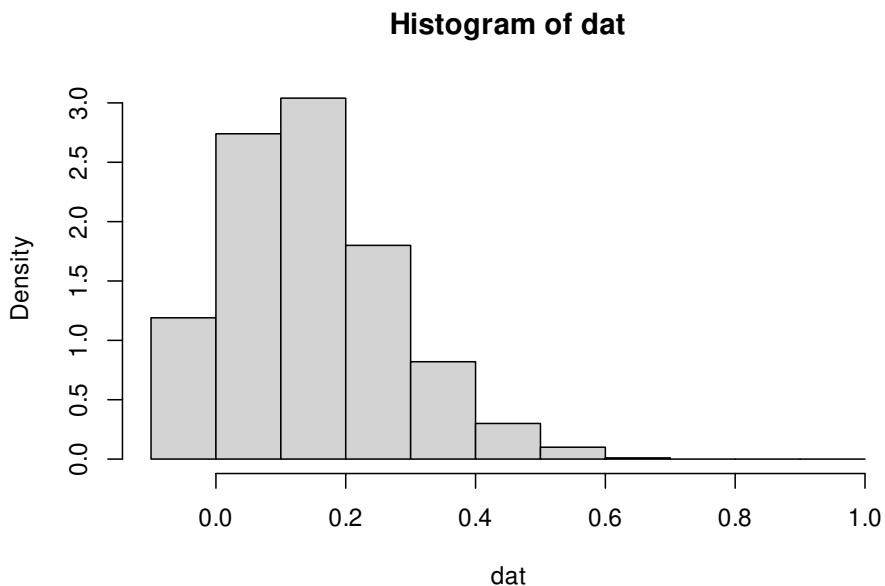
dat <- replicate(10^3, expr = {
 x = sample(0:1, size = 10, replace = TRUE, prob = c(0.8, 0.2))
 sum(x)/10
})

成功概率 p = 0.2 每个样本量 10
dat <- rbinom(n = 10^3, size = 10, prob = 0.2)/10
table(dat)

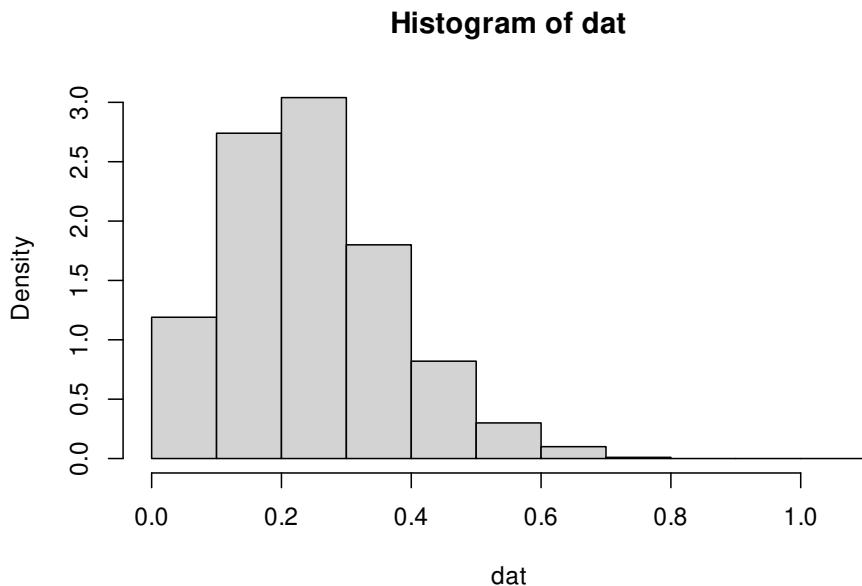
dat
```

```
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7
119 274 304 180 82 30 10 1

分布图 y 轴是密度
right = TRUE 区间形式 (a,b] 左开右闭
hist(dat, probability = T, breaks = seq(from = -0.1, to = 1, by = 0.1))
```

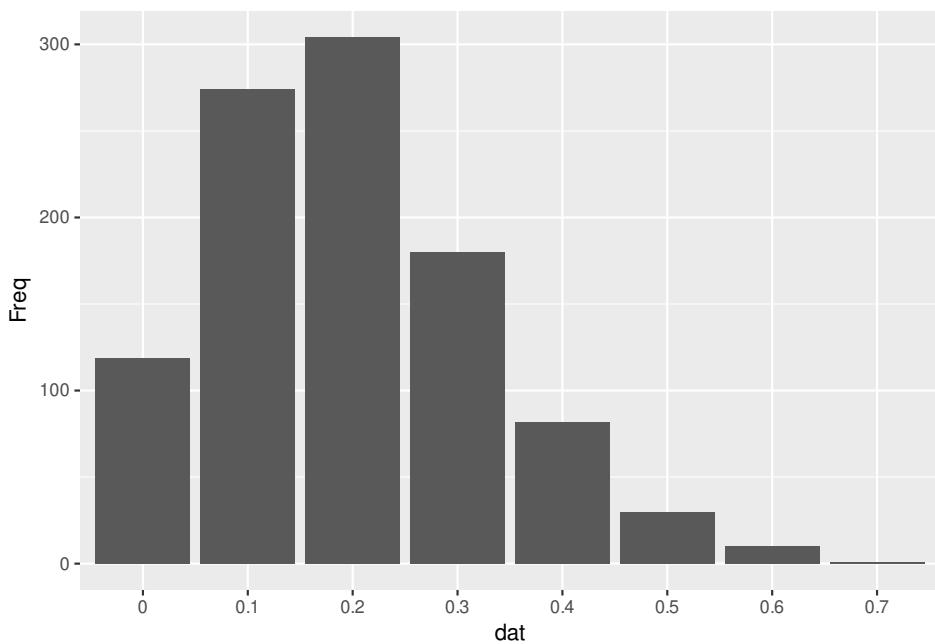


```
0.2^10 左闭右开区间
hist(dat, probability = T, breaks = seq(from = 0, to = 1.1, by = 0.1),
right = FALSE, xlim = c(0, 1.1))
```

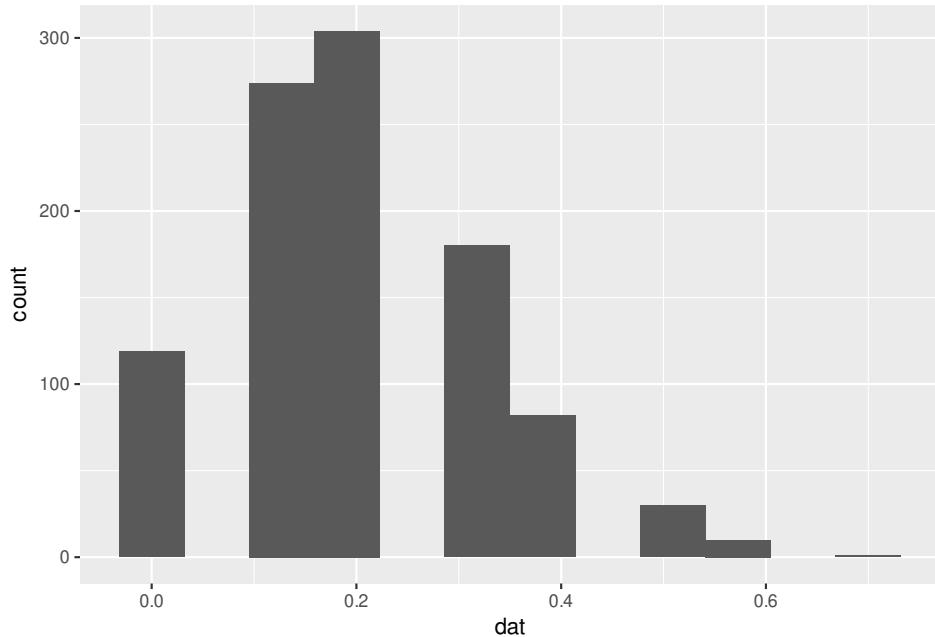


```
分布
library(ggplot2)
library(magrittr)
这个图里面会不会隐含什么信息，分布是怎样的？
二项展开有关系吗
dat1 <- as.data.frame(table(dat))

ggplot(data = dat1, aes(x = dat, y = Freq)) +
 geom_col()
```



```
ggplot(as.data.frame(dat), aes(x = dat)) +
 geom_histogram(bins = 12)
```





### 16.3 最小角回归

1. Efron, Bradley and Hastie, Trevor and Johnstone, Iain and Tibshirani, Robert. 2004. Least angle regression. *The Annals of Statistics*. 32(2): 407–499. <https://doi.org/10.1214/009053604000000067>.

方差缩减技术，修偏技术

### 16.4 刀切法

1. Efron, B. 1979. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*. 7(1):1–26. <https://doi.org/10.1214/aos/1176344552>

### 16.5 重抽样

### 16.6 Delta 方法



## 第十七章 假设检验

The Earth is Round ( $p < 0.05$ )

— Jacob Cohen [[Cohen, 1994](#)]

```
x = seq(from = -4, to = 8, length.out = 193)
y1 = dnorm(x, mean = 3, sd = 1)
y2 = dnorm(x, mean = 2, sd = 1.5)
library(magrittr)
hline <- function(y = 0, color = "red") {
 list(
 type = "line",
 x0 = 0,
 x1 = 1,
 xref = "paper",
 y0 = y,
 y1 = y,
 line = list(color = color, dash = 'dash', width = .5)
)
}

vline <- function(x = 0, color = "red") {
 list(
 type = "line",
 x0 = x,
 x1 = x,
 yref = "paper",
 y0 = 0,
 y1 = 1,
```

```
 line = list(color = color, dash = 'dash', width = .5)
)
}

plotly::plot_ly(
 x = x, y = y1,
 type = "scatter", mode = "lines",
 fill = "tozeroy", fillcolor = "rgba(92, 184, 92, 0.2)",
 text = ~ paste0(
 "x: ", x, "
",
 "y: ", round(y1, 3), "
"
),
 hoverinfo = "text",
 name = plotly::TeX("\mathcal{N}(3,1^2)"),
 line = list(shape = "spline", color = "#5CB85C")
) %>%
 plotly::add_trace(
 x = x, y = y2,
 type = "scatter", mode = "lines",
 fill = "tozeroy", fillcolor = "rgba(91, 192, 222, 0.2)",
 text = ~ paste0(
 "x: ", x, "
",
 "y: ", round(y2, 3), "
"
),
 hoverinfo = "text",
 name = plotly::TeX("\mathcal{N}(2, 1.5^2)"),
 line = list(shape = "spline", color = "#5BC0DE")
) %>%
 plotly::add_segments(
 x = 2,
 y = 0.28,
 xend = 3,
 yend = 0.28,
 line = list(color = "black"),
 showlegend = F
) %>%
```

```

plotly::add_annotations(
 x = 2.5, y = 0.3,
 showarrow = F, font = list(size = 24),
 text = plotly::TeX("d")
) %>%
plotly::add_annotations(
 x = 0, y = 1 / sqrt(2 * pi),
 font = list(size = 100), showarrow = F,
 text = plotly::TeX("\frac{1}{\sqrt{2\pi}}")
) %>%
plotly::add_annotations(
 x = 0, y = 1 / (1.5 * sqrt(2 * pi)),
 font = list(size = 100), showarrow = F,
 text = plotly::TeX("\frac{1}{1.5\sqrt{2\pi}}")
) %>%
plotly::layout(
 shapes = list(
 hline(y = 1 / sqrt(2 * pi), color = "#F27B0C"),
 hline(y = 1 / (1.5 * sqrt(2 * pi)), color = "#F27B0C"),
 vline(x = 3, color = "#F27B0C"),
 vline(x = 2, color = "#F27B0C")
),
 xaxis = list(showgrid = F, title = plotly::TeX("x")),
 yaxis = list(showgrid = F, title = plotly::TeX("f(x)")),
 legend = list(x = 0.8, y = 1, orientation = "v")
) %>%
plotly::config(displayModeBar = FALSE, mathjax = "cdn")

```

R. A. Fisher 将抽样分布、参数估计和假设检验列为统计推断的三个中心内容，可见假设检验的重要地位

呈现常见检验的公式，将手写代码和 R 内置函数计算结果进行比较，每一组原假设和备择假设要说明对应的 R 函数和及其参数设置，尽量理论和代码并重，最后结合实际的数据予以解释说明。

Jacob Cohen 实际谈的是更加深刻的问题。开篇介绍为什么需要假设检验，做检验和不做检验有什么区别？杨灿老师在[讨论帖](#)提出检验的作用和实际应用问题



有了均值和方差，为什么还要位置参数和尺度参数？为了更一般地描述问题，扩展范围。

[Summary and Analysis of Extension Program Evaluation in R](#) 介绍了各类假设检验方法

### The IQUIT R video series

假设检验，实验 A 和 B 的区分度适用于在线服务的 A/B 测试方法论 <http://www.fengjunchen.com/>

统计分布的检验

从心理学和可视化的角度谈 Cohen's d

Bootstrap 方法和置换/秩检验（Permutation Test）的入门读物

非平衡的 A/B 试验设计 Optimal unbalanced design for A/B test

Wilcoxon (WMWU) test sensitivity 检验的灵敏性

从抛硬币到 P 值和统计显著性

一分钟学会 A/B 测试

`rstatix` 包提供了一个简明的管道友好的框架，和 `tidyverse` 的设计哲学保持一致，支持常见的统计检验，如 T 检验，Wilcoxon 检验，方差分析，Kruskal-Wallis 检验，相关性分析，并将结果整理成干净的数据框形式，以方便可视化。

<https://github.com/pieces201020/AB-Test-Sample-Size-Calculator> 又一个样本量计算器

## 17.1 Ansari-Bradley 检验 `ansari.test`

Ansari-Bradley 检验目的是检验两样本的尺度参数是否有显著性差异

尺度参数可以理解为方差  $\sigma^2$

位置参数可以理解为均值  $\mu$

```
usage(ansari.test)
ansari.test(x, ...)
usage("ansari.test.default")
Default S3 method:
ansari.test(x, y, alternative = c("two.sided", "less", "greater"), exact = NULL,
```



```
conf.int = FALSE, conf.level = 0.95, ...)
usage("ansari.test.formula")
S3 method for class 'formula'
ansari.test(formula, data, subset, na.action, ...)
```



## 17.2 Bartlett 检验 **bartlett.test**

`ansari.test` 和 `mood.test` 是基于秩的两样本尺度参数显著性差异检验，是非参数检验。

**Bartlett 检验：**检验各个组的方差是否有显著性差异，即方差齐性检验。

`var.test` 和 `bartlett.test` 都属于参数检验，用于检验方差齐性问题，前者考虑正态总体下方差齐性检验，后者没有对总体的分布形式做限定。

```
usage(bartlett.test)
bartlett.test(x, ...)
usage("bartlett.test.default")
Default S3 method:
bartlett.test(x, g, ...)
usage("bartlett.test.formula")
S3 method for class 'formula'
bartlett.test(formula, data, subset, na.action, ...)
```

## 17.3 二项检验 **binom.test**

比例  $p$  的检验，做  $n$  次独立试验，样本  $X_1, \dots, X_n \sim b(1, p)$ ，事件发生的总次数  $\sum_{i=1}^n X_i$

函数 `binom.test` 用来检验伯努利试验中成功概率  $p$  和给定概率  $p_0$  的关系，属于精确检验。

编程手动实现一个，再调用函数计算，比较结果

```
模拟一组样本
x <- sample(x = c(0, 1), size = 100, replace = TRUE, prob = c(0.8, 0.2))
```

二项分布中成功概率的检验



```
binom.test(sum(x), n = 100, p = 0.5)

##
Exact binomial test
##
data: sum(x) and 100
number of successes = 26, number of trials = 100, p-value = 1.667e-06
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
0.1773944 0.3573121
sample estimates:
probability of success
0.26
```

检验成功概率  $p$  是否等于 0.5, P 值  $6.148 \times 10^{-11}$  结论是拒绝原假设

```
binom.test(sum(x), n = 100, p = 0.2)
```

```
##
Exact binomial test
##
data: sum(x) and 100
number of successes = 26, number of trials = 100, p-value = 0.1344
alternative hypothesis: true probability of success is not equal to 0.2
95 percent confidence interval:
0.1773944 0.3573121
sample estimates:
probability of success
0.26
```

检验成功概率  $p$  是否等于 0.2, P 值 0.7081 结论是不能拒绝原假设

二项检验 [Clopper and Pearson, 1934]

```
usage(binom.test)
```

```
binom.test(x, n, p = 0.5, alternative = c("two.sided", "less", "greater"),
conf.level = 0.95)
```



## 17.4 时间序列独立性检验 **Box.test**

计算 Box-Pierce 或 Ljung-Box 检验统计量来检查给定时间序列的独立性假设。

```
usage(Box.test)
```

Box.test(x, lag = 1, type = c("Box-Pierce", "Ljung-Box"), fitdf = 0)

## 17.5 皮尔逊卡方检验 **chisq.test**

用于计数数据的皮尔逊卡方检验：列联表独立性检验和拟合优度检验

chisq.test  $\chi^2$  检验：列联表检验和拟合优度检验

```
usage(chisq.test)
```

```
chisq.test(x, y = NULL, correct = TRUE, p = rep(1/length(x), length(x)),
 rescale.p = FALSE, simulate.p.value = FALSE, B = 2000)
```

## 17.6 费舍尔精确检验 **fisher.test**

固定边际的情况下，检验列联表行和列之间的独立性

```
usage(fisher.test)
```

```
fisher.test(x, y = NULL, workspace = 2e+05, hybrid = FALSE,
 hybridPars = c(expect = 5, percent = 80, Emin = 1), control = list(),
 or = 1, alternative = "two.sided", conf.int = TRUE, conf.level = 0.95,
 simulate.p.value = FALSE, B = 2000)
```

## 17.7 方差齐性检验 **fligner.test**

Fligner-Killeen (中位数) 检验各个组的样本方差是不是一致的，也是方差齐性检验

```
usage(fligner.test)
fligner.test(x, ...)
usage("fligner.test.default")
```



```
Default S3 method:
fligner.test(x, g, ...)
usage("fligner.test.formula")
S3 method for class 'formula'
fligner.test(formula, data, subset, na.action, ...)
```

## 17.8 Friedman 秩和检验 **friedman.test**

Friedman 秩和检验

Performs a Friedman rank sum test with unreplicated blocked data.

```
usage(friedman.test)
friedman.test(y, ...)
usage("friedman.test.default")
Default S3 method:
friedman.test(y, groups, blocks, ...)
usage("friedman.test.formula")
S3 method for class 'formula'
friedman.test(formula, data, subset, na.action, ...)
```

## 17.9 Kruskal-Wallis 秩和检验 **kruskal.test**

Kruskal-Wallis 秩和检验

```
usage(kruskal.test)
kruskal.test(x, ...)
usage("kruskal.test.default")
Default S3 method:
kruskal.test(x, g, ...)
usage("kruskal.test.formula")
S3 method for class 'formula'
kruskal.test(formula, data, subset, na.action, ...)
```



## 17.10 同分布检验 **ks.test**

Lilliefors 检验<sup>1</sup> 和单样本的 ks 检验的关系

As to whether you can do a **Lilliefors test** for several groups, that depends entirely on your ability to understand what the underlying question would be (see Adams D 1979).

— Knut M. Wittkowski<sup>2</sup>

Kolmogorov-Smirnov 检验：单样本或两样本的同分布检验

```
usage(ks.test)
```

```
ks.test(x, y, ..., alternative = c("two.sided", "less", "greater"),
 exact = NULL)
```

## 17.11 Cochran-Mantel-Haenszel 卡方检验

### **mantelhaen.test**

用于计数数据的 Cochran-Mantel-Haenszel 卡方检验

Performs a Cochran-Mantel-Haenszel chi-squared test of the null that two nominal variables are conditionally independent in each stratum, assuming that there is no three-way interaction.

```
usage(mantelhaen.test)
```

```
mantelhaen.test(x, y = NULL, z = NULL,
 alternative = c("two.sided", "less", "greater"), correct = TRUE,
 exact = FALSE, conf.level = 0.95)
```

## 17.12 Mauchly 球形检验 **mauchly.test**

检验：Wishart 分布的协方差矩阵是否正比于给定的矩阵

Mauchly's Test of Sphericity

<sup>1</sup><https://personal.utdallas.edu/~herve/Abdi-Lillie2007-pretty.pdf>

<sup>2</sup><https://stat.ethz.ch/pipermail/r-help/2004-February/045597.html>



Tests whether a Wishart-distributed covariance matrix (or transformation thereof) is proportional to a given matrix.

```
usage(mauchly.test)
mauchly.test(object, ...)
usage("mauchly.test.mlm")
S3 method for class 'mlm'
mauchly.test(object, ...)
usage("mauchly.test.SSD")
S3 method for class 'SSD'
mauchly.test(object, Sigma = diag(nrow = p), T = Thin.row(proj(M) - proj(X)),
M = diag(nrow = p), X = ~0, idata = data.frame(index = seq_len(p)), ...)
```

## 17.13 McNemar 卡方检验 **mcnemar.test**

两种统计量的比较参看谢益辉的博文 [渐近理想国：McNemar 检验的两种统计量](#)

用于计数数据的 McNemar's 卡方检验

McNemar's  $\chi^2$  检验：检验二维列联表行和列的对称性

```
usage(mcnemar.test)
```

```
mcnemar.test(x, y = NULL, correct = TRUE)
```

## 17.14 Mood 方差检验 **mood.test**

检验方差

Mood's 两样本检验：检验两样本尺度参数之间的差异性

```
usage(mood.test)
mood.test(x, ...)
usage("mood.test.default")
Default S3 method:
mood.test(x, y, alternative = c("two.sided", "less", "greater"), ...)
usage("mood.test.formula")
```

云  
湘  
黄  
©

## 17.15 单因素多重比较 **oneway.test**

单因素方差分析，各个组的方差不一定相同，检验两个及以上来自正态分布的样本是否有相同的均值？

```
usage(oneway.test)
```

```
oneway.test(formula, data, subset, na.action, var.equal = FALSE)
```

**## 假定方差不等**

```
oneway.test(extra ~ group, data = sleep)
```

```
##
```

```
One-way analysis of means (not assuming equal variances)
```

```
##
```

```
data: extra and group
```

```
F = 3.4626, num df = 1.000, denom df = 17.776, p-value = 0.07939
```

**## 假定方差相等**

```
oneway.test(extra ~ group, data = sleep, var.equal = TRUE)
```

```
##
```

```
One-way analysis of means
```

```
##
```

```
data: extra and group
```

```
F = 3.4626, num df = 1, denom df = 18, p-value = 0.07919
```

**## 和线性回归结果一样**

```
anova(lm(extra ~ group, data = sleep))
```

```
Analysis of Variance Table
```

```
##
```

```
Response: extra
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
--	----	--------	---------	---------	--------

## group	1	12.482	12.4820	3.4626	0.07919 .
----------	---	--------	---------	--------	-----------

## Residuals	18	64.886	3.6048		
--------------	----	--------	--------	--	--

```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

### CO2 数据

```
coplot(uptake ~ conc | Plant, data = CO2, show.given = FALSE, type = "b")
levels(CO2$Plant) # Plant 是有序的
library(ggplot2)
library(patchwork)
p1 <- ggplot(data = CO2, aes(x = conc, y = uptake)) +
 geom_point(aes(color = Treatment)) +
 geom_line(aes(color = Treatment)) +
 facet_wrap(~Plant, ncol = 4, dir = "v")
p2 <- ggplot(data = CO2, aes(x = conc, y = uptake)) +
 geom_point(aes(color = Type)) +
 geom_line(aes(color = Type)) +
 facet_wrap(~Plant, ncol = 4, dir = "v")
p1 / p2
```

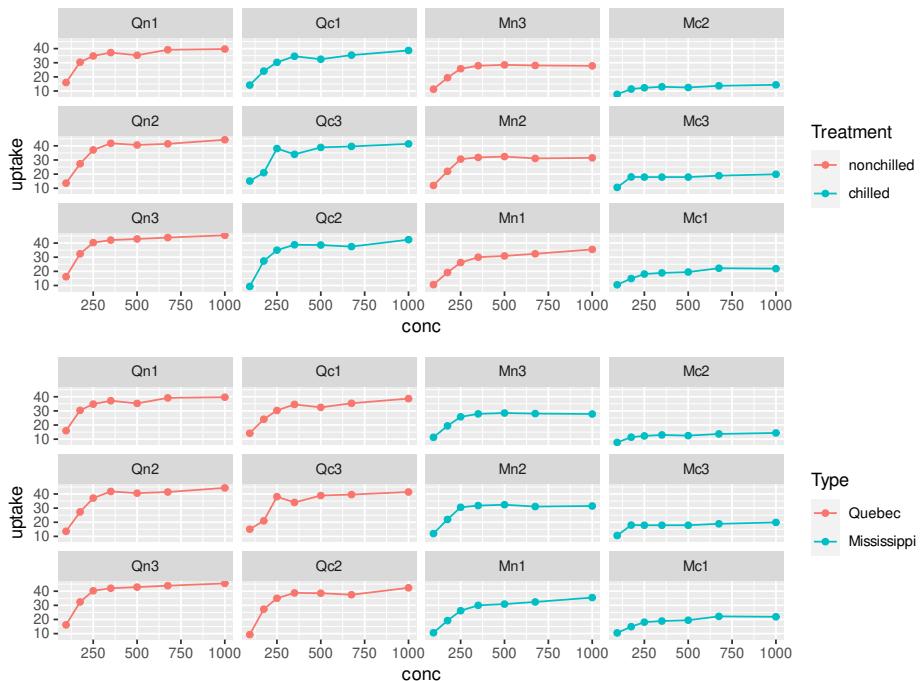


图 17.1: 草类植物吸收 CO2



## 17.16 配对样本的检验

配对样本和单样本的等价转化



### 17.16.1 配对比例检验 **pairwise.prop.test**

配对数据的比例检验

Pairwise comparisons for proportions

Calculate pairwise comparisons between pairs of proportions with correction for multiple testing

```
usage(pairwise.prop.test)
```

```
pairwise.prop.test(x, n, p.adjust.method = p.adjust.methods, ...)
```

### 17.16.2 配对 t 检验 **pairwise.t.test**

Calculate pairwise comparisons between group levels with corrections for multiple testing

```
usage(pairwise.t.test)
```

```
pairwise.t.test(x, g, p.adjust.method = p.adjust.methods, pool.sd = !paired,
 paired = FALSE, alternative = c("two.sided", "less", "greater"), ...)
```

谢益辉以配对组 t 检验谈 Cohen's d

```
pairwise.t.test(x = sleep$extra, g = sleep$group, paired = T)
```

```
##
Pairwise comparisons using paired t tests
##
data: sleep$extra and sleep$group
##
1
2 0.0028
##
P value adjustment method: holm
```



成对的 t 检验

### 17.16.3 配对 Wilcoxon 检验 pairwise.wilcox.test

Pairwise Wilcoxon Rank Sum Tests 配对的 Wilcoxon 秩和检验

Calculate pairwise comparisons between group levels with corrections for multiple testing.

```
usage(pairwise.wilcox.test)
```

```
pairwise.wilcox.test(x, g, p.adjust.method = p.adjust.methods, paired = FALSE,
...)
```

### 17.16.4 配对样本相关性检验 cor.test

配对样本的相关性检验：Pearson's 相关系数

Test for association between paired samples, using one of Pearson's product moment correlation coefficient,

Kendall's  $\tau$  检验或者 Spearman's  $\rho$  检验.

```
usage(cor.test)
```

```
cor.test(x, ...)
```

- Kendall::Kendall [McLeod, 2011]
- SuppDists::pKendall 和 SuppDists::pSpearman [Wheeler, 2020]
- pspearman::spearman.test [Savicky, 2014]

## 17.17 精确泊松检验 poisson.test

泊松分布是 1837 年由法国数学家泊松 (Poisson, 1781-1840) 首次提出  
泊松分布的参数  $\lambda (> 0)$  的精确检验

Performs an exact test of a simple null hypothesis about the rate parameter in Poisson distribution, or for the ratio between two rate parameters. 适用于单样本和两样本



```
usage(poission.test)

poission.test(x, T = 1, r = 1, alternative = c("two.sided", "less", "greater"),
conf.level = 0.95)
```



## 17.18 单位根检验 **PP.test**

时间序列平稳性检验

Phillips-Perron 的单位根检验

Computes the Phillips-Perron test for the null hypothesis that x has a unit root against a stationary alternative.

```
usage(PP.test)
```

```
PP.test(x, lshort = TRUE)
```

## 17.19 比例检验 **prop.test**

函数 **prop.test** 用来检验两组或多组二项分布的成功概率（比例）是否相等，或等于给定的值。近似检验

```
usage(prop.test)
```

```
prop.test(x, n, p = NULL, alternative = c("two.sided", "less", "greater"),
conf.level = 0.95, correct = TRUE)
```

设随机变量  $X$  服从参数为  $p$  的二项分布  $b(n, p)$ ,  $Y$  服从参数为  $\theta$  的二项分布  $b(m, \theta)$ ,  $n, m$  都假定为较大的正整数, 检验如下问题

$$H_0 : P_A \geq P_B \quad vs. \quad H_1 : P_A < P_B$$

根据中心极限定理

$$\frac{\bar{X} - \bar{Y}}{\sqrt{\frac{p(1-p)}{n} + \frac{\theta(1-\theta)}{m}}}$$



近似服从标准正态分布  $N(0, 1)$ 。如果用矩估计  $\bar{X}$  和  $\bar{Y}$  分别替代总体参数  $p$  和  $\theta$ ，构造检验统计量

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\bar{X}(1-\bar{X})}{n} + \frac{\bar{Y}(1-\bar{Y})}{m}}}$$

根据 Slutsky 定理，检验统计量  $T$  近似服从标准正态分布，当  $T$  偏大时，拒绝  $H_0$ 。该方法的优势在于当  $n, m$  比较大时，二项分布比较复杂，无法建立统计表，利用标准正态分布表来给出检验所需要的临界值，简便易行！

当  $p$  和  $\theta$  都比较小，上述方法检验效果不好，原因在于由中心极限定理对  $\bar{X}$  和  $\bar{Y}$  的正态分布近似效果不好，或者间接地导致  $\bar{X} - \bar{Y}$  的方差偏小，进而  $T$  的分辨都不好，而且当  $p, \theta$  很接近 1 时，上述现象也会产生！

下面介绍新的解决办法

上面的检验问题等价于

$$H_0 : \frac{P_A}{P_B} \geq 1 \quad vs. \quad H_1 : \frac{P_A}{P_B} < 1$$

引入检验统计量

$$T^* = \frac{\bar{X}}{\bar{Y}}$$

同样由 Slutsky 定理和中心极限定理可知， $\bar{X}/\bar{Y}$  近似服从正态分布  $N(1, \frac{1-\theta}{m\theta})$

当  $(T^* - 1)/\hat{\sigma}$  偏大时接受  $H_0$ ，临界值可通过  $N(0, \hat{\sigma}^2)$  分布表计算得到， $\hat{\sigma}^2$  是对  $\frac{1-\theta}{m\theta}$  的估计，比如取  $\hat{\sigma}^2 = \frac{1-\bar{Y}}{m} \cdot \frac{1}{\bar{Y}}$  或取  $\hat{\sigma}^2 = \frac{1-\bar{Y}}{m} \cdot \frac{1}{\bar{X}}$

由于渐近方差形如  $\frac{1-\theta}{m\theta}$ ，因而在  $\theta$  较小，渐近方差较大，克服了之前  $\bar{X} - \bar{Y}$  的方差较小的问题

$p, \theta$  很接近 1 时，我们取检验统计量

$$T^{**} = \frac{1 - \bar{Y}}{1 - \bar{X}}$$

结论和  $T^*$  类似，当  $T^{**}$  偏大时，拒绝  $H_0$ 。

两个二项总体成功概率的比较 [宋泽熙, 2011]



### 17.19.1 两个独立二项总体等价性检验

关于比例的检验问题

$$H_0 : P_A = P_B \quad vs. \quad H_1 : P_A > P_B \quad (17.1)$$

$$H_0 : P_A = P_B \quad vs. \quad H_1 : P_A < P_B \quad (17.2)$$

$H_0$  成立的情况下，暗示着两个样本来自同一总体。在这种假设设置下，拒绝原假设是不是意味着接受备择假设？如何判断样本点会落在哪个拒绝域内呢？

2009 年东南大学韦博成教授将两个独立二项总体的等价性检验应用于《红楼梦》前 80 回与后 40 回某些文风差异的统计分析 [[韦博成, 2009](#)]

### 17.19.2 不同页面的点击率问题

CTR: 点击率 Click Ratio

矩阵  $\mathbf{x}$  第一行表示页面 A 的点击情况，即 1000 次展示有 500 次点击，第二行表示页面 B 的点击情况，即 100 次展示有 80 次点击。通过统计检验的方式比较页面 A 和 B 的点击率哪个更好？

	S	F
A	500	500
B	80	20

```
(x <- matrix(c(500, 80, 500, 20), nrow = 2, ncol = 2, byrow = FALSE))
```

```
[,1] [,2]
[1,] 500 500
[2,] 80 20
等价于 prop.test(x, alternative = "two.sided", correct = TRUE)
prop.test(x) # 默认参数设置情形是双边检验
```

```
##
2-sample test for equality of proportions with continuity correction
##
```

```
data: x
X-squared = 31.632, df = 1, p-value = 1.863e-08
alternative hypothesis: two.sided
95 percent confidence interval:
-0.3898012 -0.2101988
sample estimates:
prop 1 prop 2
0.5 0.8
```

默认的假设检验问题

$$H_0 : P_A = P_B \quad vs. \quad H_1 : P_A \neq P_B$$

输出结果中 `alternative hypothesis` 表示备择假设，参数 `alternative` 指定备择假设的形式

备择假设  $P_A < P_B$  对应

```
prop.test(x, alternative = "less")

##
2-sample test for equality of proportions with continuity correction
##
data: x
X-squared = 31.632, df = 1, p-value = 9.315e-09
alternative hypothesis: less
95 percent confidence interval:
-1.0000000 -0.2237522
sample estimates:
prop 1 prop 2
0.5 0.8
```

$P$  值  $9.315 \times 10^{-9}$  结论是拒绝原假设，并且接受备择假设，即  $P_A < P_B$ ，在原假设成立的情况下，样本落入拒绝域的概率很小，小于 0.05，即在一次实验中，样本不可能落入拒绝域，应当接受原假设，因为将备择假设设为

备择假设  $P_A > P_B$

```
prop.test(x, alternative = "greater")
```

```
##
```



```
2-sample test for equality of proportions with continuity correction
##
data: x
X-squared = 31.632, df = 1, p-value = 1
alternative hypothesis: greater
95 percent confidence interval:
-0.3762478 1.0000000
sample estimates:
prop 1 prop 2
0.5 0.8
```

P 值为 1 不能拒绝原假设，在原假设成立的情况下，样本落入拒绝域的概率是 1

备择假设和原假设在这里是对立的关系

页面 A 观测到的点击率为 50% 页面 B 观测到的点击率为 80%，设置检验问题

$$H_0 : P_A = P_B \quad vs. \quad H_1 : P_A \leq P_B$$

页面点击率 A 等于 B，则备择假设页面点击率 A 不大于 B

默认启用 Yates' 连续性校正 (continuity correction, 简称 CC)

### 17.19.3 比例齐性检验

原假设四个组里面病人中吸烟的比例是相同的，备择假设是四个组里面至少有一个组的吸烟比例是不同的

```
Data from Fleiss (1981), p. 139.
H0: The null hypothesis is that the four populations from which
the patients were drawn have the same true proportion of smokers.
A: The alternative is that this proportion is different in at
least one of the populations.

smokers <- c(83, 90, 129, 70)
patients <- c(86, 93, 136, 82)
prop.test(smokers, patients)

##
```



```
4-sample test for equality of proportions without continuity
correction
##
data: smokers out of patients
X-squared = 12.6, df = 3, p-value = 0.005585
alternative hypothesis: two.sided
sample estimates:
prop 1 prop 2 prop 3 prop 4
0.9651163 0.9677419 0.9485294 0.8536585
```

Wilson 检验统计量 [Wilson, 1927] 考虑单样本比例  $p$  的区间估计问题，

Probable Inference (Usual): 可能的推断，或然推断，概率推断

在某个总体中抽取  $n$  个样本，观测到某个比率/频率  $p_0$ ，相应的标准差  $\sigma_0 = (p_0 q_0 / n)^{1/2}$ ，常见的概率推断表述是说：比率  $p$  的真值落在区间  $[p_0 - \lambda \sigma_0, p_0 + \lambda \sigma_0]$  外的概率小于等于  $P_\lambda$ ，并且随着  $\lambda$  增大， $P_\lambda$  减小。

如果使用 Tchebysheff 切比雪夫准则，我们知道  $P_\lambda$  本身小于  $1/\lambda^2$ ，但是如果使用概率表  $P_\lambda$  是概率密度曲线与坐标  $\pm \lambda \sigma_0$  之外的部分围成的面积。尽管切比雪夫准则在估计  $P_\lambda$  的时候过于保守，但是概率表给出了一个本质的估计。

严格来说，上面给出的概率推断的表述是简略的。真实概率  $p$  落在指定范围之外的机会要么是 0 要么是 1，就是说  $p$  要么在那个范围要么不在那个范围。观测的比率  $p_0$  有更大或更小的机会落在真实比率  $p$  的某个区间。观测者运气不好，观测到一个相对罕见的事件发生了，基于已有的推断理论，他会获得一个相当宽的标记。

Probable Inference (Improved):

一个更好的方式来阐述推理过程：

有某个比率  $p$  它的标准差是  $(pq/n)^{1/2} = \sigma$ ，一个观测糟糕如  $p_0$  发生的可能性，即  $p_0$  落在区间  $[p - \lambda \sigma, p + \lambda \sigma]$  是小于等于  $P_\lambda$ 。

这个表述强调了特殊观测相对于一般典型情况更容易犯的错误。

两样本比例  $p$  的检验问题。

思路需要推导，考虑如下检验问题

$$H_0 : P_A \geq P_B \quad vs. \quad H_1 : P_A < P_B$$



比例检验，未知  $p$  的情况下，且样本量有限，是 t 分布多种二项检验的办法 [Newcombe, 1998]

提示

切比雪夫不等式 Chebyshev, 1821-1894

设随机变量  $X$  的数学期望和方差都存在，则对任意常数  $\epsilon > 0$ ，有

$$P(|X - EX| \geq \epsilon) \leq \frac{Var(X)}{\epsilon^2} \quad (17.3)$$

$$P(|X - EX| \leq \epsilon) \geq 1 - \frac{Var(X)}{\epsilon^2} \quad (17.4)$$

## 17.20 比例趋势检验 `prop.trend.test`

Performs  $\chi^2$  test for trend in proportions, i.e., a test asymptotically optimal for local alternatives where the log odds vary in proportion with score. By default, score is chosen as the group numbers.

```
usage(prop.trend.test)
```

```
prop.trend.test(x, n, score = seq_along(x))
```

## 17.21 Quade 检验 `quade.test`

Quade Test

Performs a Quade test with unreplicated blocked data.

```
usage(quade.test)
quade.test(y, ...)
usage("quade.test.default")
Default S3 method:
quade.test(y, groups, blocks, ...)
usage("quade.test.formula")
S3 method for class 'formula'
quade.test(formula, data, subset, na.action, ...)
```



## 17.22 正态性检验 shapiro.test

Usually (but not always) doing tests of normality reflect a lack of understanding of the power of rank tests, and an assumption of high power for the tests (qq plots don't always help with that because of their subjectivity). When possible it's good to choose a robust method. Also, doing pre-testing for normality can affect the type I error of the overall analysis.

— Frank Harrell<sup>3</sup>

检验：拒绝原假设和接受原假设的风险，数据本身和理论的正态分布的距离，抛开 P 值

Shapiro 和 Wilk's 提出的 W 检验

Performs the Shapiro-Wilk test of normality.

```
usage(shapiro.test)
```

```
shapiro.test(x)
```

## 17.23 正态性检验 Epps-Pully 检验

The issue really comes down to the fact that the questions: “exactly normal?”, and “normal enough?” are 2 very different questions (with the difference becoming greater with increased sample size) and while the first is the easier to answer, the second is generally the more useful one.

— Greg Snow<sup>4</sup>

EP 检验对多种备择假设有较高的效率，利用样本的特征函数和正态分布的特征函数的差的模的平方产生的一个加权积分得到 EP 检验统计量 [Epps and Pulley, 1983]

提示

样本量  $n \geq 200$  EP 检验统计量  $T_{EP}$  非常接近  $n = \infty$  时  $T_{EP}$  的分位数。

<sup>3</sup><https://stat.ethz.ch/pipermail/r-help/2005-April/070508.html>

<sup>4</sup><https://stat.ethz.ch/pipermail/r-help/2009-May/390164.html>



设  $x_1, \dots, x_n$  是来自正态总体  $N(\mu, \sigma^2)$  的样本, EP 检验统计量定义为

$$T_{EP} = 1 + \frac{n}{\sqrt{3}} + \frac{2}{n} \sum_{i=2}^n \sum_{j=1}^{i-1} \exp \left\{ -\frac{(x_j - x_i)^2}{2s_*^2} \right\} - \sqrt{2} \sum_{i=1}^n \exp \left\{ -\frac{(x_i - \bar{x})^2}{4s_*^2} \right\}$$



其中  $\bar{x}, s_*^2$  就是样本均值和 (除以  $n$  的) 样本方差

提示

几个正态性检验的功效比较 <https://arxiv.org/ftp/arxiv/papers/1605/1605.06293.pdf> 和 PoweR 包 [Lafaye de Micheaux and Tran, 2016]

## 17.24 学生 t 检验 `t.test`

t 分布的推导、t 分布的形式两样本的均值检验到 Behrens-Fisher 问题  
到大规模推荐系统中的 A/B 检验

### 17.24.1 正态总体两样本的均值之差的检验

常见检验问题

$$\text{I } H_0 : \mu_1 - \mu_2 \leq 0 \quad vs. \quad H_1 : \mu_1 - \mu_2 > 0 \tag{17.5}$$

$$\text{II } H_0 : \mu_1 - \mu_2 \geq 0 \quad vs. \quad H_1 : \mu_1 - \mu_2 < 0 \tag{17.6}$$

$$\text{III } H_0 : \mu_1 - \mu_2 = 0 \quad vs. \quad H_1 : \mu_1 - \mu_2 \neq 0 \tag{17.7}$$

#### 17.24.1.1 方差 $\sigma_1^2, \sigma_2^2$ 已知

检验统计量服从标准正态分布

```
set.seed(2019)
x1 <- rnorm(100, mean = 10, sd = 2.5)
y1 <- rnorm(80, mean = 6, sd = 4.5)
u0 <- (mean(x1) - mean(y1)) / sqrt(2.5^2 / 100 + 4.5^2 / 80)
```



$$u = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}}$$

$u \sim N(0, 1)$ , 检验统计量  $u$  对应的样本值  $u_0$ , 检验的拒绝域和  $P$  值如下

$$W_1 = \{u \geq u_{1-\alpha}\}, \quad p_1 = 1 - \Phi(u_0)$$

对检验问题 I, 给定显著性水平  $\alpha = 0.05$ , 得出拒绝域  $\{u \geq 1.645\}$ , 计算样本观察值得到的检验统计量的值  $u_0 = 7.946$ , 而该值落在拒绝域, 所以拒绝原假设, 即拒绝  $\mu_1 - \mu_2 \leq 0$ , 则接受  $\mu_1 - \mu_2 > 0$ 。

# 计算拒绝域

```
qnorm(1 - 0.05)
```

```
[1] 1.644854
```

# 计算 P 值

```
1 - pnorm(u0)
```

```
[1] 9.992007e-16
```

### 17.24.1.2 方差 $\sigma_1^2 = \sigma_2^2 = \sigma^2$ 未知

检验统计量服从自由度为  $m + n - 2$  的 t 分布

```
set.seed(2019)
x1 <- rnorm(100, mean = 10, sd = 4.5)
y1 <- rnorm(80, mean = 6, sd = 4.5)
s_w <- sqrt(1 / (100 + 80 - 2) * ((100 - 1) * var(x1) + (80 - 1) * var(y1)))
t0 <- (mean(x1) - mean(y1)) / (s_w * sqrt(1 / 100 + 1 / 80))
```

样本观察值  $t_0 = 6.6816 > t_{0.95}(100 + 80 - 2) = 1.653$  落在拒绝域内, 对于检验问题 I 我们要拒绝原假设

# 临界值: 0.95 分位点对应的分位数

```
qt(1 - 0.05, df = 100 + 80 - 2)
```

```
[1] 1.653459
```



```
p 值
1 - pt(t0, df = 100 + 80 - 2, lower.tail = TRUE)
```

## [1] 1.461666e-10

利用 R 内置的 t.test() 函数计算

```
t.test(x = x1, y = y1, alternative = "greater", var.equal = TRUE)
```

```
##
Two Sample t-test
##
data: x1 and y1
t = 6.6816, df = 178, p-value = 1.462e-10
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
3.249227 Inf
sample estimates:
mean of x mean of y
9.669997 5.352296
```

与线性回归比较

```
dat <- data.frame(
 value = c(x1, y1),
 group = c(rep("x1", length(x1)), rep("y1", length(y1)))
)
fit <- lm(value ~ 1 + I(group == "y1"), data = dat)
fit <- lm(value ~ 0 + I(group == "y1"), data = dat) # 无截距项
summary(fit)

##
Call:
lm(formula = value ~ 1 + I(group == "y1"), data = dat)
##
Residuals:
Min 1Q Median 3Q Max
-11.2282 -3.0198 -0.2959 3.0161 12.1921
##
Coefficients:
```



```
Estimate Std. Error t value Pr(>|t|)
(Intercept) 9.6700 0.4308 22.446 < 2e-16 ***
I(group == "y1")TRUE -4.3177 0.6462 -6.682 2.92e-10 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.308 on 178 degrees of freedom
Multiple R-squared: 0.2005, Adjusted R-squared: 0.196
F-statistic: 44.64 on 1 and 178 DF, p-value: 2.923e-10
```

注意

lm 回归和 t 检验的差别，回归系数第二行，t 统计量为 -6.682，P 值为 2.92e-10，前者是因为截距项，后者是因为双边检验（模型系数显著性检验是和 0 比较），所以有 2 倍的关系。直观解释详见 [翻译：常见统计检验的本质都是线性模型（或：如何教统计学）](#)

两样本方差不齐、样本量严重不等，在大样本和小样本情况下的比较，[t 检验方差不齐有多重要](#)

#### 17.24.1.3 方差 $\sigma_1^2/\sigma_2^2$ 已知

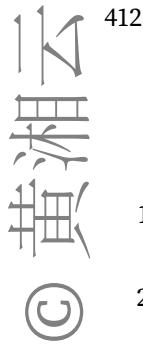
方差比  $c = \sigma_1^2/\sigma_2^2$  已知

#### 17.24.1.4 方差 $\sigma_1^2/\sigma_2^2$ 未知

英国统计学家 William Sealy Gosset (1876-1937) 于 1908 年在杂志《Biometrics》上以笔名 Student 发表论文《The probable error of a mean》["Student", 1908]，论文中展示了独立同正态分布的样本  $x_1, \dots, x_n \stackrel{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$  的样本方差  $s^2$  和样本标准差  $s$  的抽样分布，根据均值和标准差不相关的性质导出 t 分布，宣告 t 分布的诞生，因其在小样本领域的突出贡献，W. S. Gosset 进入世纪名人录 [Heyde et al., 2001]

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$$

$$\frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t(n-1)$$



$$E(s^2) = \sigma^2, \quad Var(s^2) = \frac{2\sigma^4}{n-1}$$

1. 两样本的样本量很大，总体方差未知，检验两样本均值的显著性检验，极限分布是正态， $u$  检验
2. 两个样本的样本量不是很大，总体方差也未知，检验两样本均值的显著性检验，即著名的 Behrens-Fisher 问题，Welsh 在 1938 年提出近似服从自由度为  $\ell$  的  $t$  分布。

Egon Pearson 接过他父亲 Karl Pearson 的职位，担任伦敦大学学院的高爾頓统计教授

许宝驥在 Jerzy Neyman 和 Egon Pearson 主编的杂志《Statistical Research Memoirs》发表第一篇关于 Behrens-Fisher 问题的论文

这里提及许宝驥 (Pao-Lu Hsu) 的贡献 [HSU, 1938]，

陈家鼎和郑忠国一起整理了许宝驥的生平事迹和学术成就，见《许宝驥先生的生平和学术成就》。

1998 年关于 Behrens-Fisher 问题的综述 [Kim and Cohen, 1998]

钟开涞 (Kai-Lai Chung) 将许宝驥的论文集整理出版 [HSU, 1983]

`t.test()` 提供单样本和两样本的检验

```
usage(t.test)
S3 method for class 'test'
t(x, ...)
usage("t.test.default")
Default S3 method:
t.test(x, y = NULL, alternative = c("two.sided", "less", "greater"), mu = 0,
 paired = FALSE, var.equal = FALSE, conf.level = 0.95, ...)
usage("t.test.formula")
S3 method for class 'formula'
t.test(formula, data, subset, na.action, ...)
```

学生睡眠数据 `sleep` 见图 17.2

两个样本的 Welch's  $t$  检验，总体方差未知，样本量也不大，两样本均值差的显著性检验

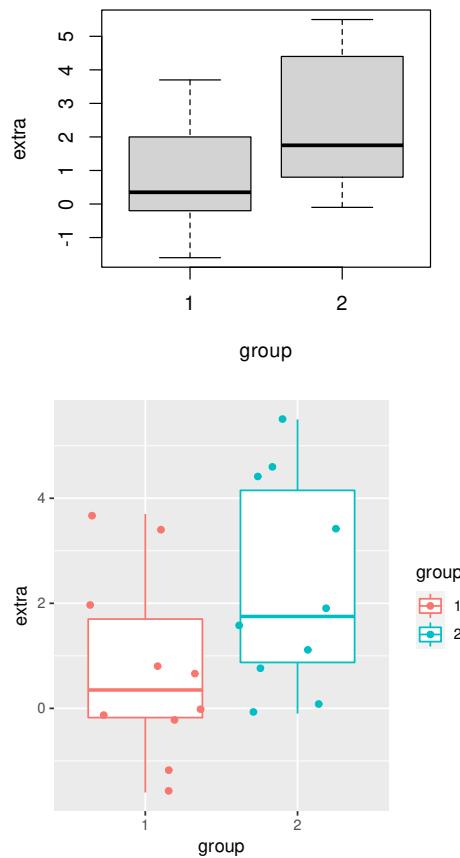


图 17.2: 学生睡眠数据 sleep



```
等价于 with(sleep, t.test(extra[group == 1], extra[group == 2]))
t.test(extra ~ group, data = sleep)
```

##  
## Welch Two Sample t-test  
##

```
data: extra by group
t = -1.8608, df = 17.776, p-value = 0.07939
alternative hypothesis: true difference in means between group 1 and group 2 is not equal to zero
95 percent confidence interval:
-3.3654832 0.2054832
sample estimates:
mean in group 1 mean in group 2
0.75 2.33
```

实际上睡眠数据是配对的，我们可以做配对数据的检验

```
数据变形操作，长格式变为宽格式
sleep2 <- reshape(sleep,
 direction = "wide",
 idvar = "ID", timevar = "group"
)
R 4.0.0
t.test(Pair(extra.1, extra.2) ~ 1, data = sleep2)
```

```

Paired t-test

data: Pair(extra.1, extra.2)
t = -4.0621, df = 9, p-value = 0.002833
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-2.4598858 -0.7001142
sample estimates:
mean of the differences
-1.58
```

注意

函数 `t.test()` 和 `wilcox.test()` 的公式接口要求 R 版本在 4.0.0 及以上

### 17.24.2 办公软件里的 T 检验

以 MacOS 上的 Numbers 表格软件为例, 如图17.3所示, 首先打开 Numbers 软件, 新建工作表, 输入两组数值, 然后点击空白处, 再从顶部导航栏找到「插入」菜单, 「公式」选项, 点击扩展选项「新建公式」, 在弹出的会话条里输入 `TTEST`, 依次选择第一组, 第二组值, 检验类型和样本类型, 最后点击确认, 即可得到两样本 T 检验的 P 值结果。

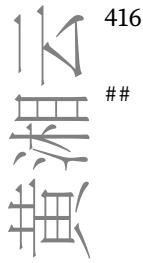


图 17.3: MacOS 的办公软件 Numbers 做两样本 T 检验

微软 Excel 办公软件也提供 T 检验计算器, 和 MacOS 系统上的 Numbers 办公软件类似, 它提供 `T.TEST` 函数, 计算结果也一样, 此处从略。R 软件自带 `t.test()` 函数, 也是用于做 T 检验, 如下:

```
t.test(x = c(3, 4, 5, 8, 9, 1, 2, 4, 5), y = c(6, 19, 3, 2, 14, 4, 5, 17, 1))

##
Welch Two Sample t-test
##
data: c(3, 4, 5, 8, 9, 1, 2, 4, 5) and c(6, 19, 3, 2, 14, 4, 5, 17, 1)
t = -1.3622, df = 10.255, p-value = 0.2023
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-8.767183 2.100516
sample estimates:
mean of x mean of y
```



```
4.555556 7.888889
```

## 17.25 方差比检验 **var.test**



**TeachingDemos** 的 `sigma.test()` 方差检验，适用于正态总体，它对非正态性很敏感。

F 检验：来自正态总体的两个样本的方差比较

```
usage(var.test)
var.test(x, ...)
usage("var.test.default")
Default S3 method:
var.test(x, y, ratio = 1, alternative = c("two.sided", "less", "greater"),
 conf.level = 0.95, ...)
usage("var.test.formula")
S3 method for class 'formula'
var.test(formula, data, subset, na.action, ...)
```

## 17.26 Wilcoxon 秩和检验 **wilcox.test**

单样本 Wilcoxon 秩和检验，两样本 Wilcoxon 符号秩检验，也叫 Mann-Whitney 检验

Wilcoxon Rank Sum and Signed Rank Tests

Performs one- and two-sample Wilcoxon tests on vectors of data; the latter is also known as ‘Mann-Whitney’ test.

```
usage(wilcox.test)
wilcox.test(x, ...)
usage("wilcox.test.default")
Default S3 method:
wilcox.test(x, y = NULL, alternative = c("two.sided", "less", "greater"),
 mu = 0, paired = FALSE, exact = NULL, correct = TRUE, conf.int = FALSE,
 conf.level = 0.95, tol.root = 1e-04, digits.rank = Inf, ...)
usage("wilcox.test.formula")
```



```
S3 method for class 'formula'
wilcox.test(formula, data, subset, na.action, ...)

• coin::wilcox_test for exact, asymptotic and Monte Carlo conditional p-values, including in the presence of ties.
```

coin 包 [Hothorn et al., 2008] 提供大量基于秩的检验

### 17.26.1 ROC 曲线和 wilcox.test 检验的关系

<https://github.com/xrobin/pROC/wiki/FAQ---Frequently-asked-questions#can-i-test-if-a-single-roc-curve-is-significantly-different-from-0.5>

ROC 曲线越往左上角拱越好，AUC 是 ROC 曲线下的面积，所以 AUC 指标越接近 1 越好。

对每个标签的预测概率指定服从均匀分布，相当于随机猜测，所以最后 ROC 会接近对角线，而且样本量越大越接近，AUC 会越来越接近 0.5

再往深一点就是研究一下 R 内置的排序算法，因为计算 AUC 最核心的步骤是排序。`order` 函数默认的排序方法是 `auto` 即当数据量较小的时候，自动选择 `radix` 排序，当数据量比较大的时候，自动选择 `shell` 排序<sup>5</sup>

```
模拟一些数据
set.seed(2019) # 设置随机数种子
N <- 10^5 # 样本量
sim_dat <- cbind.data.frame(
 pred = runif(N),
 label = rbinom(N, size = 1, prob = 0.95)
)

计算 auc 的函数
dat is a data.frame as input return AUC value
comp_auc <- function(dat, show_roc = TRUE) {
 # order label by predicted probability
 dat <- dat[order(dat$pred, dat$label, decreasing = TRUE),]
```

<sup>5</sup>radix 排序翻译过来叫桶排序或基数排序，详细描述见 ?sort



```

total samples
n_total <- length(dat$label)

number of positive label 1
n_pos <- sum(dat$label)

number of negative label 0
n_neg <- n_total - n_pos

calculate TPR and FPR
tpr <- cumsum(dat$label) / n_pos
fpr <- (1:n_total - cumsum(dat$label)) / n_neg

calculate auc
auc <- 0
for (i in 1:(n_total - 1)) {
 auc <- auc + (fpr[i + 1] - fpr[i]) * tpr[i]
}
show ROC curve or not?
if (show_roc) {
 plot(fpr, tpr, type = "l")
}
auc
}

comp_auc(dat = sim_dat, show_roc = FALSE)

```

## [1] 0.5015558

模拟一个逻辑回归模型测试自编 AUC 计算程序和 R 包 pROC 计算结果

```

set.seed(2018)
N <- 10^4 # 样本量
x <- rnorm(N)
beta_0 <- 0.5
beta_1 <- 0.3
eta <- beta_0 + beta_1 * x
模拟数据集

```



```
dat <- data.frame(x = x, y = rbinom(N, 1, prob = exp(eta) / (1 + exp(eta))))

数据集分隔
is_train <- sample(1:nrow(dat), N * 0.7)
train <- dat[is_train,]
test <- dat[-is_train,]

模型拟合
fit <- glm(y ~ x, data = train, family = binomial(link = "logit"))

预测
y_pred <- predict(fit, newdata = test, type = "response")

dat2 <- data.frame(pred = y_pred, label = test$y)

计算 auc
comp_auc(dat = dat2, show_roc = FALSE)
```

## [1] 0.5850287

对比 R 包 pROC 的计算结果是一致的

```
pROC::auc(test$y, y_pred)
```

计算一下运行时间

```
100 万样本
system.time(comp_auc(dat = dat2, show_roc = FALSE))

user system elapsed
0.001 0.000 0.002
```

更多关于 auc 计算的讨论见统计之都论坛帖 <https://d.cosx.org/d/419436>，我感觉这个问题最后会归结到排序问题。

```
https://stat.ethz.ch/pipermail/r-help/2005-April/069217.html
trap.rule <- function(x, y) sum(diff(x) * (y[-1] + y[-length(y)])) / 2
```

## 17.27 3 + 1 统计检验

Wald 检验，似然比检验/ Wilks 检验，得分检验/Rao 检验，梯度检验

Unfortunately, this is one of those situations where as far as I can tell

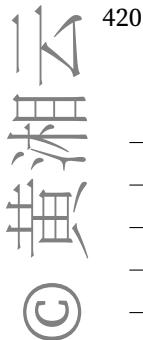


表 17.2: 伯克利大学各个院系的录取人数

Admit	Gender	DeptA	DeptB	DeptC	DeptD	DeptE	DeptF
Admitted	Male	512	353	120	138	53	22
Rejected	Male	313	207	205	279	138	351
Admitted	Female	89	17	202	131	94	24
Rejected	Female	19	8	391	244	299	317

all of the real statisticians are out there playing with large data sets where the small-sample corrections are not so important and leaving the rest of us to figure it out for ourselves ...

— Ben Bolker<sup>6</sup>

## 17.28 经典案例

### 17.28.1 1973 年加州大学伯克利分校的学生招生

录取人数按院系和性别分类统计，研究目标是各个院系在录取学生的时候是否有性别歧视？统计数据见表 17.2

```
as.data.frame(UCBAdmissions) %>%
 reshape(.,
 v.names = "Freq", idvar = c("Admit", "Gender"),
 timevar = "Dept", direction = "wide", sep = "")
) %>%
 knitr::kable(.,
 caption = "伯克利大学各个院系的录取人数",
 row.names = FALSE, col.names = gsub("(Freq)", "Dept", names(.)),
 align = "c"
)

plot(UCBAdmissions, col = "lightblue", border = "white")
library(ggmosaic)
ggplot(data = as.data.frame(UCBAdmissions)) +
 geom_mosaic(aes(weight = Freq, x = product(Gender, Admit), fill = Dept)) +
```

<sup>6</sup><https://stat.ethz.ch/pipermail/r-sig-mixed-models/2011q4/017392.html>

```
coord_flip() +
theme_minimal() +
labs(x = "Admit", y = "Gender")
```

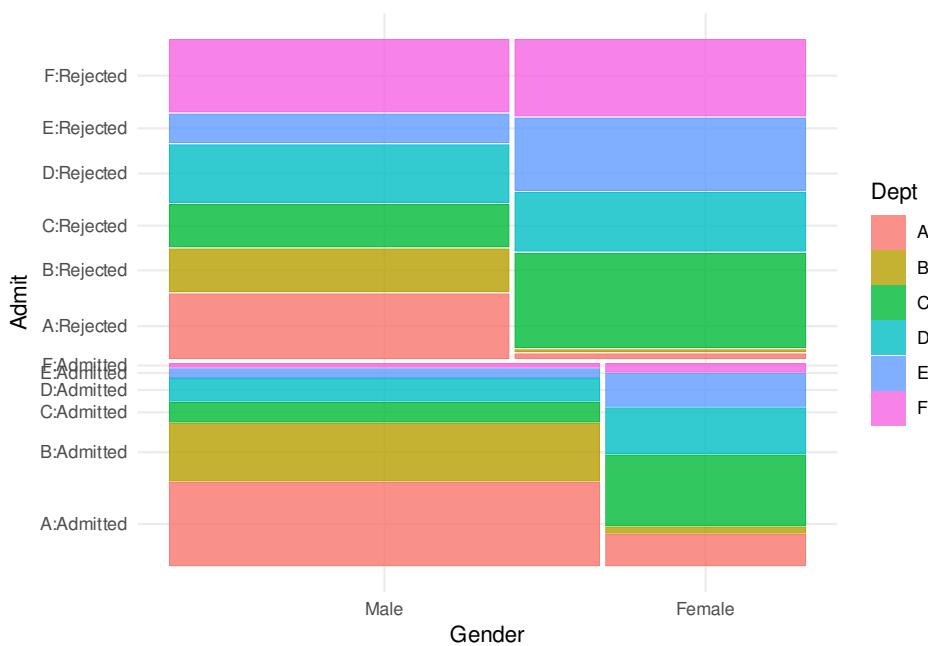


图 17.4: UCBAdmissions 马赛克图

### 17.28.2 1976~1977 年美国佛罗里达州的凶杀案件中被告肤色和死刑判决的关系

被告	被害人	判死	不判死
白人	白人	19	132
	黑人	0	9
黑人	白人	11	32
	黑人	6	97

### 17.28.3 统计专业学生的头发和眼睛的颜色

HairEyeColor 是一个 table 类型的数据对象，和数组的关系 array



```

class(HairEyeColor)
[1] "table"
str(HairEyeColor)

'table' num [1:4, 1:4, 1:2] 32 53 10 3 11 50 10 30 10 25 ...
- attr(*, "dimnames")=List of 3
..$ Hair: chr [1:4] "Black" "Brown" "Red" "Blond"
..$ Eye : chr [1:4] "Brown" "Blue" "Hazel" "Green"
..$ Sex : chr [1:2] "Male" "Female"

apply(HairEyeColor, c(1, 2), sum)

Eye
Hair Brown Blue Hazel Green
Black 68 20 15 5
Brown 119 84 54 29
Red 26 17 14 14
Blond 7 94 10 16

plot(HairEyeColor, col = "lightblue", border = "white")
library(ggmosaic)
ggplot(data = as.data.frame(HairEyeColor)) +
 geom_mosaic(aes(weight = Freq, x = product(Hair, Eye), fill = Sex)) +
 theme_minimal() +
 labs(x = "Hair", y = "Eye")

```

## 17.29 运行环境

```

sessionInfo()

R version 4.1.0 (2021-05-18)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 20.04.2 LTS
##
Matrix products: default
BLAS: /usr/lib/x86_64-linux-gnublas/libblas.so.3.9.0

```

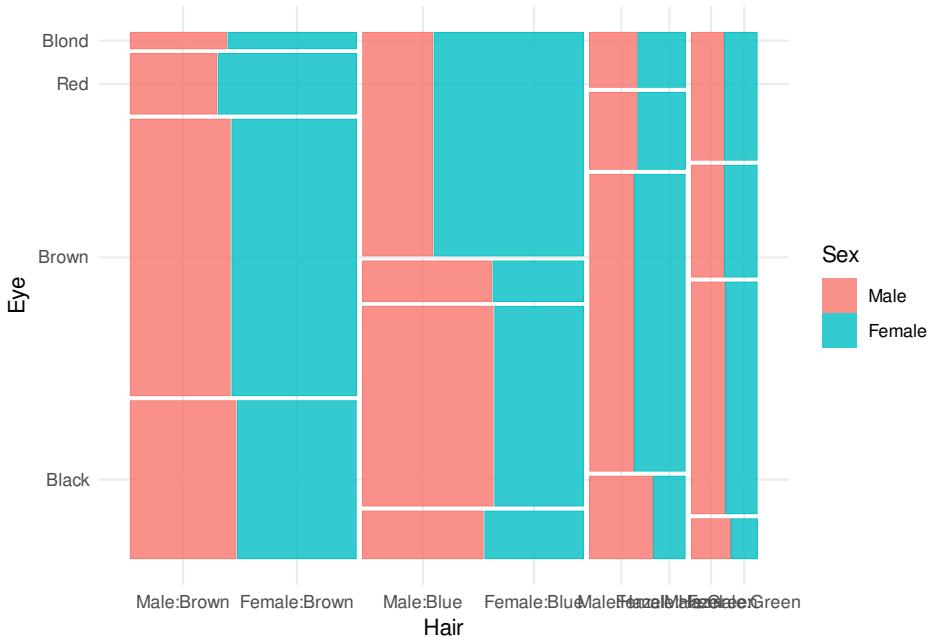


图 17.5: 头发、眼睛颜色和性别的比例

```
LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.9.0
##
locale:
[1] LC_CTYPE=en_US.UTF-8 LC_NUMERIC=C
[3] LC_TIME=en_US.UTF-8 LC_COLLATE=en_US.UTF-8
[5] LC_MONETARY=en_US.UTF-8 LC_MESSAGES=en_US.UTF-8
[7] LC_PAPER=en_US.UTF-8 LC_NAME=C
[9] LC_ADDRESS=C LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
attached base packages:
[1] stats graphics grDevices utils datasets methods base
##
other attached packages:
[1] ggmosaic_0.3.3 patchwork_1.1.1 ggplot2_3.3.5 magrittr_2.0.1
[5] formatR_1.11
##
loaded via a namespace (and not attached):
```

```
[1] Rcpp_1.0.7 plyr_1.8.6 pillar_1.6.2 compiler_4.1.0
[5] tools_4.1.0 digest_0.6.27 viridisLite_0.4.0 jsonlite_1.7.2
[9] evaluate_0.14 lifecycle_1.0.0 tibble_3.1.3 gtable_0.3.0
[13] pkgconfig_2.0.3 rlang_0.4.11 DBI_1.1.1 ggrepel_0.9.1
[17] yaml_2.2.1 xfun_0.24 httr_1.4.2 withr_2.4.2
[21] stringr_1.4.0 dplyr_1.0.7 knitr_1.33 htmlwidgets_1.5.3
[25] generics_0.1.0 vctrs_0.3.8 grid_4.1.0 tidyselect_1.1.1
[29] data.table_1.14.0 glue_1.4.2 R6_2.5.0 plotly_4.9.4.1
[33] fansi_0.5.0 rmarkdown_2.9 bookdown_0.22 tidyverse_1.1.3
[37] farver_2.1.0 purrr_0.3.4 productplots_0.1.1 scales_1.1.1
[41] ellipsis_0.3.2 htmltools_0.5.1.1 assertthat_0.2.1 colorspace_2.0-2
[45] labeling_0.4.2 utf8_1.2.2 stringi_1.7.3 lazyeval_0.2.2
[49] munsell_0.5.0 crayon_1.4.1
```



## 第十八章 功效分析

CRAN 上有很多功效计算和分析的 R 包，我们针对不同的混合效应模型和统计检验，提供对应的 R 实现。

**MKpower** 包提供 Welch 和 Hsu (许宝驥) t 检验、Wilcoxon 秩和检验、符号秩检验的功效分析和样本量计算，经验功效和第一类错误的计算方法是蒙特卡罗模拟。**Superpower** 基于模拟的方法分析三因素方差分析实验设计的功效，开发者写了本书介绍，详见 <https://aaroncaldwell.us/SuperpowerBook/>，也开发了两个 Shiny 应用。**powerlmm** 可用于计算两、三个水平的纵向多水平/线性混合效应模型的功效。**pwrAB** Welch 两样本 t 检验的功效分析，常用于 A/B 测试。**Metin Bulus** 开发 **PowerUpR** 计算响应变量是连续型的多水平随机对照实验统计功效，最小可检测的效应大小，最小样本量要求。**simr** 通过模拟方法分析广义线性混合效应模型的功效。**WebPower** 提供相关性、比例、t 检验、单因素方差分析、两因素方差分析、线性回归、逻辑回归、泊松回归、纵向数据分析、结构方程模型和多水平模型等的功效分析，详见网站 <https://webpower.psychstat.org/>，包含书籍和功效分析的工具。**PowerAnalysisIL** 功效分析的 shiny 应用 <http://daniellakens.blogspot.com/2015/01/always-use-welchs-t-test-instead-of.html>。

此外，还有 **lmerTest** [Kuznetsova et al., 2017] 和 **lmtest** [Zeileis and Hothorn, 2002]。试验设计 [茆诗松 et al., 2004] 可以视为一种组织形式，包括各类检验，R 语言实战 [Kabacoff, 2015] 作者 Robert I. Kabacoff 创建了网站 **Quick-R**，实战这本书第 10 章功效分析主要基于 **pwr** 包来介绍，Jacob Cohen 的著作《Statistical Power Analysis for the Behavioral Sciences》第二版 [Cohen, 1988]

<https://powerandsamplesize.com/> 功效和样本量计算器

```
library(pwr)
library(Matrix)
library(lme4)
```

**pbkrtest** 提供 parametric bootstrap test、Kenward-Roger-type F-test、Satterthwaite-

云  
湘  
黄  
④

type F-test 用于线性混合效应模型， parametric bootstrap test 用于广义线性混合效应模型

## 18.1 方差分析检验的功效

`power.anova.test()` 计算平衡的单因素方差分析检验的功效

```
usage(power.anova.test)
```

```
power.anova.test(groups = NULL, n = NULL, between.var = NULL, within.var = NULL,
 sig.level = 0.05, power = NULL)
```

```
power.anova.test(
 groups = 4, # 4 个组
 between.var = 1, # 组间方差为 1
 within.var = 3, # 组内方差为 3
 power = 0.95 # 1 - 犯第二类错误的概率
)
```

```
##
Balanced one-way analysis of variance power calculation
##
groups = 4
n = 18.18245
between.var = 1
within.var = 3
sig.level = 0.05
power = 0.95
##
NOTE: n is number in each group
```

## 18.2 比例检验的功效

`power.prop.test()` 计算两样本比例检验的功效

```
usage(power.prop.test)
```

```
power.prop.test(n = NULL, p1 = NULL, p2 = NULL, sig.level = 0.05, power = NULL,
```

```
alternative = c("two.sided", "one.sided"), strict = FALSE,
tol = .Machine$double.eps^0.25)
```

功效可以用来计算实验所需要的样本量，检验统计量的功效越大/高，检验方法越好，实验所需要的样本量越少

```
p1 >= p2 的检验 单边和双边检验
power.prop.test(
 p1 = .65, p2 = 0.6, sig.level = .05,
 power = 0.90, alternative = "one.sided"
)

Two-sample comparison of proportions power calculation

n = 1603.846
p1 = 0.65
p2 = 0.6
sig.level = 0.05
power = 0.9
alternative = one.sided

NOTE: n is number in *each* group
power.prop.test(
 p1 = .65, p2 = 0.6, sig.level = .05,
 power = 0.90, alternative = "two.sided"
)

Two-sample comparison of proportions power calculation

n = 1968.064
p1 = 0.65
p2 = 0.6
sig.level = 0.05
power = 0.9
alternative = two.sided
##
```



## NOTE: n is number in \*each\* group

**pwr** 包 `pwr.2p.test()` 函数提供了类似 `power.prop.test()` 函数的功能

```
library(pwr)
明确 $p_1 > p_2$ 的检验
单边检验拆分更加明细，分为大于和小于
pwr.2p.test(
 h = ES.h(p1 = 0.65, p2 = 0.6),
 sig.level = 0.05, power = 0.9, alternative = "greater"
)
```

```
##
Difference of proportion power calculation for binomial distribution (arcsine transform)
##
h = 0.1033347
n = 1604.007
sig.level = 0.05
power = 0.9
alternative = greater
##
NOTE: same sample sizes
```

已知两样本的样本量不等，检验  $H_0: p_1 = p_2$   $H_1: p_1 \neq p_2$  的功效

```
library(pwr)
pwr.2p2n.test(
 h = 0.30, n1 = 80, n2 = 245,
 sig.level = 0.05, alternative = "greater"
)
```

```
##
difference of proportion power calculation for binomial distribution (arcsine transform)
##
h = 0.3
n1 = 80
n2 = 245
sig.level = 0.05
power = 0.7532924
alternative = greater
```

```

NOTE: different sample sizes
```

$h$  表示两个样本的差异，计算得到的功效是 0.75

### 18.3 t 检验的功效

`power.t.test()` 计算单样本或两样本的 t 检验的功效，或者根据功效计算参数，如样本量

Cohen's d 单样本/配对 t 检验的功效分析

```
n = 30 # 样本量（只是一个例子）
x = seq(0, 12, 0.01)
library(ggplot2)
dat <- data.frame(xx = x/sqrt(n), yy = 2 * (1 - pt(x, n - 1)))
ggplot(data = dat, aes(x = xx, y = yy)) +
 geom_line() +
 geom_vline(xintercept = c(0.01, 0.2, 0.5, 0.8, 1.2, 2), linetype = 2) +
 theme_minimal() +
 labs(x = "d = t / sqrt(n)", y = "2 * (1 - pt(x, n - 1))")

usage(power.t.test)
```

```
power.t.test(n = NULL, delta = NULL, sd = 1, sig.level = 0.05, power = NULL,
 type = c("two.sample", "one.sample", "paired"),
 alternative = c("two.sided", "one.sided"), strict = FALSE,
 tol = .Machine$double.eps^0.25)
```

```
power.t.test(
 n = 100, delta = 2.2,
 sd = 1, sig.level = 0.05,
 type = "two.sample",
 alternative = "two.sided"
)
```

```

Two-sample t test power calculation
##
```

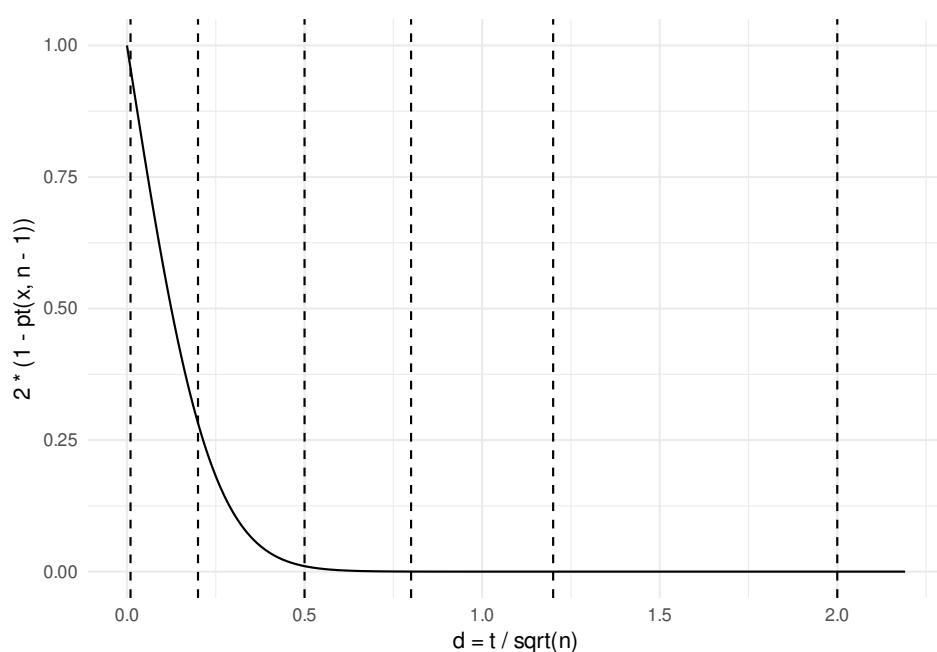


图 18.1: t 检验的功效

```
n = 100
delta = 2.2
sd = 1
sig.level = 0.05
power = 1
alternative = two.sided
##
NOTE: n is number in *each* group
```

表 18.1: 函数 `power.t.test()` 的参数表

参数	含义
<code>n</code>	每个组的样本量
<code>delta</code>	两个组的均值之差
<code>sd</code>	标准差, 默认值 1
<code>sig.level</code>	显著性水平, 默认是 0.05 (犯第 I 类错误的概率)
<code>power</code>	检验的功效 (1 - 犯第 II 类错误的概率)
<code>type</code>	t 检验的类型 "two.sample" 两样本、"one.sample" 单样本或 "paired" 配对样本

参数	含义
alternative	单边或双边检验，取值为 "two.sided" 或 "one.sided"

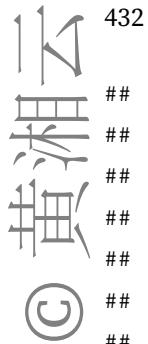
参数 `n`, `delta`, `power`, `sd` 和 `sig.level` 必须有一个值为 `NULL`，为 `NULL` 的参数是由其它参数决定的。

Jacob Cohen 提出的 Cohen's d 和 Cohen's f 详见书籍 [Cohen, 1988]，他的代表性文章，地球是圆的 [Cohen, 1994]

```
前面 t 检验和方差分析检验的等价功效计算
library(pwr)
pwr.t.test(
 d = 2.2 / 6.4,
 n = 100,
 sig.level = 0.05,
 type = "two.sample",
 alternative = "two.sided"
)

##
Two-sample t test power calculation
##
n = 100
d = 0.34375
sig.level = 0.05
power = 0.6768572
alternative = two.sided
##
NOTE: n is number in *each* group

f 是如何和上面的组间/组内方差等价指定的
pwr.anova.test(
 k = 4, # 组数
 f = 0.5,
 power = 0.95 # 检验的效
)
```



```
Balanced one-way analysis of variance power calculation
##
k = 4
n = 18.18244
f = 0.5
##
sig.level = 0.05
power = 0.95
##
NOTE: n is number in each group

with(
 aggregate(
 data = PlantGrowth, weight ~ group,
 FUN = function(x) c(dist_mean = mean(x), dist_sd = sd(x))
),
 cbind.data.frame(weight, group)
)
```

注意

R 3.5.0 以后，函数 `aggregate` 的参数 `drop` 默认设置为 `TRUE` 表示扔掉未用来分组的变量，聚合返回的是一个矩阵类型的数据对象。

### ggsignif 添加显著性注释

```
library(ggplot2)
library(ggsignif)

ggplot(data = PlantGrowth, aes(x = group, y = weight)) +
 geom_boxplot() +
 geom_signif(comparisons = list(c("ctrl", "trt1"), c("trt1", "trt2")),
 map_signif_level = function(p) sprintf("p = %.2g", p),
 textsize = 6, test = "t.test") +
 theme_minimal()
```

无条件  $2 \times 2$  列联表

`fisher.test` [https://en.wikipedia.org/wiki/Fisher's\\_exact\\_test](https://en.wikipedia.org/wiki/Fisher's_exact_test)

`Exact` [https://en.wikipedia.org/wiki/Barnard's\\_test](https://en.wikipedia.org/wiki/Barnard's_test) `exact.test` `power.exact.test`

exact2x2

## 18.4 运行环境

sessionInfo()

```
R version 4.1.0 (2021-05-18)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 20.04.2 LTS
##
Matrix products: default
BLAS: /usr/lib/x86_64-linux-gnublas/libblas.so.3.9.0
LAPACK: /usr/lib/x86_64-linux-gnulapack/liblapack.so.3.9.0
##
locale:
[1] LC_CTYPE=en_US.UTF-8 LC_NUMERIC=C
[3] LC_TIME=en_US.UTF-8 LC_COLLATE=en_US.UTF-8
[5] LC_MONETARY=en_US.UTF-8 LC_MESSAGES=en_US.UTF-8
[7] LC_PAPER=en_US.UTF-8 LC_NAME=C
[9] LC_ADDRESS=C LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
attached base packages:
[1] stats graphics grDevices utils datasets methods base
##
other attached packages:
[1] ggpplot2_3.3.5 lme4_1.1-27.1 Matrix_1.3-4 pwr_1.3-0 formatR_1.11
[6] magrittr_2.0.1
##
loaded via a namespace (and not attached):
[1] Rcpp_1.0.7 compiler_4.1.0 pillar_1.6.2 nloptr_1.2.2.2
[5] tools_4.1.0 boot_1.3-28 digest_0.6.27 evaluate_0.14
[9] lifecycle_1.0.0 tibble_3.1.3 gtable_0.3.0 nlme_3.1-152
[13] lattice_0.20-44 pkgconfig_2.0.3 rlang_0.4.11 DBI_1.1.1
[17] yaml_2.2.1 xfun_0.24 withr_2.4.2 dplyr_1.0.7
```

```
[21] stringr_1.4.0 knitr_1.33 generics_0.1.0 vctrs_0.3.8
[25] tidyselect_1.1.1 grid_4.1.0 glue_1.4.2 R6_2.5.0
[29] fansi_0.5.0 rmarkdown_2.9 bookdown_0.22 minqa_1.2.4
[33] farver_2.1.0 purrrr_0.3.4 scales_1.1.1 htmltools_0.5.1.1
[37] ellipsis_0.3.2 MASS_7.3-54 splines_4.1.0 assertthat_0.2.1
[41] colorspace_2.0-2 labeling_0.4.2 utf8_1.2.2 stringi_1.7.3
[45] munsell_0.5.0 crayon_1.4.1
```



## 第十九章 试验设计

```
library(magrittr)
library(ggplot2)
```

提示

我想不少人初次见到本章题目首先疑惑的可能是到底是试验还是实验？这里做一下说明，实验的意思是带有验证性的目的，已经有结果了，做实验验证某个规律，常常用在物理、化学的课堂里，学生做实验验证自由落体运动、做实验测量重力加速度等等。试验的意思是人为设定一系列操作步骤去探索未知，不确定结果如何，试一试。

试验设计（Design of Experiment，简称 DOE）是一个应用性很强的学科领域，R. A. Fisher 曾在农业站做实验验证孟德尔的豌豆实验结果。

Vikneswaran 提供了一份书籍 Berger and Maurer [2002] 的补充材料 – An R companion to “Experimental Design”，目前 Paul Berger 的这本书已经迭代到第二版 [Berger et al., 2018]，2015 年 Paul Berger 出版了新书《Improving the User Experience through Practical Data Analytics: Gain Meaningful Insight and Increase Your Bottom Line》[Fritz and Berger, 2015] 颇具应用性，结合产品用户体验来谈试验设计。

Bill Venables 开发的 `conf.design` 是试验设计领域的核心 R 包，CRAN 官网上试验设计视图 <https://cran.r-project.org/view=ExperimentalDesign> 可以让我们对试验设计这个领域有一个粗略的了解。

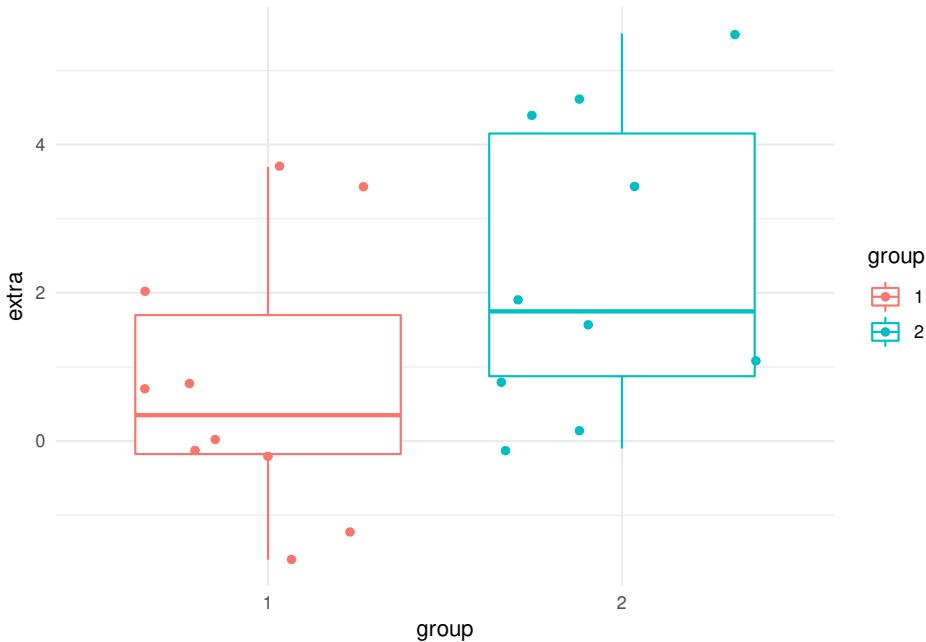
推荐读者使用贴合 R 语言的试验设计入门书《Design and Analysis of Experiments with R》[Lawson, 2014]，作者提供相应的 R 包 `daewr` 打包了该书的数据和代码。另外，推荐的读物是《Statistics for Experimenters: Design, Innovation, and Discovery》[Box et al., 2005] 和《Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing》[Kohavi et al., 2020]。

另一个和试验设计紧密相关的话题是敏感性分析，推荐 Devin Incerti 的敏感性分析系列博客 <https://devinincerti.com/blog.html>，R 包 `sensitivity` 提供 140+ 页的手册，功能非常强，模型的全局敏感性分析，`SWATplusR` SWAT 分析法和 R 语言结合。



## 19.1 学生睡眠质量

```
ggplot(data = sleep, aes(x = group, y = extra, color = group)) +
 geom_boxplot() +
 geom_jitter() +
 theme_minimal()
```



## 19.2 驱虫喷雾的效果

InsectSprays 数据集 [Beall, 1942] 来源于农业实验，记录了不同杀虫剂的效果，即杀虫剂过后，单位实验区域内虫子的数量，如图19.1所示，横轴表示杀虫剂种类，纵轴表示虫子数量。

```
ggplot(data = InsectSprays, aes(x = spray, y = count, color = spray)) +
 geom_boxplot() +
 geom_jitter() +
 theme_minimal()
```

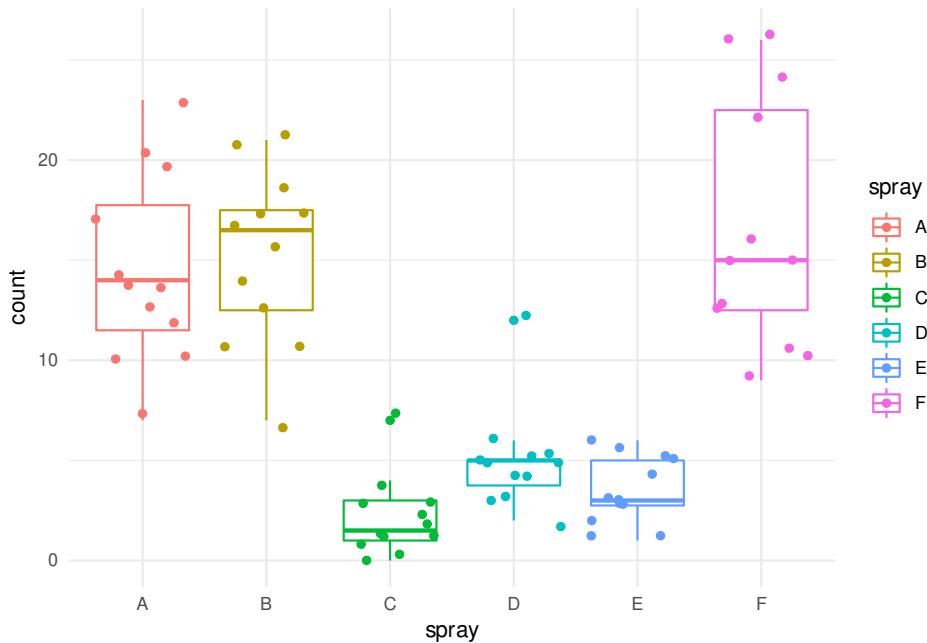


图 19.1: 不同杀虫剂的效果

先创建一个 `aov` 对象，把它命名为 `mod1`，见下方

```
mod1 <- aov(count ~ spray, data = InsectSprays)
```

第一个参数告诉 R `count` 是响应变量，`spray` 是协变量，第二个参数告诉 R 去对像 `InsectSprays` 中寻找这些变量。下面把分析结果以一种漂亮的格式打印出来

```
summary(mod1)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)						
## spray	5	2669	533.8	34.7	<2e-16 ***						
## Residuals	66	1015	15.4								
## ---											
## Signif. codes:	0	'***'	0.001	'**'	0.01	'*'	0.05	'. '	0.1	' '	1

表格中的条目是很容易理解的，比如最右边的列表示 P 值。如果我们想做固定



显著性水平下的检验，比如  $\alpha = 0.075$  时的 F 统计量的值，

```
qf(0.075, 5, 66, lower.tail = F)
```

```
[1] 2.110783
```

上面的命令是说  $F(5, 66)$  分布的 0.075 分位点，最后一个参数很关键，因为默认情况下 R 计算下分位点，详情见 ?qf。

方差分析做了三个假设

1. 残差  $\epsilon_{ij}$  是相互独立的随机变量；
2. 残差  $\epsilon_{ij}$  服从正态分布；
3. 残差  $\epsilon_{ij}$  均值为 0，方差是固定的常数。

假设 1 和 3 通过图来检验，假设 2 通过 QQ 图来检验。值得一提的是 mod1 对象除了打印出来，还有很多方法

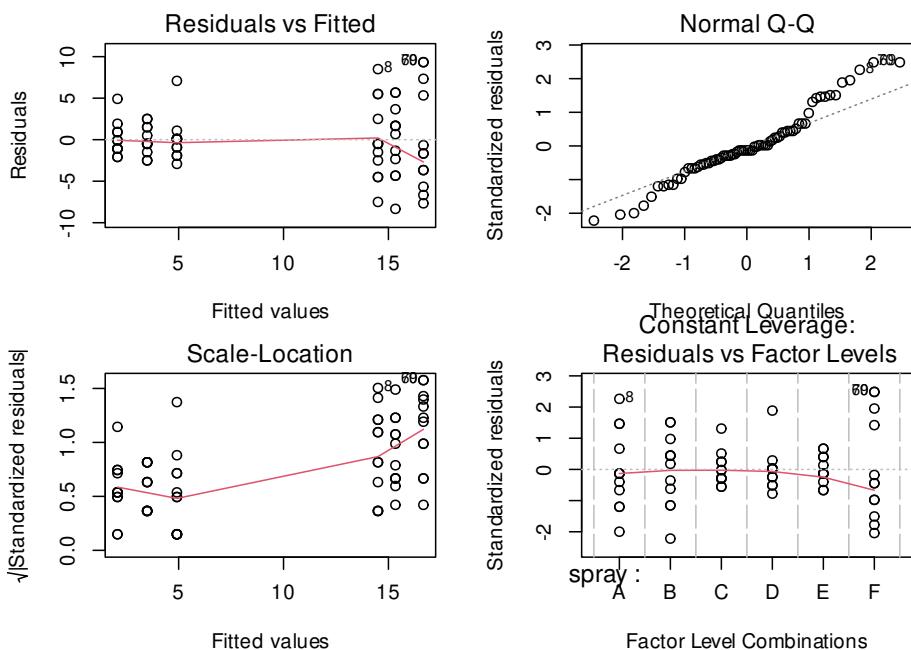
```
names(mod1)
```

```
[1] "coefficients" "residuals" "effects" "rank"
[5] "fitted.values" "assign" "qr" "df.residual"
[9] "contrasts" "xlevels" "call" "terms"
[13] "model"
```

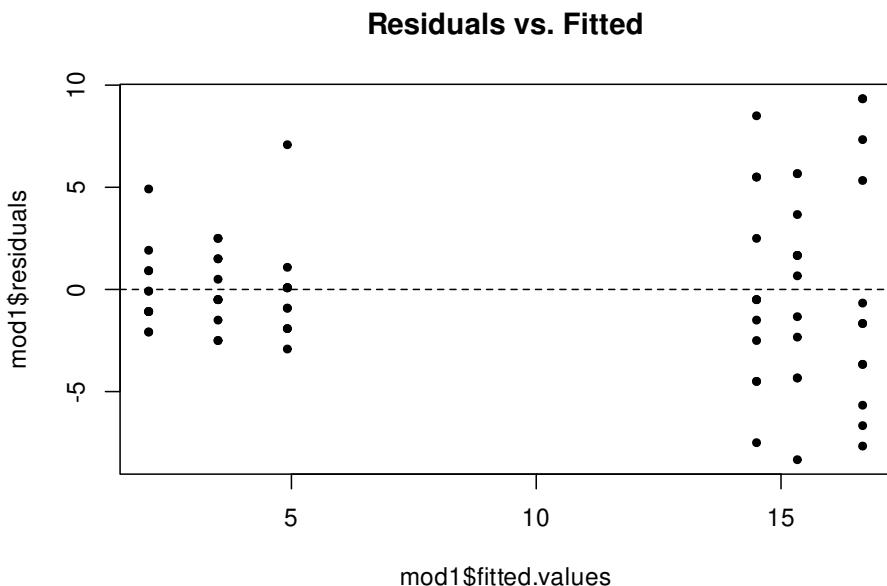
比如获取残差，考虑到篇幅，这里仅显示前 10 个

```
head(mod1$residuals, 10)
```

```
1 2 3 4 5 6 7 8 9 10
-4.5 -7.5 5.5 -0.5 -0.5 -2.5 -4.5 8.5 2.5 5.5
par(mar = c(4, 4, 2, 2), mflow = c(2, 2))
plot(mod1)
```

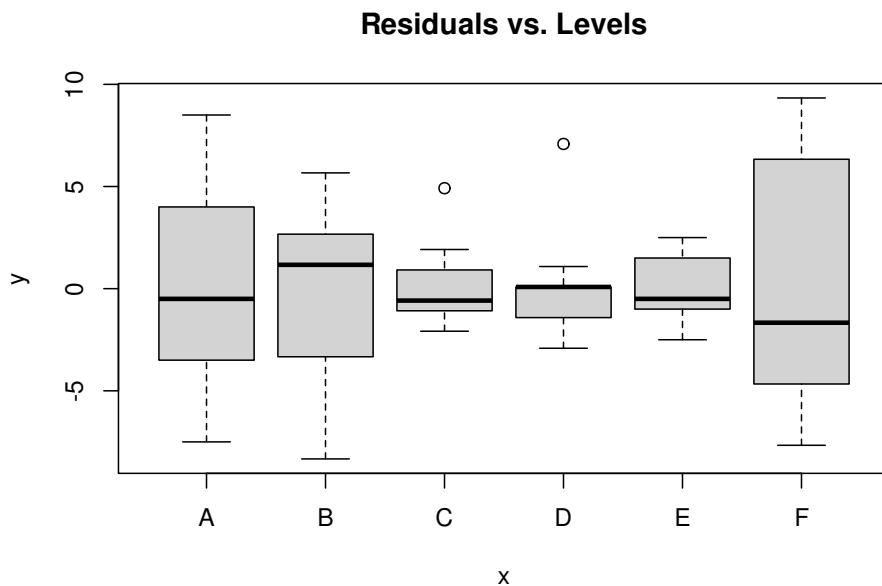


```
plot(mod1$fitted.values, mod1$residuals, main = "Residuals vs. Fitted", pch = 20)
abline(h = 0, lty = 2)
```

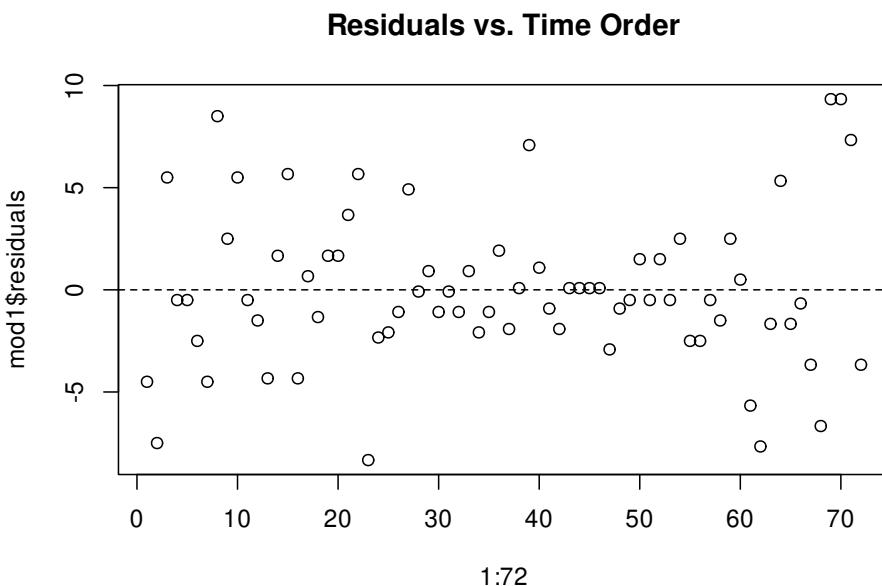


黄湘云

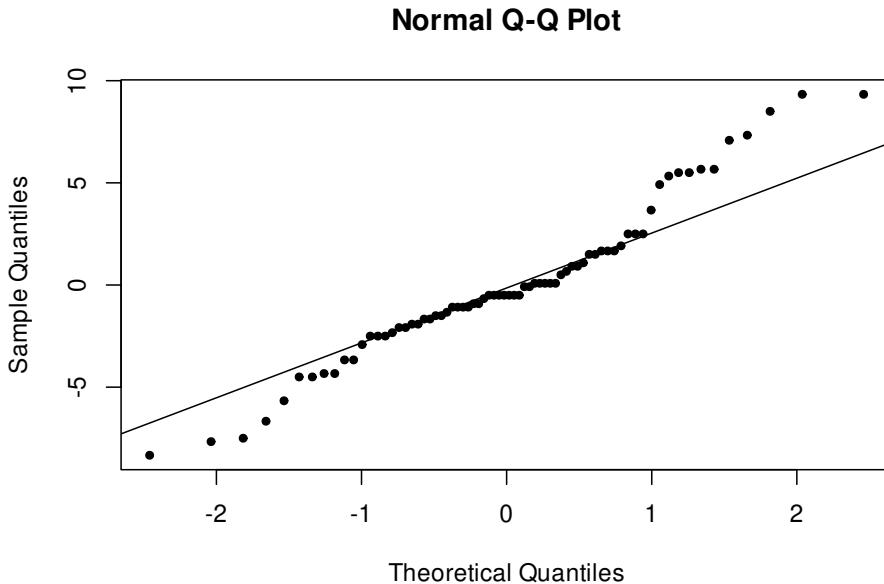
```
plot(mod1$model$spray, mod1$residuals, main = "Residuals vs. Levels")
```



```
plot(1:72, mod1$residuals, main = "Residuals vs. Time Order")
abline(h = 0, lty = 2)
```



```
qqnorm(mod1$residuals, pch = 20)
qqline(mod1$residuals)
```



如果上面的假设显著失效，我们要采用非参数检验

```
mod2 <- kruskal.test(count ~ spray, data = InsectSprays)
mod2

Kruskal-Wallis rank sum test

data: count by spray
Kruskal-Wallis chi-squared = 54.691, df = 5, p-value = 1.511e-10
```

计算给定水平下的置信区间，构造置信水平为 95% 的区间

$$\bar{X} \pm t_{1-\alpha/2}(s/\sqrt{n})$$

以 A 号杀虫剂为例，

```
xbar = mean(InsectSprays[InsectSprays$spray == "A", "count"])
t_crit <- qt(0.025, mod1$df.residual, lower.tail = F)
s <- sqrt(sum((mod1$residuals)^2) / mod1$df.residual)
```



```
n <- sum(InsectSprays$spray == "A")
最后置信区间的上下限
c(xbar - t_crit * (s/ sqrt(n)), xbar + t_crit * (s/ sqrt(n)))
[1] 12.23958 16.76042
```

比较 A 号和 C 号杀虫剂的效果，计算两个均值差的置信区间

$$\bar{X}_1 - \bar{X}_2 \pm t_{1-\alpha/2}(s/\sqrt{1/n_1 + 1/n_2})$$

```
n1 <- sum(InsectSprays$spray == "A")
n2 <- sum(InsectSprays$spray == "C")

x1bar = mean(InsectSprays[InsectSprays$spray == "A", "count"])
x2bar = mean(InsectSprays[InsectSprays$spray == "C", "count"])
```

代入公式即可计算得到置信区间

```
(x1bar - x2bar) - t_crit * s * sqrt(1/ n1 + 1/n2)
```

```
[1] 9.219948
(x1bar - x2bar) + t_crit * s * sqrt(1/ n1 + 1/n2)
```

```
[1] 15.61339
```

Fisher's 最小显著性检验 (Fisher's Least Significant Difference Test) 即

```
t_crit * s * sqrt(1/ n1 + 1/n2)
```

```
[1] 3.196719
```

Tukey's Honestly Significant Difference Test 主要测量成对实验的误差比率，假定每个水平下的实验次数是相等的，只需将上面的 aov 对象传递给函数 TukeyHSD()

```
mod3 <- TukeyHSD(mod1, ordered = TRUE)
```

```
mod3
```

```
Tukey multiple comparisons of means
95% family-wise confidence level
factor levels have been ordered
##
Fit: aov(formula = count ~ spray, data = InsectSprays)
```



```

$spray
diff lwr upr p adj
E-C 1.4166667 -3.282742 6.116075 0.9488669
D-C 2.8333333 -1.866075 7.532742 0.4920707
A-C 12.4166667 7.717258 17.116075 0.0000000
B-C 13.2500000 8.550591 17.949409 0.0000000
F-C 14.5833333 9.883925 19.282742 0.0000000
D-E 1.4166667 -3.282742 6.116075 0.9488669
A-E 11.0000000 6.300591 15.699409 0.0000000
B-E 11.8333333 7.133925 16.532742 0.0000000
F-E 13.1666667 8.467258 17.866075 0.0000000
A-D 9.5833333 4.883925 14.282742 0.0000014
B-D 10.4166667 5.717258 15.116075 0.0000002
F-D 11.7500000 7.050591 16.449409 0.0000000
B-A 0.8333333 -3.866075 5.532742 0.9951810
F-A 2.1666667 -2.532742 6.866075 0.7542147
F-B 1.3333333 -3.366075 6.032742 0.9603075
```

其中，`diff` 表示均值之差，`lwr` 和 `upr` 表示置信区间的上下限，`p adj` 是对应的。检查一下，看看哪些置信区间包含 0，包含 0 的表示不显著，从第三行来看，A 和 C 之间差别显著。之前计算过 A、C 均值，均值之差即

```
(x1bar - x2bar)
```

```
[1] 12.41667
```

在误差比率  $\alpha = 0.05$  的情况下，如果你想手动计算 HSD 值

```
q_crit <- qtukey(p = 0.05, nmeans = length(mod1$xlevels[[1]]), df = mod1$df.residual)
mod1$df.residual 是 6
hsd <- q_crit * s / sqrt(6)
hsd
```

```
[1] 6.645967
```

将模型结果 `mod3` 用图画出来，见下图

```
plot(mod3)
```

关于多重比较请见 Frank Bretz, Torsten Hothorn, Peter Westfall 的书《Multiple Comparisons Using R》及配套 R 包 `multcomp`，该 R 包现由 Torsten Hothorn 维

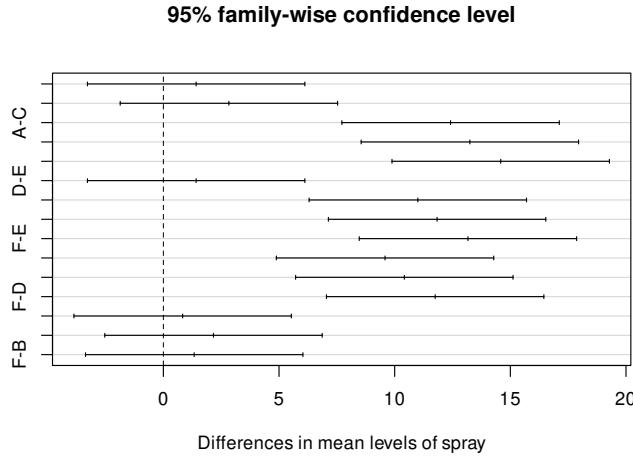


图 19.2: 成对显著性水平

护，他还维护了一个由数据集构成的 R 包 [TH.data](#)，我们后续章节也会用到。

### 19.3 重复数不等的多重比较

Tukey 的检验方法要求各个组的重复数相等，而方差分析的重复数不等时，我们需要用如下方法

$$\frac{(\bar{y}_{i\cdot} - \bar{y}_{j\cdot}) - (\mu_i - \mu_j)}{\sqrt{\frac{1}{m_i} + \frac{1}{m_j}} \hat{\sigma}} \sim t(f_e)$$

$$c_{ij} = \sqrt{(r-1)F_{1-\alpha}(r-1, f_e)(\frac{1}{m_i} + \frac{1}{m_j})\hat{\sigma}^2}$$

$\hat{\sigma}^2 = S_e/f_e$  是  $\sigma^2$  无偏估计。

$$y_{ij} = \mu + a_i + \epsilon_{ij}, \quad i = 1, 2, \dots, r, \quad j = 1, 2, \dots, m_i. \quad \sum_{i=1}^r m_i a_i = 0,$$

其中， $\epsilon_{ij}$  相互独立，服从  $\mathcal{N}(0, \sigma^2)$ .

$$f_e = n - r, S_e = \sum_{i=1}^r \sum_{j=1}^{m_i} (y_{ij} - \bar{y}_{i\cdot})^2 = S_T - S_A$$

## 19.4 不同地区的草类植物吸收二氧化碳的情况

通过观察不同地区的草类植物吸收二氧化碳的情况，研究植物的耐寒性

```
ggplot(data = CO2, aes(x = conc, y = uptake, color = Type, shape = Treatment)) +
 geom_point() +
 geom_line() +
 facet_wrap(~Plant, ncol = 3) +
 theme_minimal() +
 labs(x = "conc (mL/L)", y = "uptake (umol/m^2 sec)")
```

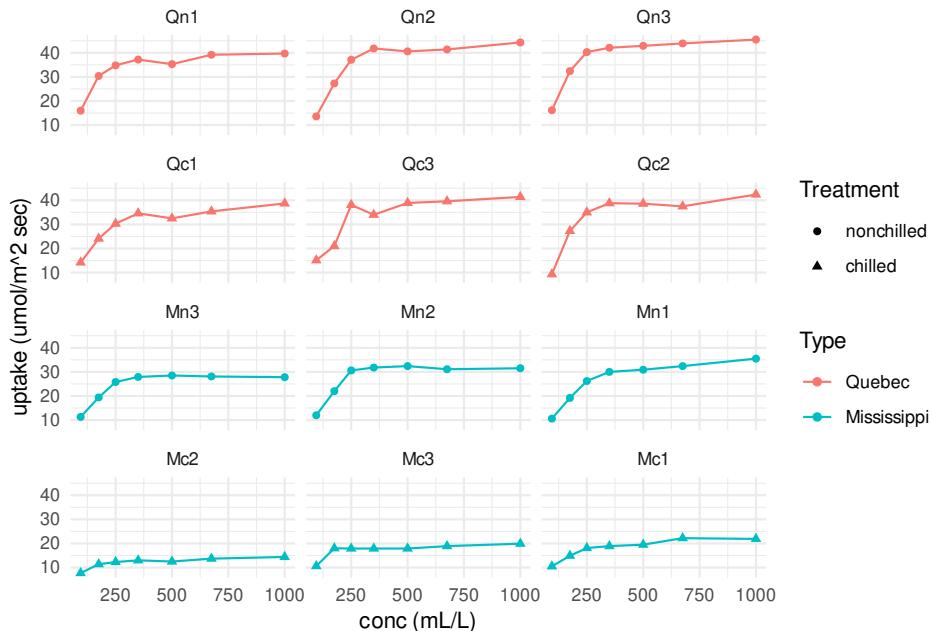


图 19.3: 草类植物吸收二氧化碳的量

## 19.5 果园喷雾剂的效力

评估喷雾杀虫剂在果园的效果

```
data("OrchardSprays")
```

## 19.6 验证孟德尔的豌豆实验结果

R. A. Fisher 在农业站做实验验证孟德尔的豌豆实验结果

```
data("npk")
```

- ④ 豌豆产量和氮 (nitrogen, N) 磷酸盐 (phosphate, P) 钾盐 (potassium, K) 的关系

## 第二十章 线性模型

There's probably some examples, but there are some examples of people using `solve(t(X) %*% W %*% X) %*% W %*% Y` to compute regression coefficients, too.

— Thomas Lumley <sup>1</sup>

### 20.1 方差分析

I was profoundly disappointed when I saw that S-PLUS 4.5 now provides “Type III” sums of squares as a routine option for the summary method for aov objects. I note that it is not yet available for multistratum models, although this has all the hallmarks of an oversight (that is, a bug) rather than common sense seeing the light of day. When the decision was being taken of whether to include this feature, “because the FDA requires it” a few of my colleagues and I were consulted and our reply was unhesitatingly a clear and unequivocal “No”, but it seems the FDA and SAS speak louder and we were clearly outvoted.

— Bill Venables <sup>2</sup>

方差分析、A/B Test 和多重比较多用于互联网数据 lme 的特例

### 20.2 单因素方差分析

chickwts 不同的喂食方式对体重的影响

---

<sup>1</sup><https://stat.ethz.ch/pipermail/r-help/2006-March/101596.html>

<sup>2</sup>来源于 [Exegeses on Linear Models](#)

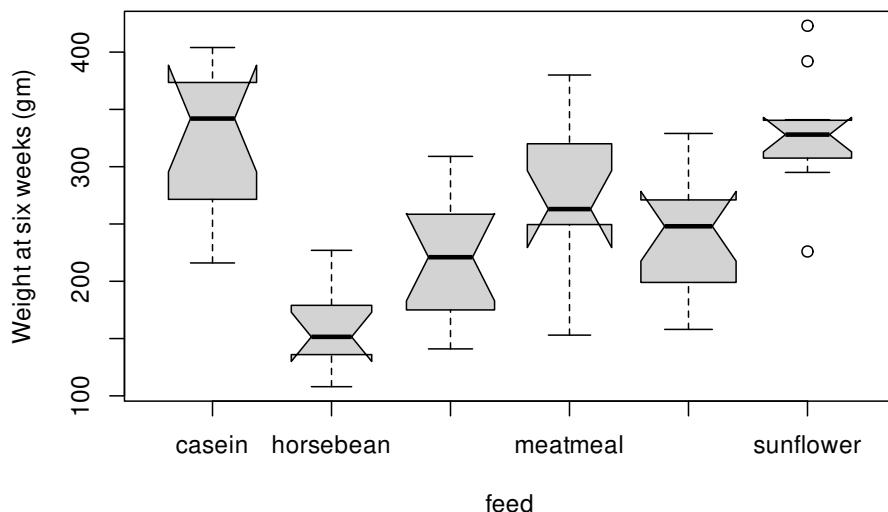
湘潭黃

(C)

```
boxplot(weight ~ feed, data = chickwts, col = "lightgray",
 varwidth = TRUE, notch = TRUE, main = "chickwt data",
 ylab = "Weight at six weeks (gm)")

Warning in (function (z, notch = FALSE, width = NULL, varwidth = FALSE, : some
notches went outside hinges ('box'): maybe set notch=FALSE
```

**chickwt data**



```
anova(fm1 <- lm(weight ~ feed, data = chickwts))
```

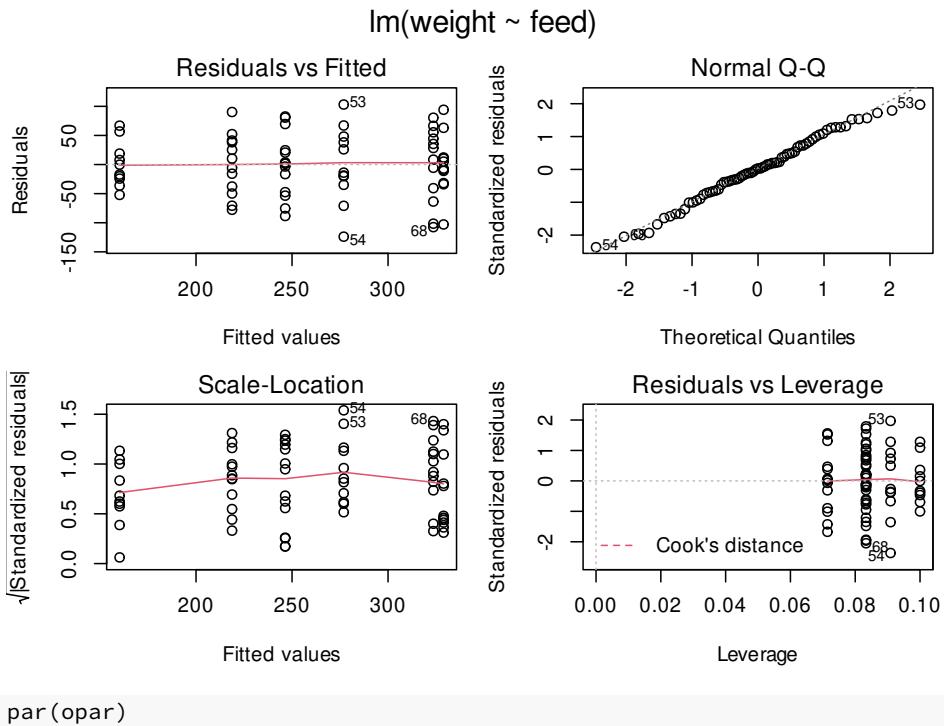
```
Analysis of Variance Table

Response: weight

Df Sum Sq Mean Sq F value Pr(>F)
feed 5 231129 46226 15.365 5.936e-10 ***
Residuals 65 195556 3009

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

opar <- par(mfrow = c(2, 2), oma = c(0, 0, 1.1, 0),
 mar = c(4.1, 4.1, 2.1, 1.1))
plot(fm1)
```



sleep

```
Student's paired t-test 成对样本的 t 检验
with(sleep,
 t.test(extra[group == 1],
 extra[group == 2], paired = TRUE))
```

```
##
Paired t-test
##
data: extra[group == 1] and extra[group == 2]
t = -4.0621, df = 9, p-value = 0.002833
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-2.4598858 -0.7001142
sample estimates:
mean of the differences
-1.58
```

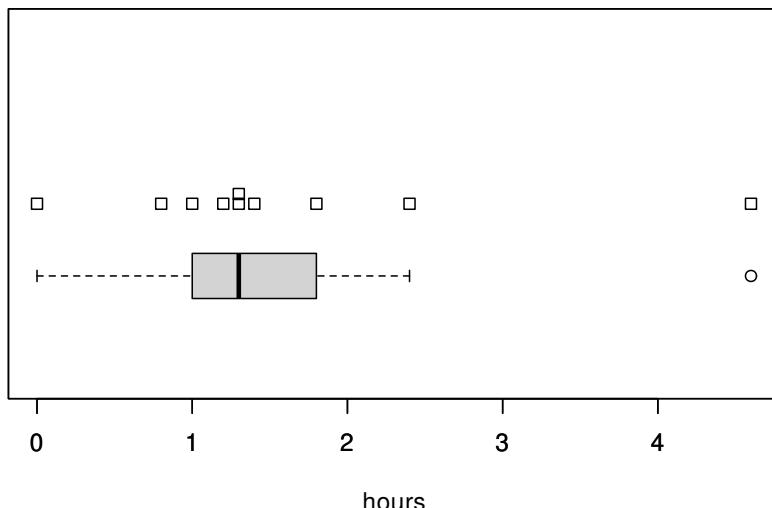
云  
湘  
黄  
©

```
The sleep *prolongations*
sleep1 <- with(sleep, extra[group == 2] - extra[group == 1])
summary(sleep1)

Min. 1st Qu. Median Mean 3rd Qu. Max.
0.00 1.05 1.30 1.58 1.70 4.60

stripchart(sleep1, method = "stack", xlab = "hours",
 main = "Sleep prolongation (n = 10)")
boxplot(sleep1, horizontal = TRUE, add = TRUE,
 at = .6, pars = list(boxwex = 0.5, staplewex = 0.25))
```

**Sleep prolongation (n = 10)**



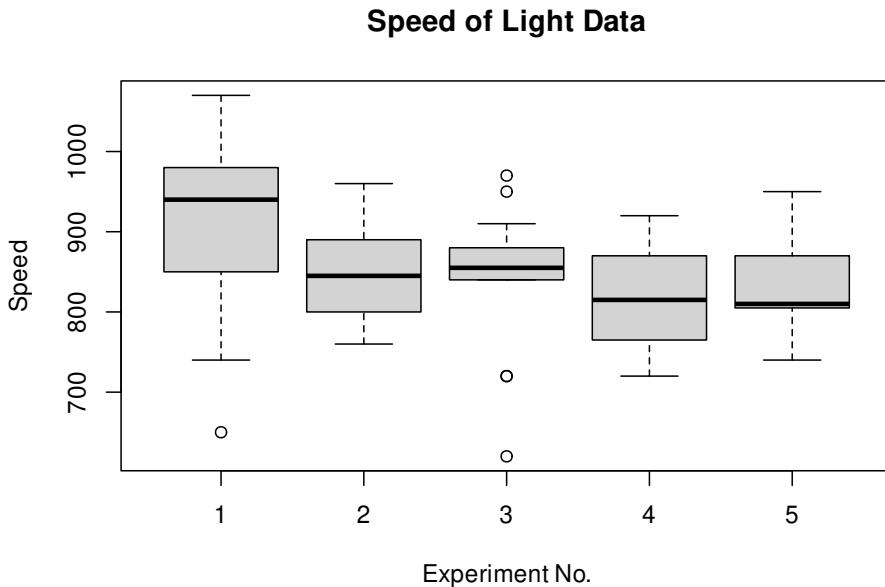
另一个关于测量光速的例子，带分类变量的

```
michelson <- transform(morley,
 Expt = factor(Expt), Run = factor(Run))
xtabs(~ Expt + Run, data = michelson) # 5 x 20 balanced (two-way)
```

```
Run
Expt 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
1
2 1
```

```
3 1
4 1
5 1

plot(Speed ~ Expt, data = michelson,
 main = "Speed of Light Data", xlab = "Experiment No.")
```



```
fm <- aov(Speed ~ Run + Expt, data = michelson)
summary(fm)
```

```
Df Sum Sq Mean Sq F value Pr(>F)
Run 19 113344 5965 1.105 0.36321
Expt 4 94514 23629 4.378 0.00307 **
Residuals 76 410166 5397

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
fm0 <- update(fm, . ~ . - Run)
anova(fm0, fm)
```

```
Analysis of Variance Table

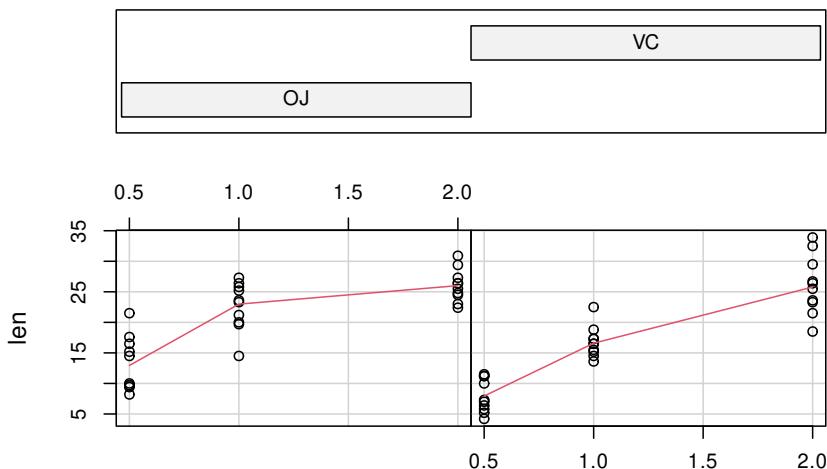
Model 1: Speed ~ Expt
```

```
Model 2: Speed ~ Run + Expt
Res.Df RSS Df Sum of Sq F Pr(>F)
1 95 523510
2 76 410166 19 113344 1.1053 0.3632
```

ToothGrowth 维生素 C 对牙齿增长的关系

```
coplot(len ~ dose | supp, data = ToothGrowth, panel = panel.smooth,
 xlab = "ToothGrowth data: length vs dose, given type of supplement")
```

Given : supp



ToothGrowth data: length vs dose, given type of supplement

## 20.3 双因素方差分析

?lm mlm

## 20.4 多因素方差分析

MANOVA.RM 和 ffmanova 包处理多因素方差分析

## 20.5 核学习

基于核的机器学习算法 [kernlab](#)

David Meyer 基于 [libsvm](#) 开发了 [e1071](#) 包，基于核方法实现了非线性回归分类算法

线性模型、逻辑回归模型、多项逻辑回归模型、神经网络、朴素贝叶斯、分类回归树等模型和算法借助 Shiny 整合在一起  
<https://radiantrstats.github.io/docs/> 和 <http://radiantrstats.github.io/radiantr.model/>

## 20.6 通用机器学习

表 20.1: R 包之间的不一致性，计算预测分类的概率的语法

函数	R 包	代码
lda	MASS	<code>predict(obj)</code>
glm	stats	<code>predict(obj, type = "response")</code>
gbm	gbm	<code>predict(obj, type = "response", n.trees)</code>
mda	mda	<code>predict(obj, type = "posterior")</code>
rpart	rpart	<code>predict(obj, type = "prob")</code>
Weka	RWeka	<code>predict(obj, type = "probability")</code>
logitboost	LogitBoost	<code>predict(obj, type = "raw", nIter)</code>
pamr.train	pamr	<code>pamr.predict(obj, type = "posterior")</code>

## 20.7 理论基础

$$Y = X\beta + \epsilon \quad (20.1)$$

$$X^\top Y = X^\top X\beta \quad (20.2)$$

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y \quad (20.3)$$

$$\hat{Y} = X(X^\top X)^{-1} X^\top Y \quad (20.4)$$

$$\hat{\sigma}^2 = \frac{\|Y - \hat{Y}\|_2}{n - rk(X)} \quad (20.5)$$

$$= \frac{\|(I - X(X^\top X)^{-1} X^\top)Y\|_2}{n - rk(X)} \quad (20.6)$$

$$= \frac{Y^\top (I - X(X^\top X)^{-1} X^\top)Y}{n - rk(X)} \quad (20.7)$$

## 20.8 多重多元线性回归

参考 John Fox 和 Sanford Weisberg 的著作 [Fox and Weisberg, 2019]  
附录<sup>3</sup>

多个响应变量和协变量<sup>4</sup>

多重多元线性回归 multiply linear regression lm R 版本 3.6 以上 PR#17407

```
fit_mtcars <- lm(cbind(mpg, qsec) ~ 1, data = mtcars, offset = cbind(wt, wt * 2))
summary(fit_mtcars)
```

```
Response mpg :
##
Call:
lm(formula = mpg ~ 1, data = mtcars, offset = cbind(wt, wt *
2))
##
Residuals:
Min 1Q Median 3Q Max
-11.897 -4.947 -1.316 2.984 15.192
```

<sup>3</sup><https://socialsciences.mcmaster.ca/jfox/Books/Companion/appendices.html>

<sup>4</sup><https://data.library.virginia.edu/getting-started-with-multivariate-multiple-regression/>



```

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.873 1.219 13.85 8.1e-15 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.893 on 31 degrees of freedom

Response qsec :

Call:
lm(formula = qsec ~ 1, data = mtcars, offset = cbind(wt, wt *
2))

Residuals:
Min 1Q Median 3Q Max
-4.6842 -2.0793 -0.1693 2.2693 5.1857

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.4142 0.5076 22.49 <2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.871 on 31 degrees of freedom
```

## 20.9 回归诊断

包括线性模型和广义线性模型

Regression Deletion Diagnostics ?influence.measures

```
library(extrafont) # 注册字体 CM Roman 到 PDF 设备
data(anscombe)
```



```

form <- paste(paste0("y", seq(4)), paste0("x", seq(4)), sep = "~") # form <- sprintf('y
fit <- lapply(form, lm, data = anscombe)
par(mfrow = c(2, 2), mar = 0.1 + c(4, 4, 1, 1), oma = c(0, 0, 2, 0), family = "CM Roman")
for (i in 1:4) {
 plot(as.formula(form[i]),
 data = anscombe, col = "black",
 pch = 19, cex = 1.2,
 xlim = c(3, 19), ylim = c(3, 13),
 xlab = as.expression(substitute(bold(x[i]), list(i = i))),
 ylab = as.expression(substitute(bold(y[i]), list(i = i)))
)
 abline(fit[[i]], col = "red", lwd = 2)
 text(7, 12, bquote(bold(R)^2 == .(round(summary(fit[[i]])$r.squared, 3))))
}
mtext("Anscombe's 4 Regression data sets", outer = TRUE, cex = 1.2)

library(ggplot2)
library(patchwork)
data("anscombe")

form <- sprintf('y%d ~ x%d', 1:4, 1:4)
fit <- lapply(form, lm, data = anscombe)

plot_lm <- function(i) {
 annotate_texts <- c("", "nonlinearity", "outlier", "influential point")
 p <- ggplot(data = anscombe, aes_string(x = paste0("x", i), y = paste0("y", i))) +
 geom_point() +
 geom_abline(intercept = coef(fit[[i]])[1], slope = coef(fit[[i]])[2], color = "red")
 theme_minimal() +
 labs(
 x = substitute(bold(x[a]), list(a = i)), y = substitute(bold(y[b]), list(b = i)),
 title = bquote(bold(R)^2 == .(round(summary(fit[[i]])$r.squared, 3)))
)
 p + annotate("text", x = 12, y = 11, label = annotate_texts[i])
}

}

```

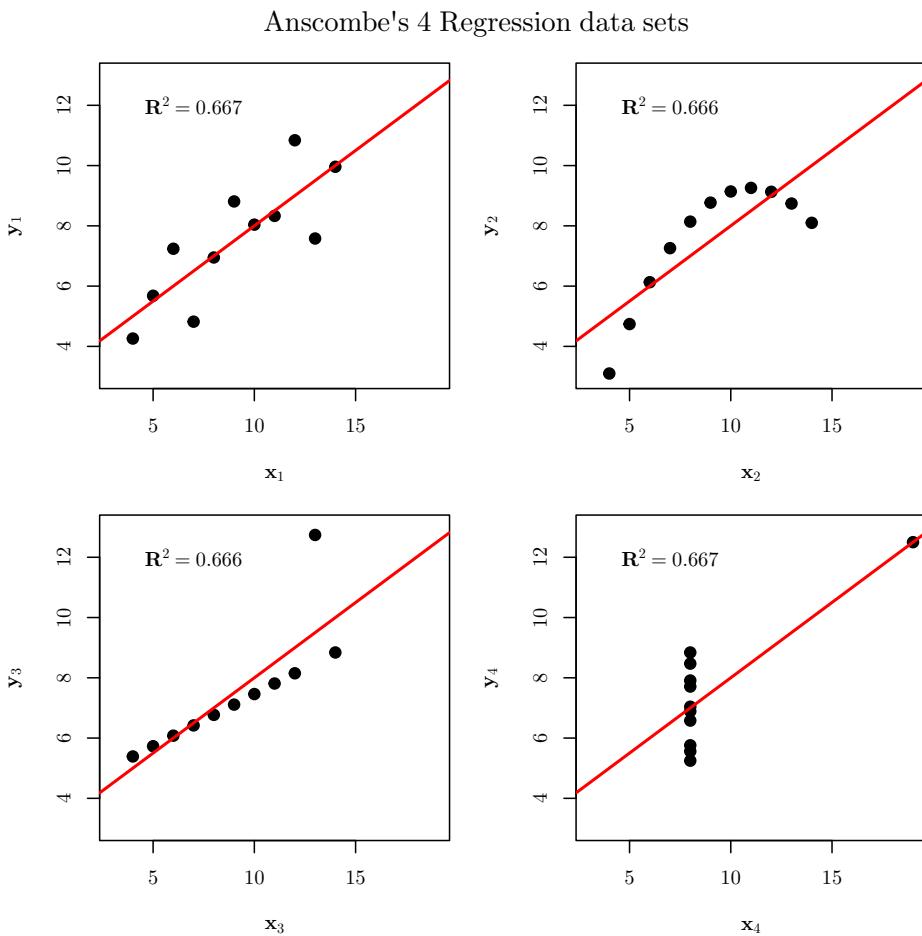


图 20.1: 模型诊断很重要

◎ 黄湘云

```
Reduce("++", lapply(1:4, plot_lm))
```

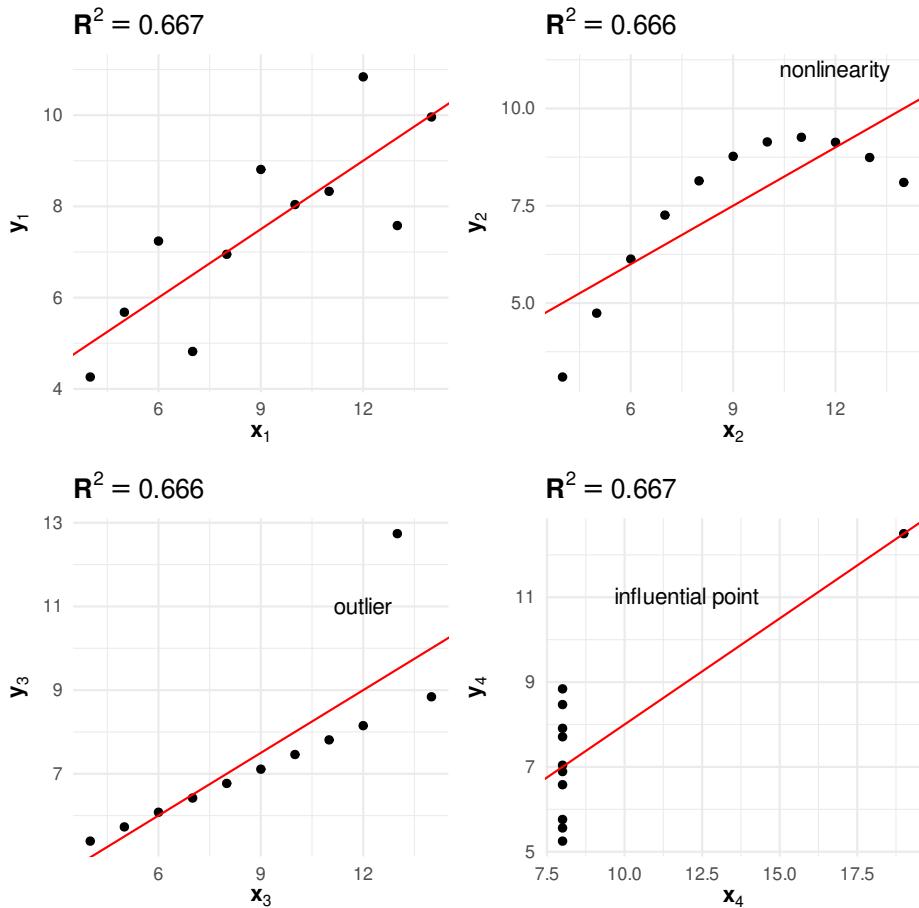


图 20.2: 线性模型可能在欺骗你

## 20.10 1977 年美国人口普查

```
state_data <- data.frame(state.x77, row.names = state.abb)
fit_state <- lm(Life.Exp ~ ., data = state_data)
summary(fit_state)
```

##



```
Call:
lm(formula = Life.Exp ~ ., data = state_data)

Residuals:
Min 1Q Median 3Q Max
-1.48895 -0.51232 -0.02747 0.57002 1.49447

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.094e+01 1.748e+00 40.586 < 2e-16 ***
Population 5.180e-05 2.919e-05 1.775 0.0832 .
Income -2.180e-05 2.444e-04 -0.089 0.9293
Illiteracy 3.382e-02 3.663e-01 0.092 0.9269
Murder -3.011e-01 4.662e-02 -6.459 8.68e-08 ***
HS.Grad 4.893e-02 2.332e-02 2.098 0.0420 *
Frost -5.735e-03 3.143e-03 -1.825 0.0752 .
Area -7.383e-08 1.668e-06 -0.044 0.9649

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7448 on 42 degrees of freedom
Multiple R-squared: 0.7362, Adjusted R-squared: 0.6922
F-statistic: 16.74 on 7 and 42 DF, p-value: 2.534e-10
step(fit_state)
```

## 20.11 石油岩石样品的测量

```
data(rock)
```

多元线性回归

## 20.12 1888 年瑞士生育率分析

1888 年瑞士生育率和社会经济指标数据，各个指标都是百分比的形式，探索性分析

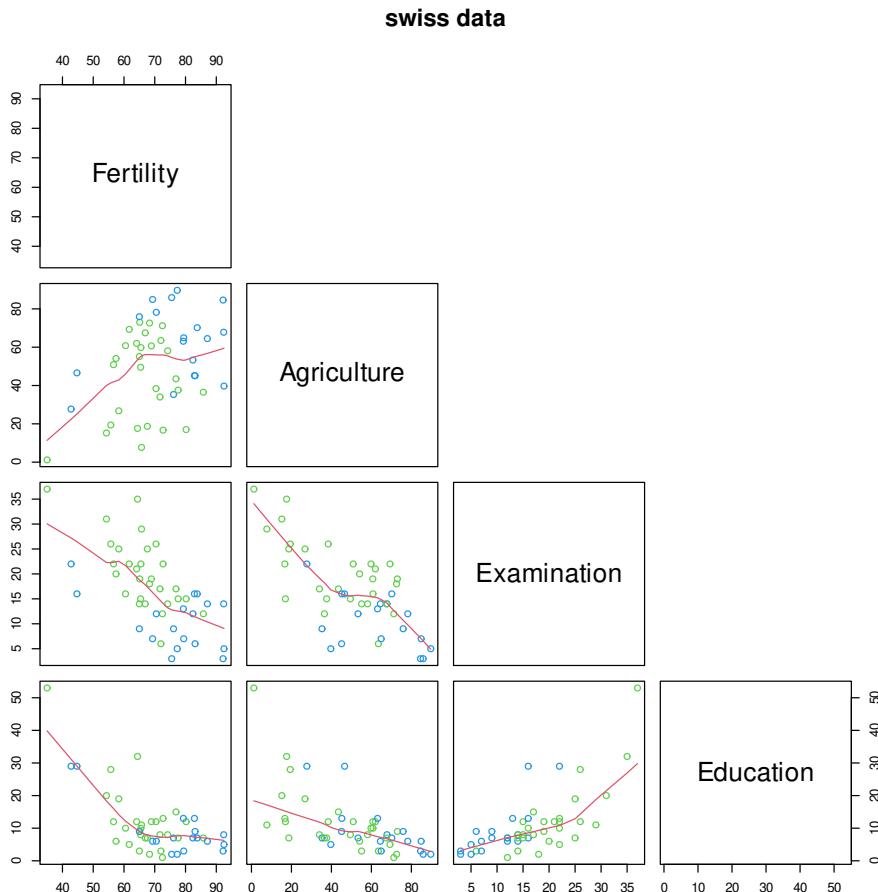


图 20.3: 1888 年瑞士生育率和社会经济指标的关系

```
fit_swiss <- lm(Fertility ~ . - 1, data = swiss)

summary(fit_swiss)

Call:
lm(formula = Fertility ~ . - 1, data = swiss)
```



```

Residuals:
Min 1Q Median 3Q Max
-16.8358 -6.3606 -0.5603 6.0585 23.3203

Coefficients:
Estimate Std. Error t value Pr(>|t|)
Agriculture 0.11100 0.07424 1.495 0.14233
Examination 0.44406 0.31435 1.413 0.16514
Education -0.70674 0.25009 -2.826 0.00719 **
Catholic 0.11707 0.04860 2.409 0.02046 *
Infant.Mortality 2.98366 0.31683 9.417 6.53e-12 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.893 on 42 degrees of freedom
Multiple R-squared: 0.9828, Adjusted R-squared: 0.9807
F-statistic: 478.8 on 5 and 42 DF, p-value: < 2.2e-16
anova(fit_swiss)
```

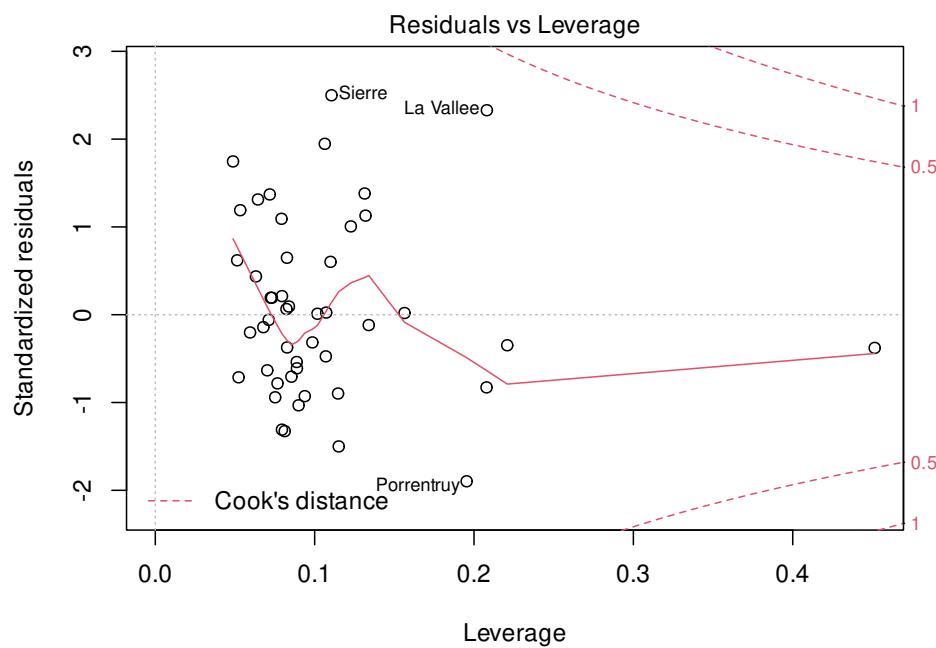
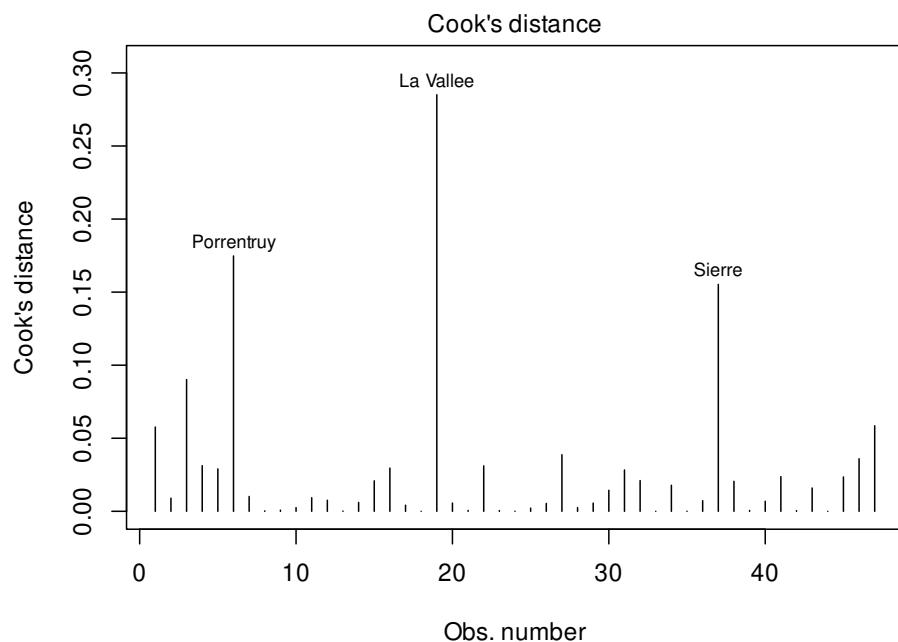
```
Analysis of Variance Table

Response: Fertility
Df Sum Sq Mean Sq F value Pr(>F)
Agriculture 1 204039 204039 2084.6865 < 2.2e-16 ***
Examination 1 16781 16781 171.4556 < 2.2e-16 ***
Education 1 24 24 0.2454 0.6229
Catholic 1 4782 4782 48.8556 1.504e-08 ***
Infant.Mortality 1 8680 8680 88.6858 6.528e-12 ***
Residuals 42 4111 98

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Cook 距离 ?plot.lm

```
par(mar = c(4, 4, 2, 2))
plot(fit_swiss, which = 4, sub.caption = "")
```





```
X <- as.matrix(swiss[, setdiff(names(swiss), "Fertility")])
Y <- as.matrix(swiss[, "Fertility"])
beta 的估计
(beta_hat <- solve(a = crossprod(X, X), b = crossprod(X, Y)))

[1]
Agriculture 0.1110005
Examination 0.4440591
Education -0.7067362
Catholic 0.1170662
Infant.Mortality 2.9836617

Y 的预测 MSE 残差平方和
sigma2_hat <- (t(Y) %*% (diag(rep(1, dim(X)[1])) - X %*% solve(crossprod(X)) %*% t(Y)))
RMSE
sqrt(sigma2_hat)

[1] 9.893187
```

## 20.13 Intercountry Life-Cycle Savings Data 1960-1970

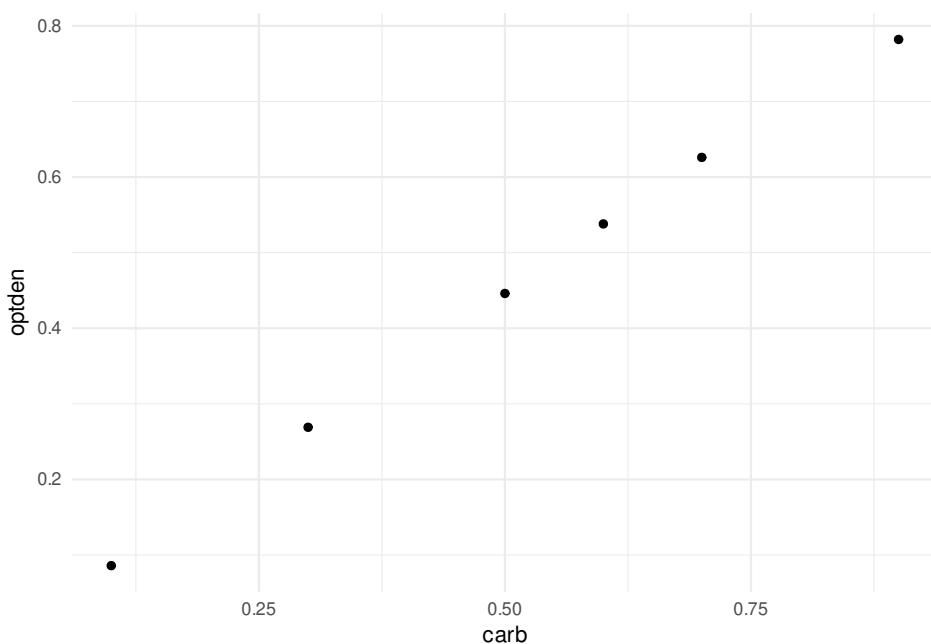
```
data("LifeCycleSavings")
```

## 20.14 Longley's Economic Regression Data 1947-1962

```
data("longley")
```

## 20.15 甲醛的测定

```
ggplot(data = Formaldehyde, aes(x = carb, y = optden)) +
 geom_point() +
 theme_minimal()
```



## 20.16 迈克尔逊光速数据分析

1879 年迈克尔逊光速测量数据，记录了五次实验，每次试验测量 20 次光速，得到表格 20.2

```
reshape(
 data = morley, v.names = "Speed", idvar = "Expt",
 timevar = "Run", direction = "wide", sep = ""
) %>%
 knitr::kable(.,
 caption = "迈克尔逊光速数据",
 row.names = FALSE, col.names = gsub("(Speed)", "", names(.)),
 align = "c"
)
```

数据集 morley 中光速 Speed 已经编码过了，原始观测速度减去了 299000 (km/sec)，为了展示方便

```
ggplot(data = morley, aes(x = Expt, y = Speed, group = Expt)) +
 geom_boxplot() +
```

表 20.2: 迈克尔逊光速数据

Expt	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	850	740	900	1070	930	850	950	980	980	880	1000	980	930	650
2	960	940	960	940	880	800	850	880	900	840	830	790	810	880
3	880	880	880	860	720	720	620	860	970	950	880	910	850	870
4	890	810	810	820	800	770	760	740	750	760	910	920	890	860
5	890	840	780	810	760	810	790	810	820	850	870	870	810	740

```
geom_jitter() +
theme_minimal() +
labs(x = "Expt", y = "Speed (km/sec)")
```

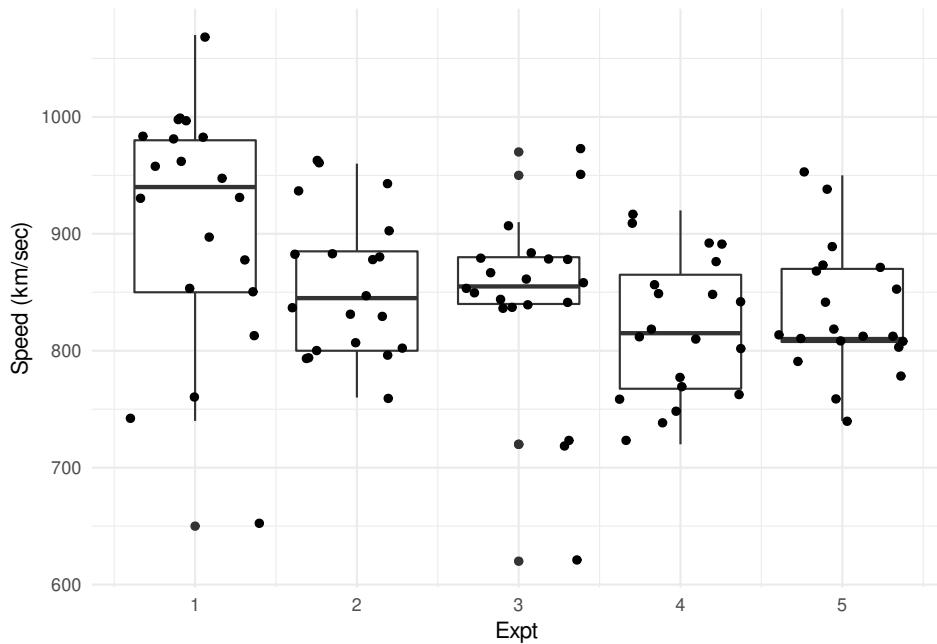


图 20.4: 1879 年迈克尔逊光速实验数据

## 20.17 不同喂食方式对小鸡体重的影响 I

```
ggplot(data = chickwts, aes(x = feed, y = weight, color = feed)) +
 geom_boxplot() +
 geom_jitter() +
 theme_minimal()
```

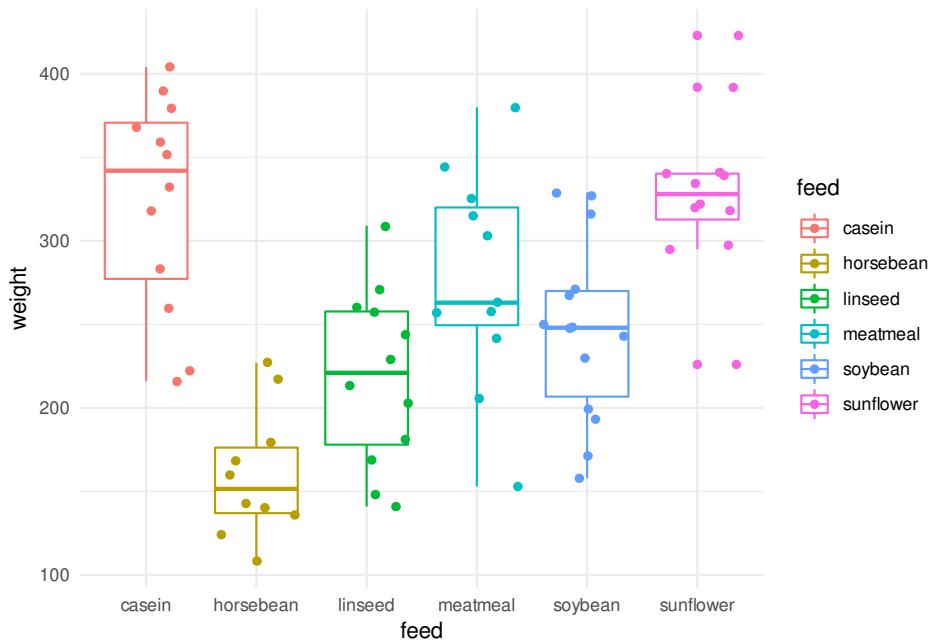
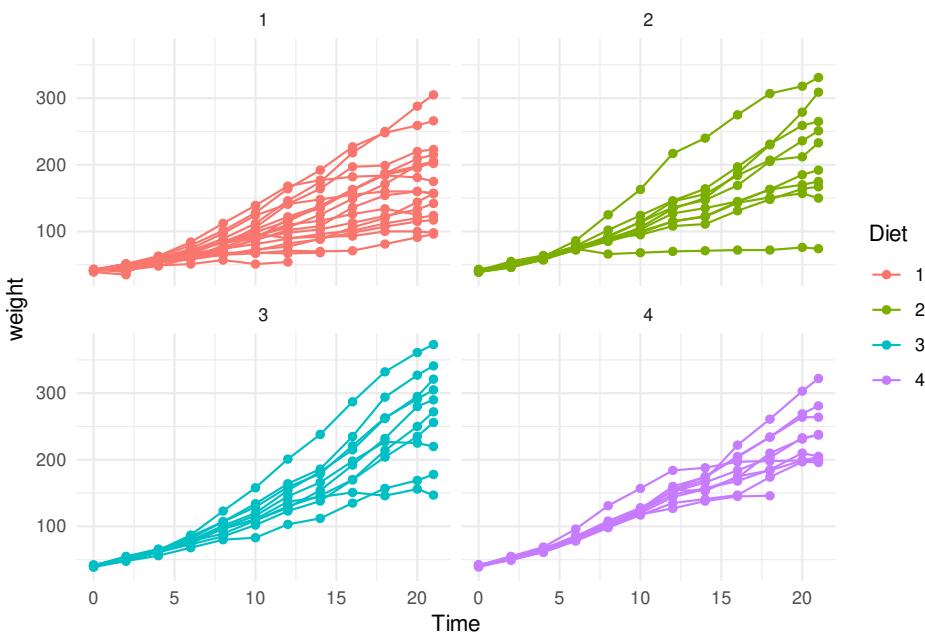


图 20.5: 不同喂食方式对小鸡的影响

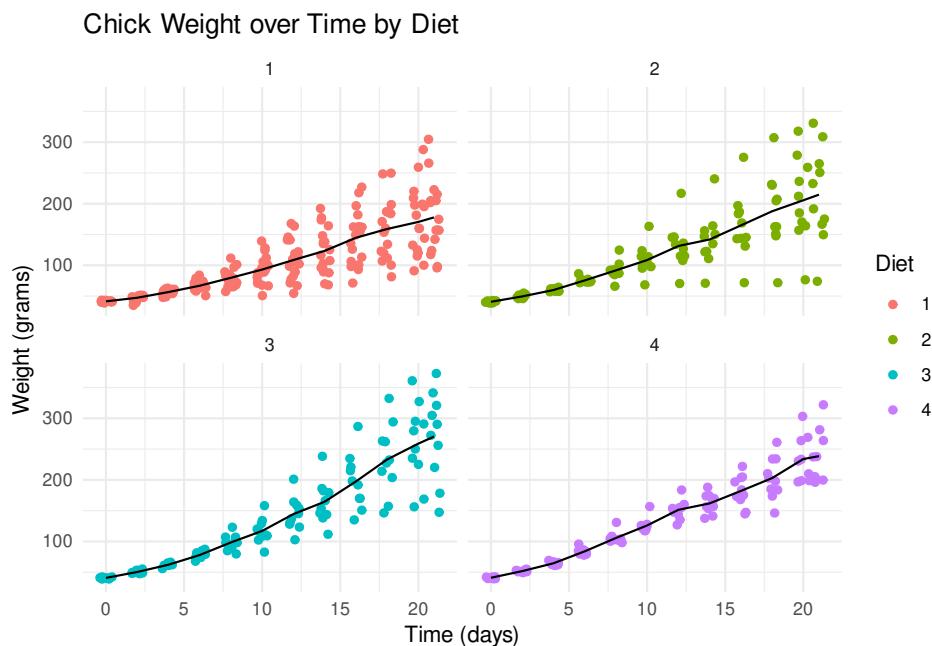
## 20.18 不同喂食方式对小鸡体重的影响 II

```
ggplot(data = ChickWeight, aes(x = Time, y = weight, group = Chick, color = Diet)) +
 geom_point() +
 geom_line() +
 facet_wrap(~Diet) +
 theme_minimal()
```



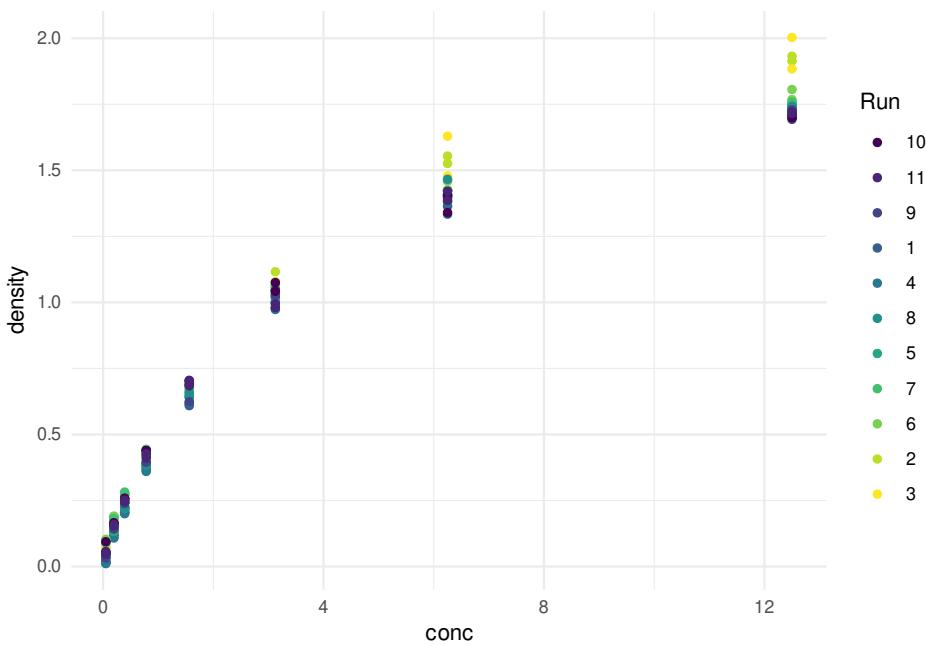
添加趋势线

```
ggplot(data = ChickWeight,
 aes(x = Time, y = weight, group = Diet, colour = Diet)) +
 facet_wrap(~Diet) +
 geom_jitter() +
 stat_summary(fun = "mean", geom = "line", colour = "black") +
 theme_minimal() +
 labs(
 title = "Chick Weight over Time by Diet",
 x = "Time (days)",
 y = "Weight (grams)"
)
```



## 20.19 酶的酶联免疫吸附测定

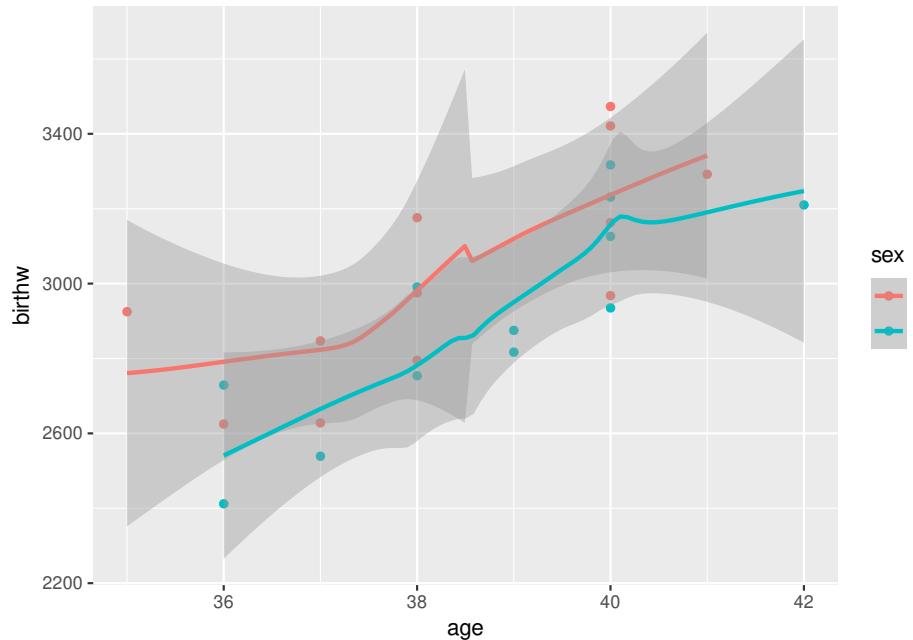
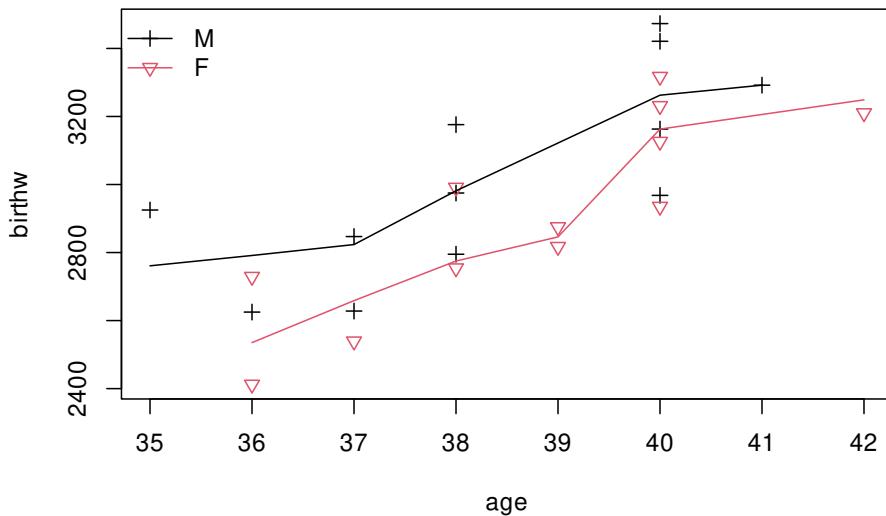
```
ggplot(data = DNase, aes(x= conc, y= density, color = Run)) +
 geom_point() +
 theme_minimal()
```



## 20.20 婴儿的体重随年龄的变化情况

BirthWeight 数据集记录了婴儿的体重随年龄的变化情况，年龄以周为单位计，体重以克为单位计。

### Dobson's Birth Weight Data



性别和年龄两个变量，分别是离散型的分类变量和连续型的变量

# 带截距项和不带截距项

```
summary(l1 <- lm(birthw ~ sex + age), correlation = TRUE)
```



```

Call:
lm(formula = birthw ~ sex + age)

Residuals:
Min 1Q Median 3Q Max
-257.49 -125.28 -58.44 169.00 303.98

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -1610.28 786.08 -2.049 0.0532 .
sexF -163.04 72.81 -2.239 0.0361 *
age 120.89 20.46 5.908 7.28e-06 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 177.1 on 21 degrees of freedom
Multiple R-squared: 0.64, Adjusted R-squared: 0.6057
F-statistic: 18.67 on 2 and 21 DF, p-value: 2.194e-05

Correlation of Coefficients:
(Intercept) sexF
sexF 0.07
age -1.00 -0.12

anova(l1)

Analysis of Variance Table

Response: birthw
Df Sum Sq Mean Sq F value Pr(>F)
sex 1 76163 76163 2.4279 0.1341
age 1 1094940 1094940 34.9040 7.284e-06 ***
Residuals 21 658771 31370

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

湘云  
◎黃

```
与带交互项的模型比较
summary(li <- lm(birthw ~ sex + sex:age), correlation = TRUE)

##
Call:
lm(formula = birthw ~ sex + sex:age)
##
Residuals:
Min 1Q Median 3Q Max
-246.69 -138.11 -39.13 176.57 274.28
##
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -1268.67 1114.64 -1.138 0.268492
sexF -872.99 1611.33 -0.542 0.593952
sexM:age 111.98 29.05 3.855 0.000986 ***
sexF:age 130.40 30.00 4.347 0.000313 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
Residual standard error: 180.6 on 20 degrees of freedom
Multiple R-squared: 0.6435, Adjusted R-squared: 0.59
F-statistic: 12.03 on 3 and 20 DF, p-value: 0.000101
##
Correlation of Coefficients:
(Intercept) sexF sexM:age
sexF -0.69
sexM:age -1.00 0.69
sexF:age 0.00 -0.72 0.00

anova(li, l1)

Analysis of Variance Table

##
Model 1: birthw ~ sex + sex:age
Model 2: birthw ~ sex + age
Res.Df RSS Df Sum of Sq F Pr(>F)
```

```
1 20 652425
2 21 658771 -1 -6346.2 0.1945 0.6639

类似, 只是使用 glm 命令来拟合而已
summary(zi <- glm(birthw ~ sex + age, family = gaussian()))

Call:
glm(formula = birthw ~ sex + age, family = gaussian())

Deviance Residuals:
Min 1Q Median 3Q Max
-257.49 -125.28 -58.44 169.00 303.98

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -1610.28 786.08 -2.049 0.0532 .
sexF -163.04 72.81 -2.239 0.0361 *
age 120.89 20.46 5.908 7.28e-06 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 31370.04)

Null deviance: 1829873 on 23 degrees of freedom
Residual deviance: 658771 on 21 degrees of freedom
AIC: 321.39

Number of Fisher Scoring iterations: 2

anova(zi)

Analysis of Deviance Table

Model: gaussian, link: identity

Response: birthw

```

## Terms added sequentially (first to last)

##

##

	Df	Deviance	Resid. Df	Resid. Dev
## NULL			23	1829873
## sex	1	76163	22	1753711
## age	1	1094940	21	658771

```
summary(z.o4 <- update(zi, subset = -4))
summary(zz <- update(zi, birthw ~ sex + age + sex:age))
```

##

## Call:

```
glm(formula = birthw ~ sex + age + sex:age, family = gaussian())
```

##

## Deviance Residuals:

##	Min	1Q	Median	3Q	Max
##	-246.69	-138.11	-39.13	176.57	274.28

##

## Coefficients:

##	Estimate	Std. Error	t value	Pr(> t )
## (Intercept)	-1268.67	1114.64	-1.138	0.268492
## sexF	-872.99	1611.33	-0.542	0.593952
## age	111.98	29.05	3.855	0.000986 ***
## sexF:age	18.42	41.76	0.441	0.663893

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## (Dispersion parameter for gaussian family taken to be 32621.23)

##

## Null deviance: 1829873 on 23 degrees of freedom

## Residual deviance: 652425 on 20 degrees of freedom

## AIC: 323.16

##

## Number of Fisher Scoring iterations: 2

```
anova(zi, zz)
```

## Analysis of Deviance Table



```

Model 1: birthw ~ sex + age
Model 2: birthw ~ sex + age + sex:age
Resid. Df Resid. Dev Df Deviance
1 21 658771
2 20 652425 1 6346.2
```

## 20.21 火炬松树的生长情况

表 20.3 记录了 14 颗火炬树种子的生长情况

```
reshape(Loblolly, idvar = "Seed", timevar = "age",
 v.names = "height", direction = "wide", sep = "") %>%
knitr:::kable(.,
 caption = "火炬松树的高度 (英尺) 随时间 (年) 的变化",
 row.names = FALSE, col.names = gsub("(height)", "", names(.)),
 align = "c"
)
```

图 20.6 火炬树种子基本决定了树的长势，不同种子预示最后的高度，并且在生长期也是很稳定地生长

```
p <- ggplot(data = Loblolly, aes(x = age, y = height, color = Seed)) +
 geom_point() +
 geom_line() +
 theme_minimal() +
 labs(x = "age (yr)", y = "height (ft)")
p

library(gganimate)
p + transition_reveal(age)
```

表 20.3: 火炬松树的高度 (英尺) 随时间 (年) 的变化

Seed	3	5	10	15	20	25
301	4.51	10.89	28.72	41.74	52.70	60.92
303	4.55	10.92	29.07	42.83	53.88	63.39
305	4.79	11.37	30.21	44.40	55.82	64.10
307	3.91	9.48	25.66	39.07	50.78	59.07
309	4.81	11.20	28.66	41.66	53.31	63.05
311	3.88	9.40	25.99	39.55	51.46	59.64
315	4.32	10.43	27.16	40.85	51.33	60.07
319	4.57	10.57	27.90	41.13	52.43	60.69
321	3.77	9.03	25.45	38.98	49.76	60.28
323	4.33	10.79	28.97	42.44	53.17	61.62
325	4.38	10.48	27.93	40.20	50.06	58.49
327	4.12	9.92	26.54	37.82	48.43	56.81
329	3.93	9.34	26.08	37.79	48.31	56.43
331	3.46	9.05	25.85	39.15	49.12	59.49

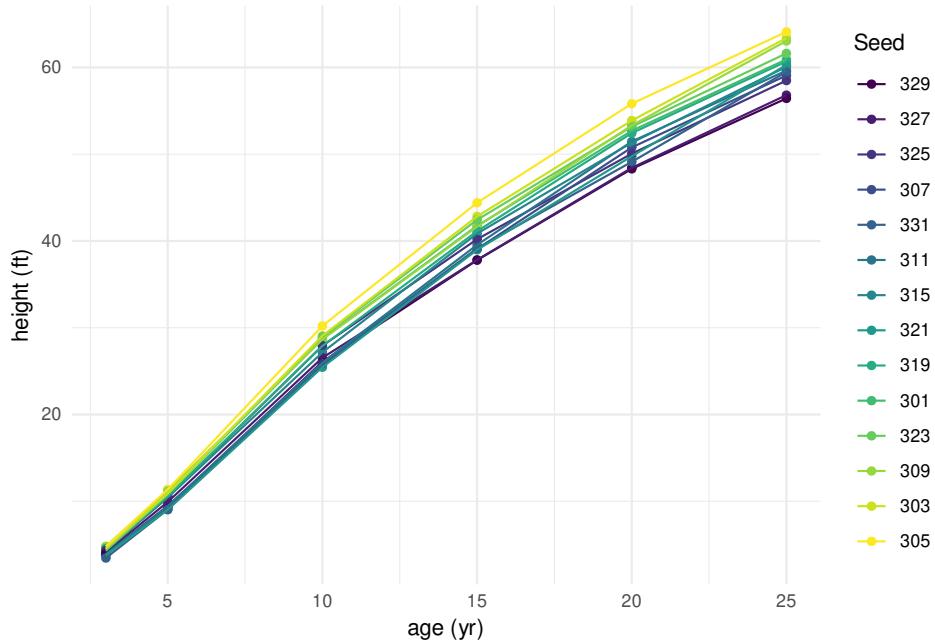
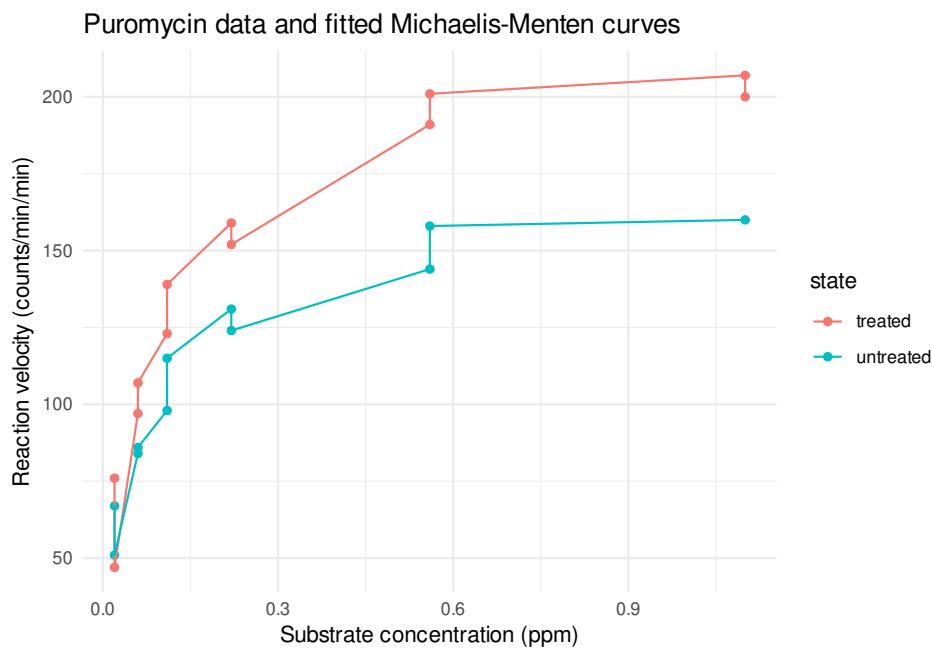


图 20.6: 不同火炬树的生长情况

## 20.22 酶促反应的反应速率

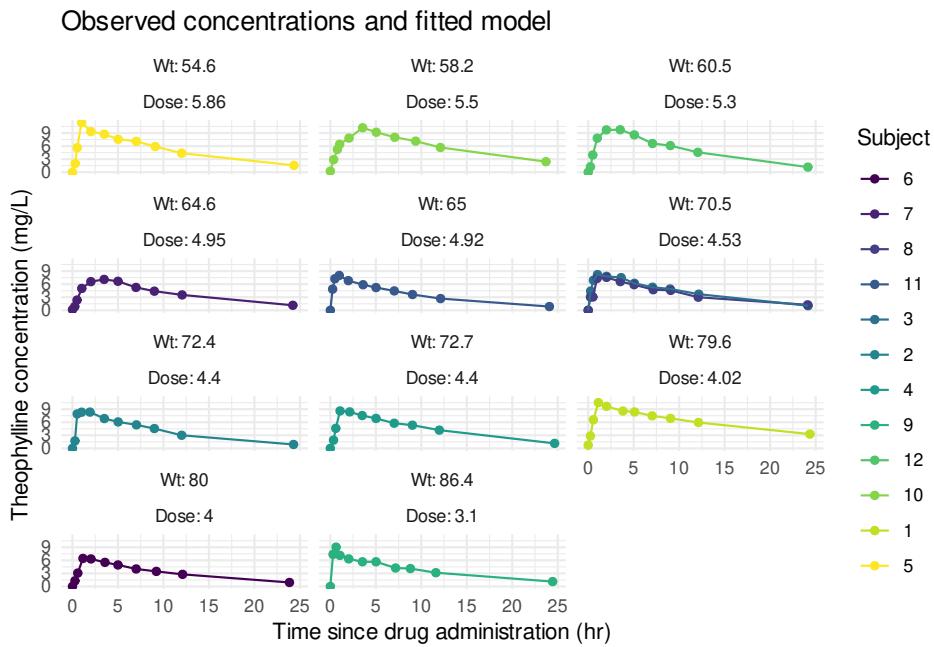
Puromycin 酶促反应的反应速度，模型拟合

```
ggplot(data = Puromycin, aes(x = conc, y = rate, color = state)) +
 geom_point() +
 geom_line() +
 theme_minimal() +
 labs(
 x = "Substrate concentration (ppm)",
 y = "Reaction velocity (counts/min/min)",
 title = "Puromycin data and fitted Michaelis-Menten curves"
)
```



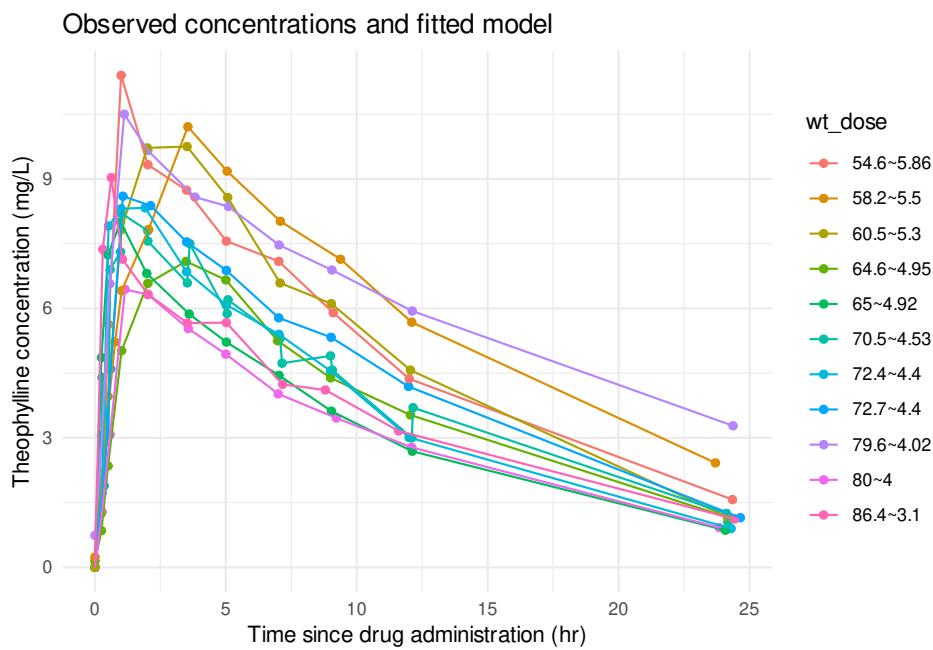
## 20.23 茶碱的药代动力学

```
ggplot(data = Theoph, aes(x = Time, y = conc, color = Subject)) +
 geom_point() +
 geom_line() +
 facet_wrap(Wt ~ Dose, ncol = 3, labeller = "label_both") +
 theme_minimal() +
 labs(
 x = "Time since drug administration (hr)",
 y = "Theophylline concentration (mg/L)",
 title = "Observed concentrations and fitted model"
)
```

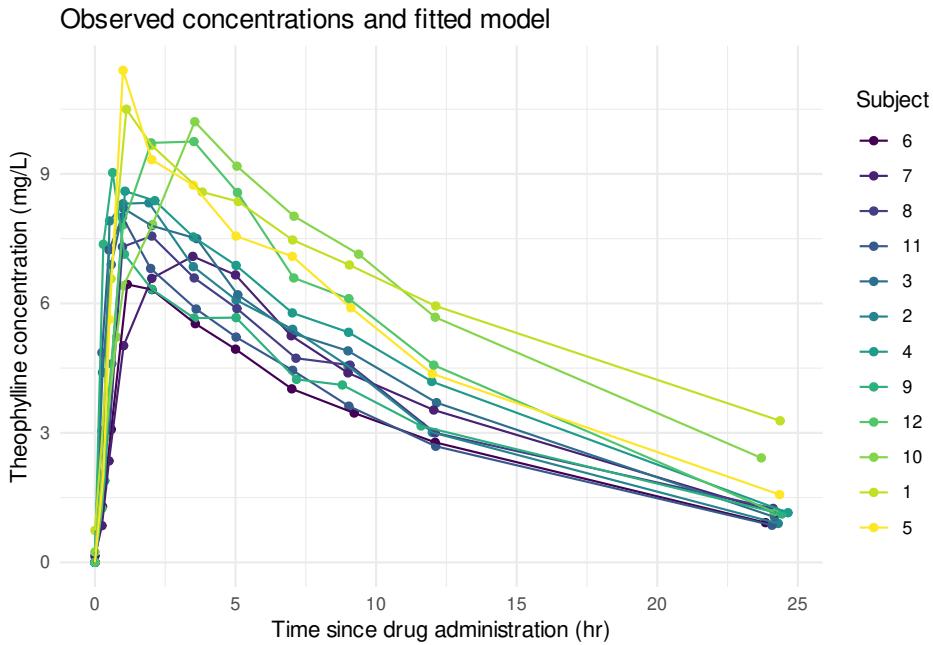


Theoph %>%

```
transform(., wt_dose = paste(Wt, Dose, sep = "~")) %>%
 ggplot(., aes(x = Time, y = conc, color = wt_dose)) +
 geom_point() +
 geom_line() +
 theme_minimal() +
 labs(
 x = "Time since drug administration (hr)",
 y = "Theophylline concentration (mg/L)",
 title = "Observed concentrations and fitted model"
)
```



```
ggplot(data = Theoph, aes(x = Time, y = conc, color = Subject)) +
 geom_point() +
 geom_line() +
 theme_minimal() +
 labs(
 x = "Time since drug administration (hr)",
 y = "Theophylline concentration (mg/L)",
 title = "Observed concentrations and fitted model"
)
```



## 20.24 本章总结

模型永远没完，总是需要自己去构造符合自己需求的模型及其实现，只有自己能够实现，才能在海洋中遨游

This is a bit like asking how should I tweak my sailboat so I can explore the ocean floor.

— Roger Koenker<sup>5</sup>

## 20.25 运行环境

```
sessionInfo()
R version 4.1.0 (2021-05-18)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 20.04.2 LTS
```

<sup>5</sup><https://stat.ethz.ch/pipermail/r-help/2013-May/354311.html>

```

Matrix products: default
BLAS: /usr/lib/x86_64-linux-gnublas/libblas.so.3.9.0
LAPACK: /usr/lib/x86_64-linux-gnulapack/liblapack.so.3.9.0

locale:
[1] LC_CTYPE=en_US.UTF-8 LC_NUMERIC=C
[3] LC_TIME=en_US.UTF-8 LC_COLLATE=en_US.UTF-8
[5] LC_MONETARY=en_US.UTF-8 LC_MESSAGES=en_US.UTF-8
[7] LC_PAPER=en_US.UTF-8 LC_NAME=C
[9] LC_ADDRESS=C LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] stats graphics grDevices utils datasets methods base

other attached packages:
[1] patchwork_1.1.1 extrafont_0.17 gganimate_1.0.7 ggplot2_3.3.5
[5] magrittr_2.0.1

loaded via a namespace (and not attached):
[1] progress_1.2.2 tidyselect_1.1.1 xfun_0.24 purrr_0.3.4
[5] splines_4.1.0 lattice_0.20-44 colorspace_2.0-2 vctrs_0.3.8
[9] generics_0.1.0 htmltools_0.5.1.1 viridisLite_0.4.0 yaml_2.2.1
[13] mgcv_1.8-36 utf8_1.2.2 rlang_0.4.11 pillar_1.6.2
[17] glue_1.4.2 withr_2.4.2 DBI_1.1.1 tweenr_1.0.2
[21] lifecycle_1.0.0 stringr_1.4.0 munsell_0.5.0 gtable_0.3.0
[25] evaluate_0.14 labeling_0.4.2 knitr_1.33 fansi_0.5.0
[29] gifski_1.4.3-1 Rttf2pt1_1.3.9 Rcpp_1.0.7 scales_1.1.1
[33] farver_2.1.0 hms_1.1.0 digest_0.6.27 stringi_1.7.3
[37] bookdown_0.22 dplyr_1.0.7 grid_4.1.0 tools_4.1.0
[41] tibble_3.1.3 crayon_1.4.1 extrafontdb_1.0 pkgconfig_2.0.3
[45] ellipsis_0.3.2 Matrix_1.3-4 prettyunits_1.1.1 assertthat_0.2.1
[49] rmarkdown_2.9 R6_2.5.0 nlme_3.1-152 compiler_4.1.0
```

## 第二十一章 广义线性模型

It's not meant for sampling weights. It's meant for precision weights. How best to include sampling weights in mixed models is a research problem at the moment, but you can rely on getting the wrong answer if you just use the `weights = argument`.

— Thomas Lumley<sup>1</sup>

一般广义线性模型理论参考文献 An Introduction to Generalized Linear Models [Dobson and Barnett, 2018] 和 Generalized Linear Models [McCullagh and Nelder, 1989]，逻辑回归模型主要参考 Applied Logistic Regression [Hosmer and Lemeshow, 2000] 和 Discrete Choice Methods with Simulation [Train, 2009]。

简单线性模型 (Linear Models, 简称 LM)，`stats::lm()` 函数可以拟合线性模型，而一般线性模型 (General Linear Models, 简称 GLM) 允许线性模型方差非齐性、存在相关关系，甚至可以扩展到线性混合效应模型，将线性回归模型，方差分析模型，协方差分析模型统一地看待，一般要采用广义最小二乘 (Generalized Least Squares, 简称 GLS) 拟合，`nlme::gls()` 函数实现广义最小二乘拟合线性模型，类似地，`nlme::gnls()` 函数实现广义最小二乘拟合非线性模型。`glm2::glm2()` 补充 `glm()`，提供更加稳定的拟合方法，适应于 `glm()` 不收敛的情况，而 `fastglm::fastglm()` 主要是加快 `glm()` 求解效率，收敛效果也比 `glm()` 和 `glm2()` 好。

`glmnet` 包是处理广义线性模型的事实标准。其官网见 <https://glmnet.stanford.edu/>，而 `glmnetUtils` 补充公式接口，适用于弹性网络回归，交叉验证筛选  $\alpha$  参数等。`glmpath` 包实现 path-following 算法用于带 L1 正则项的广义线性模型和 Cox 比例风险模型。`Boom` 和 `BoomSpikeSlab` 包实现 MCMC 算法用于 Spike 和 Slab 回归，而 `spikeslab` 包进一步实现预测和变量选择 [Ishwaran and Rao, 2005]。`Cyclops` 包实现 Cyclic coordinate descent 算法用于逻辑回归、泊松回归和生存

<sup>1</sup><https://stat.ethz.ch/pipermail/r-help/2012-January/301501.html>



分析，适用于大规模正则回归 large scale regularized regressions，达到百万级别的观测和特征变量，交叉验证自动选择超参数，独立变量稀疏表示，用剖面似然估计某个变量的置信区间。[plsRglm](#) 包实现偏最小二乘回归方法用于广义线性模型。[biglm](#)、[speedglm](#) 和 [bigReg](#) 用于处理大数据集的回归，求解限制内存的 GLM [biglmm](#)。[cglm](#) 估计带聚类数据的条件 GLM 的回归系数和发散参数。

[MGLM](#) 拟合多个响应变量的广义线性回归模型（多重 GLM）。[robmixglm](#) 响应变量扩展到混合分布的情形，实现稳健 GLM 回归估计。[ClusterBootstrap](#) 实现自主法估计带聚类数据的 GLM。[lcpm](#) 和 [oglmx](#) 处理有序输出的回归。[gmln](#)、[mlogit](#) [Train, 2009] 和 [mnlogit](#) [Hasan et al., 2016] 处理多项逻辑回归。[pscl](#) 包 (Political Science Computational Laboratory) 可以处理贝叶斯 IRT 模型，zero-inflated 零膨胀模型，广义线性模型的拟合优度度量。

## 21.1 介绍

模型结构，模型种类，参数估计办法，相当于综述

响应变量分别服从二项分布、多项分布、对数正态分布、泊松分布、伽马分布

## 21.2 理论基础

分两个段落分别介绍指数族和 GLM

$$f(y; \theta, \phi) = \exp[(a(y)b(\theta) + c(\theta))/f(\phi) + d(y, \phi)]$$

泊松分布 (with  $\lambda \rightarrow \theta, x \rightarrow y$ ) ( $\phi = 1$ ):

$$\begin{aligned} f(y, \theta) &= \exp(-\theta)\theta^y/(y!) \\ &= \exp \left( \underbrace{y}_{a(y)} \underbrace{\log \theta}_{b(\theta)} + \underbrace{(-\theta)}_{c(\theta)} + \underbrace{(-\log(y!))}_{d(y)} \right) \end{aligned} \tag{21.1}$$

### 21.2.1 岭回归

Geometry and properties of generalized ridge regression in high dimensions <http://web.ccs.miami.edu/~hishwaran/papers/IR.conmath2014.pdf>



这篇文章借助三维几何图形展示高维情形下的广义岭回归

### 21.2.2 Lasso

glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models  
<https://glmnet.stanford.edu>

### 21.2.3 最优子集回归

bestglm: Best Subset GLM and Regression Utilities

### 21.2.4 偏最小二乘回归

pls 包 [Mevik and Wehrens, 2007] 实现了偏最小二乘回归 (partial least squares regression, PLS) 和主成分回归 (principal component regression, PCR), 详见主页 <https://mevik.net/work/software/pls.html> 帮助文档的质量较高, 是比较完整全面的。

- several algorithms: the traditional orthogonal scores (NIPALS) PLS algorithm, kernel PLS, wide kernel PLS, Simpls and PCR through svd
- supports multi-response models (aka PLS2)
- flexible cross-validation
- Jackknife variance estimates of regression coefficients
- extensive and flexible plots: scores, loadings, predictions, coefficients, (R)MSEP, R<sup>2</sup>, correlation loadings
- formula interface, modelled after lm(), with methods for predict, print, summary, plot, update, etc.
- extraction functions for coefficients, scores and loadings
- MSEP, RMSEP and R<sup>2</sup> estimates
- multiplicative scatter correction (MSC)

## 21.3 吸烟喝酒和食道癌的关系

存在有序分类数据



酒精的作用 effects of alcohol, tobacco and interaction, age-adjusted 数据集描述  
见 `help(esoph)`

```
head(esoph)
```

```
agegp alcgp tobgp ncases ncontrols
1 25-34 0-39g/day 0-9g/day 0 40
2 25-34 0-39g/day 10-19 0 10
3 25-34 0-39g/day 20-29 0 6
4 25-34 0-39g/day 30+ 0 5
5 25-34 40-79 0-9g/day 0 27
6 25-34 40-79 10-19 0 7
```

```
str(esoph)
```

```
'data.frame': 88 obs. of 5 variables:
$ agegp : Ord.factor w/ 6 levels "25-34"<"35-44"<...: 1 1 1 1 1 1 1 1 ...
$ alcgp : Ord.factor w/ 4 levels "0-39g/day"<"40-79"<...: 1 1 1 2 2 2 2 3 3 ...
$ tobgp : Ord.factor w/ 4 levels "0-9g/day"<"10-19"<...: 1 2 3 4 1 2 3 4 1 2 ...
$ ncases : num 0 0 0 0 0 0 0 0 0 ...
$ ncontrols: num 40 10 6 5 27 7 4 7 2 1 ...

p1 <- ggplot(data = esoph, aes(x = agegp, y = ncases / ncontrols, color = agegp)) +
 geom_boxplot(show.legend = FALSE) +
 geom_jitter(show.legend = FALSE) +
 theme_minimal()

p2 <- ggplot(data = esoph, aes(x = alcgp, y = ncases / ncontrols, color = alcgp)) +
 geom_boxplot(show.legend = FALSE) +
 geom_jitter(show.legend = FALSE) +
 theme_minimal()

p3 <- ggplot(data = esoph, aes(x = tobgp, y = ncases / ncontrols, color = tobgp)) +
 geom_boxplot(show.legend = FALSE) +
 geom_jitter(show.legend = FALSE) +
 theme_minimal()

bottom_row <- plot_grid(p2, p3, labels = c('B', 'C'), label_size = 12)

Warning: Removed 12 rows containing non-finite values (stat_boxplot).
```

```
Warning: Removed 12 rows containing missing values (geom_point).

Warning: Removed 12 rows containing non-finite values (stat_boxplot).

Warning: Removed 12 rows containing missing values (geom_point).
plot_grid(p1, bottom_row, labels = c('A', ''), label_size = 12, ncol = 1)

Warning: Removed 12 rows containing non-finite values (stat_boxplot).

Warning: Removed 12 rows containing missing values (geom_point).
```

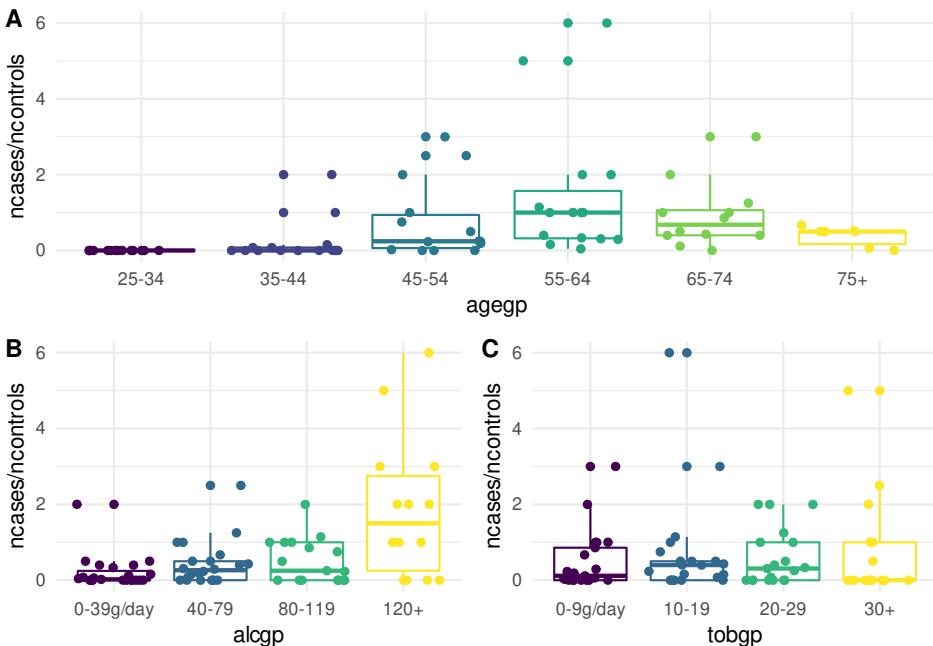


图 21.1: 吸烟喝酒和食道癌的关系

```
fit_esoph_glm <- glm(cbind(ncases, ncontrols) ~ agegp + tobgp * alcgp,
 data = esoph, family = binomial(link = "logit"))

library(Rcpp)
fit_esoph_brm <- brms:::brm(ncases | trials(ncases + ncontrols) ~ agegp + tobgp * a
```



## 21.4 自然流产和人工流产后的不育

```

help(infert)
head(infert)

education age parity induced case spontaneous stratum pooled.stratum
1 0-5yrs 26 6 1 1 2 1 3
2 0-5yrs 42 1 1 1 0 2 1
3 0-5yrs 39 6 2 1 0 3 4
4 0-5yrs 34 4 2 1 0 4 2
5 6-11yrs 35 3 1 1 1 5 32
6 6-11yrs 36 4 2 1 1 6 36

str(infert)

'data.frame': 248 obs. of 8 variables:
$ education : Factor w/ 3 levels "0-5yrs","6-11yrs",...: 1 1 1 1 2 2 2 2 2 ...
$ age : num 26 42 39 34 35 36 23 32 21 28 ...
$ parity : num 6 1 6 4 3 4 1 2 1 2 ...
$ induced : num 1 1 2 2 1 2 0 0 0 0 ...
$ case : num 1 1 1 1 1 1 1 1 1 1 ...
$ spontaneous: num 2 0 0 0 1 1 0 0 1 0 ...
$ stratum : int 1 2 3 4 5 6 7 8 9 10 ...
$ pooled.stratum: num 3 1 4 2 32 36 6 22 5 19 ...

```

存在无序分类变量

```

infert_glm_1 <- glm(case ~ spontaneous + induced,
 data = infert, family = binomial()
)
summary(infert_glm_1)

##
Call:
glm(formula = case ~ spontaneous + induced, family = binomial(),
data = infert)
##
Deviance Residuals:
Min 1Q Median 3Q Max

```

```
-1.6678 -0.8360 -0.5772 0.9030 1.9362
##
Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.7079 0.2677 -6.380 1.78e-10 ***
spontaneous 1.1972 0.2116 5.657 1.54e-08 ***
induced 0.4181 0.2056 2.033 0.042 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
(Dispersion parameter for binomial family taken to be 1)
##
Null deviance: 316.17 on 247 degrees of freedom
Residual deviance: 279.61 on 245 degrees of freedom
AIC: 285.61
##
Number of Fisher Scoring iterations: 4
```

考慮其他潛在的因素

```
infert_glm_2 <- glm(case ~ age + parity + education + spontaneous + induced,
 data = infert, family = binomial()
)
summary(infert_glm_2)

##
Call:
glm(formula = case ~ age + parity + education + spontaneous +
induced, family = binomial(), data = infert)
##
Deviance Residuals:
Min 1Q Median 3Q Max
-1.7603 -0.8162 -0.4956 0.8349 2.6536
##
Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.14924 1.41220 -0.814 0.4158
age 0.03958 0.03120 1.269 0.2046
```

云 湘 黄

```

parity -0.82828 0.19649 -4.215 2.49e-05 ***
education6-11yrs -1.04424 0.79255 -1.318 0.1876
education12+ yrs -1.40321 0.83416 -1.682 0.0925 .
spontaneous 2.04591 0.31016 6.596 4.21e-11 ***
induced 1.28876 0.30146 4.275 1.91e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
(Dispersion parameter for binomial family taken to be 1)
##
Null deviance: 316.17 on 247 degrees of freedom
Residual deviance: 257.80 on 241 degrees of freedom
AIC: 271.8
##
Number of Fisher Scoring iterations: 4

```

实际上应该使用条件逻辑回归，调用 **survival** 包

```

library(survival)
infert_glm_3 <- clogit(case ~ spontaneous + induced + strata(stratum),
 data = infert
)
summary(infert_glm_3)

```

```

Call:
coxph(formula = Surv(rep(1, 248L), case) ~ spontaneous + induced +
strata(stratum), data = infert, method = "exact")
##
n= 248, number of events= 83
##
coef exp(coef) se(coef) z Pr(>|z|)
spontaneous 1.9859 7.2854 0.3524 5.635 1.75e-08 ***
induced 1.4090 4.0919 0.3607 3.906 9.38e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
exp(coef) exp(-coef) lower .95 upper .95
spontaneous 7.285 0.1373 3.651 14.536

```

```
induced 4.092 0.2444 2.018 8.298
##
Concordance= 0.776 (se = 0.044)
Likelihood ratio test= 53.15 on 2 df, p=3e-12
Wald test = 31.84 on 2 df, p=1e-07
Score (logrank) test = 48.44 on 2 df, p=3e-11
```

## 21.5 细菌数据集

流感嗜血杆菌的细菌与中耳炎患儿

```
data(bacteria, package = "MASS")

惩罚拟似然
fit_glmpql <- MASS::glmmPQL(y ~ trt + I(week > 2),
 random = ~ 1 | ID, verbose = FALSE,
 family = binomial, data = bacteria
)
summary(fit_glmpql)

Linear mixed-effects model fit by maximum likelihood
Data: bacteria
AIC BIC logLik
NA NA NA
##
Random effects:
Formula: ~1 | ID
(Intercept) Residual
StdDev: 1.410637 0.7800511
##
Variance function:
Structure: fixed weights
Formula: ~invwt
##
Fixed effects: y ~ trt + I(week > 2)
Value Std.Error DF t-value p-value
(Intercept) 3.412014 0.5185033 169 6.580506 0.0000
trtdrug -1.247355 0.6440635 47 -1.936696 0.0588
```

云  
湘  
黄  
◎

```

trtdrug+ -0.754327 0.6453978 47 -1.168779 0.2484
I(week > 2)TRUE -1.607257 0.3583379 169 -4.485311 0.0000
Correlation:
(Intr) trtdrg trtdr+
trtdrug -0.598
trtdrug+ -0.571 0.460
I(week > 2)TRUE -0.537 0.047 -0.001
##
Standardized Within-Group Residuals:
Min Q1 Med Q3 Max
-5.1985361 0.1572336 0.3513075 0.4949482 1.7448845
##
Number of Observations: 220
Number of Groups: 50

拉普拉斯近似
fit_glmer <- lme4::glmer(y ~ trt + I(week > 2) + (1 | ID),
 family = binomial, data = bacteria
)
summary(fit_glmer)

```

```

Generalized linear mixed model fit by maximum likelihood (Laplace
Approximation) [glmerMod]
Family: binomial (logit)
Formula: y ~ trt + I(week > 2) + (1 | ID)
Data: bacteria
##
AIC BIC logLik deviance df.resid
202.3 219.2 -96.1 192.3 215
##
Scaled residuals:
Min 1Q Median 3Q Max
-4.5615 0.1359 0.3022 0.4217 1.1276
##
Random effects:
Groups Name Variance Std.Dev.
ID (Intercept) 1.543 1.242

```



```
Number of obs: 220, groups: ID, 50
##
Fixed effects:
Estimate Std. Error z value Pr(>|z|)
(Intercept) 3.5479 0.6958 5.099 3.41e-07 ***
trtdrug -1.3667 0.6770 -2.019 0.043516 *
trtdrug+ -0.7826 0.6831 -1.146 0.251926
I(week > 2)TRUE -1.5985 0.4759 -3.359 0.000783 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
Correlation of Fixed Effects:
(Intr) trtdrg trtdr+
trtdrug -0.593
trtdrug+ -0.537 0.487
I(wk>2)TRUE -0.656 0.126 0.064
```

## 21.6 研究婴儿出生体重低的相关危险因素

在线性回归的基础上，响应变量是离散的类别，且无序 [Hasan et al., 2016]

birthwt 数据是 1986 年在马萨诸塞州斯普林菲尔德的 Baystate 医疗中心收集的，用于研究婴儿出生体重低的相关危险因素

```
加载数据
library(MASS)
data(birthwt, package = "MASS")
查看 birthwt 数据集 `help(birthwt)`
head(birthwt)

low age lwt race smoke ptl ht ui ftv bwt
85 0 19 182 2 0 0 0 1 0 2523
86 0 33 155 3 0 0 0 0 3 2551
87 0 20 105 1 1 0 0 0 1 2557
88 0 21 108 1 1 0 0 1 2 2594
89 0 18 107 1 1 0 0 1 0 2600
91 0 21 124 3 0 0 0 0 0 2622
```

黄湘云

```
str(birthwt)

'data.frame': 189 obs. of 10 variables:
$ low : int 0 0 0 0 0 0 0 0 0 ...
$ age : int 19 33 20 21 18 21 22 17 29 26 ...
$ lwt : int 182 155 105 108 107 124 118 103 123 113 ...
$ race : int 2 3 1 1 1 3 1 3 1 1 ...
$ smoke: int 0 0 1 1 1 0 0 0 1 1 ...
$ ptl : int 0 0 0 0 0 0 0 0 0 ...
$ ht : int 0 0 0 0 0 0 0 0 0 ...
$ ui : int 1 0 0 1 1 0 0 0 0 0 ...
$ ftv : int 0 3 1 2 0 0 1 1 1 0 ...
$ bwt : int 2523 2551 2557 2594 2600 2622 2637 2663 2665 ...
```

`low` 表示婴儿出生体重小于 2.5kg, `age` 表示母亲的年龄 (年), `lwt` 母亲最后一次月经期间的体重 (磅), `race` 母亲的种族 (1 = 白人, 2 = 黑人, 3 = 其他)。, `smoke` 怀孕期间的吸烟状况, `ptl` 以前早产的次数, `ht` 高血压病史, `ui` 子宫过敏, `ftv` 妊娠头三个月的医生就诊次数, `bwt` 出生体重 (克)

```
with(birthwt, tapply(lwt, ui, var))

0 1
940.8472 783.7196

t.test(lwt ~ ui, data = birthwt, var.equal = TRUE)

##
Two Sample t-test
##
data: lwt by ui
t = 2.1138, df = 187, p-value = 0.03586
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to zero
95 percent confidence interval:
0.8753389 25.3544748
sample estimates:
mean in group 0 mean in group 1
131.7578 118.6429

t.test(lwt ~ ui, data = birthwt)
```

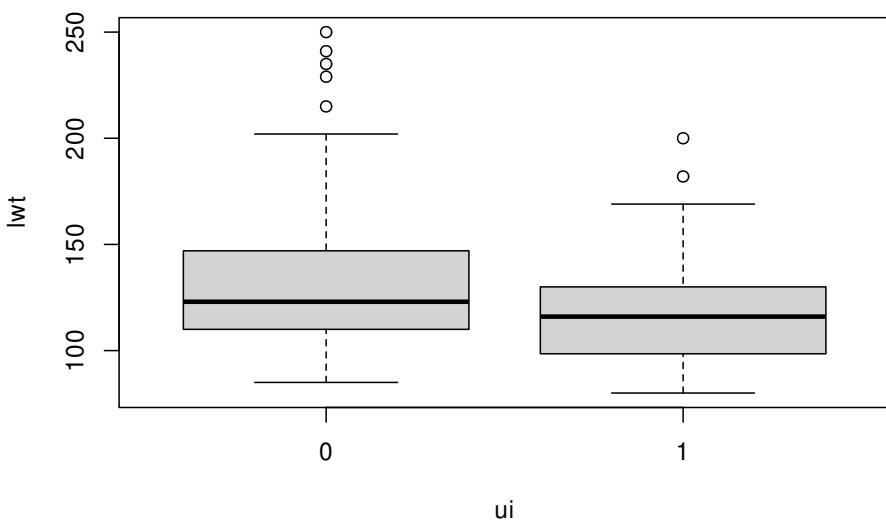
```

Welch Two Sample t-test

data: lwt by ui
t = 2.2547, df = 39.163, p-value = 0.02982
alternative hypothesis: true difference in means between group 0 and group 1 is not
95 percent confidence interval:
1.351128 24.878685
sample estimates:
mean in group 0 mean in group 1
131.7578 118.6429

birthwt$ui <- as.factor(birthwt$ui)
library(lattice)
bwplot(lwt ~ ui, data = birthwt, pch = "|")

boxplot(lwt ~ ui, data = birthwt)
```



```
重新编码，数据预处理，方便代入模型
bwt <- with(birthwt, {
 race <- factor(race, labels = c("white", "black", "other"))})
```



```

ptd <- factor(ptl > 0)
ftv <- factor(ftv)
levels(ftv)[-c(1:2)] <- "2+" # 除了前两个水平外，其余的都编码为 2+
data.frame(
 low = factor(low), age, lwt, race, smoke = (smoke > 0),
 ptd, ht = (ht > 0), ui = (ui > 0), ftv
)
}

查看编码后的数据
head(bwt)

low age lwt race smoke ptd ht ui ftv
1 0 19 182 black FALSE FALSE FALSE TRUE 0
2 0 33 155 other FALSE FALSE FALSE FALSE 2+
3 0 20 105 white TRUE FALSE FALSE FALSE 1
4 0 21 108 white TRUE FALSE FALSE TRUE 2+
5 0 18 107 white TRUE FALSE FALSE TRUE 0
6 0 21 124 other FALSE FALSE FALSE FALSE 0

str(bwt)

'data.frame': 189 obs. of 9 variables:
$ low : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 ...
$ age : int 19 33 20 21 18 21 22 17 29 26 ...
$ lwt : int 182 155 105 108 107 124 118 103 123 113 ...
$ race : Factor w/ 3 levels "white","black",...: 2 3 1 1 1 3 1 3 1 1 ...
$ smoke: logi FALSE FALSE TRUE TRUE TRUE FALSE ...
$ ptd : Factor w/ 2 levels "FALSE","TRUE": 1 1 1 1 1 1 1 1 1 1 ...
$ ht : logi FALSE FALSE FALSE FALSE FALSE FALSE ...
$ ui : logi TRUE FALSE FALSE TRUE TRUE FALSE ...
$ ftv : Factor w/ 3 levels "0","1","2+": 1 3 2 3 1 1 2 2 2 1 ...

广义线性模型拟合，二项逻辑回归，响应变量为婴儿出生的体重，以 2.5kg 为界，

它被编码成二分类变量 0 或 1

options(contrasts = c("contr.treatment", "contr.poly"))
glm(formula = low ~ ., family = binomial, data = bwt)

##

```

```
Call: glm(formula = low ~ ., family = binomial, data = bwt)
##
Coefficients:
(Intercept) age lwt raceblack raceother smokeTRUE
0.82302 -0.03723 -0.01565 1.19241 0.74068 0.75553
ptdTRUE htTRUE uiTRUE ftv1 ftv2+
1.34376 1.91317 0.68020 -0.43638 0.17901
##
Degrees of Freedom: 188 Total (i.e. Null); 178 Residual
Null Deviance: 234.7
Residual Deviance: 195.5 AIC: 217.5
```

多项逻辑回归

```
library(nnet)
(bwt.mu <- multinom(formula = low ~ ., data = bwt))

weights: 12 (11 variable)
initial value 131.004817
iter 10 value 98.029803
final value 97.737759
converged

Call:
multinom(formula = low ~ ., data = bwt)
##
Coefficients:
(Intercept) age lwt raceblack raceother smokeTRUE
0.82320102 -0.03723828 -0.01565359 1.19240391 0.74065606 0.75550487
ptdTRUE htTRUE uiTRUE ftv1 ftv2+
1.34375901 1.91320116 0.68020207 -0.43638470 0.17900392
##
Residual Deviance: 195.4755
AIC: 217.4755

summary(bwt.mu)

Call:
multinom(formula = low ~ ., data = bwt)
##
```



```
Coefficients:
Values Std. Err.
(Intercept) 0.82320102 1.24476766
age -0.03723828 0.03870437
lwt -0.01565359 0.00708079
raceblack 1.19240391 0.53598076
raceother 0.74065606 0.46176615
smokeTRUE 0.75550487 0.42503626
ptdTRUE 1.34375901 0.48063449
htTRUE 1.91320116 0.72076133
uiTRUE 0.68020207 0.46434974
ftv1 -0.43638470 0.47941107
ftv2+ 0.17900392 0.45639129
##
Residual Deviance: 195.4755
AIC: 217.4755
```

计算 Z 分数和 P 值

```
z <- summary(bwt.mu)$coefficients / summary(bwt.mu)$standard.errors
z
```

```
(Intercept) age lwt raceblack raceother smokeTRUE
0.6613291 -0.9621210 -2.2107121 2.2247140 1.6039635 1.7775069
ptdTRUE htTRUE uiTRUE ftv1 ftv2+
2.7958023 2.6544170 1.4648486 -0.9102516 0.3922159
```

```
p <- (1 - pnorm(abs(z), 0, 1)) * 2
p
```

```
(Intercept) age lwt raceblack raceother smokeTRUE
0.508401310 0.335988847 0.027055777 0.026100443 0.108722092 0.075484881
ptdTRUE htTRUE uiTRUE ftv1 ftv2+
0.005177106 0.007944557 0.142962228 0.362689827 0.694898695
```

模型解释

## 21.7 哥本哈根住房状况调查

响应变量是离散类别，且存在强弱，等级，大小之分

调用函数 MASS::polr()

数据集 housing 哥本哈根住房状况调查中的次数分布表，Sat 住户对目前居住环境的满意程度，是一个有序的因子变量，Infl 住户对物业管理的感知影响程度，Type 租赁住宿类型，如塔楼、中庭、公寓、露台，Cont 联系居民可与其他居民联系(低、高)，Freq 每个类中的居民人数，调查的人数

```
data("housing", package = "MASS")
```

```
查看数据 help(housing)
```

```
head(housing)
```

```
Sat Infl Type Cont Freq
1 Low Low Tower Low 21
2 Medium Low Tower Low 21
3 High Low Tower Low 28
4 Low Medium Tower Low 34
5 Medium Medium Tower Low 22
6 High Medium Tower Low 36
```

```
str(housing)
```

```
'data.frame': 72 obs. of 5 variables:
$ Sat : Ord.factor w/ 3 levels "Low"<"Medium"<...: 1 2 3 1 2 3 1 2 3 1 ...
$ Infl: Factor w/ 3 levels "Low","Medium",...: 1 1 1 2 2 2 3 3 3 1 ...
$ Type: Factor w/ 4 levels "Tower","Apartment",...: 1 1 1 1 1 1 1 1 1 2 ...
$ Cont: Factor w/ 2 levels "Low","High": 1 1 1 1 1 1 1 1 1 1 ...
$ Freq: int 21 21 28 34 22 36 10 11 36 61 ...
```

居民对居住环境满意度 Sat 三个等级的有序回归

```
options(contrasts = c("contr.treatment", "contr.poly"))
```

```
house.plr <- MASS::polr(Sat ~ Infl + Type + Cont, weights = Freq, data = housing)
```

```
Call:
```

```
MASS::polr(formula = Sat ~ Infl + Type + Cont, data = housing,
```

```
weights = Freq)
```

```


Coefficients:
InflMedium InflHigh TypeApartment TypeAtrium TypeTerrace
0.5663937 1.2888191 -0.5723501 -0.3661866 -1.0910149
ContHigh
0.3602841
##
Intercepts:
Low|Medium Medium|High
-0.4961353 0.6907083
##
Residual Deviance: 3479.149
AIC: 3495.149

```

再计算一下 P 值，置信区间

```

ctable <- coef(summary(house.plr))
p <- pnorm(abs(ctable[, "t value"]), lower.tail = FALSE) * 2
ctable <- cbind(ctable, "p value" = p)
confidence intervals 计算置信区间
ci <- confint(house.plr)
exp(coef(house.plr))

```

```

InflMedium InflHigh TypeApartment TypeAtrium TypeTerrace
1.7619017 3.6284990 0.5641979 0.6933734 0.3358754
ContHigh
1.4337368

```

```

OR and CI
exp(cbind(OR = coef(house.plr), ci))

```

```

OR 2.5 % 97.5 %
InflMedium 1.7619017 1.4356845 2.1639915
InflHigh 3.6284990 2.8319659 4.6626461
TypeApartment 0.5641979 0.4462124 0.7121941
TypeAtrium 0.6933734 0.5114084 0.9398410
TypeTerrace 0.3358754 0.2492277 0.4514276
ContHigh 1.4337368 1.1892931 1.7296674

```

模型解释



参考文档 `help(housing)` 包含泊松回归、多项回归、比例风险模型，以及 <https://www.analyticsvidhya.com/blog/2016/02/multinomial-ordinal-logistic-regression/>

好好看文档 `help(housing)` 和对应的参考书籍，把原理弄清楚

有序因子变量是如何实现编码的

## 21.8 癫痫病发作次数

纵向数据 [Thall and Vail, 1990]，考虑了过度发散 overdispersion 异方差 heteroscedasticity 观测不独立

数据集 `epil` 记录癫痫发作的次数及病人的特征，下面是数据建模分析过程

```
data(epil, package = "MASS")
fit_glm_epil <- glm(y ~ lbase * trt + lage + V4,
 family = poisson,
 data = epil
)
summary(fit_glm_epil)

fit_glmm_epil<- MASS::glmmPQL(y ~ lbase * trt + lage + V4,
 random = ~ 1 | subject,
 family = poisson, data = epil
)
summary(fit_glmm_epil)

fit_glmm_lme4 <- lme4::glmer(y ~ lbase * trt + lage + V4 + (1 | subject),
 family = poisson, data = epil
)
summary(fit_glmm_lme4)

fit_glmm_glmmtmb <- glmmTMB::glmmTMB(y ~ lbase * trt + lage + V4 + (1 | subject),
 data = epil, family = poisson, REML = TRUE
) # REML 估计
summary(fit_glmm_glmmtmb)
```



```
https://github.com/drizopoulos/GLMMadaptive
fit_glmm_glmmadaptive <- GLMMadaptive::mixed_model(
 fixed = y ~ lbase * trt + lage + V4,
 random = ~ 1 | subject, data = epil,
 family = poisson()
)
summary(fit_glmm_glmmadaptive)
```

## 21.9 对数线性模型

当响应变量  $Y$  服从对数正态分布的时候，广义线性模型具化为对数线性模型，**gllm** 包 [Espeland and Hui, 1987]

## 21.10 泊松回归模型

加载数据

```
data(beall.webworms, package = "agridat")
```

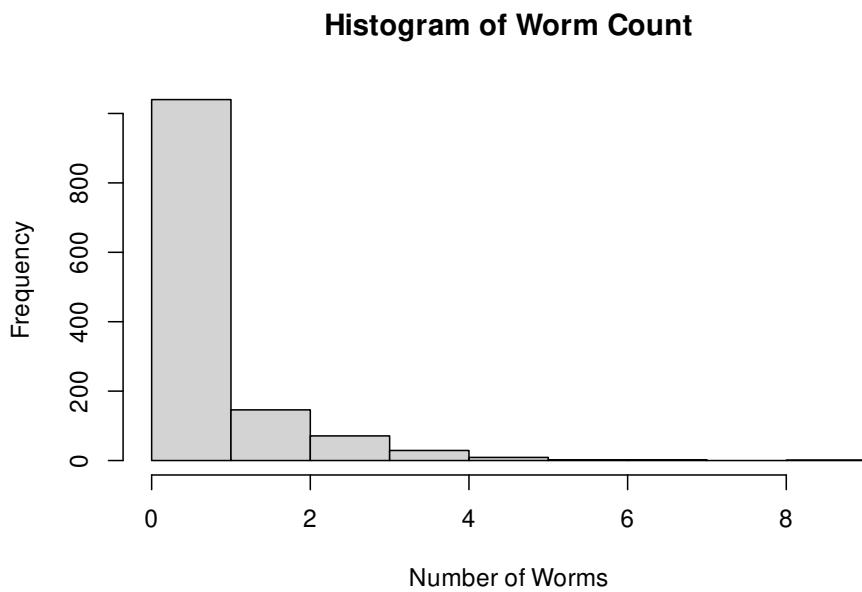
查看数据

```
head(beall.webworms)
```

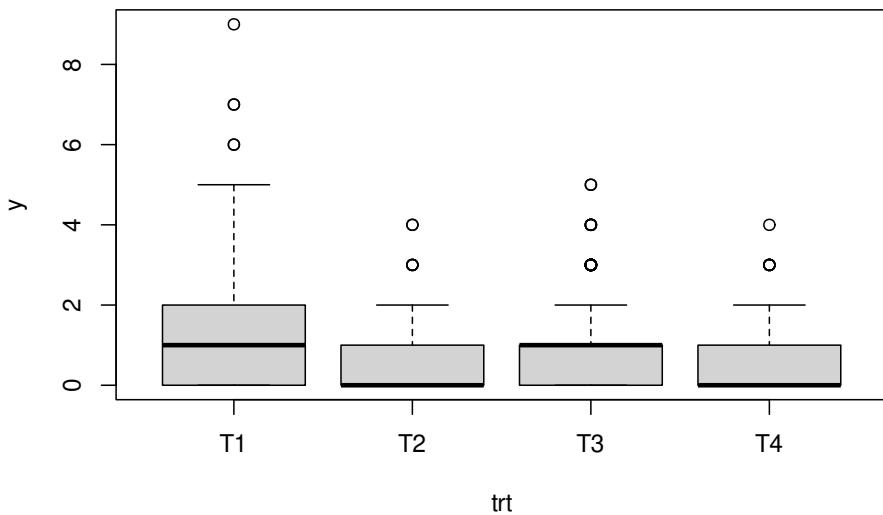
```
row col y block trt spray lead
1 1 1 1 B1 T1 N N
2 2 1 0 B1 T1 N N
3 3 1 1 B1 T1 N N
4 4 1 3 B1 T1 N N
5 5 1 6 B1 T1 N N
6 6 1 0 B2 T1 N N
```

描述响应变量的分布

```
hist(beall.webworms$y, main = "Histogram of Worm Count", xlab = "Number of Worms")
```



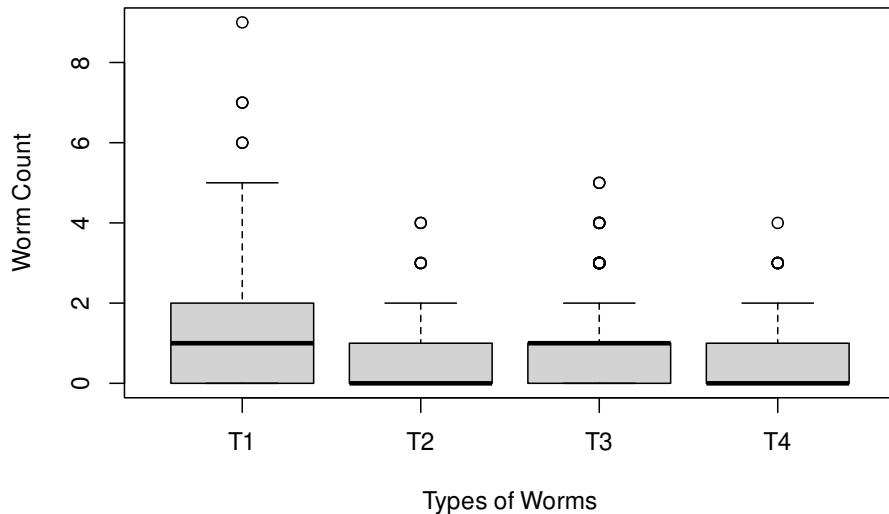
```
boxplot(y ~ trt, data = beall.webworms)
```



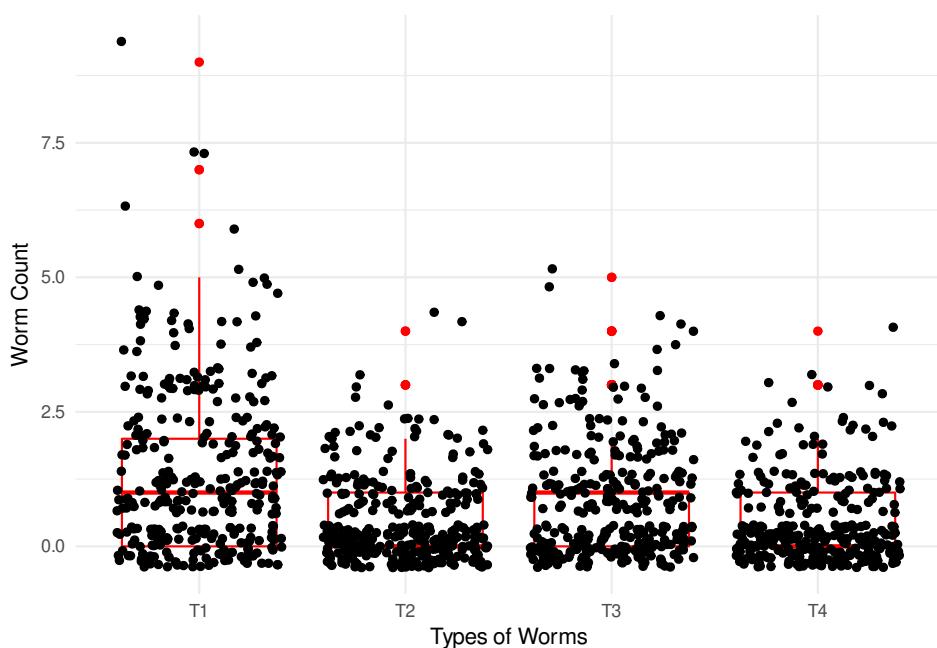
抖动图

云  
湘  
黄  
④

```
plot(y ~ trt, data = beall.webworms, xlab = "Types of Worms", ylab = "Worm Count")
```



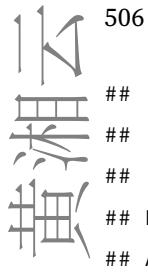
```
ggplot(beall.webworms, aes(trt, y)) +
 geom_boxplot(colour = "red") +
 geom_jitter() +
 labs(x = "Types of Worms", y = "Worm Count") +
 theme_minimal()
```



```
pois.mod <- glm(y ~ trt, data = beall.webworms, family = "poisson")
summary(pois.mod)
```

```
##
Call:
glm(formula = y ~ trt, family = "poisson", data = beall.webworms)
##
Deviance Residuals:
Min 1Q Median 3Q Max
-1.6733 -1.0046 -0.9081 0.6141 4.2771
##
Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.33647 0.04688 7.177 7.12e-13 ***
trtT2 -1.02043 0.09108 -11.204 < 2e-16 ***
trtT3 -0.49628 0.07621 -6.512 7.41e-11 ***
trtT4 -1.22246 0.09829 -12.438 < 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```



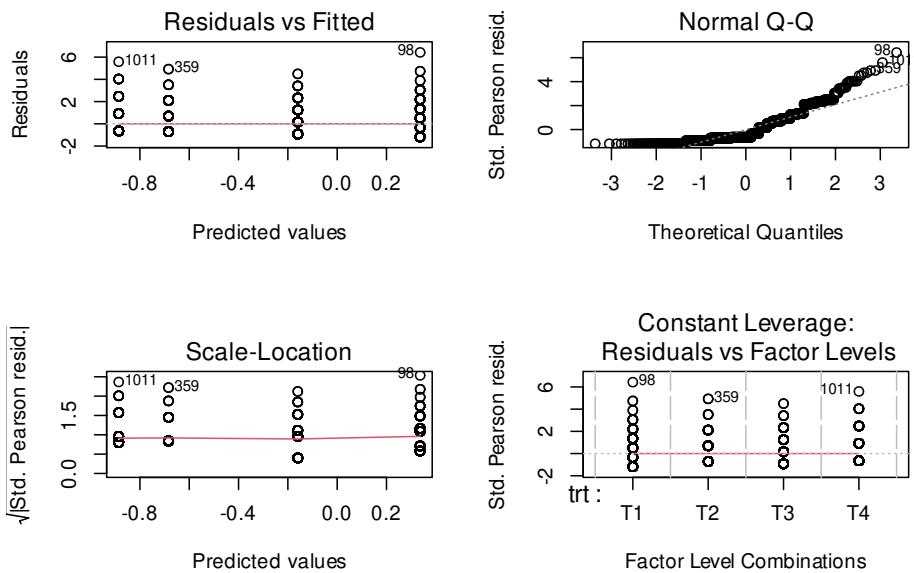
```
(Dispersion parameter for poisson family taken to be 1)
##
Null deviance: 1955.9 on 1299 degrees of freedom
Residual deviance: 1720.4 on 1296 degrees of freedom
AIC: 3125.5
##
Number of Fisher Scoring iterations: 6
```

模型系数 T2 的解释，这里 GLM 使用了对数联系函数 (log link function)，因此 -1.02 是对数变换后的值，T2 的系数实际是 0.3605949，实际意义是相对于 T1，T2 类型的蠕虫数量是 T1 的 0.3605949 倍

The first valuable information is related to the residuals of the model, which should be symmetrical as for any normal linear model. From this output we can see that minimum and maximum, as well as the first and third quartiles, are similar, so this assumption is confirmed. Then we can see that the variable trt (i.e. treatment factor) is highly significant for the model, with very low p-values. The statistical test in this case is not a t-test, as in the output of the function lm, but a Wald Test ([Wald Test](#)). This test computes the probability that the coefficient is 0, if the p is significant it means the chances the coefficient is zero are very low so the variable should be included in the model since it has an effect on y.

Another important information is the deviance, particularly the residual deviance. As a general rule, this value should be lower or in line than the residuals degrees of freedom for the model to be good. In this case the fact that the residual deviance is high (even though not dramatically) may suggests the explanatory power of the model is low. We will see below how to obtain p-value for the significance of the model.

```
par(mfrow = c(2, 2))
plot(poiss.mod)
```



```
predict(pois.mod, newdata = data.frame(trt = c("T1", "T2")))
```

```
1 2
0.3364722 -0.6839588
```

模型的 P 值

```
1 - pchisq(deviance(pois.mod), df.residual(pois.mod))
```

```
[1] 1.709743e-14
```

模型选择

```
pois.mod2 <- glm(y ~ block + spray * lead, data = beall.webworms, family = "poisson")
```

两模型的 AIC 比较

```
AIC(pois.mod, pois.mod2)
```

```
df AIC
pois.mod 4 3125.478
pois.mod2 16 3027.438
```

假设响应变量 Y 服从泊松分布，意味着随机变量 Y 的期望和方差相等

```
mean(beall.webworms$y)
```

```
[1] 0.7923077
```

```
var(beall.webworms$y)
```

```
[1] 1.290164
```

实际上方差比均值大，这种情况称之为过度发散 (overdispersed)，分布应该修正为拟 (似然) 泊松分布

```
pois.mod3 <- glm(y ~ trt, data = beall.webworms, family = c("quasipoisson"))
summary(pois.mod3)
```

```
##
Call:
glm(formula = y ~ trt, family = c("quasipoisson"), data = beall.webworms)
##
Deviance Residuals:
Min 1Q Median 3Q Max
-1.6733 -1.0046 -0.9081 0.6141 4.2771
##
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.33647 0.05457 6.166 9.32e-10 ***
trtT2 -1.02043 0.10601 -9.626 < 2e-16 ***
trtT3 -0.49628 0.08870 -5.595 2.69e-08 ***
trtT4 -1.22246 0.11440 -10.686 < 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
(Dispersion parameter for quasipoisson family taken to be 1.35472)
##
Null deviance: 1955.9 on 1299 degrees of freedom
Residual deviance: 1720.4 on 1296 degrees of freedom
AIC: NA
##
Number of Fisher Scoring iterations: 6
```

计算得知发散参数 (dispersion parameter) 是 1.35472，可见数据 Y 并不是发散得

离谱，泊松分布可能仍然是对这个数据的合理假设

AER 包是书籍 Applied Econometrics with R 的配套材料 [Kleiber and Zeileis, 2008]，可用于直接检验发散参数是否大于 1

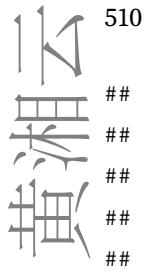
```
AER::dispersiontest(pois.mod, alternative="greater")
```

如果数据真的过度离散，就应该使用负二项分布作为响应变量的拟合分布，拟合它就采用 MASS 包 [Venables and Ripley, 2002] 提供的 glm.nb 函数

```
NB.mod1 <- MASS::glm.nb(y ~ trt, data = beall.webworms)
summary(NB.mod1)
```

```
##
Call:
MASS::glm.nb(formula = y ~ trt, data = beall.webworms, init.theta = 2.004130573,
link = log)
##
Deviance Residuals:
Min 1Q Median 3Q Max
-1.4572 -0.9488 -0.8660 0.5340 2.7698
##
Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.33647 0.06110 5.507 3.65e-08 ***
trtT2 -1.02043 0.10661 -9.572 < 2e-16 ***
trtT3 -0.49628 0.09423 -5.267 1.39e-07 ***
trtT4 -1.22246 0.11283 -10.834 < 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
(Dispersion parameter for Negative Binomial(2.0041) family taken to be 1)
##
Null deviance: 1442.7 on 1299 degrees of freedom
Residual deviance: 1275.3 on 1296 degrees of freedom
AIC: 3053
##
Number of Fisher Scoring iterations: 1
##
```



```
Theta: 2.004
Std. Err.: 0.325
##
2 x log-likelihood: -3042.969
```



## 两个模型的方差分析

```
anova(pois.mod, pois.mod2, test = "Chisq")
```

```
Analysis of Deviance Table
##
Model 1: y ~ trt
Model 2: y ~ block + spray * lead
Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1 1296 1720.4
2 1284 1598.4 12 122.04 < 2.2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

从方差分析比较的结果来看，P 值告诉我们，两模型是显著不同的，由上面对两模型的 AIC 计算结果来看，模型 `pois.mod2` 比模型 `pois.mod` 要好，模型的 AIC 值越小，表明拟合得越准确。

## 第二十二章 案例研究

提升回归模型的 10 个提示 [10 quick tips to improve your regression modeling](#)

`tidymodels` 和 `easystats` 都是基于 `tidyverse` [Wickham et al., 2019] 的统计模型套件, `strengejacke`、`mlr3verse` 目的和 `tidymodels` 差不多, 都是提供做数据建模的完整解决方案, 区别在于它不基于 `tidyverse` 这套东西。

`easystats` 包含 `insight` [Lüdecke et al., 2019] 和 `bayestestR` [Makowski et al., 2019] 等共 9 个 R 包, `tidymodels` 也包含差不多量的 R 包。

`rms` Regression Modeling Strategies

`modelsummary` 整理模型输出, 提供丰富的格式输出, 如 PDF, Text/Markdown, LaTeX, MS Word, RTF, JPG, and PNG.

[DrWhy](#)

R for Data Science Online Learning Community 在线学习社区以 `tidytuesday` 闻名遐迩。

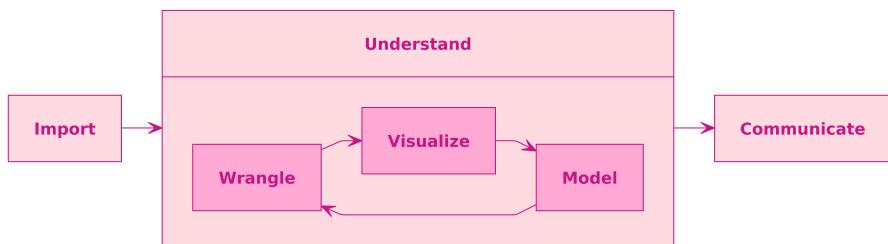


图 22.1: 模型

统计建模: 两种文化 [Breiman, 2001]



这些案例来自 Kaggle、Tudescday 或者自己找的数据集，而不是论文里，或者 R 包里的小数据集，应该更加真实，贴近实际问题，考虑更多细节

## 22.1 统计学家生平

世纪统计学家 100 位统计学家，寿命的影响因素，关联分析，图展示数据本身的注明每位统计学家所在的年代经历的重大事件，如欧洲中世纪霍乱，第二次世界大战，文化大革命，用图形来讲故事，展现数据可视化的魅力，参考文献 [Johnson and Kotz, 1997]

## 22.2 R 语言发展历史

R 语言发展历史和现状，用图来表达

## 22.3 不同实验条件下植物生长情况

PlantGrowth 数据集收集自 Annette J. Dobson 所著书籍《An Introduction to Statistical Modelling》[Dobson, 1983] 第 2 章第 2 节的案例 — 研究植物在两种不同试验条件下的生长情况，植物通过光合作用吸收土壤的养分和空气中的二氧化碳，完成积累，故以植物的干重来刻画植物的生长情况，首先将几乎相同的种子随机地分配到实验组和对照组，基于完全随机实验设计（completely randomized experimental design），经过预定的时间后，将植物收割，干燥并称重，结果如表 22.1 所示

```
do.call("cbind", lapply(split(PlantGrowth, f = PlantGrowth$group), subset, select = "weight"))
或者
library(magrittr)
split(PlantGrowth, f = PlantGrowth$group) %>% # 分组
 lapply(., subset, select = "weight") %>% # 计算
 Reduce("cbind", .) %>% # 合并
 setNames(., levels(PlantGrowth$group)) %>% # 重命名 `colnames<-`(.,
 levels(PlantGrowth$group))
 knitr::kable(.,
```

表 22.1: 不同生长环境下植物的干重

	1	2	3	4	5	6	7	8	9	10
ctrl	4.17	5.58	5.18	6.11	4.50	4.61	5.17	4.53	5.33	5.14
trt1	4.81	4.17	4.41	3.59	5.87	3.83	6.03	4.89	4.32	4.69
trt2	6.31	5.12	5.54	5.50	5.37	5.29	4.92	6.15	5.80	5.26

```

caption = "不同生长环境下植物的干重", row.names = TRUE,
align = "c"
)

```

设立对照组（控制组）ctrl 和实验组 trt1 和 trt2，比较不同的处理方式对植物干重的影响

```
summary(PlantGrowth)
```

```

weight group
Min. :3.590 ctrl:10
1st Qu.:4.550 trt1:10
Median :5.155 trt2:10
Mean :5.073
3rd Qu.:5.530
Max. :6.310

```

每个组都有 10 颗植物，生长情况如图22.2所示

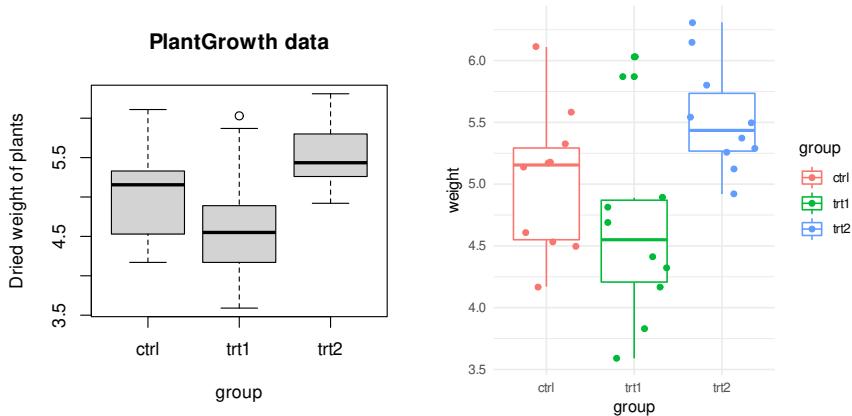


图 22.2: 植物干重



实验条件 trt1 和 trt2 对植物生长状况有显著的影响，为了量化这种影响，建立线性回归模型

```

fit_sublm <- lm(weight ~ group,
 data = PlantGrowth,
 subset = group %in% c("ctrl", "trt1")
)
anova(fit_sublm)

Analysis of Variance Table
##
Response: weight
Df Sum Sq Mean Sq F value Pr(>F)
group 1 0.6882 0.68820 1.4191 0.249
Residuals 18 8.7292 0.48496

summary(fit_sublm)

##
Call:
lm(formula = weight ~ group, data = PlantGrowth, subset = group %in%
c("ctrl", "trt1"))
##
Residuals:
Min 1Q Median 3Q Max
-1.0710 -0.4938 0.0685 0.2462 1.3690
##
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.0320 0.2202 22.850 9.55e-15 ***
grouptrt1 -0.3710 0.3114 -1.191 0.249

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##
Residual standard error: 0.6964 on 18 degrees of freedom
Multiple R-squared: 0.07308, Adjusted R-squared: 0.02158
F-statistic: 1.419 on 1 and 18 DF, p-value: 0.249

```

下面再通过检验的方式比较实验组和对照组相比，是否有显著作用

```
控制组和实验组1比较
t.test(weight ~ group, data = PlantGrowth, subset = group %in% c("ctrl", "trt1"))

##
Welch Two Sample t-test
##
data: weight by group
t = 1.1913, df = 16.524, p-value = 0.2504
alternative hypothesis: true difference in means between group ctrl and group trt1
95 percent confidence interval:
-0.2875162 1.0295162
sample estimates:
mean in group ctrl mean in group trt1
5.032 4.661

控制组和实验组2比较
t.test(weight ~ group, data = PlantGrowth, subset = group %in% c("ctrl", "trt2"))

##
Welch Two Sample t-test
##
data: weight by group
t = -2.134, df = 16.786, p-value = 0.0479
alternative hypothesis: true difference in means between group ctrl and group trt2
95 percent confidence interval:
-0.98287213 -0.00512787
sample estimates:
mean in group ctrl mean in group trt2
5.032 5.526
```

检验结果表明，实验条件 trt2 会对植物生长产生显著效果，而实验条件 trt1 不会。在假定同方差的情况下，建立线性回归模型，同时考虑实验条件 trt1 和 trt2

```
模型拟合
fit_lm <- lm(weight ~ group, data = PlantGrowth)

模型输出
summary(fit_lm)
```

云  
湘  
黄  
④

```


Call:
lm(formula = weight ~ group, data = PlantGrowth)

Residuals:
Min 1Q Median 3Q Max
-1.0710 -0.4180 -0.0060 0.2627 1.3690

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.0320 0.1971 25.527 <2e-16 ***
grouptrt1 -0.3710 0.2788 -1.331 0.1944
grouptrt2 0.4940 0.2788 1.772 0.0877 .

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6234 on 27 degrees of freedom
Multiple R-squared: 0.2641, Adjusted R-squared: 0.2096
F-statistic: 4.846 on 2 and 27 DF, p-value: 0.01591

方差分析
anova(fit_lm)

Analysis of Variance Table

Response: weight
Df Sum Sq Mean Sq F value Pr(>F)
group 2 3.7663 1.8832 4.8461 0.01591 *
Residuals 27 10.4921 0.3886

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

参数估计
coef(summary(fit_lm))

Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.032 0.1971284 25.526514 1.936575e-20
grouptrt1 -0.371 0.2787816 -1.330791 1.943879e-01

```

表 22.2: 线性回归的输出

	估计值	标准差	t 统计量	P 值
$\alpha$	5.032	0.1971	25.5265	0.0000
$\beta_1$	-0.371	0.2788	-1.3308	0.1944
$\beta_2$	0.494	0.2788	1.7720	0.0877

```
grouptrt2 0.494 0.2787816 1.771996 8.768168e-02
```

模型输出整理成表 22.2 所示

还可以将模型转化为数学公式

```
理论模型
```

```
equatiomatic::extract_eq(fit_lm)
```

$$\text{weight} = \alpha + \beta_1(\text{group}_{\text{trt1}}) + \beta_2(\text{group}_{\text{trt2}}) + \epsilon$$

```
拟合模型
```

```
equatiomatic::extract_eq(fit_lm, use_coefs = TRUE)
```

$$\widehat{\text{weight}} = 5.03 - 0.37(\text{group}_{\text{trt1}}) + 0.49(\text{group}_{\text{trt2}})$$

进一步地，我们在线性模型的基础上考虑每个实验组有不同的方差，先做方差齐性检验。

```
bartlett.test(weight ~ group, data = PlantGrowth)
```

```
##
```

```
Bartlett test of homogeneity of variances
```

```
##
```

```
data: weight by group
```

```
Bartlett's K-squared = 2.8786, df = 2, p-value = 0.2371
```

```
fligner.test(weight ~ group, data = PlantGrowth)
```

```
##
```

```
Fligner-Killeen test of homogeneity of variances
```

```
##
```

```
data: weight by group
```



```
Fligner-Killeen:med chi-squared = 2.3499, df = 2, p-value = 0.3088
```

检验的结果显示，可以认为三个组的方差没有显著差异，但我们还是考虑每个组有不同的方差，看看放开假设能获得多少提升，后续会发现，从对数似然的角度来看，实际提升量很小，只有 7.72%



上面同时比较多个总体的方差，会发现方差没有显著差异，那么接下来在假定方差齐性的条件下，比较均值的差异是否显著？

# 参数检验，假定异方差

```
oneway.test(weight ~ group, data = PlantGrowth, var.equal = FALSE)
```

```
##
```

```
One-way analysis of means (not assuming equal variances)
```

```
##
```

```
data: weight and group
```

```
F = 5.181, num df = 2.000, denom df = 17.128, p-value = 0.01739
```

# 参数检验，假定方差齐性

```
oneway.test(weight ~ group, data = PlantGrowth, var.equal = TRUE)
```

```
##
```

```
One-way analysis of means
```

```
##
```

```
data: weight and group
```

```
F = 4.8461, num df = 2, denom df = 27, p-value = 0.01591
```

# 非参数检验

```
kruskal.test(weight ~ group, data = PlantGrowth)
```

```
##
```

```
Kruskal-Wallis rank sum test
```

```
##
```

```
data: weight by group
```

```
Kruskal-Wallis chi-squared = 7.9882, df = 2, p-value = 0.01842
```

检验结果显示它们的均值是有显著差异的！

# 固定效应模型

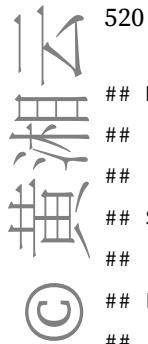
```
fit_gls <- nlme::gls(weight ~ 1,
weights = nlme::varIdent(form = ~ 1 | group),
```

```
 data = PlantGrowth, method = "REML"
)
summary(fit_gls)

Generalized least squares fit by REML
Model: weight ~ 1
Data: PlantGrowth
AIC BIC logLik
70.48628 75.95547 -31.24314
##
Variance function:
Structure: Different standard deviations per stratum
Formula: ~1 | group
Parameter estimates:
ctrl trt1 trt2
1.0000000 1.5825700 0.9230865
##
Coefficients:
Value Std.Error t-value p-value
(Intercept) 5.199759 0.1162421 44.73214 0
##
Standardized residuals:
Min Q1 Med Q3 Max
-1.74647988 -0.91870713 -0.07591108 0.60676033 2.03987301
##
Residual standard error: 0.5896195
Degrees of freedom: 30 total; 29 residual

随机效应模型
fit_lme <- nlme::lme(weight ~ 1, random = ~ 1 | group, data = PlantGrowth)
summary(fit_lme)

Linear mixed-effects model fit by REML
Data: PlantGrowth
AIC BIC logLik
67.44473 71.54662 -30.72237
##
```



```
Random effects:
Formula: ~1 | group
(Intercept) Residual
StdDev: 0.3865976 0.6233746

Fixed effects: weight ~ 1
Value Std.Error DF t-value p-value
(Intercept) 5.073 0.2505443 27 20.24792 0

Standardized Within-Group Residuals:
Min Q1 Med Q3 Max
-1.854449795 -0.688750457 0.006389611 0.406096866 2.059729645

Number of Observations: 30
Number of Groups: 3
```

$\sigma_i^2 = \text{Var}(\epsilon_{ij}), i = 1, 2, 3$  表示第  $i$  组的方差，

$$y_{ij} = \mu + \epsilon_{ij}, i = 1, 2, 3$$

其中  $\mu$  是固定的未知参数，我们和之前假定同方差情形下的模型比较一下，现在异方差情况下模型提升的情况，从对数似然的角度来看

```
logLik(fit_lm)
'log Lik.' -26.80952 (df=4)
logLik(fit_lm, REML = TRUE)
'log Lik.' -29.00481 (df=4)
logLik(fit_gls)
'log Lik.' -31.24314 (df=4)
logLik(fit_lme)
'log Lik.' -30.72237 (df=3)
```

进一步地，我们考虑两水平模型，认为不同的实验组其均值和方差都不一样，检验三样本均值是否相等？

$\mu_1 = \mu_2 = \mu_3$  检验，这里因为每组的样本量都一样，因此考虑 Turkey 的 T 法检验，检验均值是否有显著差别，实际上这里因为实验组数量只有 2 个，可以两两比对，如前所述。但是这里我们想扩展一下，考虑多组比较的问题。

和上面用 `gls` 拟合的模型是一致的。

$$y_{ij} = \mu_i + \epsilon_{ij}, \quad (22.1)$$

$$\mu_i = \mu_\theta + \xi_i. \quad i = 1, \dots, 3; \quad j = 1, \dots, 10. \quad (22.2)$$

其中  $\mu_i$  是随机的未知变量，服从均值为  $\mu_\theta$  方差为  $Var(\xi_i) = \tau^2$  的正态分布

我们用 **MASS** 包提供的 `glmmPQL()` 函数拟合该数据集

```
fit_lme_pql <- MASS::glmmPQL(weight ~ 1,
 random = ~ 1 | group, verbose = FALSE,
 family = gaussian(), data = PlantGrowth
)
summary(fit_lme_pql)

Linear mixed-effects model fit by maximum likelihood
Data: PlantGrowth
AIC BIC logLik
NA NA NA
##
Random effects:
Formula: ~1 | group
(Intercept) Residual
StdDev: 0.2944234 0.6233746
##
Variance function:
Structure: fixed weights
Formula: ~invwt
Fixed effects: weight ~ 1
Value Std.Error DF t-value p-value
(Intercept) 5.073 0.2080656 27 24.38174 0
##
Standardized Within-Group Residuals:
Min Q1 Med Q3 Max
-1.922640850 -0.734727623 0.004564386 0.405111223 1.991538416
##
Number of Observations: 30
```



```
Number of Groups: 3
```

我们再借助 **brms** 包从贝叶斯的角度来分析数据，并建模

```
贝叶斯模型
fit_brm <- brms:::brm(weight ~ group, data = PlantGrowth)
参考 https://www.xiangyunhuang.com.cn/2019/05/normal-hierarchical-model/
library(Rcpp)
fit_lme_brm <- brms:::brm(weight ~ 1 + (1 | group),
 data = PlantGrowth, family = gaussian(),
 refresh = 0, seed = 2019
)
summary(fit_lme_brm)
```

## 22.4 橘树生长情况

`Orange` 数据集包含三个变量，记录了加利福尼亚南部的一个小树林中的五棵橘树的生长情况，在 **datasets** 包里，数据集保存为 `c("nfnGroupedData", "nfGroupedData", "groupedData", "data.frame")` 类型的数据，同时具有着四个类的特点。

- `Tree`: 有序的指示变量，根据 5 棵橘树的最大直径划分，测量值很可能是根据林务员常用的“胸围周长”
- `age`: 橘树的树龄，自 1968 年 12 月 31 日起按天计算
- `circumference`: 橘树树干的周长，单位是毫米

查看部分数据的情况

```
head(Orange)
```

```
Grouped Data: circumference ~ age | Tree
Tree age circumference
1 1 118 30
2 1 484 58
3 1 664 87
4 1 1004 115
5 1 1231 120
6 1 1372 142
```



查看变量的属性

```
str(Orange)
```

```
Classes 'nfnGroupedData', 'nfGroupedData', 'groupedData' and 'data.frame': 35 ob
$ Tree : Ord.factor w/ 5 levels "3" < "1" < "5" < "2" < ...: 2 2 2 2 2 2 2 4 4 4 ...
$ age : num 118 484 664 1004 1231 ...
$ circumference: num 30 58 87 115 120 142 145 33 69 111 ...
- attr(*, "formula")=Class 'formula' language circumference ~ age | Tree
.. .- attr(*, ".Environment")=<environment: R_EmptyEnv>
- attr(*, "labels")=List of 2
..$ x: chr "Time since December 31, 1968"
..$ y: chr "Trunk circumference"
- attr(*, "units")=List of 2
..$ x: chr "(days)"
..$ y: chr "(mm)"
```

说明 5 棵树之间的大小关系是  $3 < 1 < 5 < 2 < 4$ ，这里的数字 1, 2, 3, 4, 5 只是对树的编号，第一次测量时树的大小关系在 R 内用有序因子来表示。

```
levels(Orange$Tree)
```

```
[1] "3" "1" "5" "2" "4"
```

表 22.3 记录了 5 颗橘树自 1968 年 12 月 31 日以来的生长情况

```
aggregate(data = Orange, circumference ~ age, FUN = mean)
library(magrittr)
reshape(
 data = Orange, v.names = "circumference", idvar = "Tree",
 timevar = "age", direction = "wide", sep = ""
) %>%
 knitr::kable(.,
 caption = "躯干周长（毫米）随时间（天）的变化",
 row.names = FALSE, col.names = gsub("(circumference)", "", names(.)),
 align = "c"
)
```

图 22.3 以直观的方式展示 5 颗橘树的生长变化，相比于表 22.3 我们能更加明确读取数据中的变化

表 22.3: 躯干周长(毫米)随时间(天)的变化

Tree	118	484	664	1004	1231	1372	1582
1	30	58	87	115	120	142	145
2	33	69	111	156	172	203	203
3	30	51	75	108	115	139	140
4	32	62	112	167	179	209	214
5	30	49	81	125	142	174	177

```
library(ggplot2)
p <- ggplot(data = Orange, aes(x = age, y = circumference, color = Tree)) +
 geom_point() +
 geom_line() +
 theme_minimal() +
 labs(x = "age (day)", y = "circumference (mm)")
p
```

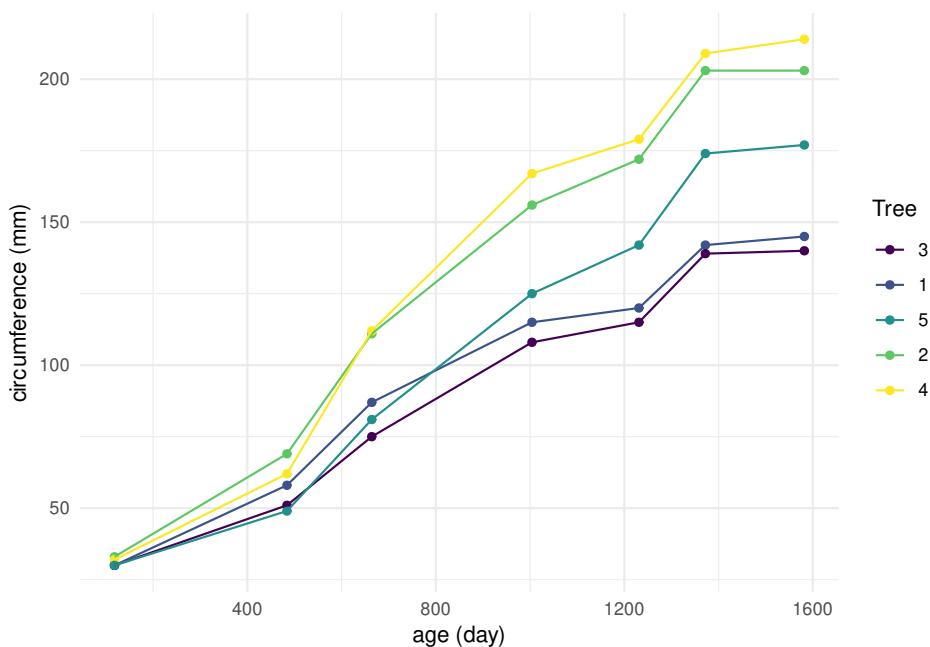


图 22.3: 橘树生长模型

```
library(gganimate)
p + transition_reveal(age)
```

## 第二十三章 数据探索

### DataExplorer

[DALEX](#) 提供探索性模型分析，支持 `mlr`、`caret`、`keras`、`h2o` 和 `xgboost` 等一系列统计建模分析的 R 包。

[breakDown](#) Model Agnostic Explainers for Individual Predictions

## 第二十四章 生存分析

The fact that some people murder doesn't mean we should copy them.  
And murdering data, though not as serious, should also be avoided.

— Frank E. Harrell<sup>1</sup>

R 软件内置了 **survival** 包，它是实现生存分析的核心 R 包。文档见 <https://cran.r-project.org/package=survival> 相关书籍见 [Terry M. Therneau and Patricia M. Grambsch \[2000\]](#)

**survminer** 竟然严重依赖 **ggpubr** 包，**ggpubr** 包曾被 **ggtree** 的作者余光创严重吐槽过。**ggfortify** 包大大扩展了 **ggplot2** 包的 **autoplot()** 函数，使得它适应各种模型对象的自动绘图。

### 24.1 急性粒细胞白血病生存数据

```
library(survival)
leukemia.surv <- survfit(Surv(time, status) ~ x, data = aml)
library(ggfortify)
autoplot(leukemia.surv, data = aml) +
 theme_minimal()
```

<sup>1</sup><https://stat.ethz.ch/pipermail/r-help/2005-July/075649.html>

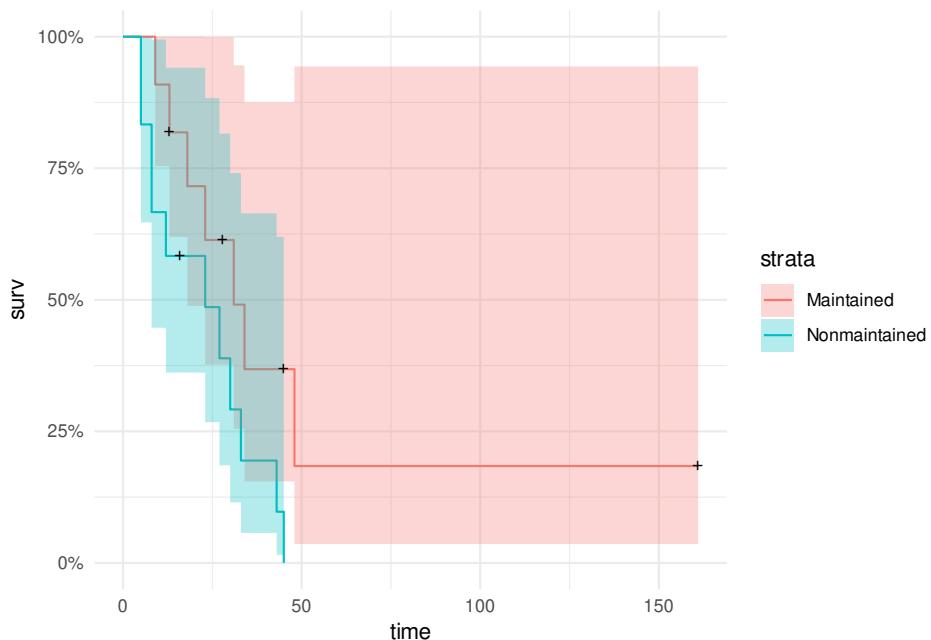


图 24.1: 急性粒细胞白血病生存数据



## 第二十五章 时序分析

```
library(formatR)
`%>%` <- magrittr::`%>%`
library(ggplot2)
library(ggfortify) # **ggfortify** 包提供的 `autoplot()` 函数可以根据数据对象的不
library(highcharter)
library(dygraphs)
library(robustbase) # Robust Statistics
library(timeDate) # 日期处理
library(timeSeries) # 序列处理
library(fPortfolio) # 投资组合
library(prophet) # 时间序列预测
https://github.com/business-science/timetk
library(timetk) # 处理时间序列数据的工具箱
```

首先介绍时序数据对象及操作, 处理时序数据的工具, 包括时序图、相关图、平稳性检验, 相关检验, 之后才是时序建模。`timeDate` `timeSeries` 是处理日期和时间序列的 R 包, 有专门的官网 <https://www.rmetrics.org/>, 扩展到时间序列、组合优化、金融市场、投资管理等一系列书籍, 非常值得一看。此外, 北大李东风老师的[金融时间序列分析讲义](#)是这方面非常好的中文参考材料。David R. Brillinger 在 1975 年出版的书《Time Series: Data Analysis and Theory》[Brillinger, 2001] 是经典著作, 我们可以从时间序列分析的综述上开始入手, 比如从 ARIMA 过渡到异方差和非高斯分布 [https://mason.gmu.edu/~jgentle/talks/CompFin\\_Tutorial.pdf](https://mason.gmu.edu/~jgentle/talks/CompFin_Tutorial.pdf), <https://www.stat.berkeley.edu/~brill/Papers/encysbs.pdf> 和 ARCH or GARCH 的综述 [http://public.econ.duke.edu/~boller/Papers/glossary\\_arch.pdf](http://public.econ.duke.edu/~boller/Papers/glossary_arch.pdf) , 宾州州立大学开设的 Applied Time Series Analysis 课程 <https://newonlinecourses.science.psu.edu/stat510/>, 以及《Time Series Analysis and Its Applications With R Examples》已经出到第四版了, 和 R 语言结合, 理论和应用结合



<https://www.stat.pitt.edu/stoffer/tsa4/>。从时间序列中寻找规律，这样才是真的数据建模，从数据到模型，而不是相反 **Finding Patterns in Time Series**，识别金融时间序列的模式和统计规律。现在工业界做时序分析和预测的工具，如 facebook 出品的 **prophet**，微软收集了一些时间序列预测的最佳实战案例 <https://github.com/microsoft/forecasting>



**forecastML** 自回归模型结合机器学习方法。

**CausalImpact** 借助贝叶斯分析方法推断时间序列中的因果关系，比如广告促销带来的点击效果。

**robustbase [Maronna et al., 2006]** 提供稳健统计方法。

**prophet** 基于可加模型的时间序列预测

**AnomalyDetection** 时间序列数据中的异常值检测

## 25.1 时序数据

以数据集 **AirPassengers** 为例说明一下 R 内置的存储时间序列数据的数据结构 – **ts** 数据对象。函数 **class()**、**mode()** 和 **str()** 分别可以查看其数据类型、存储类型和数据结构。

```
数据类型
class(AirPassengers)

[1] "ts"

存储类型
mode(AirPassengers)

[1] "numeric"

数据结构
str(AirPassengers)

Time-Series [1:144] from 1949 to 1961: 112 118 132 129 121 ...
..`attr`='list[1:3]'

查看该数据集开始和结束的时间点

c(start(AirPassengers), end(AirPassengers))

[1] 1949 1 1960 12
```



数据集 AirPassengers 在以上时间区间的划分

```
time(AirPassengers)
```

```
Jan Feb Mar Apr May Jun Jul Aug
1949 1949.000 1949.083 1949.167 1949.250 1949.333 1949.417 1949.500 1949.583
1950 1950.000 1950.083 1950.167 1950.250 1950.333 1950.417 1950.500 1950.583
1951 1951.000 1951.083 1951.167 1951.250 1951.333 1951.417 1951.500 1951.583
1952 1952.000 1952.083 1952.167 1952.250 1952.333 1952.417 1952.500 1952.583
1953 1953.000 1953.083 1953.167 1953.250 1953.333 1953.417 1953.500 1953.583
1954 1954.000 1954.083 1954.167 1954.250 1954.333 1954.417 1954.500 1954.583
1955 1955.000 1955.083 1955.167 1955.250 1955.333 1955.417 1955.500 1955.583
1956 1956.000 1956.083 1956.167 1956.250 1956.333 1956.417 1956.500 1956.583
1957 1957.000 1957.083 1957.167 1957.250 1957.333 1957.417 1957.500 1957.583
1958 1958.000 1958.083 1958.167 1958.250 1958.333 1958.417 1958.500 1958.583
1959 1959.000 1959.083 1959.167 1959.250 1959.333 1959.417 1959.500 1959.583
1960 1960.000 1960.083 1960.167 1960.250 1960.333 1960.417 1960.500 1960.583
Sep Oct Nov Dec
1949 1949.667 1949.750 1949.833 1949.917
1950 1950.667 1950.750 1950.833 1950.917
1951 1951.667 1951.750 1951.833 1951.917
1952 1952.667 1952.750 1952.833 1952.917
1953 1953.667 1953.750 1953.833 1953.917
1954 1954.667 1954.750 1954.833 1954.917
1955 1955.667 1955.750 1955.833 1955.917
1956 1956.667 1956.750 1956.833 1956.917
1957 1957.667 1957.750 1957.833 1957.917
1958 1958.667 1958.750 1958.833 1958.917
1959 1959.667 1959.750 1959.833 1959.917
1960 1960.667 1960.750 1960.833 1960.917
```

期初和期末的周期

```
tsp(AirPassengers)
```

```
[1] 1949.000 1960.917 12.000
```

函数 `diff()` 实现差分算子，默认参数 `lag = 1`, `differences = 1` 表示延迟期数为 1 的一阶差分。

黄  
湘  
云

# 差分前

AirPassengers

```
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
1949 112 118 132 129 121 135 148 148 136 119 104 118
1950 115 126 141 135 125 149 170 170 158 133 114 140
1951 145 150 178 163 172 178 199 199 184 162 146 166
1952 171 180 193 181 183 218 230 242 209 191 172 194
1953 196 196 236 235 229 243 264 272 237 211 180 201
1954 204 188 235 227 234 264 302 293 259 229 203 229
1955 242 233 267 269 270 315 364 347 312 274 237 278
1956 284 277 317 313 318 374 413 405 355 306 271 306
1957 315 301 356 348 355 422 465 467 404 347 305 336
1958 340 318 362 348 363 435 491 505 404 359 310 337
1959 360 342 406 396 420 472 548 559 463 407 362 405
1960 417 391 419 461 472 535 622 606 508 461 390 432
```

# 差分后

diff(AirPassengers)

```
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
1949 6 14 -3 -8 14 13 0 -12 -17 -15 14
1950 -3 11 15 -6 -10 24 21 0 -12 -25 -19 26
1951 5 5 28 -15 9 6 21 0 -15 -22 -16 20
1952 5 9 13 -12 2 35 12 12 -33 -18 -19 22
1953 2 0 40 -1 -6 14 21 8 -35 -26 -31 21
1954 3 -16 47 -8 7 30 38 -9 -34 -30 -26 26
1955 13 -9 34 2 1 45 49 -17 -35 -38 -37 41
1956 6 -7 40 -4 5 56 39 -8 -50 -49 -35 35
1957 9 -14 55 -8 7 67 43 2 -63 -57 -42 31
1958 4 -22 44 -14 15 72 56 14 -101 -45 -49 27
1959 23 -18 64 -10 24 52 76 11 -96 -56 -45 43
1960 12 -26 28 42 11 63 87 -16 -98 -47 -71 42
```

# 延迟一期的二阶差分

diff(AirPassengers, lag = 1, differences = 2)

```
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
```



```
1949 8 -17 -5 22 -1 -13 -12 -5 2 29
1950 -17 14 4 -21 -4 34 -3 -21 -12 -13 6 45
1951 -21 0 23 -43 24 -3 15 -21 -15 -7 6 36
1952 -15 4 4 -25 14 33 -23 0 -45 15 -1 41
1953 -20 -2 40 -41 -5 20 7 -13 -43 9 -5 52
1954 -18 -19 63 -55 15 23 8 -47 -25 4 4 52
1955 -13 -22 43 -32 -1 44 4 -66 -18 -3 1 78
1956 -35 -13 47 -44 9 51 -17 -47 -42 1 14 70
1957 -26 -23 69 -63 15 60 -24 -41 -65 6 15 73
1958 -27 -26 66 -58 29 57 -16 -42 -115 56 -4 76
1959 -4 -41 82 -74 34 28 24 -65 -107 40 11 88
1960 -31 -38 54 14 -31 52 24 -103 -82 51 -24 113
```

## 25.2 时序图

美国纽黑文自 1912 年至 1971 年的年平均气温变化见图 25.1。

```
plot(nhtemp, main = "美国纽黑文的年平均气温", family = "source-han-sans-cn")

构造多个 ts 序列
tmp <- ts(
 data = data.frame(
 pay_one = rnorm(20),
 pay_two = rnorm(20),
 pay_three = rnorm(20)
),
 start = c(1961, 1), frequency = 12
)

plot(tmp, main = "pay cnt")
```

美国纽黑文的年平均气温

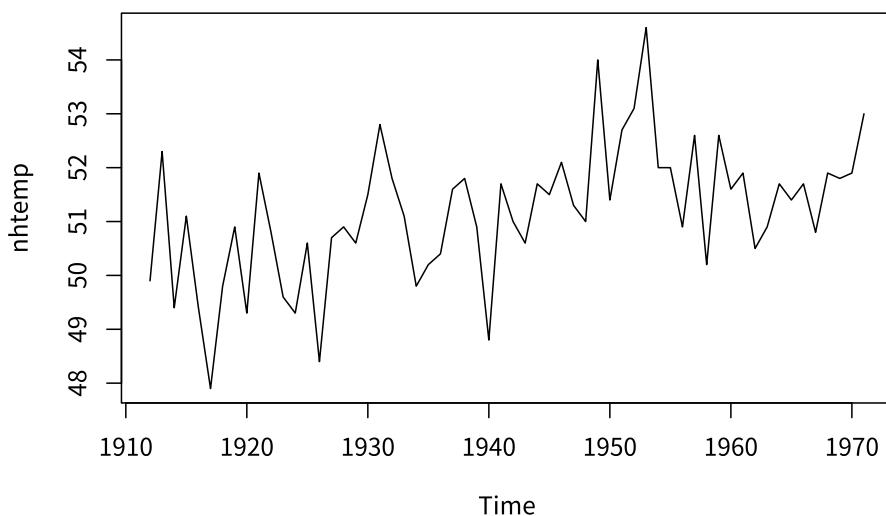
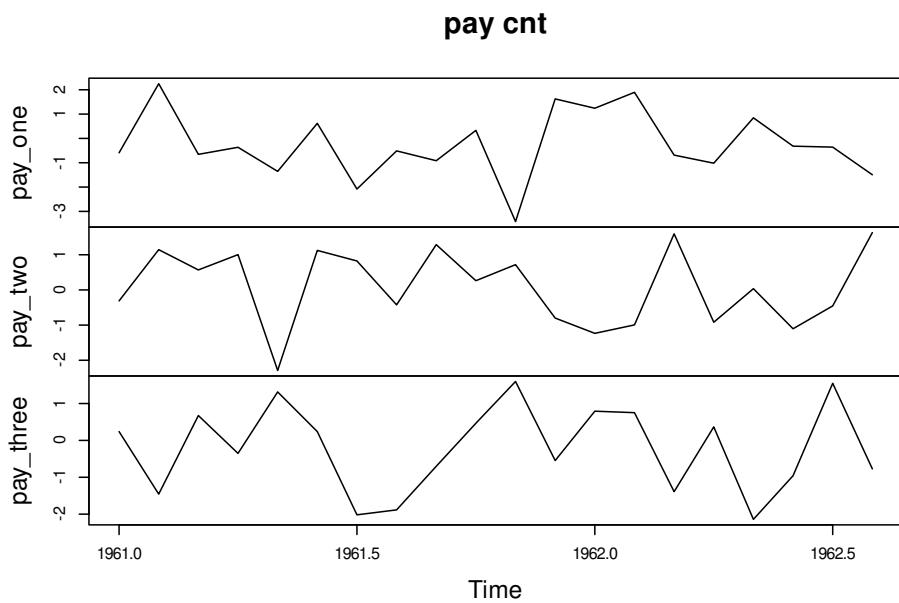
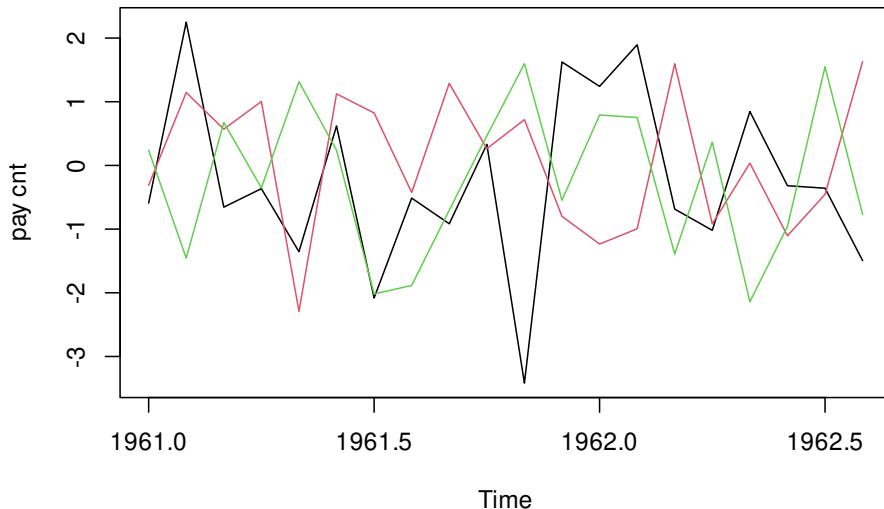


图 25.1: 美国纽黑文的年平均气温, 单位: 华氏温度

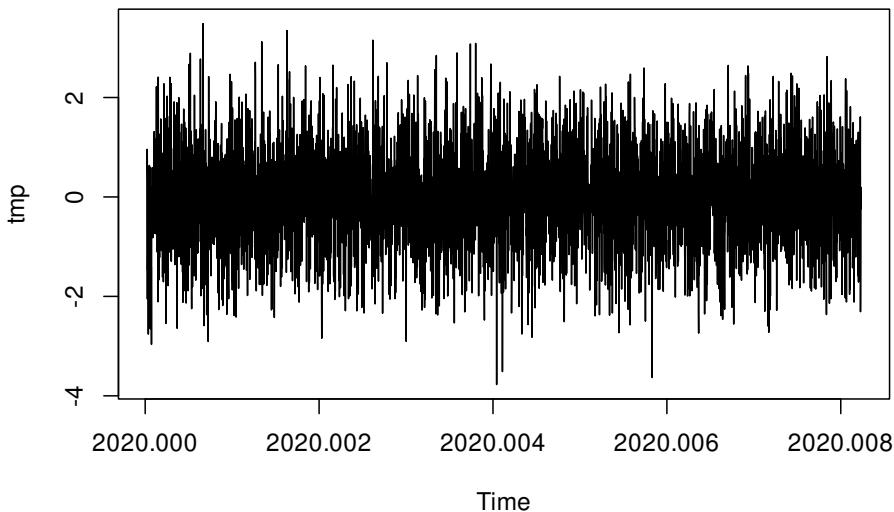


```
plot(tmp, plot.type = "single", col = 1:3, ylab = "pay cnt")
```



## 25.3 基本概念

```
从某个完整的一天开始统计数据
分钟级 ts 数据
time_min <- format(seq.POSIXt(
 from = as.POSIXct("2020-01-01 00:00"),
 to = as.POSIXct("2020-01-01 23:59"), by = "1 min"
)
 ,
 format = "%H:%M"
)
tmp = ts(data = rnorm(1440 * 3), start = c(2020, 12),
 frequency = 24*60*365.25, names = "访问量")
plot(tmp)
```



frequency: the number of observations per unit of time.

frequency 里面乘以 365.25 而不是 365 是因为每隔 4 年出现一次 366 天，多出的这一天分摊到每一年。frequency 表示单位时间内发生的次数，ts 对象的时间基准为 1 年，所以，frequency = 4 表示一年出现四次，依此类推。关于季节性周期的说明 <https://robjhyndman.com/hyndtsight/seasonal-periods/>。

序列长度一样，但是周期不一样，这里的单位时间指的是 1 年

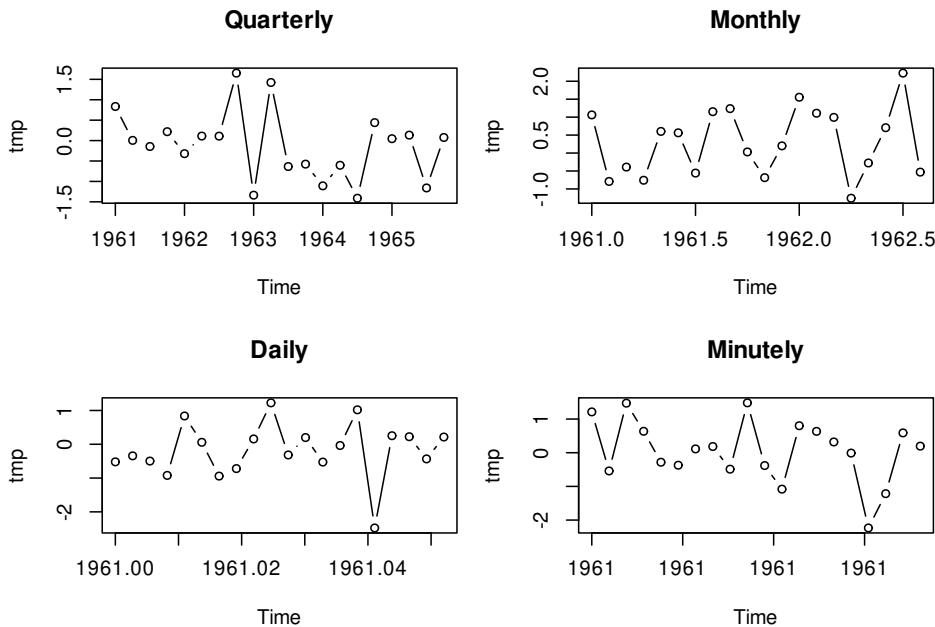
```
季数据
op = par(mfrow = c(2,2), mar = c(4,4,4,1))
tmp = ts(rnorm(20), start = c(1961, 1), frequency = 4)
plot(tmp, main = "Quarterly", type = "b")

月数据
tmp = ts(rnorm(20), start = c(1961, 1), frequency = 12) # 自然时间周期是一年，每月采样
plot(tmp, main = "Monthly", type = "b")

日数据
tmp = ts(rnorm(20), start = c(1961, 1), frequency = 365.25)
plot(tmp, main = "Daily", type = "b")

分钟数据
tmp = ts(rnorm(20), start = c(1961, 1), frequency = 24*60*365.25)
```

```
plot(tmp, main = "Minutely", type = "b")
```



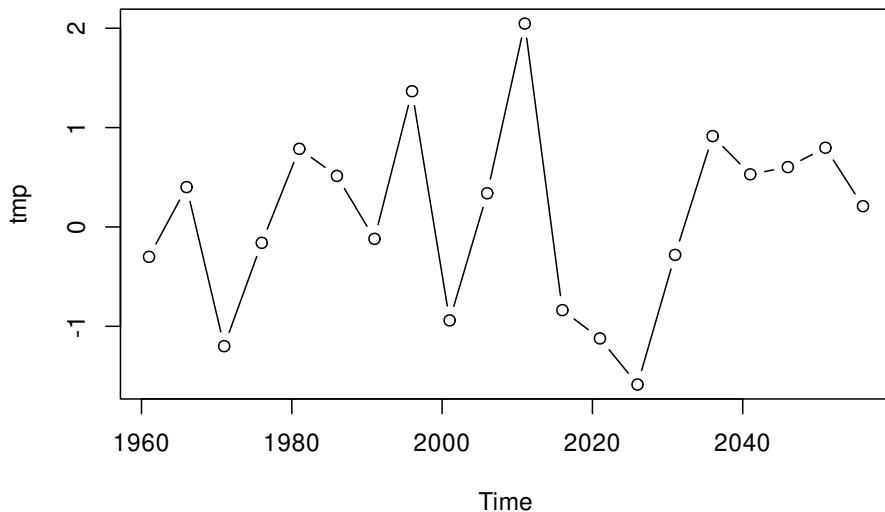
```
par(op)
```

默认情况下，自然时间周期是一年，每月采样。那可不可以设置自然时间周期是一周，每天采样呢？当然可以，只是 Base R 暂不支持，其实表达数据粒度的能力没有变化，以年或周为基准，都可以表达上面的季、月、日、分钟数据。

deltat 和 frequency 只需提供一个参数值即可 deltat = 1/12 和 frequency = 12 表示同样的含义。

R 4.0.0 开始，frequency 不必是整数，还可以是小数，frequency = .2 表示每 5 个时间单位抽样一次，根据周期 T 和频率 f 的关系  $T = 1/f$

```
tmp = ts(rnorm(20), start = c(1961, 1), frequency = .2)
plot(tmp, type = "b")
```



ts 和 seq 构造时间向量的关系是什么？

```
seq(from = 1961, to = 2056, by = 5)
```

```
[1] 1961 1966 1971 1976 1981 1986 1991 1996 2001 2006 2011 2016 2021 2026 2031
[16] 2036 2041 2046 2051 2056
```

即每隔 5 年抽样一次，采一个数据点

```
ts(rnorm(20), start = c(1961, 1), frequency = 365.25/7)
```

```
Time Series:
Start = 1961
End = 1961.36413415469
Frequency = 52.1785714285714
[1] -2.2608077 -0.4609664 -1.6837146 0.1486230 -0.1673218 1.9559586
[7] -0.4381615 0.7410979 0.7498425 1.1168765 -1.6807615 -0.1288980
[13] 0.8818968 1.0251742 -1.0492848 -1.6525099 -0.2998529 0.3871222
[19] -0.5864519 0.4392477
```

周数据，一周采一个点，采了 20 个点



## 25.4 时序检验

参数的计算公式，实现的 R 代码

- Applies linear filtering to a univariate time series or to each series separately of a multivariate time series. 过滤

一元时间序列的线性过滤，或者对多元时间序列的单个序列分别做线性过滤

$$y[i] = x[i] + f[1] * y[i - 1] + \dots + f[p] * y[i - p]$$

$$y[i] = f[1] * x[i + o] + \dots + f[p] * x[i + o - (p - 1)]$$

其中  $o$  代表 offset

介绍 FFT 算法细节

不同的方法对时间序列平滑的影响 FFT 快速傅里叶变换算法

```
usage(stats::filter)
```

```
filter(x, filter, method = c("convolution", "recursive"), sides = 2L,
circular = FALSE, init = NULL)
```

- `filter()` 时间序列线性过滤
- `fft()` 快速离散傅里叶变换

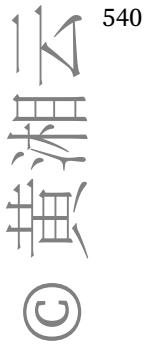
## 25.5 指数平滑

### 25.6 Holt-Winters

可加 Holt-Winters [Winters, 1960, Holt, 2004] 预测函数，周期长度为  $p$

$$\hat{Y}[t + h] = a[t] + h * b[t] + s[t - p + 1 + (h - 1) \mod p]$$

其中  $a[t], b[t], s[t]$  由以下决定



$$a[t] = \alpha(Y[t] - s[t-p]) + (1-\alpha)(a[t-1] + b[t-1]) \quad (25.1)$$

$$b[t] = \beta(a[t] - a[t-1]) + (1-\beta)b[t-1] \quad (25.2)$$

$$s[t] = \gamma(Y[t] - a[t]) + (1-\gamma)s[t-p] \quad (25.3)$$

可乘 Holt-Winters

$$\hat{Y}[t+h] = (a[t] + h * b[t]) * s[t-p+1 + (h-1) \bmod p]$$

其中  $a[t], b[t], s[t]$  由如下决定

$$a[t] = \alpha(Y[t]/s[t-p]) + (1-\alpha)(a[t-1] + b[t-1]) \quad (25.4)$$

$$b[t] = \beta(a[t] - a[t-1]) + (1-\beta)b[t-1] \quad (25.5)$$

$$s[t] = \gamma(Y[t]/a[t]) + (1-\gamma)s[t-p] \quad (25.6)$$

`HoltWinters()` 用 Shiny App / 动画的形式展示  $\alpha, \beta, \gamma$  三个参数对模型预测的影响，参数的确定通过最小化预测均方误差

```
Seasonal Holt-Winters
(m <- HoltWinters(co2))
plot(m)
plot(fitted(m))

p <- predict(m, 50, prediction.interval = TRUE)
plot(m, p)

(m <- HoltWinters(AirPassengers, seasonal = "mult"))
plot(m)

指数平滑 Exponential Smoothing
m2 <- HoltWinters(x, gamma = FALSE, beta = FALSE)
lines(fitted(m2)[,1], col = 3)
```

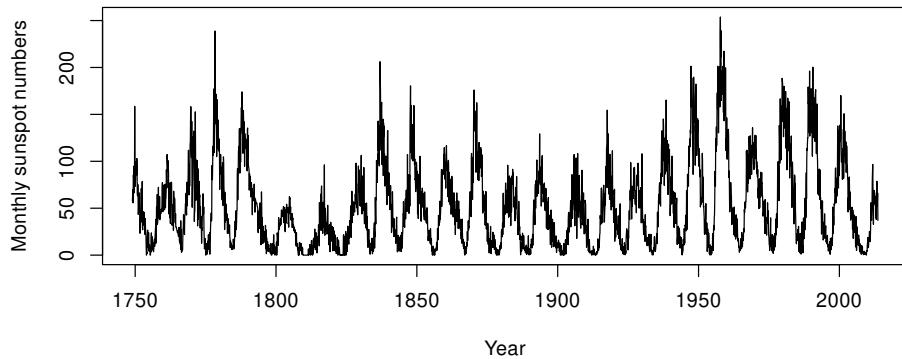
## 25.7 1749-2013 年太阳黑子数据

再从官网拿到最近的数据

```
plot(sunspot.month, xlab = "Year", ylab = "Monthly sunspot numbers",
 main = "Monthly mean relative sunspot numbers from 1749 to 2013")

autoplot(sunspot.month,
 main = "Monthly mean relative sunspot numbers from 1749 to 2013",
 xlab = "Year", ylab = "Monthly sunspot numbers"
)
```

Monthly mean relative sunspot numbers from 1749 to 2013



Monthly mean relative sunspot numbers from 1749 to 2013

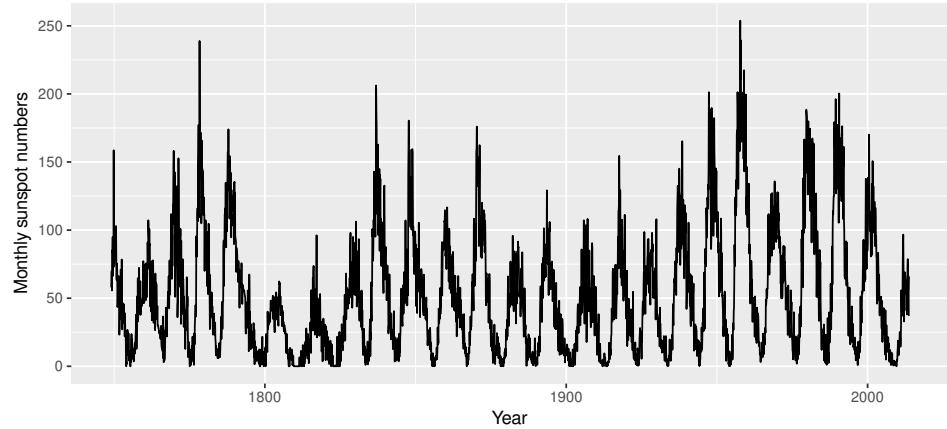
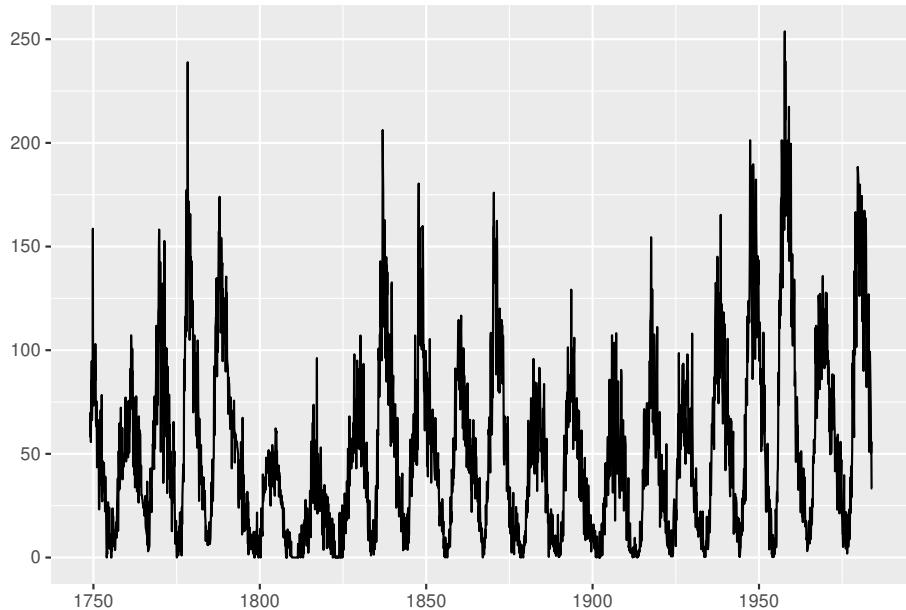


图 25.2: 时序图: 太阳黑子月均数量

```
autoplot(sunspots)
```



```
autoplot(sunspot.year, xlab = "Year", ylab = "Yearly Sunspot Data, 1700-1988") +
 theme_minimal()
```

```
library(dygraphs)
hw <- HoltWinters(sunspot.month)
predicted <- predict(hw, n.ahead = 72, prediction.interval = TRUE)

dygraph(predicted, main = "Predicted sunspot numbers") %>%
 dyAxis("x", drawGrid = FALSE) %>%
 dySeries(c("lwr", "fit", "upr"), label = "sunspot") %>%
 dyOptions(colors = hcl.colors(3))

par(family = "source-han-sans-cn")
plot(sunspot.month, col = "black")
lines(sunspots, col = "red")
legend("topright", legend = c("1749 至今", "1749-1983"), col = c("black", "red"), lty =
```

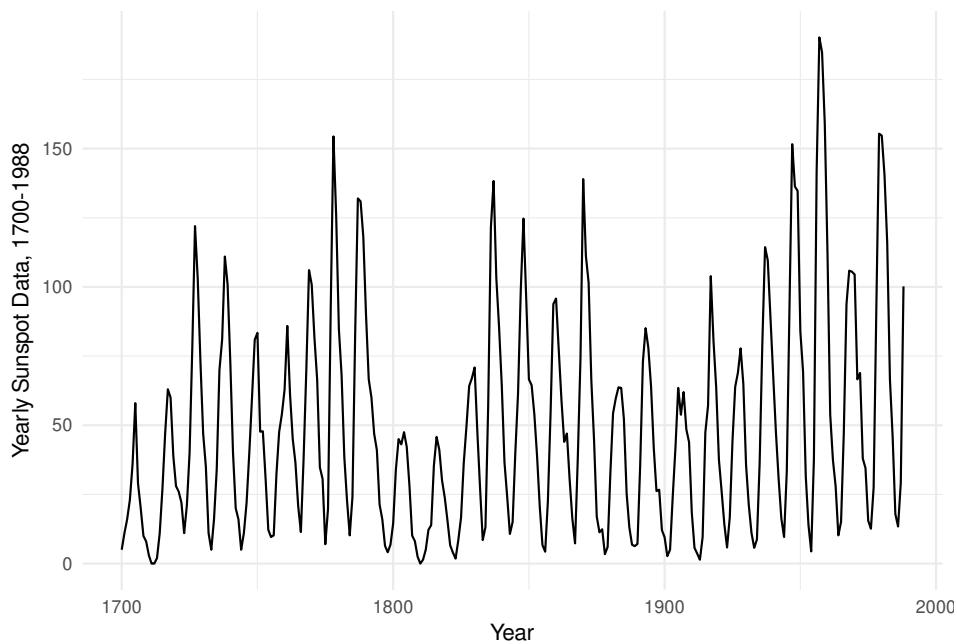


图 25.3: 太阳黑子数量年平均时序图

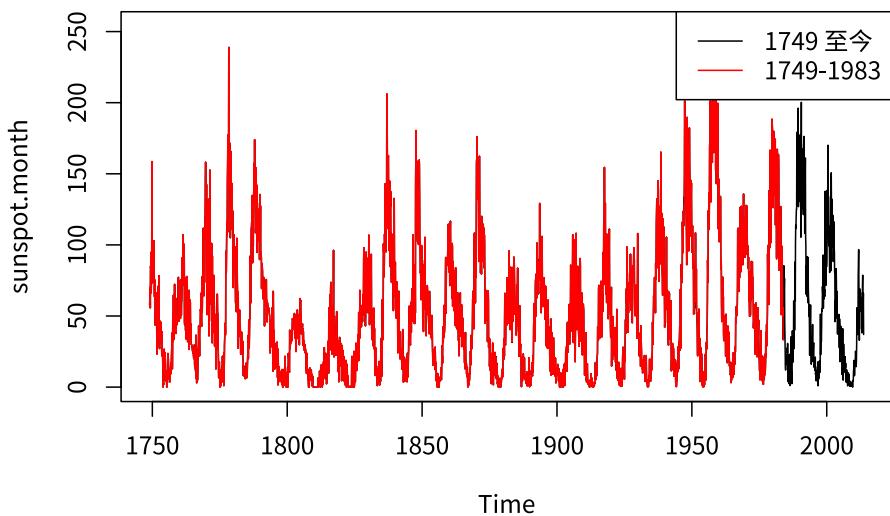


图 25.4: 月均太阳黑子数

## 25.8 1991-1998 年欧洲主要股票市场日闭市价格指数

```
matplot(time(EuStockMarkets), EuStockMarkets,
 main = "",
 xlab = "Date", ylab = "closing prices",
 pch = 17, type = "l", col = 1:4
)
legend("topleft", colnames(EuStockMarkets), pch = 17, lty = 1, col = 1:4)
```

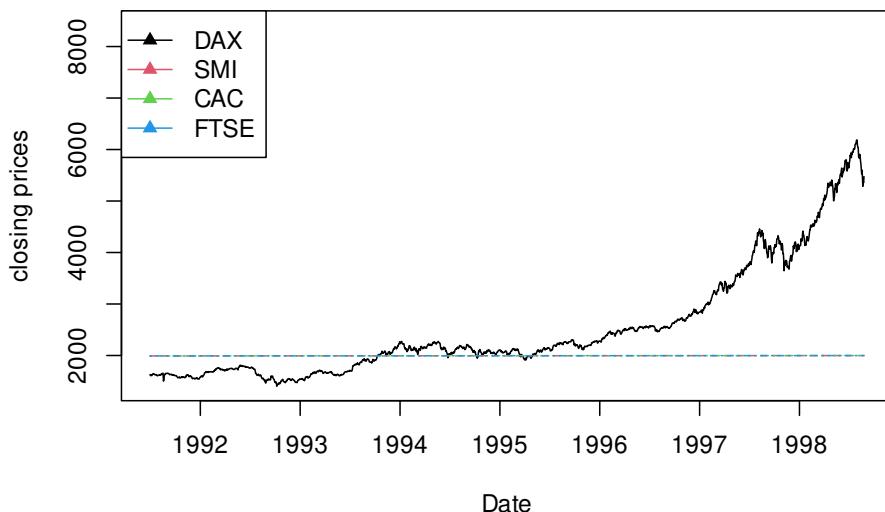
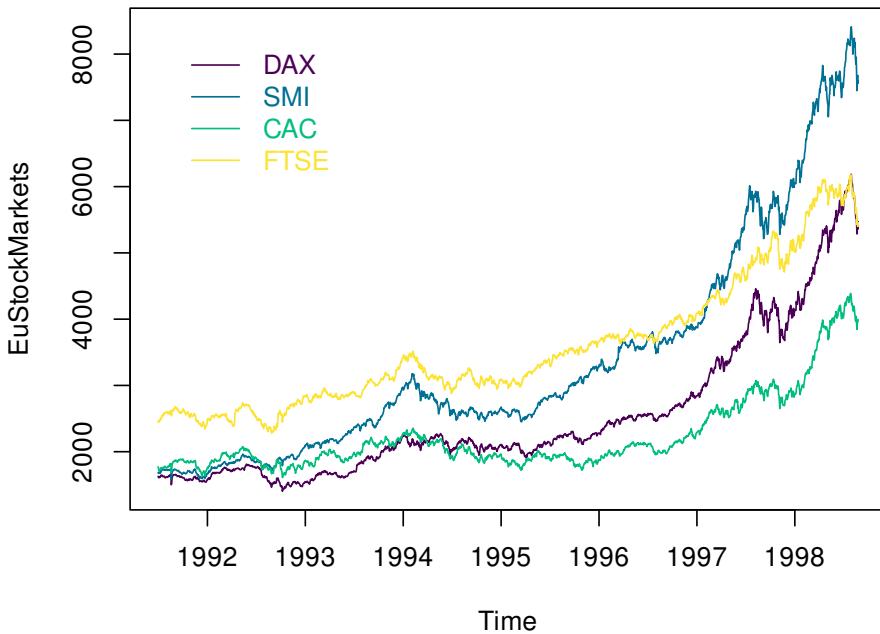


图 25.5: 1991-1998 年间欧洲主要股票市场日闭市价格指数图德国 DAX (Ibis), Switzerland SMI, 法国 CAC 和英国 FTSE

```
考虑收集加入最新的数据 1991~1998年的数据
plot(EuStockMarkets, plot.type = "single", col = hcl.colors(4))
legend("topleft", colnames(EuStockMarkets),
 col = hcl.colors(4), text.col = hcl.colors(4), lty = 1,
 box.col = NA, inset = 0.05
)
```



## 25.9 自回归模型

ar()

## 25.10 移动平均模型

arima()

## 25.11 自回归移动平均模型

arima() ARIMA

## 25.12 自回归条件异方差模型

自回归条件异方差模型 ARCH

## 25.13 广义自回归条件异方差模型

广义自回归条件异方差模型 (Generalized Autoregressive Conditional Heteroskedasticity, 简称 GARCH )

## 25.14 其它特征的时间序列

```
plot(JohnsonJohnson)
plot(AirPassengers)
plot(nottem)
plot(lynx)
```

## 25.15 港股走势

美团、阿里巴巴在香港上市

```
美团
meituan <- quantmod:::getSymbols("3690.HK", auto.assign = FALSE, src = "yahoo", from = '2019-01-01')
阿里
ali <- quantmod:::getSymbols("9988.HK", auto.assign = FALSE, src = "yahoo", from = '2019-01-01')
京东
sw <- quantmod:::getSymbols("9618.HK", auto.assign = FALSE, src = "yahoo", from = '2019-01-01')
腾讯
tx <- quantmod:::getSymbols("0700.HK", auto.assign = FALSE, src = "yahoo", from = '2019-01-01')

如何共 x 轴, 右对齐
plot(as.ts(meituan[, "3690.HK.Close"]), col = "orange", ylab = "股价")
lines(as.ts(alix[, "9988.HK.Close"])), col = "springgreen4")
lines(as.ts(sw[, "9618.HK.Close"])), col = "purple4")
```

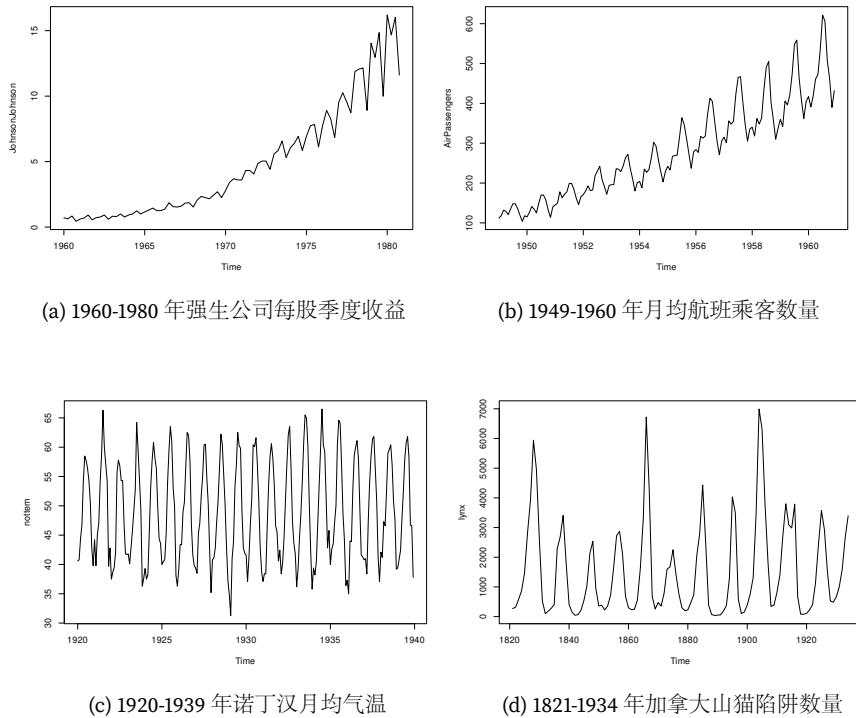


图 25.6: 时间序列: 非平稳、周期性、非线性



```
lines(as.ts(tx[, "0700.HK.Close"]), col = "lightsteelblue4")
legend("topright", col = c("Orange", "springgreen4", "purple4", "lightsteelblue4"),
 lty = 1, legend = c("美团", "阿里", "京东", "腾讯")))
```



## 25.16 美股走势

拼多多、京东、阿里巴巴、51Talk 在美股上市

```
拼多多
pdd <- quantmod::getSymbols("PDD", auto.assign = FALSE, src = "yahoo")
京东
jd <- quantmod::getSymbols("JD", auto.assign = FALSE, src = "yahoo")
阿里巴巴
baba <- quantmod::getSymbols("BABA", auto.assign = FALSE, src = "yahoo")
51Talk
coe <- quantmod::getSymbols("COE", auto.assign = FALSE, src = "yahoo", from = '2016-06-
```

## 25.17 51Talk 股价走势

Joshua M. Ulrich 开发维护的 `quantmod` 包可以下载国内外股票市场的数据

51Talk 于 2016 年 6 月 10 日在美国纽交所上市，股票代码 COE，2020 年 1 月 22 日，武汉封城，受新冠肺炎病毒影响，政府停课不停学的号召，线下教育纷纷转线上，线上教育的春天来临，股价开始回升到发行价的水平，在公司将资源转变为能力后，预期公司股价继续翻倍，回到理性的水平。

```
coe <- quantmod::getSymbols("COE", auto.assign = FALSE, src = "yahoo", from = '2016-06-
```

读者可以从雅虎财经获取数据源 <https://finance.yahoo.com/>

```
plot(coe[, "COE.Close"],
 subset = "2016-06-30/2021-06-30",
 col = "Orange", main = "COE Stock Close Price"
)
```

COE 股价变化趋势见下图，包含开盘价 Open、最低价 Low、最高价 High、闭市价 Close 和调整价 Adjust 和交易额 Volume



图 25.7: 51Talk 公司上市以来的股价走势

```
autoplot(coe)
```

## 25.18 运行环境

```
sessionInfo()
```

```
R version 4.1.0 (2021-05-18)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 20.04.2 LTS
##
Matrix products: default
BLAS: /usr/lib/x86_64-linux-gnublas/libblas.so.3.9.0
LAPACK: /usr/lib/x86_64-linux-gnulapack/liblapack.so.3.9.0
##
locale:
[1] LC_CTYPE=en_US.UTF-8 LC_NUMERIC=C
```

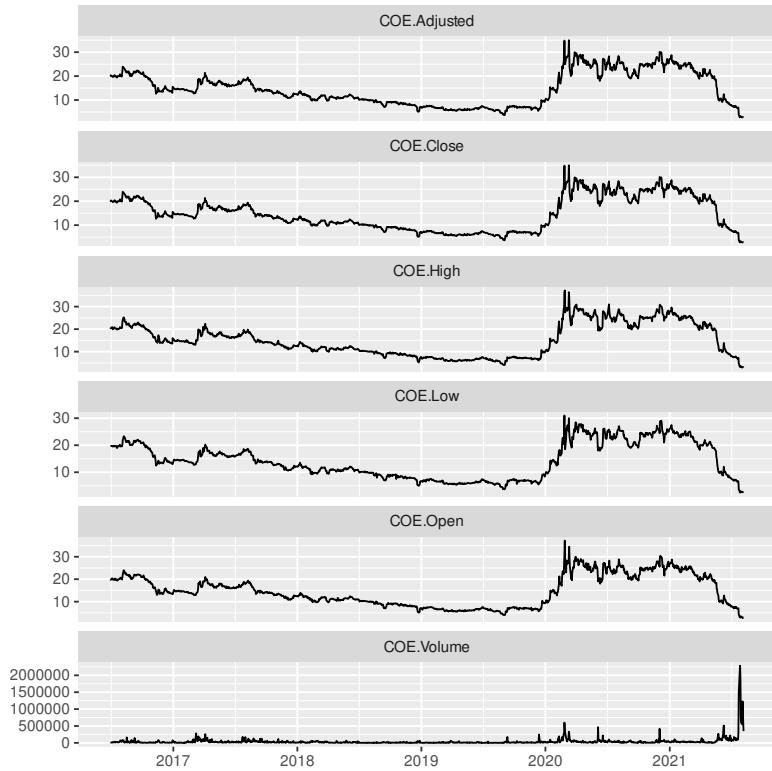


图 25.8: CEO 股价变化趋势



```
[3] LC_TIME=en_US.UTF-8 LC_COLLATE=en_US.UTF-8
[5] LC_MONETARY=en_US.UTF-8 LC_MESSAGES=en_US.UTF-8
[7] LC_PAPER=en_US.UTF-8 LC_NAME=C
[9] LC_ADDRESS=C LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
attached base packages:
[1] stats graphics grDevices utils datasets methods base
##
other attached packages:
[1] dygraphs_1.1.1.6 highcharter_0.8.2 ggfortify_0.4.12 ggplot2_3.3.5
[5] formatR_1.11
##
loaded via a namespace (and not attached):
[1] zoo_1.8-9 tidyselect_1.1.1 xfun_0.24 purrr_0.3.4
[5] lattice_0.20-44 colorspace_2.0-2 vctrs_0.3.8 generics_0.1.0
[9] htmltools_0.5.1.1 yaml_2.2.1 utf8_1.2.2 rlang_0.4.11
[13] pillar_1.6.2 glue_1.4.2 withr_2.4.2 DBI_1.1.1
[17] TTR_0.24.2 lifecycle_1.0.0 quantmod_0.4.18 stringr_1.4.0
[21] munsell_0.5.0 gtable_0.3.0 htmlwidgets_1.5.3 evaluate_0.14
[25] labeling_0.4.2 knitr_1.33 curl_4.3.2 fansi_0.5.0
[29] broom_0.7.9 xts_0.12.1 Rcpp_1.0.7 scales_1.1.1
[33] backports_1.2.1 showtext_0.9-3 jsonlite_1.7.2 sysfonts_0.8.4
[37] farver_2.1.0 gridExtra_2.3 digest_0.6.27 stringi_1.7.3
[41] showtextdb_3.0 rlist_0.4.6.1 bookdown_0.22 dplyr_1.0.7
[45] grid_4.1.0 tools_4.1.0 magrittr_2.0.1 tibble_3.1.3
[49] crayon_1.4.1 tidyverse_1.1.3 pkgconfig_2.0.3 ellipsis_0.3.2
[53] data.table_1.14.0 lubridate_1.7.10 assertthat_0.2.1 rmarkdown_2.9
[57] R6_2.5.0 igraph_1.2.6 compiler_4.1.0
```



## 第二十六章 空间分析

`maps` 是 `cartography` 的继任者，它更加友好、轻量和稳健。

`choroplethr` 简化创建 thematic maps 的过程。

`ggmap` 依赖 `RgoogleMaps` 就不介绍了

`mapdeck` 支持调用 GPU 渲染 `deck.gl` MIT 协议

`googleway`

Edzer Pebesma

- UseR2020 [Analyzing and visualising spatial and spatiotemporal data cubes - Part I](#)
- UseR2019 [UseR! 2019 Spatial workshop part I](#) [UseR! 2019 Spatial workshop part II](#)
- UseR2017 [Spatial Data in R: New Directions](#)
- UseR2016 [Handling and Analyzing Spatial, Spatiotemporal and Movement Data](#)

```
library(sp)
library(RColorBrewer)
library(raster)
library(lattice)
library(latticeExtra)
library(terra) #
library(rasterVis) # https://oscarperpinan.github.io/rastervis/
https://oscarperpinan.github.io/rastervis/FAQ.html
library(sf)
library(sfarrow) # https://github.com/wcjochem/sfarrow
library(arrow) # 列式存储
```



```
library(rgdal) # 要替换掉
library(highcharter) # 要替换掉

library(RgoogleMaps)
library(mapdeck)
library(mapsf)
```

## 26.1 冈比亚儿童疟疾

冈比亚地形

```
sp_path <- "data/" # 存储临时地形文件
if (!dir.exists(sp_path)) dir.create(sp_path, recursive = TRUE)
Gambia 海拔数据
gambia_alt <- raster:::getData(name = "alt", country = "GMB", mask = TRUE, path = sp_path)
Gambia 市级行政边界数据
gambia_map <- raster:::getData("GADM", country = "GMB", level = 2, path = sp_path)
绘制冈比亚地形
rasterVis:::levelplot(gambia_alt,
 margin = FALSE,
 main = "Elevation",
 colorkey = list(
 space = "top",
 labels = list(at = seq(from = -5, to = 65, by = 10)),
 axis.line = list(col = "black")
),
 par.settings = list(
 axis.line = list(col = "transparent")
),
 scales = list(draw = FALSE),
 col.regions = hcl.colors,
 at = seq(-5, 65, len = 101)
) +
 latticeExtra:::layer(sp::sp.polygons(gambia_map, lwd = 1.5))
```

rgdal 包可以实现坐标变换

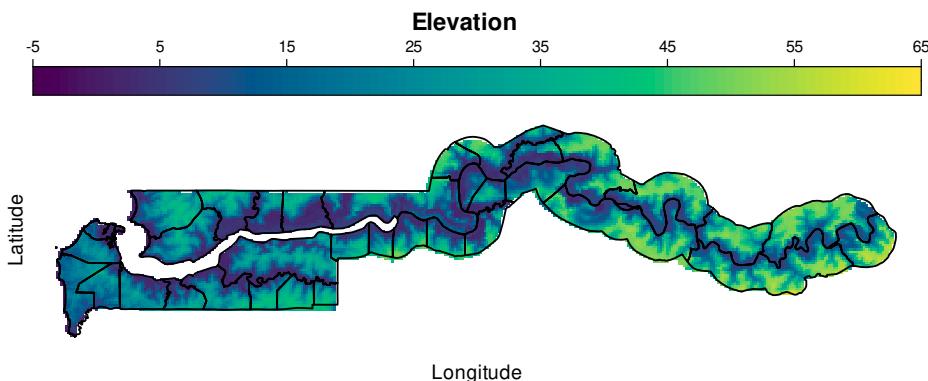


图 26.1: 冈比亚地形海拔数据

```
加载数据
data(gambia, package = "geoR")
坐标变换
library(sp)
sps <- SpatialPoints(gambia[, c("x", "y")],
proj4string = CRS("+proj=utm +zone=28"))
)
spst <- spTransform(sps, CRS("+proj=longlat +datum=WGS84"))
gambia[, c("x", "y")] <- coordinates(spst)
聚合数据
gambia_agg <- aggregate(
 formula = cbind(pos, netuse, treated) ~ x + y + green + phc,
 data = gambia, FUN = function(x) sum(x) / length(x)
)
抽取指定位置的海拔数据
raster::extract(gambia_alt, gambia[, c("x", "y")])
```

$Y \sim b(1, p)$  每个人检验结果，就是感染 1 或是没有感染 0，感染率  $p$  的建模分析，个体水平

```
library(highcharter)
hchart(gambia_agg, "bubble", hcaes(x = x, y = y, fill = pos, size = pos),
```



```
maxSize = "5%", name = "Gambia", showInLegend = FALSE
) %>%
 hc_yAxis(title = list(text = "Latitude")) %>%
 hc_xAxis(title = list(text = "Longitude"), labels = list(align = "center")) %>%
 hc_colorAxis(
 stops = color_stops(colors = hcl.colors(palette = "Plasma", n = 10))
) %>%
 hc_tooltip(
 pointFormat = "({point.x:.2f}, {point.y:.2f})
 Size: {point.z:.2f}"
)

gm_data <- download_map_data("https://code.highcharts.com/mapdata/countries/gm/gm-all.js")
get_data_from_map(gm_data)

hcmap("countries/gm/gm-all.js") %>%
 hc_title(text = "Gambia")

data("USArrests", package = "datasets")
data("usgeojson") # 加载地图数据 地图数据的结构

USArrests <- transform(USArrests, state = rownames(USArrests))

highchart() %>%
 hc_title(text = "Violent Crime Rates by US State") %>%
 hc_subtitle(text = "Source: USArrests data") %>%
 hc_add_series_map(usgeojson, USArrests,
 name = "Murder arrests (per 100,000)",
 value = "Murder", joinBy = c("woename", "state"),
 dataLabels = list(
 enabled = TRUE,
 format = "{point.properties.postalcode}"
)
) %>%
 hc_colorAxis(stops = color_stops()) %>%
 hc_legend(valueDecimals = 0, valueSuffix = "%") %>%
 hc_mapNavigation(enabled = TRUE)
```



highcharter 包含三个数据集分别是: worldgeojson 世界地图(国家级)、usgeojson 美国地图(州级)、uscountygeojson 美国地图(城镇级)。其它地图数据见 <https://code.highcharts.com/mapdata/>。

```
添加地图数据
hcmap(map = "countries/cn/custom/cn-all-sar-taiwan.js") %>%
 hc_title(text = "中国地图")

library(mapdeck)
多边形
mapdeck() %>%
 add_polygon(
 data = spatialwidget::widget_melbourne,
 fill_colour = "SA2_NAME",
 palette = "spectral"
)

mapdeck(location = c(145, -37.8), zoom = 10) %>%
 add_geojson(
 data = mapdeck::geojson
)
```

## 26.2 运行环境

```
sessionInfo()

R version 4.1.0 (2021-05-18)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 20.04.2 LTS
##
Matrix products: default
BLAS: /usr/lib/x86_64-linux-gnublas/libblas.so.3.9.0
LAPACK: /usr/lib/x86_64-linux-gnulapack/liblapack.so.3.9.0
##
locale:
[1] LC_CTYPE=en_US.UTF-8 LC_NUMERIC=C
```



```
[3] LC_TIME=en_US.UTF-8 LC_COLLATE=en_US.UTF-8
[5] LC_MONETARY=en_US.UTF-8 LC_MESSAGES=en_US.UTF-8
[7] LC_PAPER=en_US.UTF-8 LC_NAME=C
[9] LC_ADDRESS=C LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
attached base packages:
[1] stats graphics grDevices utils datasets methods base
##
other attached packages:
[1] sfarrow_0.4.0 sf_1.0-2 rasterVis_0.50.3
[4] terra_1.3-4 latticeExtra_0.6-29 lattice_0.20-44
[7] raster_3.4-13 RColorBrewer_1.1-2 sp_1.4-5
##
loaded via a namespace (and not attached):
[1] zoo_1.8-9 tidyselect_1.1.1 xfun_0.24 purrrr_0.3.4
[5] vctrs_0.3.8 generics_0.1.0 htmltools_0.5.1.1 viridisLite_0.4.0
[9] yaml_2.2.1 utf8_1.2.2 rlang_0.4.11 e1071_1.7-8
[13] hexbin_1.28.2 pillar_1.6.2 glue_1.4.2 DBI_1.1.1
[17] jpeg_0.1-9 lifecycle_1.0.0 stringr_1.4.0 codetools_0.2-18
[21] evaluate_0.14 knitr_1.33 parallel_4.1.0 class_7.3-19
[25] fansi_0.5.0 Rcpp_1.0.7 KernSmooth_2.23-20 classInt_0.4-3
[29] png_0.1-7 digest_0.6.27 stringi_1.7.3 bookdown_0.22
[33] dplyr_1.0.7 grid_4.1.0 rgdal_1.5-23 tools_4.1.0
[37] magrittr_2.0.1 proxy_0.4-26 tibble_3.1.3 crayon_1.4.1
[41] pkgconfig_2.0.3 ellipsis_0.3.2 assertthat_0.2.1 rmarkdown_2.9
[45] R6_2.5.0 units_0.7-2 compiler_4.1.0
```

## 第二十七章 空间建模

```
library(geoR)
library(INLA)
library(leaflet)
library(highcharter)
```

### 27.1 西非眼线虫病

loaloa 眼线虫病，人群感染，村庄水平，响应变量服从二项分布  $Y \sim b(n, p)$ ，每个村庄感染的人数  $Y_i \sim b(n_i, p_i)$  其中  $n_i$  是第  $i$  个村庄调查的人数， $p_i$  是观测的感染率

```
data("loaloa", package = "PrevMap")
hcmap(map = "countries/cm/cm-all.js") %>%
 hc_title(text = "喀麦隆及其周边地区眼线虫病流行度")
```

### 27.2 运行环境

```
sessionInfo()

R version 4.1.0 (2021-05-18)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 20.04.2 LTS
##
Matrix products: default
```



```
BLAS: /usr/lib/x86_64-linux-gnublas/libblas.so.3.9.0
LAPACK: /usr/lib/x86_64-linux-gnulapack/liblapack.so.3.9.0
##
locale:
[1] LC_CTYPE=en_US.UTF-8 LC_NUMERIC=C
[3] LC_TIME=en_US.UTF-8 LC_COLLATE=en_US.UTF-8
[5] LC_MONETARY=en_US.UTF-8 LC_MESSAGES=en_US.UTF-8
[7] LC_PAPER=en_US.UTF-8 LC_NAME=C
[9] LC_ADDRESS=C LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
attached base packages:
[1] parallel stats graphics grDevices utils datasets methods
[8] base
##
other attached packages:
[1] highcharter_0.8.2 geoR_1.8-1 INLA_21.02.23 sp_1.4-5
[5] foreach_1.5.1 Matrix_1.3-4
##
loaded via a namespace (and not attached):
[1] zoo_1.8-9 tidyselect_1.1.1 xfun_0.24
[4] purrr_0.3.4 splines_4.1.0 lattice_0.20-44
[7] tcltk_4.1.0 vctrs_0.3.8 generics_0.1.0
[10] htmltools_0.5.1.1 yaml_2.2.1 utf8_1.2.2
[13] rlang_0.4.11 pillar_1.6.2 glue_1.4.2
[16] DBI_1.1.1 TTR_0.24.2 lifecycle_1.0.0
[19] quantmod_0.4.18 stringr_1.4.0 htmlwidgets_1.5.3
[22] codetools_0.2-18 evaluate_0.14 knitr_1.33
[25] curl_4.3.2 fansi_0.5.0 broom_0.7.9
[28] xts_0.12.1 Rcpp_1.0.7 backports_1.2.1
[31] jsonlite_1.7.2 digest_0.6.27 stringi_1.7.3
[34] rlist_0.4.6.1 bookdown_0.22 dplyr_1.0.7
[37] splancs_2.01-42 grid_4.1.0 tools_4.1.0
[40] magrittr_2.0.1 tibble_3.1.3 crayon_1.4.1
[43] tidyr_1.1.3 pkgconfig_2.0.3 MASS_7.3-54
[46] ellipsis_0.3.2 data.table_1.14.0 RandomFieldsUtils_0.5.3
```

```
[49] RandomFields_3.3.8 lubridate_1.7.10 assertthat_0.2.1
[52] rmarkdown_2.9 iterators_1.0.13 R6_2.5.0
[55] igraph_1.2.6 compiler_4.1.0
```



## 第二十八章 贝叶斯模型

`LaplaceDemon` 支持常见模型的贝叶斯推断, 具体可见[网站 \[Statisticat and LLC., 2021\]](#), `shinystan` 借助 `rstan` 打包了一些 `stan` 编写的统计模型, 提供模型评估的功能。相比于 `rstan`, `brms` 支持了更加广泛的模型, `shinybrms` 类似 `shinystan` 提供可视化的 `shiny` 前端, 方便用户调用模型和评估效果。`rstanarm` 基于 `stan` 语言重写了 `arm` 里的模型, 和 `brms` 一样, 提供类似 `lme4` 的公式语法, 和 Base R 内置的函数 `lm()` 和 `glm()` 保持一致, 降低用户学习成本。

`cmdstanr` 相比于 `rstan` 将会更加轻量, 更快地将 CmdStan 的新功能融入进来, 方便用户滚动升级, 相比于 `rstan` 包, `cmdstanr` 包的一个巨大优势是和 Stan 软件的更新分离。做贝叶斯计算的软件框架还包括 JAGS 和 WinBUGS, 苏毓松开发的 R2jags 包 [[Su and Yajima, 2020](#)] 是 JAGS 的 R 接口。

### TMB

- 书籍: [Richard McElreath](#) 为《Statistical Rethinking》写的 `rethinking` 包, 参考 Derek S. Young [[Young, 2017](#)] 和 Michael H. Kutner 等 [[Kutner et al., 2005](#)]
- 论文: An Introduction to Inductive Statistical Inference: from Parameter Estimation to Decision-Making <https://arxiv.org/abs/1808.10173v2> 固定效应/随机效应广义线性模型: 多水平各种模型回归, 模型结构和 Stan 代码
- 课程: 线性模型的内容主要分为四大块, 分别是线性回归模型、方差分析模型、协方差分析模型和线性混合效应模型。国外 David Pollard 的线性模型 [课程内容](#)
- 会议: Paul-Christian Bürkner 在 Stan 大会上介绍 `brms` 和 `rstanarm` <https://github.com/InsuranceDataScience/StanWorkshop2018>

## 28.1 软件配置

从 GitHub 下载最新版的源码包 <https://github.com/stan-dev/cmdstan/releases/latest>, 编译二进制版本

```
tar -xzf /Users/xiangyun/Desktop/cmdstan-2.26.0.tar.gz -C /opt/
cd cmdstan-2.26.0
make build
```

设置环境变量 `CMDSTAN` 指向 CmdStan 安装路径, 加载 `cmdstanr` 包会自动检测和加载

```
Sys.setenv(CMDSTAN="/opt/cmdstan-2.26.0")
```

还可以设置环境变量 `CMDSTANR_NO_VER_CHECK=TRUE`, 让 `cmdstanr` 不要检查 CmdStan 版本状态, 是不是最新版, 比如本书将固定下 CmdStan 版本为 2.26.0

`cmdstanr` 当前还在开发中, 安装方式如下

```
remotes::install_github('stan-dev/cmdstanr')
或者
install.packages("cmdstanr", repos = c("https://mc-stan.org/r-packages/", getOption("re
```

另有一篇博文介绍在 Windows 系统上安装 `cmdstanr` 的过程, 这里不做展开。

```
rstan
brms
rstanarm
remotes::install_github('rmcelreath/rethinking')
```

## 28.2 正态分布

我们以估计正态分布参数为例说明贝叶斯估计方法

$$Y \sim \mathcal{N}(\mu, \sigma^2)$$

已知  $y_1, y_2, \dots, y_n$  是来自正态总体  $\mathcal{N}(\mu, \sigma^2)$  的一个样本, 我们需要估计这个正态分布模型的参数  $\mu$  和  $\sigma^2$ 。

最大似然估计, 简单推导过程, 计算代码; 再讲 `stan` 的计算步骤



```
library(cmdstanr)
mod <- cmdstan_model(stan_file = "code/normal_dist.stan", compile = TRUE)
```

打包观测数据，初始化待估参数值，指定链条数，其中 `dataList` 必须与 `stan` 代码中数据块声明保持一致（如变量名称，长度），每条链使用不同的初始值，选择合适的初始值可以有效地提高收敛的速度。

```
数据准备
set.seed(20190427)
设置参数
mu <- 10
sd <- 2
样本量
nobs <- 500
nchains <- 4
生成随机数
y <- rnorm(n = nobs, mean = mu, sd = sd)
给每条链设置不同的参数初始值
inits_data <- lapply(1:nchains, function(i) {
 list(
 mu = runif(1, min(y), max(y)),
 sigma = runif(1, 1, 10)
)
})
```

将参数初值代入模型，抽样，获取参数的后验分布

```
normal_fit <- mod$sample(
 data = list(
 N = nobs,
 y = y
),
 init = inits_data,
 iter_warmup = 1000, # 每条链预处理迭代次数
 iter_sampling = 2000, # 每条链总迭代次数
 chains = nchains, # 马尔科夫链的数目
 parallel_chains = 1, # 指定 CPU 核心数，可以给每条链分配一个
 show_messages = FALSE, # 不显示迭代的中间过程
```

湘  
黄  
云

```

refresh = 0, # 不显示采样的进度
seed = 20190425 # 设置随机数种子, 不要使用 set.seed() 函数
)

Running MCMC with 4 sequential chains...
##
Chain 1 finished in 0.1 seconds.
Chain 2 finished in 0.0 seconds.
Chain 3 finished in 0.0 seconds.
Chain 4 finished in 0.0 seconds.
##
All 4 chains finished successfully.
Mean chain execution time: 0.0 seconds.
Total execution time: 0.6 seconds.

```

检查收敛性，Rhat 决定收敛性，所有待估参数的 Rhat 必须小于 1.1，同时有效样本数量 n\_eff 除以抽样总数 N 必须小于 0.001，否则收敛性是值得怀疑的。拟合结果及解释如下：

```
模型参数估计结果
```

```
normal_fit$cmdstan_summary()
```

```

Inference for Stan model: normal_dist_model
4 chains: each with iter=(2000,2000,2000,2000); warmup=(0,0,0,0); thin=(1,1,1,1); 8000
##
Warmup took (0.012, 0.012, 0.013, 0.012) seconds, 0.049 seconds total
Sampling took (0.039, 0.036, 0.035, 0.033) seconds, 0.14 seconds total
##
Mean MCSE StdDev 5% 50% 95% N_Eff N_Eff/s R_hat
##
lp__ -602 1.7e-02 1.0 -604 -601 -601 3591 25114 1.0
accept_stat__ 0.92 3.3e-03 0.11 0.69 0.96 1.0 1.1e+03 7.4e+03 1.0e+00
stepsize__ 0.88 6.9e-02 0.098 0.73 0.90 1.0 2.0e+00 1.4e+01 1.5e+13
treedepth__ 1.9 1.2e-01 0.56 1.0 2.0 3.0 2.1e+01 1.5e+02 1.1e+00
n_leapfrog__ 3.7 3.7e-01 1.8 1.0 3.0 7.0 2.4e+01 1.7e+02 1.0e+00
divergent__ 0.00 nan 0.00 0.00 0.00 0.00 nan nan nan
energy__ 603 2.5e-02 1.4 601 602 605 3.3e+03 2.3e+04 1.0e+00
##
```



```
mu 10 1.2e-03 0.092 9.9 10 10 5732 40084 1.00
sigma 2.0 7.7e-04 0.064 1.9 2.0 2.1 6885 48146 1.00
##
Samples were drawn using hmc with nuts.
For each parameter, N_Eff is a crude measure of effective sample size,
and R_hat is the potential scale reduction factor on split chains (at
convergence, R_hat=1).
```

调用 `draws` 方法 `normal_fit$draws()`, 获得一个由 **posterior** 构造的 `draws_array` 对象,

```
draws_array <- normal_fit$draws()
str(draws_array)
```

```
'draws_array' num [1:2000, 1:4, 1:3] -601 -601 -602 -601 -601 ...
- attr(*, "dimnames")=List of 3
..$ iteration: chr [1:2000] "1" "2" "3" "4" ...
..$ chain : chr [1:4] "1" "2" "3" "4"
..$ variable : chr [1:3] "lp__" "mu" "sigma"
```

采样结果可以直接传递给 **bayesplot** 包, 绘制参数的后验分布和马尔科夫链蒙特卡罗采样的轨迹图 (trace plot)。

```
library(bayesplot)
mcmc_trace(normal_fit$draws(c("mu", "sigma")))
mcmc_hist(normal_fit$draws(c("mu", "sigma")))
```

## 28.3 高斯过程

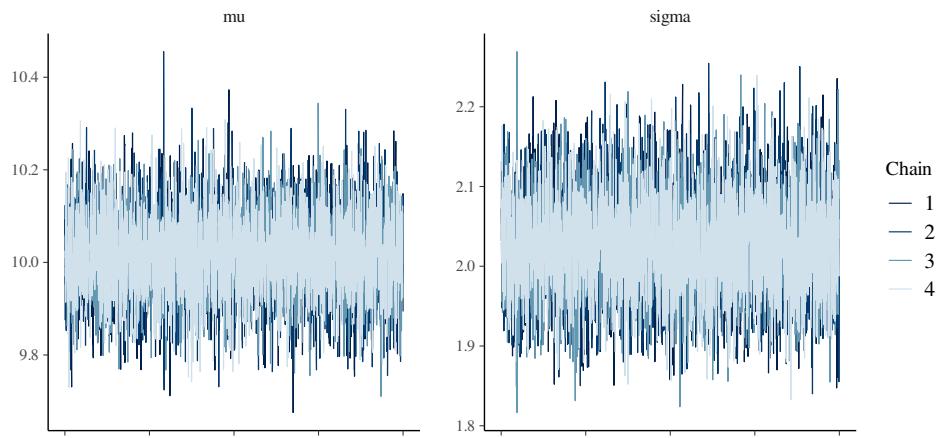
模拟高斯过程例子来自 Stan 参考手册 [[Stan Development Team, 2019](#)]

```
mod <- cmdstan_model(stan_file = "code/normal_gp.stan")
```

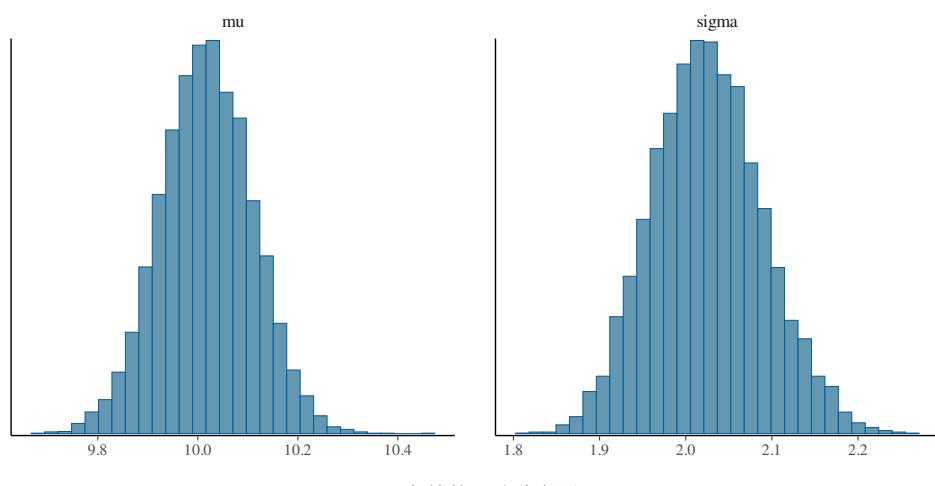
stan 库内置了核函数为二次幂指数的实现, 因此可以直接调用 `cov_exp_quad` 函数计算协方差矩阵

```
mod <- cmdstan_model(stan_file = "code/compat_gp.stan")
```

以 MASS 的 topo 数据集引出高斯过程回归模型问题复杂性



(a) 参数的轨迹图



(b) 参数的后验分布图

图 28.1: 参数  $\mu, \sigma$  的迭代轨迹图和后验分布图



## 28.4 分层正态模型

Multilevel Models 多水平模型、Hierarchical Models 层次模型

### 28.4.1 schools 数据

```
模型编译
mod <- cmdstan_model(stan_file = "code/eight_schools.stan")

模型拟合
eight_schools_fit <- mod$sample(
 data = list(# 观测数据
 J = 8,
 y = c(28, 8, -3, 7, -1, 1, 18, 12),
 sigma = c(15, 10, 16, 11, 9, 11, 10, 18)
),
 iter_warmup = 1000, # 每条链预处理迭代次数
 iter_sampling = 2000, # 每条链总迭代次数
 chains = 4, # 马尔科夫链的数目
 parallel_chains = 1, # 指定 CPU 核心数，可以给每条链分配一个
 show_messages = FALSE, # 不显示迭代的中间过程
 refresh = 0, # 不显示采样的进度
 seed = 20190425 # 设置随机数种子，不要使用 set.seed() 函数
)

Running MCMC with 4 sequential chains...
##
Chain 1 finished in 0.1 seconds.
Chain 2 finished in 0.1 seconds.
Chain 3 finished in 0.1 seconds.
Chain 4 finished in 0.1 seconds.
##
All 4 chains finished successfully.
Mean chain execution time: 0.1 seconds.
Total execution time: 0.5 seconds.
```

模型拟合结果

```

eight_schools_fit$cmdstan_summary()

Inference for Stan model: eight_schools_model
4 chains: each with iter=(2000,2000,2000,2000); warmup=(0,0,0,0); thin=(1,1,1,1); 8000
##
Warmup took (0.020, 0.023, 0.020, 0.023) seconds, 0.086 seconds total
Sampling took (0.060, 0.072, 0.073, 0.076) seconds, 0.28 seconds total
##
Mean MCSE StdDev 5% 50% 95% N_Eff N_Eff/s R_hat
##
lp__ -4.0e+01 5.4e-02 2.7 -44 -3.9e+01 -36 2447 8709 1.00
accept_stat__ 0.88 1.5e-02 0.20 0.40 0.96 1.0 1.8e+02 6.5e+02 1.0e+00
stepsize__ 0.34 3.2e-02 0.045 0.28 0.33 0.41 2.0e+00 7.1e+00 1.8e+13
treedepth__ 3.5 1.8e-01 0.54 3.0 4.0 4.0 8.4e+00 3.0e+01 1.1e+00
n_leapfrog__ 12 1.3e+00 4.0 7.0 15 15 9.9e+00 3.5e+01 1.1e+00
divergent__ 0.00 nan 0.00 0.00 0.00 0.00 nan nan nan
energy__ 45 7.2e-02 3.5 39 44 51 2.4e+03 8.6e+03 1.0e+00
##
mu 8.0e+00 8.1e-02 5.0 0.015 7.9e+00 17 3886 13829 1.0
tau 6.6e+00 1.0e-01 5.6 0.48 5.3e+00 17 3064 10904 1.0
eta[1] 3.9e-01 1.1e-02 0.95 -1.2 4.2e-01 1.9 7716 27458 1.00
eta[2] -4.0e-04 9.5e-03 0.88 -1.4 2.1e-04 1.4 8427 29989 1.00
eta[3] -2.0e-01 1.0e-02 0.94 -1.7 -2.0e-01 1.4 8319 29607 1.0
eta[4] -3.0e-02 9.7e-03 0.88 -1.5 -2.8e-02 1.4 8220 29253 1.0
eta[5] -3.7e-01 1.0e-02 0.88 -1.8 -3.9e-01 1.1 7355 26174 1.0
eta[6] -2.2e-01 9.8e-03 0.90 -1.7 -2.5e-01 1.3 8439 30032 1.0
eta[7] 3.5e-01 1.0e-02 0.88 -1.1 3.7e-01 1.8 7255 25817 1.0
eta[8] 5.4e-02 1.1e-02 0.93 -1.5 6.6e-02 1.6 7356 26177 1.00
theta[1] 1.1e+01 1.1e-01 8.4 0.10 1.0e+01 27 5668 20170 1.00
theta[2] 7.9e+00 6.8e-02 6.4 -2.5 7.9e+00 18 8940 31816 1.0
theta[3] 6.2e+00 9.3e-02 7.7 -7.5 6.6e+00 18 6961 24772 1.0
theta[4] 7.7e+00 7.1e-02 6.5 -3.0 7.7e+00 18 8515 30303 1.0
theta[5] 5.0e+00 7.0e-02 6.4 -6.4 5.5e+00 14 8218 29244 1.0
theta[6] 6.1e+00 7.3e-02 6.7 -5.7 6.5e+00 16 8504 30262 1.00
theta[7] 1.1e+01 8.4e-02 6.8 0.92 1.0e+01 23 6537 23263 1.00
theta[8] 8.5e+00 1.0e-01 7.8 -3.8 8.2e+00 22 5904 21009 1.0

```

```

Samples were drawn using hmc with nuts.
For each parameter, N_Eff is a crude measure of effective sample size,
and R_hat is the potential scale reduction factor on split chains (at
convergence, R_hat=1).
```

4条马尔可夫链，19个变量，2000次迭代，轨迹数据如下

```
eight_schools_fit$draws()
```

```
A draws_array: 2000 iterations, 4 chains, and 19 variables
, , variable = lp_--

chain
iteration 1 2 3 4
1 -42 -39 -38 -40
2 -37 -40 -37 -42
3 -40 -38 -38 -43
4 -39 -39 -40 -43
5 -38 -45 -38 -40

, , variable = mu

chain
iteration 1 2 3 4
1 9.0 -3.696 8.9 1.9
2 5.9 13.895 1.5 13.5
3 6.2 0.013 4.2 1.5
4 12.0 2.365 10.6 15.5
5 5.8 -7.633 5.2 12.2

, , variable = tau

chain
iteration 1 2 3 4
1 3.20 10.3 7.93 4.11
2 9.70 1.9 10.70 0.65
3 0.38 6.3 4.31 1.18
```

云湘黄

```
4 7.19 9.2 0.43 11.03
5 4.06 5.6 3.36 8.74
##
, , variable = eta[1]
##
chain
iteration 1 2 3 4
1 1.10 1.72 1.23 1.10
2 -0.30 -0.98 1.19 -0.74
3 -0.31 1.28 0.34 -0.85
4 0.97 1.53 -0.86 2.02
5 1.04 0.75 1.46 1.24
##
... with 1995 more iterations, and 15 more variables
```

提取参数  $\mu$  的四条迭代点列

```
eight_schools_fit$draws("mu")
```

```
A draws_array: 2000 iterations, 4 chains, and 1 variables
, , variable = mu
##
chain
iteration 1 2 3 4
1 9.0 -3.696 8.9 1.9
2 5.9 13.895 1.5 13.5
3 6.2 0.013 4.2 1.5
4 12.0 2.365 10.6 15.5
5 5.8 -7.633 5.2 12.2
##
... with 1995 more iterations
```

`eight_schools_fit` 是一个 R6 对象，包含整个模型信息

```
class(eight_schools_fit)
```

```
[1] "CmdStanMCMC" "CmdStanFit" "R6"
str(eight_schools_fit)
```

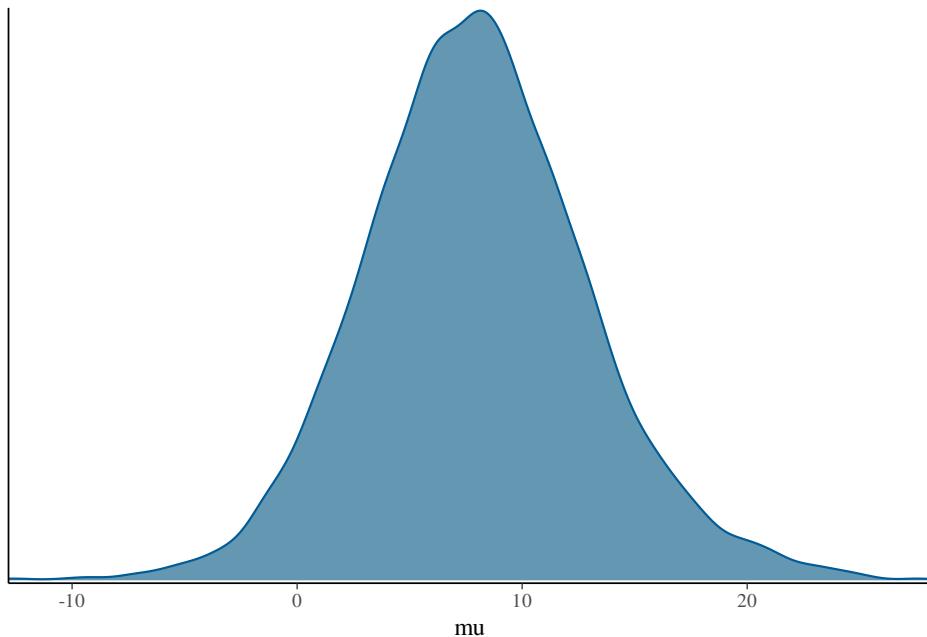
```
Classes 'CmdStanMCMC', 'CmdStanFit', 'R6' <CmdStanMCMC>
```

```
Inherits from: <CmdStanFit>
Public:
clone: function (deep = FALSE)
cmdstan_diagnose: function ()
cmdstan_summary: function (flags = NULL)
data_file: function ()
draws: function (variables = NULL, inc_warmup = FALSE, format = getOption("cmdstan"
init: function ()
initialize: function (runset)
inv_metric: function (matrix = TRUE)
latent_dynamics_files: function (include_failed = FALSE)
loo: function (variables = "log_lik", r_eff = TRUE, ...)
lp: function ()
metadata: function ()
num_chains: function ()
num_procs: function ()
output: function (id = NULL)
output_files: function (include_failed = FALSE)
print: function (variables = NULL, ..., digits = 2, max_rows = getOption("cmdstan"
profile_files: function (include_failed = FALSE)
profiles: function ()
return_codes: function ()
runset: CmdStanRun, R6
sampler_diagnostics: function (inc_warmup = FALSE, format = getOption("cmdstan"
save_data_file: function (dir = ".", basename = NULL, timestamp = TRUE, random = F
save_latent_dynamics_files: function (dir = ".", basename = NULL, timestamp = T
save_object: function (file, ...)
save_output_files: function (dir = ".", basename = NULL, timestamp = TRUE, random = F
save_profile_files: function (dir = ".", basename = NULL, timestamp = TRUE, random = F
summary: function (variables = NULL, ...)
time: function ()
Private:
draws_: -42.0125 -37.337 -40.2068 -38.9824 -38.4394 -38.7721 -35 ...
init_: NULL
inv_metric_: list
metadata_: list
```

```
read_csv_: function (variables = NULL, sampler_diagnostics = NULL, format = getOption("readr.format"))
sampler_diagnostics_: 3 3 4 3 4 3 3 3 3 2 3 3 3 3 3 4 3 3 3 3 2 3 3 3 3 3 3 ...
warmup_draws_: NULL
warmup_sampler_diagnostics_: NULL
```

模型诊断：查看迭代点列的平稳性

(C) `mcmc_dens(eight_schools_fit$draws(c("mu")))`



分层线性模型之生长曲线模型 [Gelfand et al., 1990]

#### 28.4.2 rats 数据

贝叶斯分层图

```
数据准备
modified code from https://github.com/stan-dev/example-models/tree/master/bugs_examples/R世俗
N <- 30
T <- 5
y <- structure(c(
 151, 145, 147, 155, 135, 159, 141, 159, 177, 134,
 160, 143, 154, 171, 163, 160, 142, 156, 157, 152, 154, 139, 146,
```



```
157, 132, 160, 169, 157, 137, 153, 199, 199, 214, 200, 188, 210,
189, 201, 236, 182, 208, 188, 200, 221, 216, 207, 187, 203, 212,
203, 205, 190, 191, 211, 185, 207, 216, 205, 180, 200, 246, 249,
263, 237, 230, 252, 231, 248, 285, 220, 261, 220, 244, 270, 242,
248, 234, 243, 259, 246, 253, 225, 229, 250, 237, 257, 261, 248,
219, 244, 283, 293, 312, 272, 280, 298, 275, 297, 350, 260, 313,
273, 289, 326, 281, 288, 280, 283, 307, 286, 298, 267, 272, 285,
286, 303, 295, 289, 258, 286, 320, 354, 328, 297, 323, 331, 305,
338, 376, 296, 352, 314, 325, 358, 312, 324, 316, 317, 336, 321,
334, 302, 323, 331, 345, 333, 316, 291, 324
), .Dim = c(30, 5))
x <- c(8.0, 15.0, 22.0, 29.0, 36.0)
xbar <- 22.0

模型参数设置
chains <- 4
iter <- 1000

init <- rep(list(list(
 alpha = rep(250, 30), beta = rep(6, 30),
 alpha_c = 150, beta_c = 10,
 tausq_c = 1, tausq_alpha = 1,
 tausq_beta = 1
))), chains)

mod <- cmdstan_model(stan_file = "code/rats.stan")

rats_fit <- mod$sample(
 data = list(N = N, T = T, y = y, x = x, xbar = xbar),
 init = init,
 iter_warmup = 1000, # 每条链预处理迭代次数
 iter_sampling = 2000, # 每条链总迭代次数
 chains = chains, # 马尔科夫链的数目
 parallel_chains = 1, # 指定 CPU 核心数, 可以给每条链分配一个
 show_messages = FALSE, # 不显示迭代的中间过程
```



```
refresh = 0, # 不显示采样的进度
seed = 20190425 # 设置随机数种子, 不要使用 set.seed() 函数
)
Running MCMC with 4 sequential chains...
##
Chain 1 finished in 0.5 seconds.
Chain 2 finished in 0.6 seconds.
Chain 3 finished in 0.6 seconds.
Chain 4 finished in 0.5 seconds.
##
All 4 chains finished successfully.
Mean chain execution time: 0.6 seconds.
Total execution time: 2.6 seconds.
```

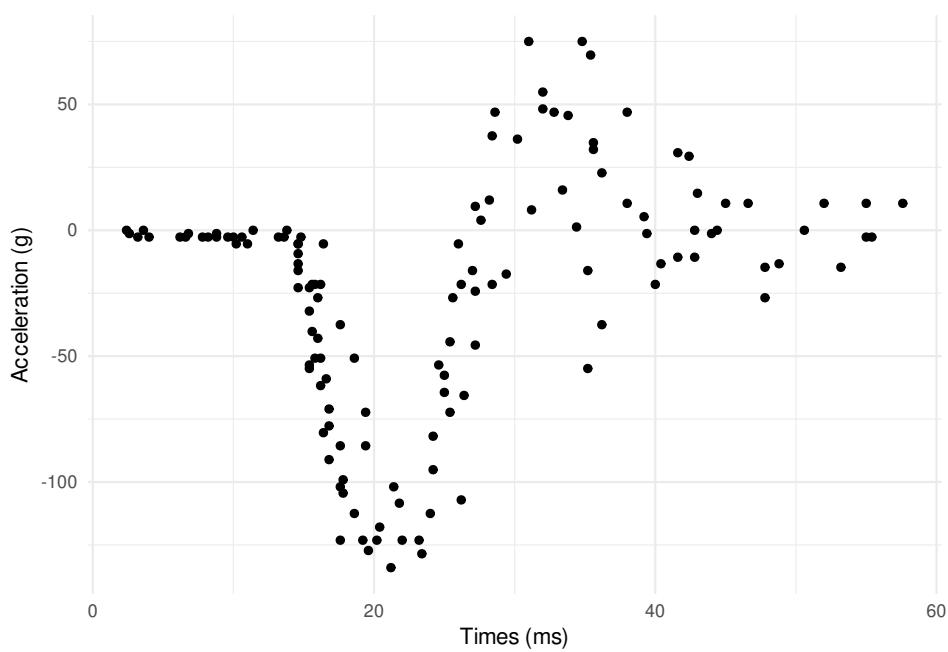
## 28.5 非线性模型

高斯过程

### 28.5.1 mcycle 数据

```
library(MASS)
library(ggplot2)

ggplot(data = mcycle, aes(x = times, y= accel)) +
 geom_point() +
 # geom_smooth() +
 labs(x = "Times (ms)", y = "Acceleration (g)") +
 theme_minimal()
```





## 第二十九章 梯度提升机

关于决策树和梯度提升的扩展包/库，近年来层出不穷。2001 年 Jerome H. Friedman 提出梯度提升机后 [Friedman, 2001]，2003 年 Greg Ridgeway 开发了 gbm 包，目前 Brandon Greenwell 在维护。[gbm](#) 实现了 Freund and Schapire's AdaBoost 算法和 Friedman 的梯度提升机。[h2o](#) 是基于 Java 平台的机器学习平台，学习材料 [h2o-tutorials](#)。基于决策树的分类和回归方法 [caret](#) 和基于模型的提升方法 <https://github.com/boost-R> 倾向统计学习，侧重各类统计模型，仅提供 R 语言接口。[xgboost](#) 目前已然成为做梯度提升的决策树的工业标准，使用案例丰富，中文帮助文档 <https://xgboost.apache.org/cn/latest/>，也提供多种语言接口。类似的还有 [compboost](#)，其它比较小众的提升库还有 [xLearn](#)。[catboost](#) 开源的基于决策树的梯度提升库，支持分类特征，提供 R 和 Python 接口，详见官网 <https://catboost.ai>。[LightGBM](#) 提供了 R 包，微软的工具主要支持 Windows 平台和 VS 编译工具。Python 接口的中文文档 <https://lightgbm.apache.org/>，顺便一提，袁进辉等人开发的[LightLDA](#) 是大规模主题建模的框架。

### 29.1 XGBoost

```
library(xgboost)
```

## 第三十章 神经网络

A big computer, a complex algorithm and a long time does not equal science.

— Robert Gentleman, SSC 2003, Halifax (June 2003)

近年来，深度学习框架越来越多，比较受欢迎的有 tensorflow、pytorch 和 mxnet，RStudio 团队也陆续给它们提供了 R 接口，tensorflow、keras 和 torch。此外，相关主题的还有 fastai。

Norm Matloff 等开发的 polyreg 包以多元多项式回归替代神经网络，Brian Ripley 开发的 nnet 包以单层前馈神经网络用于多项对数线性模型。

### 30.1 mxnet

信息 mxnet 的 R 接口不太稳定好用，安装也比较麻烦，因此，通过 reticulate 包将 Python 模块 mxnet 导入 R 环境，然后调用其函数。

mxnet 框架包含很多子模块，详见[接口文档](#)，比如 ndarray, gluon, symbol 等等，下面具体以多维数组 ndarray 为例展开。

```
导入 mxnet 中的 ndarray
nd <- reticulate::import("mxnet.ndarray", convert = FALSE)
class(nd)
```

```
[1] "python.builtin.module" "python.builtin.object"
```

zeros 是子模块 mxnet.ndarray 下的一个函数

```
x <- nd$zeros(c(3L, 4L)) # 得到 python 中的 mx.nd.array
x
```



```

[[0. 0. 0. 0.]
[0. 0. 0. 0.]
[0. 0. 0. 0.]
<NDArray 3x4 @cpu(0)>
```

将 Python 中的数据对象 `mx.nd.array` 转化为 R 中的矩阵，而数据对象 `mx.nd.array` 有 `asnumpy()` 方法

```
(m1 <- x$asnumpy()) # 得到 R 中的 matrix

[[0. 0. 0. 0.]
[0. 0. 0. 0.]
[0. 0. 0. 0.]

class(m1)

[1] "numpy.ndarray" "python.builtin.object"

m2 = matrix(data = 1:12, nrow = 3, ncol = 4, byrow = TRUE)
class(m2)

[1] "matrix" "array"
```

## 30.2 运行环境

```
sessionInfo()

R version 4.1.0 (2021-05-18)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 20.04.2 LTS

Matrix products: default
BLAS: /usr/lib/x86_64-linux-gnublas/libblas.so.3.9.0
LAPACK: /usr/lib/x86_64-linux-gnulapack/liblapack.so.3.9.0

locale:
[1] LC_CTYPE=en_US.UTF-8 LC_NUMERIC=C
[3] LC_TIME=en_US.UTF-8 LC_COLLATE=en_US.UTF-8
```



```
[5] LC_MONETARY=en_US.UTF-8 LC_MESSAGES=en_US.UTF-8
[7] LC_PAPER=en_US.UTF-8 LC_NAME=C
[9] LC_ADDRESS=C LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
attached base packages:
[1] stats graphics grDevices utils datasets methods base
##
other attached packages:
[1] reticulate_1.20
##
loaded via a namespace (and not attached):
[1] Rcpp_1.0.7 bookdown_0.22 codetools_0.2-18 lattice_0.20-44
[5] png_0.1-7 digest_0.6.27 grid_4.1.0 jsonlite_1.7.2
[9] magrittr_2.0.1 evaluate_0.14 rlang_0.4.11 stringi_1.7.3
[13] Matrix_1.3-4 rmarkdown_2.9 tools_4.1.0 stringr_1.4.0
[17] xfun_0.24 yaml_2.2.1 compiler_4.1.0 htmltools_0.5.1.1
[21] knitr_1.33
```



## 第三十一章 矩阵运算

Eigenvectors from Eigenvalues [Denton et al., 2019]

参考 `matlib` 和 `Matrix` 包, `SparseM` 更加强调稀疏矩阵的 Cholesky 分解和后退法, 矩阵取子集和 Kronecker 积。矩阵计算一般介绍参考在线书籍 Stephen Boyd and Lieven Vandenberghe 最新著作 `Introduction to Applied Linear Algebra – Vectors, Matrices, and Least Squares` [Boyd and Vandenberghe, 2018] 及其 Julia 语言实现, 矩阵分解部分参考 `Introduction to Linear Algebra, 5th Edition`、`Linear Algebra for Data Science with examples in R`

`fastmatrix`、`abind` 各种矩阵操作。

Evan Chen 的书 `An Infinitely Large Napkin` 介绍矩阵代数的内积空间、群、环、域等高级内容, 作者提供免费的电子书。

分块矩阵操作, 各类分解算法, 及其 R 实现

```
library(Matrix)
```

以 `attitude` 数据集为例介绍各种矩阵操作

```
head(attitude)
```

```
rating complaints privileges learning raises critical advance
1 43 51 30 39 61 92 45
2 63 64 51 54 63 73 47
3 71 70 68 69 76 86 48
4 61 63 45 47 54 84 35
5 81 78 56 66 71 83 47
6 43 55 49 44 54 49 34
```

rating 总体评价 complaints 处理员工投诉 privileges 不允许特权 learning 学习机

会 raises 根据表现晋升 critical 批评 advance! 进步

```

fit <- lm(rating ~ ., data = attitude)
summary(fit) # 模型是显著的，很多变量的系数不显著

Call:
lm(formula = rating ~ ., data = attitude)

Residuals:
Min 1Q Median 3Q Max
-10.9418 -4.3555 0.3158 5.5425 11.5990

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.78708 11.58926 0.931 0.361634
complaints 0.61319 0.16098 3.809 0.000903 ***
privileges -0.07305 0.13572 -0.538 0.595594
learning 0.32033 0.16852 1.901 0.069925 .
raises 0.08173 0.22148 0.369 0.715480
critical 0.03838 0.14700 0.261 0.796334
advance -0.21706 0.17821 -1.218 0.235577

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

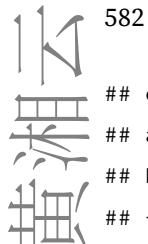
Residual standard error: 7.068 on 23 degrees of freedom
Multiple R-squared: 0.7326, Adjusted R-squared: 0.6628
F-statistic: 10.5 on 6 and 23 DF, p-value: 1.24e-05

anova(fit)

Analysis of Variance Table

Response: rating
Df Sum Sq Mean Sq F value Pr(>F)
complaints 1 2927.58 2927.58 58.6026 9.056e-08 ***
privileges 1 7.52 7.52 0.1505 0.7016
learning 1 137.25 137.25 2.7473 0.1110
raises 1 0.94 0.94 0.0189 0.8920

```



```
critical 1 0.56 0.56 0.0113 0.9163
advance 1 74.11 74.11 1.4835 0.2356
Residuals 23 1149.00 49.96

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(C) attitude_mat <- as.matrix.data.frame(attitude)
生成演示用的矩阵
demo_mat <- t(attitude_mat[, -1]) %*% attitude_mat[, -1]
```

## 31.1 矩阵乘法

```
A <- matrix(c(1, 2, 2, 3), nrow = 2)
A
[,1] [,2]
[1,] 1 2
[2,] 2 3
B <- matrix(c(1, 2, 3, 4, 5, 6), nrow = 2)
B
[,1] [,2] [,3]
[1,] 1 3 5
[2,] 2 4 6
```

通常的矩阵乘法也叫矩阵内积

```
A %*% B
[,1] [,2] [,3]
[1,] 5 11 17
[2,] 8 18 28
A ** 2
[,1] [,2]
[1,] 1 4
[2,] 4 9
```

```
A ^ 2

[,1] [,2]
[1,] 1 4
[2,] 4 9

A ** A

[,1] [,2]
[1,] 1 4
[2,] 4 27
```

## 31.2 Hadamard 积

Hadamard 积 (法国数学家 Jacques Hadamard) 也叫 Schur 积 (德国数学家 Issai Schur ) 或 entrywise 积是两个维数相同的矩阵对应元素相乘，特别地， $A^2$  表示将矩阵  $A$  的每个元素平方

$$(A \circ B)_{ij} = (A)_{ij}(B)_{ij}$$

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \circ \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} & a_{12}b_{12} & a_{13}b_{13} \\ a_{21}b_{21} & a_{22}b_{22} & a_{23}b_{23} \\ a_{31}b_{31} & a_{32}b_{32} & a_{33}b_{33} \end{bmatrix}$$

```
A^2

[,1] [,2]
[1,] 1 4
[2,] 4 9
```

## 31.3 矩阵转置

```
t(B)
```

```
[,1] [,2]
[1,] 1 2
```

◎黃湘云

## 31.4 矩阵外积

```
A %o% B # outer(A, B, FUN = "*")
```

```
, , 1, 1
##
[,1] [,2]
[1,] 1 2
[2,] 2 3
##
, , 2, 1
##
[,1] [,2]
[1,] 2 4
[2,] 4 6
##
, , 1, 2
##
[,1] [,2]
[1,] 3 6
[2,] 6 9
##
, , 2, 2
##
[,1] [,2]
[1,] 4 8
[2,] 8 12
##
, , 1, 3
##
[,1] [,2]
[1,] 5 10
```



```
[2,] 10 15
##
, , 2, 3
##
[,1] [,2]
[1,] 6 12
[2,] 12 18
```

直积/克罗内克积

```
A %x% B # kronecker(A, B, FUN = "*")
```

```
[,1] [,2] [,3] [,4] [,5] [,6]
[1,] 1 3 5 2 6 10
[2,] 2 4 6 4 8 12
[3,] 2 6 10 3 9 15
[4,] 4 8 12 6 12 18
```

## 31.5 矩阵乘方

矩阵 A 首先是一个方阵，对称性和正定性未知，n 个矩阵 A 相乘

统计之都论坛讨论如何求矩阵的乘方 <https://d.cosx.org/d/5619-svd>

```
"%^%" <- function(mat, pow) {
 if (!is.matrix(mat)) mat <- as.matrix(mat)
 stopifnot(!diff(dim(mat)))
 if (pow < 0) {
 pow <- -pow
 mat <- solve(mat)
 }
 pow <- round(pow)
 switch(pow + 1, return(diag(1, nrow(mat))), return(mat))
 get.exponents <- function(pow)
 if (pow == 0) NULL else c(k <- 2^floor(log2(pow)), get.exponents(pow - k))
 ans <- diag(nrow(mat))
 dlog2exp <- rev(-diff(c(log2(get.exponents(pow)), 0)))
 for (j in 1:length(dlog2exp)) {
```



```
 if (dlog2exp[j]) for (i in 1:dlog2exp[j]) mat <- mat %*% mat
 ans <- ans %*% mat
 }
ans
}
```



## 奇异值分解

```
s <- svd(A)
all.equal(s$u %*% diag(s$d) %*% t(s$v), A)
```

```
[1] TRUE
```

特征值及分解  $A = V\Lambda V^{-1}$  求解矩阵 A 的 n 次方

```
eigen(A)
```

```
eigen() decomposition
$values
[1] 4.236068 -0.236068
##
$vectors
[,1] [,2]
[1,] 0.5257311 -0.8506508
[2,] 0.8506508 0.5257311
```

```
eigen(A)$vectors %*% diag(eigen(A)$values) %*% solve(eigen(A)$vectors)
```

```
[,1] [,2]
[1,] 1 2
[2,] 2 3
```

```
eigen(A)$vectors %*% diag(eigen(A)$values)^3 %*% solve(eigen(A)$vectors)
```

```
[,1] [,2]
[1,] 21 34
[2,] 34 55
```

```
A %*% A %*% A
```

```
[,1] [,2]
[1,] 21 34
[2,] 34 55
```



## 31.6 矩阵求幂

```
2^A

[,1] [,2]
[1,] 2 4
[2,] 4 8

exp(A)

[,1] [,2]
[1,] 2.718282 7.389056
[2,] 7.389056 20.085537
```

`expm` 包含更多关于矩阵开方、取对数等计算

## 31.7 矩阵交叉积

交叉积  $A^\top A$

```
crossprod(A, A) # t(x) %*% y

[,1] [,2]
[1,] 5 8
[2,] 8 13

tcrossprod(A, A) # x %*% t(y)

[,1] [,2]
[1,] 5 8
[2,] 8 13
```

## 31.8 矩阵行列式

```
det(A)

[1] -1
```

`expm` 包计算矩阵  $e^A$

## 31.9 矩阵条件数

```
library(Matrix)
base:::rcond(A)

[1] 0.04

kappa(A)

[1] 21.85714

Matrix:::rcond(Matrix:::Hilbert(6))

[1] 3.439939e-08

Matrix:::rcond(A)

[1] 0.04
```

## 31.10 矩阵求逆

```
solve(A)

[,1] [,2]
[1,] -3 2
[2,] 2 -1
```

应用之线性方程组

```
B <- Hilbert(6)
b <- rowSums(B)
not inv
solve(B,b)
```

```
6 x 1 Matrix of class "dgeMatrix"
[,1]
[1,] 1
[2,] 1
[3,] 1
[4,] 1
[5,] 1
```

```
[6,] 1
inv
solve(B) %*% b

6 x 1 Matrix of class "dgeMatrix"
[,1]
[1,] 1
[2,] 1
[3,] 1
[4,] 1
[5,] 1
[6,] 1
```

Moore-Penrose generalized inverse 广义逆，如果 A 可逆则，广义逆就是逆

```
library(MASS) # ginv 来自 MASS 包
ginv(A)
```

```
[,1] [,2]
[1,] -3 2
[2,] 2 -1
```

```
A %*% ginv(A) %*% A
```

```
[,1] [,2]
[1,] 1 2
[2,] 2 3
```

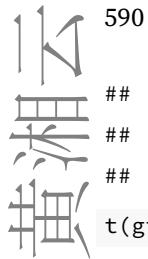
```
ginv(A) %*% A %*% ginv(A)
```

```
[,1] [,2]
[1,] -3 2
[2,] 2 -1
```

```
t(A %*% ginv(A))
```

```
[,1] [,2]
[1,] 1.000000e+00 8.881784e-16
[2,] -8.881784e-16 1.000000e+00
```

```
A %*% ginv(A)
```



```
[,1] [,2]
[1,] 1.000000e+00 -8.881784e-16
[2,] 8.881784e-16 1.000000e+00
t(ginv(A) %*% A)

④ ## [,1] [,2]
[1,] 1.000000e+00 -8.881784e-16
[2,] 8.881784e-16 1.000000e+00
ginv(A) %*% A

[,1] [,2]
[1,] 1.000000e+00 8.881784e-16
[2,] -8.881784e-16 1.000000e+00
```

## 31.11 矩阵伴随

伴随矩阵  $A * A^* = A^* * A = |A| * I, A^* = |A| * A^{-1}$

- $|A^*| = |A|^{n-1}, A \in \mathbb{R}^{n \times n}, n \geq 2$
- $(A^*)^* = |A|^{n-2} A, A \in \mathbb{R}^{n \times n}, n \geq 2$
- $(A^*)^* A$  的  $n$  次伴随是?

```
det(A)*solve(A)
```

```
[,1] [,2]
[1,] 3 -2
[2,] -2 1
```

## 31.12 矩阵范数

向量和矩阵的范数，包括 1, 2, 无穷范数，其他操作看 Matrix 包，尤其关于稀疏矩阵计算部分

1-范数 列和绝对值最大的

$\infty$ -范数 行和绝对值最大的

**Frobenius** - 范数 Euclidean 范数

$M$  - 范数 矩阵里模最大的元素，矩阵里面的元素可能含有复数，所以取模最大



2 - 范数 又称谱范数，矩阵最大的奇异值，如果是方阵，就是最大的特征值

```
norm(A, type = "1") # max(abs(colSums(A)))
```

```
[1] 5
```

```
norm(A, type = "I") # max(abs(rowSums(A)))
```

```
[1] 5
```

```
norm(A, type = "F")
```

```
[1] 4.242641
```

```
norm(A, type = "M") #
```

```
[1] 3
```

```
norm(A, type = "2") # max(svd(A)$d)
```

```
[1] 4.236068
```

显然， $1-, \infty-, M-$  的范数计算比  $F-$  范数快，函数 `norm` 默认情况下求  $1-$  范数

## 31.13 矩阵求秩

```
qr(A)$rank # or qr.default(A)$rank
```

```
[1] 2
```

## 31.14 矩阵求迹

若

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

则矩阵  $A$  的迹  $\text{tr}(A) = \sum_{i=1}^n a_{ii}$

◎ 黄湘云

```
sum(diag(A))
```

```
[1] 4
```

特殊矩阵的构造

## 31.15 单位矩阵

矩阵对角线上全是 1，其余位置都是 0

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

```
diag(rep(3))
```

```
[,1] [,2] [,3]
[1,] 1 0 0
[2,] 0 1 0
[3,] 0 0 1
```

而全 1 矩阵是所有元素都是 1 的矩阵，可以借助外积运算构造，如 3 阶全 1 矩阵

```
rep(1,3) %o% rep(1,3)
```

```
[,1] [,2] [,3]
[1,] 1 1 1
[2,] 1 1 1
[3,] 1 1 1
```

## 31.16 对角矩阵

```
diag(A) # diagonal of a matrix
```

```
[1] 1 3
```

```
diag(diag(A)) # construct a diagonal matrix
```



```
[,1] [,2]
[1,] 1 0
[2,] 0 3
```

## 31.17 上/下三角矩阵

矩阵下三角

row 和 col

```
row(A)
```

```
[,1] [,2]
[1,] 1 1
[2,] 2 2
```

```
col(A)
```

```
[,1] [,2]
[1,] 1 2
[2,] 1 2
```

```
A[row(A)]
```

```
[1] 1 3
```

```
upper.tri(A) # 矩阵上三角
```

```
[,1] [,2]
[1,] FALSE TRUE
[2,] FALSE FALSE
```

```
A[upper.tri(A)]
```

```
[1] 2
```

```
A[lower.tri(A)] <- 0 # 获得上三角矩阵
```

```
A
```

```
[,1] [,2]
[1,] 1 2
[2,] 0 3
```

云  
湘  
黄

- 下三角矩阵

```
A <- matrix(c(1, 2, 2, 3), nrow = 2)
A
```

```
[,1] [,2]
[1,] 1 2
[2,] 2 3
```

```
lower.tri(A)
```

```
[,1] [,2]
[1,] FALSE FALSE
[2,] TRUE FALSE
```

```
A[lower.tri(A)]
```

```
[1] 2
```

```
A[upper.tri(A)] <- 0 # 获得下三角矩阵
```

```
A
```

```
[,1] [,2]
[1,] 1 0
[2,] 2 3
```

```
A <- matrix(c(1, 2, 2, 3), nrow = 2)
```

```
A[row(A) < col(A)] <- 0
```

```
A
```

```
[,1] [,2]
[1,] 1 0
[2,] 2 3
```

## 31.18 稀疏矩阵

```
dn <- list(LETTERS[1:3], letters[1:5])
pointer vectors can be used, and the (i,x) slots are sorted if necessary:
使用指针构造
m <- sparseMatrix(i = c(3,1, 3:2, 2:1), p= c(0:2, 4,4,6), x = 1:6, dimnames = dn)
m
```



```
3 x 5 sparse Matrix of class "dgCMatrix"
a b c d e
A . 2 . . 6
B . . 4 . 5
C 1 . 3 . .

典型构造方式
i <- c(1,3:8); j <- c(2,9,6:10); x <- 7 * (1:7)
(AA <- sparseMatrix(i, j, x = x)) ## 8 x 10 "dgCMatrix"

8 x 10 sparse Matrix of class "dgCMatrix"
##
[1,] . 7
[2,]
[3,] 14 .
[4,] 21
[5,] 28 . . .
[6,] 35 . .
[7,] 42 .
[8,] 49
```

## 31.19 三对角矩阵

## 31.20 LU 分解

## 31.21 Schur 分解

## 31.22 Cholesky 分解

实对称正定矩阵的 Choleski 分解

```
chol(A + diag(rep(1,2)))

[,1] [,2]
[1,] 1.414214 0
[2,] 0.000000 2
```

云  
湘  
黄  
◎

```
Inverse from Choleski (or QR) Decomposition
Matrix::chol2inv(A + diag(rep(1,2)))

[,1] [,2]
[1,] 0.25 0.0000
[2,] 0.00 0.0625
```

Matrix::Cholesky 实现稀疏 Cholesky 分解

## 31.23 特征值分解

特征值分解 (Eigenvalues Decomposition) 也叫谱分解 (Spectral Decomposition)

```
eigen(A)
```

```
eigen() decomposition
$values
[1] 3 1
##
$vectors
[,1] [,2]
[1,] 0 0.7071068
[2,] 1 -0.7071068
```

## 31.24 SVD 分解

Fast truncated singular value decompositions 奇异值分解是特征值分解的推广

```
svd(A)
```

```
$d
[1] 3.6502815 0.8218544
##
$u
[,1] [,2]
[1,] -0.1601822 -0.9870875
[2,] -0.9870875 0.1601822
```

```

$v
[,1] [,2]
[1,] -0.5847103 -0.8112422
[2,] -0.8112422 0.5847103
svd(A)$d
```

```
[1] 3.6502815 0.8218544
```

邱怡轩将奇异值分解用于图像压缩 <https://cosx.org/2014/02/svd-and-image-compression> 并制作了 [Shiny App](#) 交互式演示

## 31.25 QR 分解

```
qr.default(A)

$qr
[,1] [,2]
[1,] -2.2360680 -2.683282
[2,] 0.8944272 1.341641

$rank
[1] 2

$qraux
[1] 1.447214 1.341641

$pivot
[1] 1 2

attr(,"class")
[1] "qr"
qr.X(qr.default(A))

[,1] [,2]
[1,] 1 0
```

```

[2,] 2 3
qr.Q(qr.default(A))

[,1] [,2]
[1,] -0.4472136 -0.8944272
[2,] -0.8944272 0.4472136
qr.R(qr.default(A))

[,1] [,2]
[1,] -2.236068 -2.683282
[2,] 0.000000 1.341641
qr.Q(qr.default(A)) %*% qr.R(qr.default(A))

[,1] [,2]
[1,] 1 -2.220446e-16
[2,] 2 3.000000e+00

```

用 Householder 变换 做 QR 分解 [Bates and Watts, 1988] 及其 R 语言实现 <https://rpubs.com/aaronsc32/qr-decomposition-householder>

Householder 变换是平面反射的一般情况：要计算  $N \times P$  维矩阵  $X$  的 QR 分解，我们采用 Householder 变换

$$\mathbf{H}_u = \mathbf{I} - 2\mathbf{u}\mathbf{u}^\top$$

其中  $I$  是  $N \times N$  维的单位矩阵， $u$  是  $N$  维单位向量，即  $\|\mathbf{u}\| = \sqrt{\mathbf{u}\mathbf{u}^\top} = 1$ 。则  $H_u$  是对称正交的，因为

$$\mathbf{H}_u^\top = \mathbf{I}^\top - 2\mathbf{u}\mathbf{u}^\top = \mathbf{H}_u$$

并且

$$\mathbf{H}_u^\top \mathbf{H}_u = \mathbf{I} - 4\mathbf{u}\mathbf{u}^\top + 4\mathbf{u}\mathbf{u}^\top \mathbf{u}\mathbf{u}^\top = \mathbf{I}$$

让  $\mathbf{H}_u$  乘以向量  $\mathbf{y}$ ，即

$$\mathbf{H}_u \mathbf{y} = \mathbf{y} - 2\mathbf{u}\mathbf{u}^\top \mathbf{y}$$

它是  $y$  关于垂直于过原点的  $u$  的直线的反射，只要

$$\mathbf{u} = \frac{\mathbf{y} - \|\mathbf{y}\|\mathbf{e}_1}{\|\mathbf{y} - \|\mathbf{y}\|\mathbf{e}_1\|} \quad (31.1)$$

或者

$$\mathbf{u} = \frac{\mathbf{y} + \|\mathbf{y}\|\mathbf{e}_1}{\|\mathbf{y} + \|\mathbf{y}\|\mathbf{e}_1\|} \quad (31.2)$$

其中  $\mathbf{e}_1 = (1, 0, \dots, 0)^\top$ , Householder 变换使得向量  $y$  成为  $x$  轴，在新的坐标系统中，向量  $H_u y$  的坐标为  $(\pm\|y\|, 0, \dots, 0)^\top$

举个例子

借助 Householder 变换做 QR 分解的优势：

1. 更快、数值更稳定比直接构造  $Q$ ，特别当  $N$  大于  $P$  的时候
2. 相比于存储矩阵  $Q$  的  $N^2$  个元素，Householder 变换只存储  $P$  个向量  $u_1, \dots, u_P$
3. QR 分解的真实实现，比如在 LINPACK 中，定义  $u$  的时候，公式 (31.1) 或 (31.2) 的选择基于  $y$  的第一个坐标的符号。如果坐标是负的，使用公式(31.1)，如果是正的，使用公式 (31.2)，这个做法可以使得数值计算更加稳定。

Stan 实现的 QR 分解在贝叶斯线性回归模型中的应用<sup>1</sup>

## 31.26 Jordan 分解

## 31.27 Givens 旋转

- Givens 旋转 [https://www.wikiwand.com/en/Givens\\_rotation](https://www.wikiwand.com/en/Givens_rotation)
- 帽子矩阵在统计中的应用回归与方差分析 [Hoaglin and Welsch, 1978]

---

<sup>1</sup>[https://mc-stan.org/users/documentation/case-studies/qr\\_regression.html](https://mc-stan.org/users/documentation/case-studies/qr_regression.html)



## 31.28 特殊函数

### 31.28.1 阶乘

- 阶乘  $n! = 1 \times 2 \times 3 \cdots n$
- 双阶乘  $(2n+1)!! = 1 \times 3 \times 5 \times \cdots \times (2n+1), n = 0, 1, 2, \dots$

```
factorial(5) # 阶乘
```

```
[1] 120
```

```
seq(from = 1, to = 5, length.out = 3)
```

```
[1] 1 3 5
```

```
prod(seq(from = 1, to = 5, length.out = 3)) # 连乘 双阶乘
```

```
[1] 15
```

```
seq(5)
```

```
[1] 1 2 3 4 5
```

```
cumprod(seq(5)) # 累积
```

```
[1] 1 2 6 24 120
```

```
cumsum(seq(5)) # 累和
```

```
[1] 1 3 6 10 15
```

此外还有 `cummax` 和 `cummin`

- 组合数  $C_n^k = \frac{n(n-1)(n-k+1)}{k!}$

$C_5^3 = \frac{5 \times 4 \times 3}{3 \times 2 \times 1}$

```
choose(5,3)
```

```
[1] 10
```

斯特林公式

### 31.28.2 伽马函数

$$\Gamma(x) = \int_0^\infty t^{x-1} \exp(-t) dt \quad \Gamma(n) = (n-1)!, n \in \mathbb{Z}^+$$

```
gamma(2)
[1] 1
gamma(10)
[1] 362880
gamma2 <- function(t,x){
 t^(x-1)*exp(-t)
}
integrate(gamma2, lower = 0, upper = + Inf, x = 10)

362880 with absolute error < 0.025
```

- `psigamma(x, deriv)` 表示  $\psi(x)$  的 `deriv` 阶导数

$\text{digamma}(x) \triangleq \psi(x) = \frac{d}{dx} \ln \Gamma(x) = \Gamma'(x)/\Gamma(x)$

```
例1
x <- 2
eval(deriv(~ gamma(x), "x"))/gamma(x)
```

```
[1] 1
attr(),"gradient")
x
[1,] 0.4227843
```

# 与此等价

```
psigamma(2, 0)
```

```
[1] 0.4227843
```

`digamma(x)` #  $\psi(x)$  的一阶导数

```
[1] 0.4227843
```

`trigamma(x)` #  $\psi(x)$  的二阶导数

```
[1] 0.6449341
```

# 例2

```
eval(deriv(~ psigamma(x, 1), "x"))
```

```
[1] 0.6449341
```

```

attr(,"gradient")
x
[1,] -0.4041138
与此等价
psigamma(2, 2)
[1] -0.4041138

注意与下面这个例子比较
dx2x <- deriv(~ x^3, "x")
eval(dx2x)

[1] 8
attr(,"gradient")
x
[1,] 12

```

### 31.28.3 贝塔函数

$$B(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a + b) = \int_0^1 t^{a-1}(1-t)^{b-1} dt$$

```

beta(1, 1)

[1] 1

beta(2, 3)

[1] 0.08333333

beta2 <- function(t, a, b) {
 t^(a - 1) * (1 - t)^(b - 1)
}

integrate(beta2, lower = 0, upper = 1, a = 2, b = 3)

0.08333333 with absolute error < 9.3e-16

```

### 31.28.4 贝塞尔函数

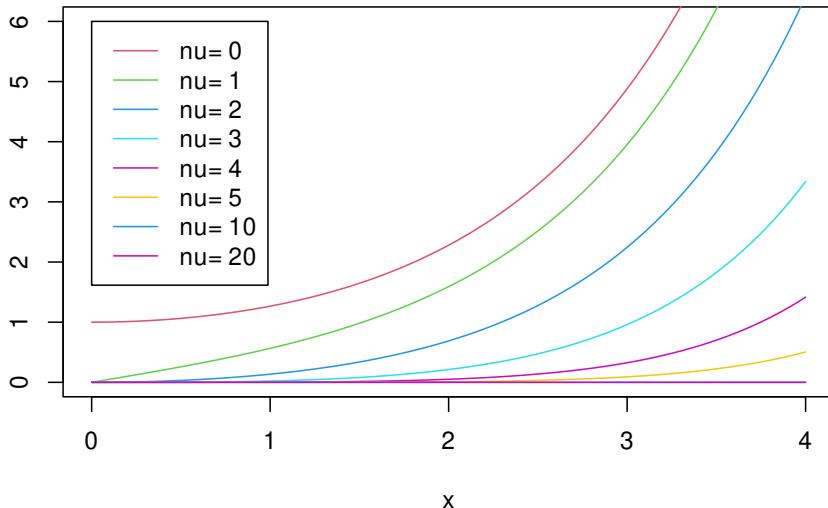
```
besselI(x, nu, expon.scaled = FALSE) # 修正的第一类
besselK(x, nu, expon.scaled = FALSE) # 修正的第二类
besselJ(x, nu) # 第一类
besselY(x, nu) # 第二类
```



- $\nu$  贝塞尔函数的阶，可以是分数
- `expon.scaled` 是否使用指数表示

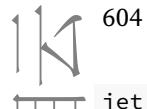
```
nus <- c(0:5, 10, 20)
x <- seq(0, 4, length.out = 501)
plot(x, x,
 ylim = c(0, 6), ylab = "", type = "n",
 main = "Bessel Functions I_nu(x)"
)
for (nu in nus) lines(x, besselI(x, nu = nu), col = nu + 2)
legend(0, 6, legend = paste("nu=", nus), col = nus + 2, lwd = 1)
```

Bessel Functions I\_nu(x)



介绍复数矩阵的计算，矩阵的指数计算，介绍一点分形

```
考虑用 ganimate 实现，去掉 caTools 依赖
library(caTools)
```



```
jet.colors <- colorRampPalette(c(
 "green", "blue", "red", "cyan", "#7FFF7F",
 "yellow", "#FF7F00", "red", "#7F0000"
))
m <- 1000 # define size
C <- complex(
 real = rep(seq(-1.8, 0.6, length.out = m), each = m),
 imag = rep(seq(-1.2, 1.2, length.out = m), m)
)
C <- matrix(C, m, m) # reshape as square matrix of complex numbers
Z <- 0 # initialize Z to zero
X <- array(0, c(m, m, 20)) # initialize output 3D array
for (k in 1:20) { # loop with 20 iterations
 Z <- Z^2 + C # the central difference equation
 X[, , k] <- exp(-abs(Z)) # capture results
}
write.gif(X, "Mandelbrot.gif", col = jet.colors, delay = 100)
```

## 第三十二章 符号计算

相比于数值计算，符号计算可以无限精度，包括微分、积分运法，求解线性、非线性方程（组），常微分、偏微分方程（组）等，R自带几个函数如`deriv()`、`D()`等可以做一些简单的微分运算，扩展包 `Ryacas` 提供 `Yacas` 核心计算引擎，`symengine` 引入 C++ 符号计算库 `SymEngine`，相比于 `Ryacas`，`symengine` 不会和 Base R 函数冲突。Python 的符号计算模块 `sympy` [Meurer et al., 2017] 不仅支持简单的四则运算，还支持微分、积分、解方程等，详见官方文档 <https://sympy.org/>。

16 年在统计之都灌水[符号计算与 R 语言](#)，相应的 Rmd 源文件放在[Github](#)上。

```
多元函数求偏导
ft <- deriv(expression(sin(x1) + sin(x2) + cos(3 * x1 * x2) + x1^2 + x2^2),
 namevec = c("x1", "x2"), function.arg = TRUE
)
隐函数求偏导
deriv(y ~ sin(cos(x) * y), namevec = c("x", "y"), function.arg = TRUE)

function (x, y)
{
.expr1 <- cos(x)
.expr2 <- .expr1 * y
.expr4 <- cos(.expr2)
.value <- sin(.expr2)
.grad <- array(0, c(length(.value), 2L), list(NULL, c("x",
"y")))
.grad[, "x"] <- -(.expr4 * (sin(x) * y))
.grad[, "y"] <- .expr4 * .expr1
attr(.value, "gradient") <- .grad
.value
```



```
}
```

下面以标准正态分布的密度函数为例，

```
NormDensity <- expression(1 / sqrt(2 * pi) * exp(-x^2 / 2))
递归的方法求高阶倒数
DD <- function(expr, name, order = 1) {
 if (order < 1) {
 stop("'order' must be >= 1")
 }
 if (order == 1) {
 D(expr, name)
 } else {
 DD(D(expr, name), name, order - 1)
 }
}
计算三阶导数
DD(NormDensity, "x", 3)
```

```
1/sqrt(2 * pi) * (exp(-x^2/2) * (2 * x/2) * (2/2) + ((exp(-x^2/2) *
(2/2) - exp(-x^2/2) * (2 * x/2) * (2 * x/2)) * (2 * x/2) +
exp(-x^2/2) * (2 * x/2) * (2/2)))
```

Deriv 可以将 R 表达式简化

```
library(Deriv)
Simplify(DD(NormDensity, "x", 3))
```

```
x * (3 - x^2) * exp(-(x^2/2))/sqrt(2 * pi)
```

即  $x(3 - x^2)e^{-x^2/2}/\sqrt{2\pi}$ ，eval() 将表达式转为函数，代入数值运算。

$$\tau(x) = \frac{(-1)^{j-1}}{\sqrt{j!}} \phi^{(j)}(x)$$

```
Tetrachoric <- function(x, j) {
 (-1)^(j - 1) / sqrt(factorial(j)) * eval(DD(NormDensity, "x", j))
}
Tetrachoric(2, 3)
```

```
[1] -0.04408344
```

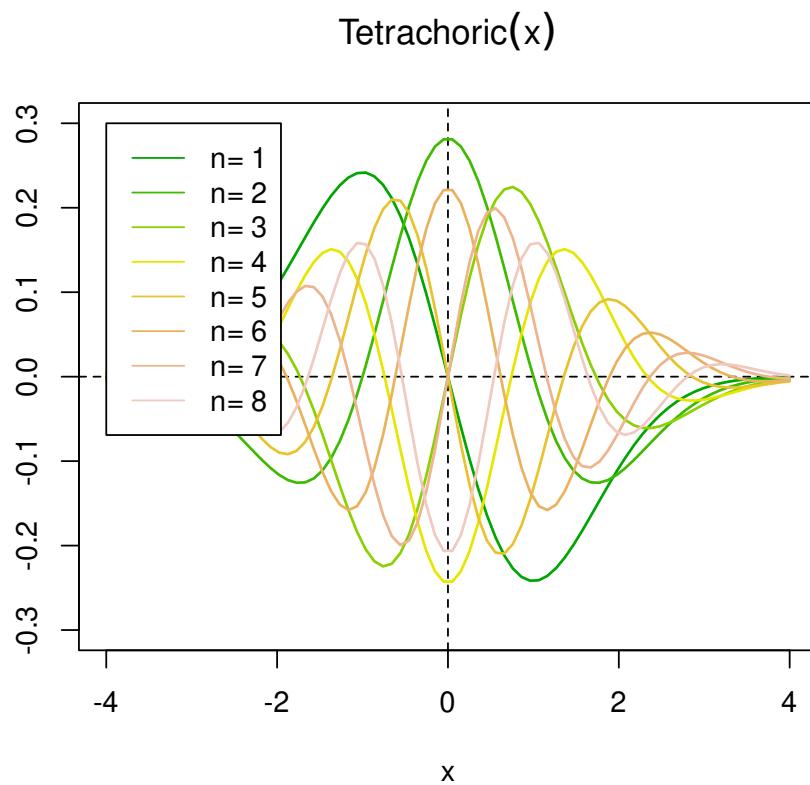


图 32.1: Tetrachoric 函数



表达式转函数

```
ExpToFun <- function(x) eval(Simplify(DD(NormDensity, "x", 4)))
ExpToFun(2)
[1] -0.2699548
```

函数求积分

```
integrate(ExpToFun, 0, pi)

-0.06192048 with absolute error < 5.8e-12
```

对函数求微分

```
fun <- function(x) x * pi / sqrt(x)
D(body(fun), 'x')
Simplify(D(body(fun), "x"))

0.5 * pi/sqrt(x)

Dfun <- function(x) {
}
body(Dfun) <- Simplify(D(body(fun), "x"))
Dfun
```

```
function (x)
0.5 * pi/sqrt(x)

Dfun(4)
```

```
[1] 0.7853982
```

下面简单介绍 symengine 的符号计算能力

```
library(symengine)
声明几个符号变量
use_vars(x, y, z)
表达式展开
expr <- (x + y + z) ^ 2L - 42L
expand(expr)

(Add) -42 + 2*x*y + 2*x*z + 2*y*z + x^2 + y^2 + z^2
```

变量替换



```
a <- S("a")
z 用 a 替换
expr <- subs(expr, z, a)
y 用 x^2 替换
expr <- subs(expr, y, x^2L)
expr
```

```
(Add) -42 + (a + x + x^2)^2
```

表达式求 2 阶偏导

```
d1_expr <- DD(expr, "x", 2)
expand(d1_expr)
```

```
(Add) 2 + 4*a + 12*x + 12*x^2
```

求解带参数  $a$  的一元二次方程

```
solutions <- solve(d1_expr, "x")
solutions
```

```
VecBasic of length 2
```

```
V(-1/2 + (-1/2)*sqrt(1 + (-1/3)*(2 + 4*a)), -1/2 + (1/2)*sqrt(1 + (-1/3)*(2 + 4*a)))
```



## 第三十三章 数值优化

R 语言提供了相当多的优化求解器，比较完整的概览见[优化视图](#)。本章介绍一些常用的优化算法及其 R 实现，涵盖线性规划、整数规划、二次规划、非线性规划等。

商业优化求解器的能力都覆盖非线性规划 (NLP)，线性 (LP)、二次 (QP) 和锥规划 (SOCP)，混合整数线性规划 (MILP)，多目标优化，最小二乘和方程求解。此外，还有很多文档介绍，[LINGO 提供用户手册](#)，[Matlab 优化工具箱 提供 Optimization 工具箱使用指南](#)，[MOSEK \(<https://www.mosek.com/>\) 提供 MOSEK 建模食谱](#)，[LocalSolver 提供基本使用手册](#)，[Gurobi 提供 Gurobi 参考手册](#)，[CPLEX Optimization Studio](#)。

开源社区有不少工具，也能求解常见的优化问题，如 Julia 的 [JuMP \(<https://jump.dev/>\)](#)，Octave (<https://www.gnu.org/software/octave/>) 内置的优化函数，Python 模块 [SciPy 提供 Optimization 优化求解器](#)，[cvxopt 凸优化求解器](#)，主要基于内点法，提供 Julia、Python、Matlab 接口，算法介绍见[锥优化 机器学习优化](#)。课程见[Optimization for Machine Learning](#)，书籍见[Convex Optimization](#)，相关综述见[Convex Optimization: Algorithms and Complexity](#)。

Berwin A. Turlach 开发的 [quadprog](#) 主要用于求解二次规划问题。Anqi Fu 开发的 [CVXR](#) 可解很多凸优化问题 [Fu et al., 2020]，详见网站 <https://cvxr.rbind.io/>，Jelmer Ypma 开发的 [nloptr](#) 可解无约束和有约束的非线性规划问题 [Ypma, 2020]，GPareto 求解多目标优化问题，帕雷托前沿优化和估计 [Binois and Picheny, 2019]。[igraph](#) 可以用来解决最短路径、最大网络流、最小生成树等图优化相关的问题。[https://palomar.home.ece.ust.hk/MAFS6010R\\_lectures/Rsession\\_solvers.html](https://palomar.home.ece.ust.hk/MAFS6010R_lectures/Rsession_solvers.html) 提供了一般的求解器介绍。ROI 包力图统一各个求解器的调用接口，打造一个优化算法的基础设施平台。Theußl et al. [2020] 详细介绍了目前优化算法发展情况及 R 社区提供的优化能力。GA 包实现了遗传算法，支持连续和离散的空间搜索，可以并行 [Scrucca, 2013, 2017]，是求解 TSP 问题的重要方法。NMOF 包实现了差分进化、遗传算法、粒子群算法、模拟退火算法等启发



式优化算法，还提供网格搜索和贪婪搜索工具，[Gilli et al. \[2019\]](#) 提供了详细的介绍。[Nash \[2014\]](#) 总结了 R 语言环境下最优化问题的最佳实践。[RcppEnsmalle](#) 数值优化通用标准的优化方法，前沿最新的优化方法，包含小批量/全批量梯度下降技术、无梯度优化器，约束优化技术。[RcppNumerical](#) 无约束数值优化，一维/多维数值积分。

谷歌开源的运筹优化工具 [or-tools](#) 提供了约束优化、线性优化、混合整数优化、装箱和背包算法、TSP (Traveling Salesman Problem)、VRP (Vehicle Routing Problem)、图算法 (最短路径、最小成本流、最大流等) 等算法和求解器。「运筹 OR 帷幄」社区开源的 [线性规划](#) 一书值得一看。

```
加载 ROI 时不要自动加载插件
Sys.setenv(ROI_LOAD_PLUGINS = FALSE)

library(lpSolve) # 线性规划求解器
library(ROI) # 优化工具箱
library(ROI.plugin.alabama) # 注册 alabama 求解器
library(ROI.plugin.nloptr) # 注册 nloptr 求解器
library(ROI.plugin.lpsolve) # 注册 lpsolve 求解器
library(ROI.plugin.quadprog) # 注册 quadprog 求解器
library(lattice) # 图形绘制
library(kernlab) # 优化问题和机器学习的关系

library(rootSolve) # 非线性方程
library(BB) # 非线性方程组
library(deSolve) # ODE 常微分方程
library(scatterplot3d) # 三维曲线图

library(shape)
library(ReacTran) # PDE 偏微分方程
library(PBSddesolve) # DAE 延迟微分方程

library(nlme) # 混合效应模型
library(nlmeODE) # ODE 应用于混合效应模型

library(Sim.DiffProc) # SDE 随机微分方程

library(nlmixr) # Population ODE modeling
```



表 33.1: ROI 插件按优化问题分类

	Linear	Quadratic	Conic	Functional
No				
Box				optimx
Linear	clp*, cbc <sup>++</sup> , glpk <sup>++</sup> , lpsolve <sup>++</sup> , msbinlp <sup>++</sup> , symphony <sup>++</sup>	ipop, quadprog*, qpoases		
Quadratic		cplex <sup>+</sup> , gurobi <sup>++</sup> , mosek <sup>++</sup> , neos <sup>+</sup>		
Conic			ecos <sup>++</sup> , scs*	
Functional				alabama, deoptim, nlminb, nloptr

\* 求解器受限于凸优化问题

+ 求解器可以处理整型约束

表 33.1 对目前的优化器按优化问题做了分类

### 33.1 线性规划

`clpAPI` 线性规划求解器。`glpk` 的两个 R 接口 - `glpkAPI` 和 `Rglpk` 提供线性规划和混合整数规划的求解能力。`lp_solve` 的两个 R 接口 - `lpSolveAPI` 和 `lpSolve` 也提供类似的能力。`ompr` 求解混合整数线性规划问题。

举个例子，如下



$$\begin{aligned} & \min_x \quad -6x_1 - 5x_2 \\ s.t. \quad & \left\{ \begin{array}{l} x_1 + 4x_2 \leq 16 \\ 6x_1 + 4x_2 \leq 28 \\ 2x_1 - 5x_2 \leq 6 \end{array} \right. \end{aligned}$$

写成矩阵形式

$$\begin{aligned} & \min_x \quad \begin{bmatrix} -6 \\ -5 \end{bmatrix}^T x \\ s.t. \quad & \begin{bmatrix} 1 & 4 \\ 6 & 4 \\ 2 & -5 \end{bmatrix} x \leq \begin{bmatrix} 16 \\ 28 \\ 6 \end{bmatrix} \end{aligned}$$

对应成 R 代码如下

```
lpSolve 添加约束条件
library(lpSolve)
目标
f.obj <- c(-6, -5)
约束
f.con <- matrix(c(1, 4, 6, 4, 2, -5), nrow = 3, byrow = TRUE)
方向
f.dir <- c("<=", "<=", "<=")
右手边
f.rhs <- c(16, 28, 6)
res <- lp("min", f.obj, f.con, f.dir, f.rhs)
res$objval
```

```
[1] -31.4
```

```
res$solution
```

```
[1] 2.4 3.4
```

## 33.2 整数规划

### 33.2.1 一般整数规划

(C)

$$\begin{aligned} & \max_x \quad 0.2x_1 + 0.6x_2 \\ s.t. \quad & \begin{cases} 5x_1 + 3x_2 \leq 250 \\ -3x_1 + 2x_2 \leq 4 \\ x_1, x_2 \geq 0, \quad x_1, x_2 \in \mathbb{Z} \end{cases} \end{aligned}$$

```
目标
f.obj <- c(0.2, 0.6)
约束
f.con <- matrix(c(5, 3, -3, 2), nrow = 2, byrow = TRUE)
方向
f.dir <- c("<=", "<=")
右手边
f.rhs <- c(250, 4)
限制两个变量都是整数
res <- lp("max", f.obj, f.con, f.dir, f.rhs, int.vec=1:2)
res$objval
```

```
[1] 29.2
```

```
res$solution
```

```
[1] 26 40
```

### 33.2.2 0-1 整数规划

$$\begin{aligned} & \max_x \quad 0.2x_1 + 0.6x_2 \\ s.t. \quad & \begin{cases} 5x_1 + 3x_2 \leq 250 \\ -3x_1 + 2x_2 \leq 4 \\ x_1, x_2 \in \{0, 1\} \end{cases} \end{aligned}$$

```
目标
f.obj <- c(0.2, 0.6)
约束
```



```
f.con <- matrix(c(5, 3, -3, 2), nrow = 2, byrow = TRUE)
方向
f.dir <- c("<=", "<=")
右手边
f.rhs <- c(250, 4)
限制两个变量都是0-1整数
res <- lp("max", f.obj, f.con, f.dir, f.rhs, int.vec=1:2, all.bin = TRUE)
res$objval

[1] 0.8

res$solution

[1] 1 1
```

### 33.2.3 混合整数规划

一部分变量要求是整数

$$\begin{aligned} \max_x \quad & 3x_1 + 7x_2 - 12x_3 \\ s.t. \quad & \begin{cases} 5x_1 + 7x_2 + 2x_3 \leq 61 \\ 3x_1 + 2x_2 - 9x_3 \leq 35 \\ x_1 + 3x_2 + x_3 \leq 31 \\ x_1, x_2 \geq 0, \quad x_2, x_3 \in \mathbb{Z}, \quad x_3 \in [-10, 10] \end{cases} \end{aligned}$$

矩阵形式如下

$$\begin{aligned} \min_x \quad & \begin{bmatrix} 3 \\ 7 \\ -12 \end{bmatrix}^T x \\ s.t. \quad & \begin{cases} \begin{bmatrix} 5 & 7 & 2 \end{bmatrix} x \leq \begin{bmatrix} 61 \\ 35 \\ 31 \end{bmatrix} \end{cases} \end{aligned}$$

```
op <- OP(
 objective = L_objective(c(3, 7, -12)),
 # 指定变量类型: 第1个变量是连续值, 第2、3个变量是整数
 types = c("C", "I", "I"),
```



```
constraints = L_constraint(
 L = matrix(c(
 5, 7, 2,
 3, 2, -9,
 1, 3, 1
), ncol = 3, byrow = TRUE),
 dir = c("<=", "<=", "<="),
 rhs = c(61, 35, 31)
),
添加约束: 第3个变量的下、上界分别是 -10 和 10
bounds = V_bound(li = 3, ui = 3, lb = -10, ub = 10, nobj = 3),
maximum = TRUE
)
op
```

```
ROI Optimization Problem:
##
Maximize a linear objective function of length 3 with
- 1 continuous objective variable,
- 2 integer objective variables,
##
subject to
- 3 constraints of type linear.
- 1 lower and 1 upper non-standard variable bound.
res <- ROI_solve(op, solver = "lpsolve")
res$solution
```

```
[1] 0.3333333 8.0000000 -2.0000000
```

```
res$objval
```

```
[1] 81
```



## 33.3 二次规划

### 33.3.1 凸二次规划

在 R 中使用 **quadprog** [Turlach, 2019] 包求解二次规划<sup>1</sup>, **quadprogXT** 包用来求解带绝对值约束的二次规划, **pracma** [Borchers, 2021] 包提供 `quadprog()` 函数就是对 **quadprog** 包的 `solve.QP()` 进行封装, 调用风格更像 Matlab。**quadprog** 包实现了 Goldfarb and Idnani (1982, 1983) 提出的对偶方法, 主要用来求解带线性约束的严格凸二次规划问题。**quadprog** 求解的二次型的形式如下:

$$\min_b -d^\top b + \frac{1}{2} b^\top D b, \quad A^\top b \geq b_0$$

```
solve.QP(Dmat, dvec, Amat, bvec, meq = 0, factorized = FALSE)
```

参数 `Dmat`、`dvec`、`Amat`、`bvec` 分别对应二次规划问题中的  $D, d, A, b_0$ 。下面举个二次规划的具体例子

$$D = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}, \quad d = (-3, 2), \quad A = \begin{bmatrix} 1 & 1 \\ -1 & 1 \\ 0 & -1 \end{bmatrix}, \quad b_0 = (2, -2, -3)$$

即目标函数

$$Q(x, y) = x^2 + y^2 - xy + 3x - 2y + 4$$

它的可行域如图33.1所示

```
plot(0, 0,
 xlim = c(-2, 5.5), ylim = c(-1, 3.5), type = "n",
 xlab = "x", ylab = "y", main = "Feasible Region")
)
polygon(c(2, 5, -1), c(0, 3, 3), border = TRUE, lwd = 2, col = "gray")
```

调用 **quadprog** 包的 `solve.QP()` 函数求解此二次规划问题

```
library(quadprog)
Dmat <- matrix(c(2, -1, -1, 2), nrow = 2, byrow = TRUE)
dvec <- c(-3, 2)
A <- matrix(c(1, 1, -1, 1, 0, -1), ncol = 2, byrow = TRUE)
```

<sup>1</sup><https://rwalk.xyz/solving-quadratic-programs-with-rs-quadprog-package/>

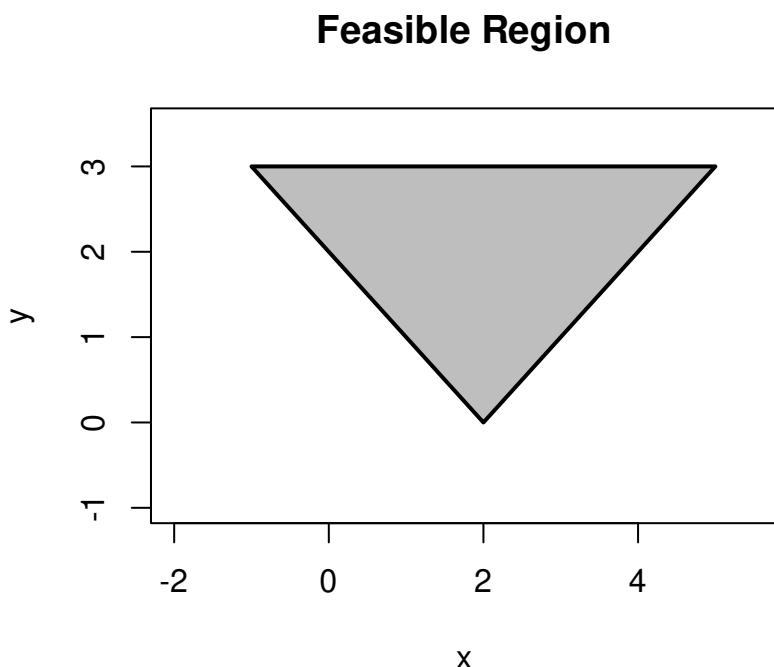


图 33.1: 可行域

```
bvec <- c(2, -2, -3)
Amat <- t(A)
sol <- solve.QP(Dmat = Dmat, dvec = dvec, Amat = Amat, bvec = bvec)
sol

$solution
[1] 0.1666667 1.8333333
##
$value
[1] -0.08333333
##
$unconstrained.solution
[1] -1.3333333 0.3333333
##
$iterations
[1] 2 0
##
$Lagrangian
[1] 1.5 0.0 0.0
##
$iact
[1] 1
```

ROI 默认的二次规划的标准形式为  $\frac{1}{2}x^\top Qx + a^\top x$ , 在传递参数值的时候注意和上面的区别。

```
library(ROI)
op <- OP(
 objective = Q_objective(Q = Dmat, L = -dvec),
 constraints = L_constraint(A, rep(">=", 3), bvec),
 maximum = FALSE # 默认求最小
)
nlp <- ROI_solve(op, solver = "nloptr.slsqp", start = c(1, 2))
nlp$objval

[1] -0.08333333
nlp$solution
```



```
[1] 0.1666667 1.8333333
```

对变量  $x$  添加整型约束，原二次规划即变成混合整数二次规划（Mixed Integer Quadratic Programming，简称 MIQP）

```
目前开源的求解器都不能处理 MIQP 问题
op <- OP(
 objective = Q_objective(Q = Dmat, L = -dvec),
 constraints = L_constraint(A, rep(">=", 3), bvec),
 types = c("I", "C"),
 maximum = FALSE # 默认求最小
)
nlp <- ROI_solve(op, solver = "nloptr.slsqp", start = c(1, 2))
nlp$objval
nlp$solution
```

在可行域上画出等高线，标记目标解的位置，图中红点表示无约束下的解，黄点表示线性约束下的解

```
qp_sol <- sol$solution # 二次规划的解
uc_sol <- sol$unconstrained.solution # 无约束情况下的解
画图
library(lattice)
x <- seq(-2, 5.5, length.out = 500)
y <- seq(-1, 3.5, length.out = 500)
grid <- expand.grid(x = x, y = y)
二次规划的目标函数
grid$z <- with(grid, x^2 + y^2 - x * y + 3 * x - 2 * y + 4)
levelplot(z ~ x * y, grid,
 cuts = 40,
 panel = function(...) {
 panel.levelplot(...)
 panel.polygon(c(2, 5, -1), c(0, 3, 3),
 border = TRUE,
 lwd = 2, col = "transparent"
)
 panel.points(
 c(uc_sol[1], qp_sol[1]),
 col = c("red", "yellow")
)
 }
)
```

```
c(uc_sol[2], qp_sol[2]),
lwd = 5, col = c("red", "yellow"), pch = 19
)
,
colorkey = TRUE,
col.regions = terrain.colors(40)
)
```

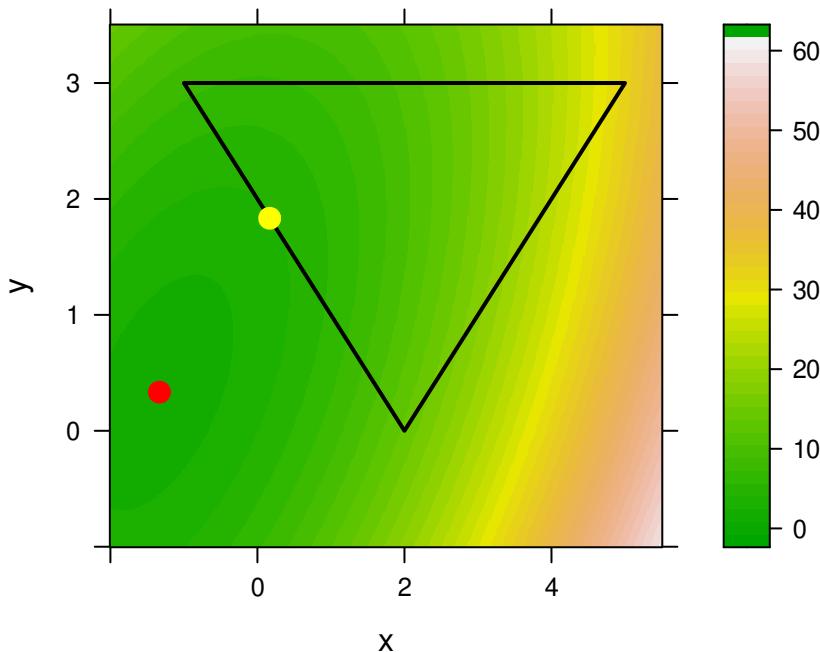
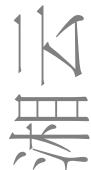


图 33.2: 无约束和有约束条件下的解

### 33.3.2 半正定二次优化

kernlab 提供基于核的机器学习方法，可用于分类、回归、聚类、异常检测、分位回归、降维等场景，包含支撑向量机、谱聚类、核 PCA、高斯过程和二次规划求解器，将优化方法用于机器学习，展示二者的关系。

R 包 kernlab 的函数 ipop() 实现内点法可以求解半正定的二次规划问题，对应



到上面的例子，就是要求  $A \geq 0$ ，而 R 包 quadprog 只能求解正定的二次规划问题，即要求  $A > 0$ 。



以二分类问题为例，采用 SMO (Sequential Minimization Optimization) 求解器，将 SVM 的二次优化问题分解。



```
library(kernlab)
set.seed(123)

x <- rbind(matrix(rnorm(120), 60, 2), matrix(rnorm(120, mean = 3), 60, 2))
y <- matrix(c(rep(1, 60), rep(-1, 60)))
svp <- ksvm(x, y, type = "C-svc")
plot(svp, data = x)
```

SVM classification plot

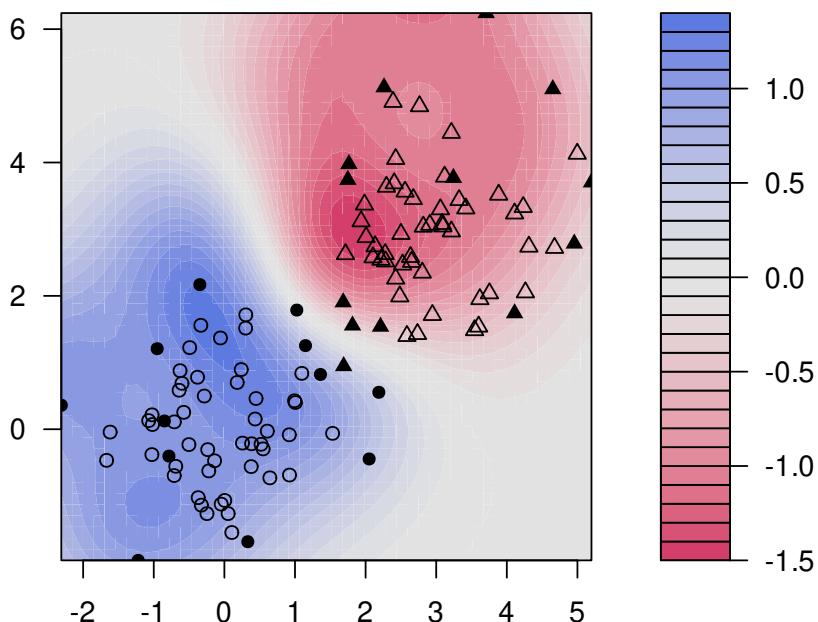


图 33.3: 二分类问题



## 33.4 非线性规划

开源的非线性优化求解器，推荐使用 `nloptr`，它支持全局优化，同时推荐 `ROI`，它有统一的接口函数。

### 33.4.1 一元非线性优化

下面考虑一个稍微复杂的一元函数优化问题，求复合函数的极值

$$g(x) = \int_0^x -\sqrt{t} \exp(-t^2) dt, \quad f(y) = \int_0^y g(s) \exp(-s) ds$$

```
g <- function(x) {
 integrate(function(t) {
 -sqrt(t) * exp(-t^2)
 }, lower = 0, upper = x$value
}

f <- function(y) {
 integrate(function(s) {
 Vectorize(g, "x")(s) * exp(-s)
 }, lower = 0, upper = y$value
}

optimize(f, interval = c(10, 100), maximum = FALSE)

$minimum
[1] 66.84459
##
$objective
[1] -0.3201572
```



## 提示

计算积分的时候，输入了一系列 `s` 值，参数是向量，而函数 `g` 只支持输入参数是单个值，`g(c(1,2))` 会报错，因此上面对函数 `g()` 用了向量化函数 `Vectorize()` 操作。

```
g(1)
[1] -0.453392
```

类似地，同时计算多个目标函数 `f(y)` 的值，也需要 `Vectorize()` 实现向量化操作。

```
Vectorize(f, "y")(c(1, 2))
[1] -0.1103310 -0.2373865
```

### 33.4.2 多元非线性无约束优化

下面这些用来测试优化算法的函数来自[维基百科](#)

#### 33.4.2.1 Himmelblau 函数

Himmelblau 函数是一个多模函数，常用于比较优化算法的优劣。

$$f(x_1, x_2) = (x_1^2 + x_2 - 11)^2 + (x_1 + x_2^2 - 7)^2$$

它在四个位置取得一样的极小值，分别是  $f(-3.7793, -3.2832) = 0$ ,  $f(-2.8051, 3.1313) = 0$ ,  $f(3, 2) = 0$ ,  $f(3.5844, -1.8481) = 0$ 。函数图像见图 33.4。

```
目标函数
fn <- function(x) {
 (x[1]^2 + x[2] - 11)^2 + (x[1] + x[2]^2 - 7)^2
}

df <- expand.grid(
 x = seq(-5, 5, length = 101),
 y = seq(-5, 5, length = 101)
)

df$fnxy = apply(df, 1, fn)
```

黄湘云

```

library(lattice)
减少三维图形的边空
lattice.options(
 layout.widths = list(
 left.padding = list(x = -.6, units = "inches"),
 right.padding = list(x = -1.0, units = "inches")
),
 layout.heights = list(
 bottom.padding = list(x = -.8, units = "inches"),
 top.padding = list(x = -1.0, units = "inches")
)
)

wireframe(
 data = df, fnxy ~ x * y,
 shade = TRUE, drape = FALSE,
 xlab = expression(x[1]),
 ylab = expression(x[2]),
 zlab = list(expression(italic(f) ~ group("(", list(x[1], x[2]), ")"))), rot = 90,
 scales = list(arrows = FALSE, col = "black"),
 par.settings = list(axis.line = list(col = "transparent")),
 screen = list(z = -240, x = -70, y = 0)
)

梯度函数
gr <- function(x) {
 numDeriv::grad(fn, c(x[1], x[2]))
}

optim(par = c(-1.2, 1), fn = fn, gr = gr, method = "BFGS")

$par
[1] -2.805118 3.131313
##
$value
[1] 2.069971e-27
##

```

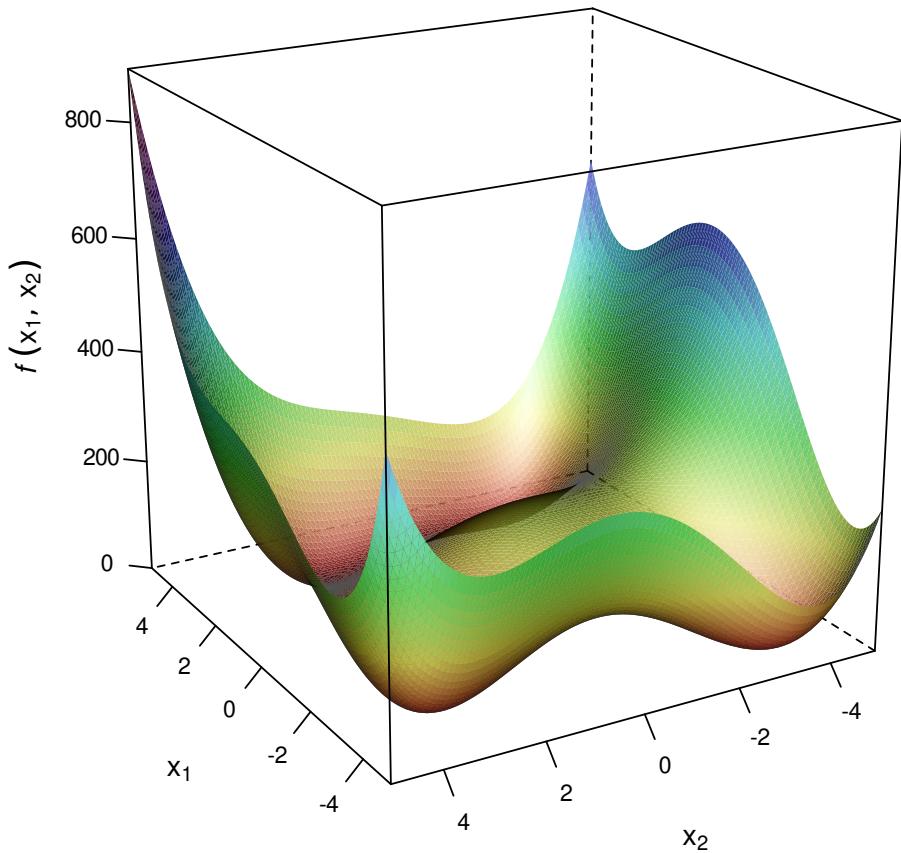


图 33.4: Himmelblau 函数图像



```
$counts
function gradient
42 15
##
$convergence
[1] 0
##
$message
NULL
```

### 33.4.2.2 Rosenbrock 函数

香蕉函数 定义如下：

$$f(x_1, x_2) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2$$

```
fn <- function(x) {
 (100 * (x[2] - x[1]^2)^2 + (1 - x[1])^2)
}

df <- expand.grid(
 x = seq(-2.5, 2.5, length = 101),
 y = seq(-2.5, 2.5, length = 101)
)
df$fnxy = apply(df, 1, fn)

wireframe(
 data = df, fnxy ~ x * y,
 shade = TRUE, drape = FALSE,
 xlab = expression(x[1]),
 ylab = expression(x[2]),
 zlab = list(expression(italic(f) ~ group("(", list(x[1], x[2]), ")"))),
 rot = 90,
 scales = list(arrows = FALSE, col = "black"),
 par.settings = list(axis.line = list(col = "transparent")),
 screen = list(z = 120, x = -70, y = 0)
)
```

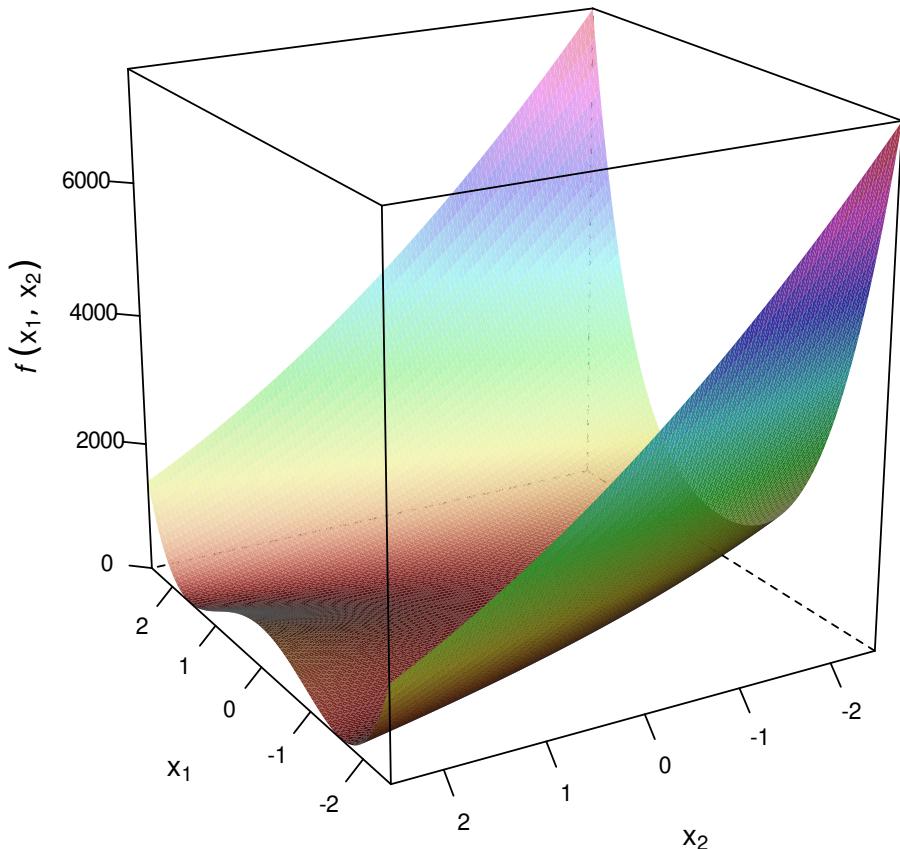


图 33.5: 香蕉函数图像



```
梯度函数
gr <- function(x) {
 numDeriv::grad(fn, c(x[1], x[2]))
}
optim(par = c(-1.2, 1), fn = fn, gr = gr, method = "BFGS")

$par
[1] 1 1
##
$value
[1] 9.595012e-18
##
$counts
function gradient
110 43
##
$convergence
[1] 0
##
$message
NULL

op <- OP(
 objective = F_objective(fn, n = 2L, G = gr),
 bounds = V_bound(ld = -3, ud = 3, nobj = 2L)
)
nlp <- ROI_solve(op, solver = "nloptr.lbfgs", start = c(-1.2, 1))
nlp$objval

[1] 1.364878e-17
nlp$solution

[1] 1 1
```



### 33.4.2.3 Ackley 函数

Ackley 函数是一个非凸函数，有大量局部极小值点，获取全局极小值点是一个比较有挑战的事。它的  $n$  维形式如下：

$$f(\mathbf{x}) = -ae^{-b\sqrt{\frac{1}{n}\sum_{i=1}^n x_i^2}} - e^{\frac{1}{n}\sum_{i=1}^n \cos(cx_i)} + a + e$$

其中， $a = 20, b = 0.2, c = 2\pi$ ，对  $\forall i = 1, 2, \dots, n$ ,  $x_i \in [-10, 10]$ ， $f(\mathbf{x})$  在  $\mathbf{x}^* = (0, 0, \dots, 0)$  取得全局最小值  $f(\mathbf{x}^*) = 0$ ，二维图像如图 33.6。

```
fn <- function(x, a = 20, b = 0.2, c = 2 * pi) {
 mean1 <- mean(x^2)
 mean2 <- mean(cos(c * x))
 -a * exp(-b * sqrt(mean1)) - exp(mean2) + a + exp(1)
}

df <- expand.grid(
 x = seq(-10, 10, length.out = 201),
 y = seq(-10, 10, length.out = 201)
)

df$fnxy = apply(df, 1, fn)

wireframe(
 data = df, fnxy ~ x * y,
 shade = TRUE, drape = FALSE,
 xlab = expression(x[1]),
 ylab = expression(x[2]),
 zlab = list(expression(italic(f) ~ group("(", list(x[1], x[2]), ")"))), rot = 90),
 scales = list(arrows = FALSE, col = "black"),
 par.settings = list(axis.line = list(col = "transparent")),
 screen = list(z = 120, x = -70, y = 0)
)
```

以 10 维的 Ackley 函数为例，先试一下普通的局部优化算法 — Nelder-Mead 算法，选择初值  $(2, 2, \dots, 2)$ ，看下效果，再与全局优化算法比较。

```
op <- OP(
 objective = F_objective(fn, n = 10L),
```

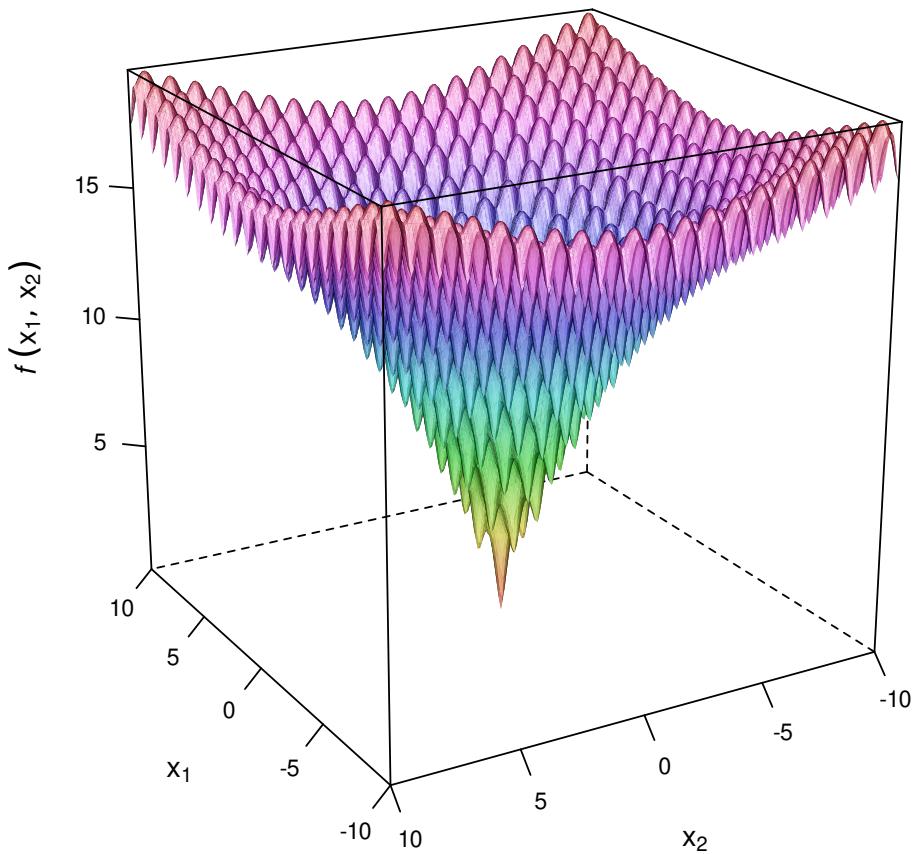


图 33.6: 二维 Ackley 函数图像

```

 bounds = V_bound(lb = -10, ub = 10, nobj = 10L)
}

nlp <- ROI_solve(op, solver = "nloptr.neldermead", start = rep(2, 10))
nlp$solution

[1] 2 2 2 2 2 2 2 2 2 2 2 2

nlp$objval

[1] 6.593599

```

可以说完全没有优化效果，已经陷入局部极小值。根据[nloptr 全局优化算法](#)的介绍，这里采用 directL 算法，因为是全局优化，不用选择初值。

```

调全局优化器
nlp <- ROI_solve(op, solver = "nloptr.directL")
nlp$solution

[1] 0 0 0 0 0 0 0 0 0 0 0 0

nlp$objval

[1] 4.440892e-16

fn(x = c(2, 2))

[1] 6.593599

fn(x = rep(2, 10))

[1] 6.593599

```

### 33.4.2.4 Radistrigin 函数

这里，还有另外一个例子，Radistrigin 函数也是多模函数

$$f(\mathbf{x}) = \sum_{i=1}^n (x_i^2 - 10 \cos(2\pi x_i) + 10)$$

```

fn <- function(x) {
 sum(x^2 - 10 * cos(2 * pi * x) + 10)
}

```



```
df <- expand.grid(
 x = seq(-4, 4, length.out = 201),
 y = seq(-4, 4, length.out = 201)
)

df$fnxy = apply(df, 1, fn)

wireframe(
 data = df, fnxy ~ x * y,
 shade = TRUE, drape = FALSE,
 xlab = expression(x[1]),
 ylab = expression(x[2]),
 zlab = list(expression(italic(f) ~ group("(", list(x[1], x[2]), ")"))), rot = 90,
 scales = list(arrows = FALSE, col = "black"),
 par.settings = list(axis.line = list(col = "transparent")),
 screen = list(z = 120, x = -65, y = 0)
)
```

设置 10 维的优化

```
op <- OP(
 objective = F_objective(fn, n = 10L),
 bounds = V_bound(ld = -50, ud = 50, nobj = 10L)
)
```

调全局优化器求解非凸优化问题

```
nlp <- ROI_solve(op, solver = "nloptr.directL")
nlp$solution
```

```
[1] 0 0 0 0 0 0 0 0 0 0
```

```
nlp$objval
```

```
[1] 0
```

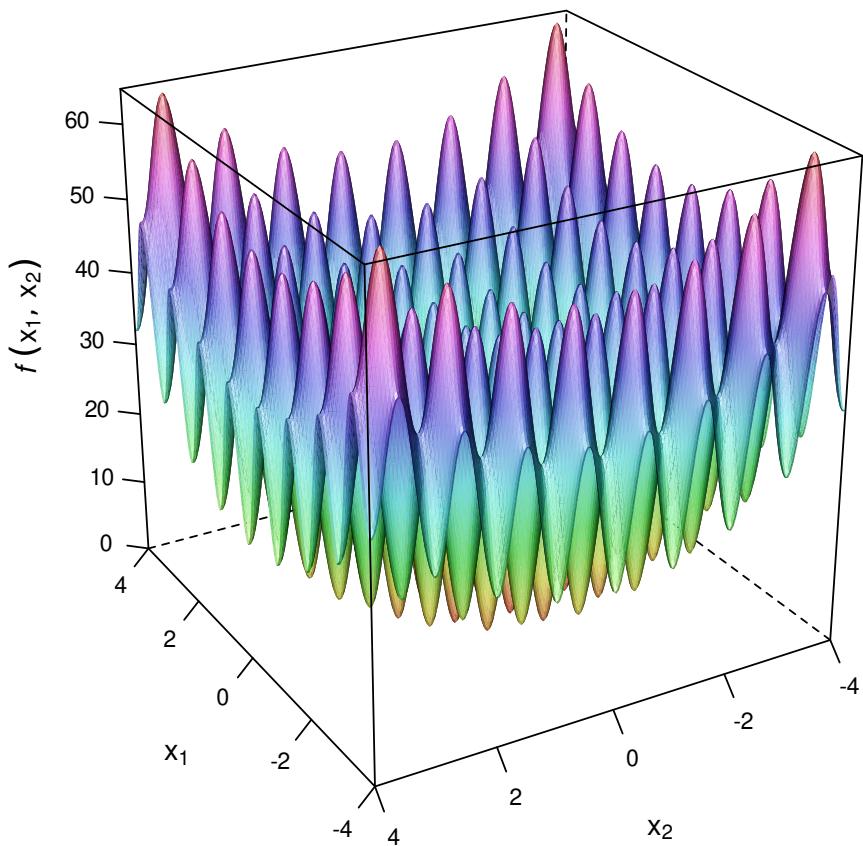


图 33.7: Radistrigin 函数



### 33.4.2.5 Schaffer 函数

$$f(x_1, x_2) = 0.5 + \frac{\sin^2(x_1^2 - x_2^2) - 0.5}{[1 + 0.001(x_1^2 + x_2^2)]^2}$$

在  $\mathbf{x}^* = (0, 0)$  处取得全局最小值  $f(\mathbf{x}^*) = 0$

```
fn <- function(x) {
 0.5 + ((sin(x[1]^2 - x[2]^2))^2 - 0.5) / (1 + 0.001*(x[1]^2 + x[2]^2))^2
}

df <- expand.grid(
 x = seq(-50, 50, length = 201),
 y = seq(-50, 50, length = 201)
)
df$fnxy = apply(df, 1, fn)

wireframe(
 data = df, fnxy ~ x * y,
 shade = TRUE, drape = FALSE,
 xlab = expression(x[1]),
 ylab = expression(x[2]),
 zlab = list(expression(italic(f) ~ group("(, list(x[1], x[2]), ")")),
 rot = 90),
 scales = list(arrows = FALSE, col = "black"),
 par.settings = list(axis.line = list(col = "transparent")),
 screen = list(z = 120, x = -70, y = 0)
)

df <- expand.grid(
 x = seq(-2, 2, length = 101),
 y = seq(-2, 2, length = 101)
)
df$fnxy = apply(df, 1, fn)

wireframe(
 data = df, fnxy ~ x * y,
 shade = TRUE, drape = FALSE,
 xlab = expression(x[1]),
```

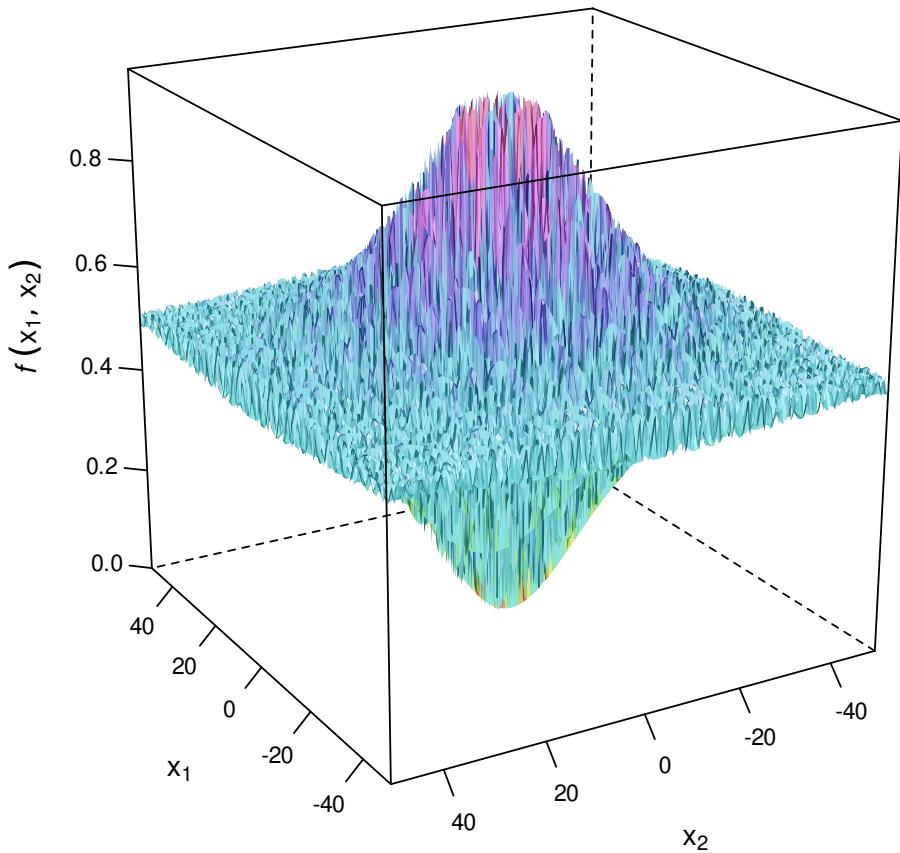


图 33.8: Schaffer 函数

```
ylab = expression(x[2]),
zlab = list(expression(italic(f) ~ group("(, list(x[1], x[2]), "))), rot = 90),
scales = list(arrows = FALSE, col = "black"),
par.settings = list(axis.line = list(col = "transparent")),
screen = list(z = 120, x = -70, y = 0)
)
```

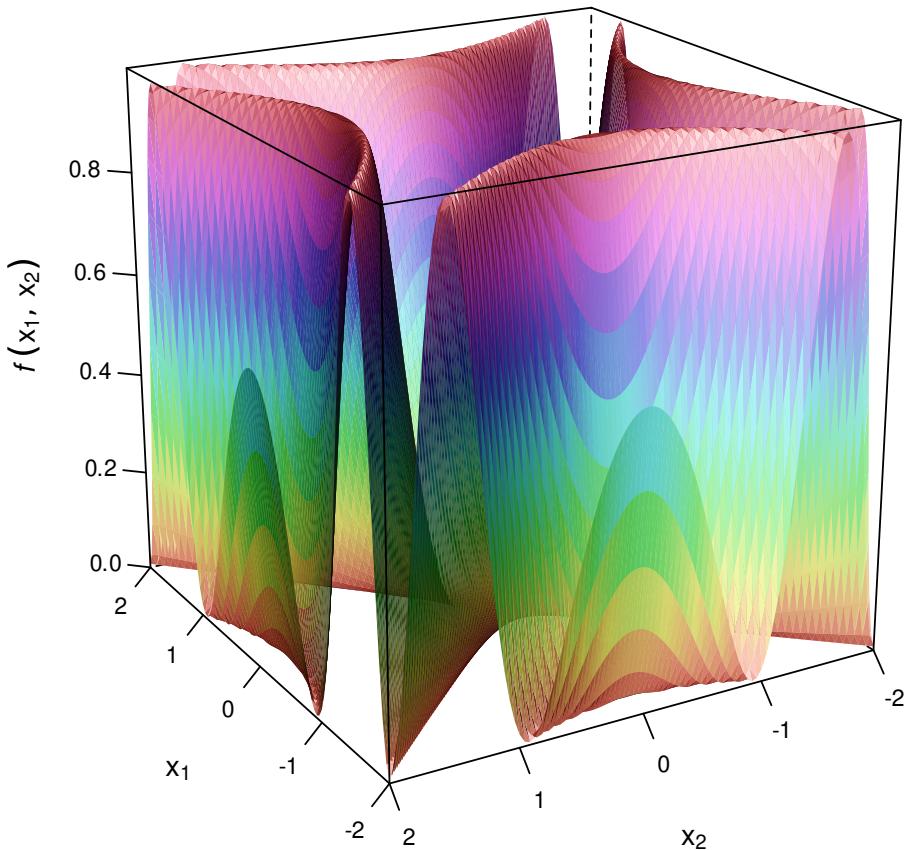


图 33.9: Schaffer 函数

#### 33.4.2.6 Hölder 函数

Hölder 桌面函数

$$f(x_1, x_2) = -|\sin(x_1) \cos(x_2) \exp\left(|1 - \frac{\sqrt{x_1^2 + x_2^2}}{\pi}|\right)|$$



在  $(8.05502, 9.66459), (-8.05502, 9.66459), (8.05502, -9.66459), (-8.05502, -9.66459)$  同时取得最小值  $-19.2085$ 。

```

fn <- function(x) {
 -abs(sin(x[1]) * cos(x[2])) * exp(abs(1 - sqrt(x[1]^2 + x[2]^2) / pi))
}

df <- expand.grid(
 x = seq(-10, 10, length = 101),
 y = seq(-10, 10, length = 101)
)
df$fnxy = apply(df, 1, fn)

wireframe(
 data = df, fnxy ~ x * y,
 shade = TRUE, drape = FALSE,
 xlab = expression(x[1]),
 ylab = expression(x[2]),
 zlab = list(expression(italic(f) ~ group("(", list(x[1], x[2]), ")"))),
 rot = 90),
 scales = list(arrows = FALSE, col = "black"),
 par.settings = list(axis.line = list(col = "transparent")),
 screen = list(z = 120, x = -60, y = 0)
)

```

### 33.4.2.7 Trid 函数

$n \geq 2$  维 Trid 函数

$$f(x) = \sum_{i=1}^n (x_i - 1)^2 - \sum_{i=2}^n x_i x_{i-1}$$

$\forall i = 1, 2, \dots, n$ ,  $f(x)$  在  $x_i = i(n+1-i)$  处取得全局极小值  $f(\mathbf{x}^*) = -n(n+4)(n-1)/6$ , 取值区间  $x \in [-n^2, n^2]$ ,  $\forall i = 1, 2, \dots, n$

```

fn <- function(x) {
 n <- length(x)
 sum((x - 1)^2) - sum(x[-1] * x[-n])
}

```

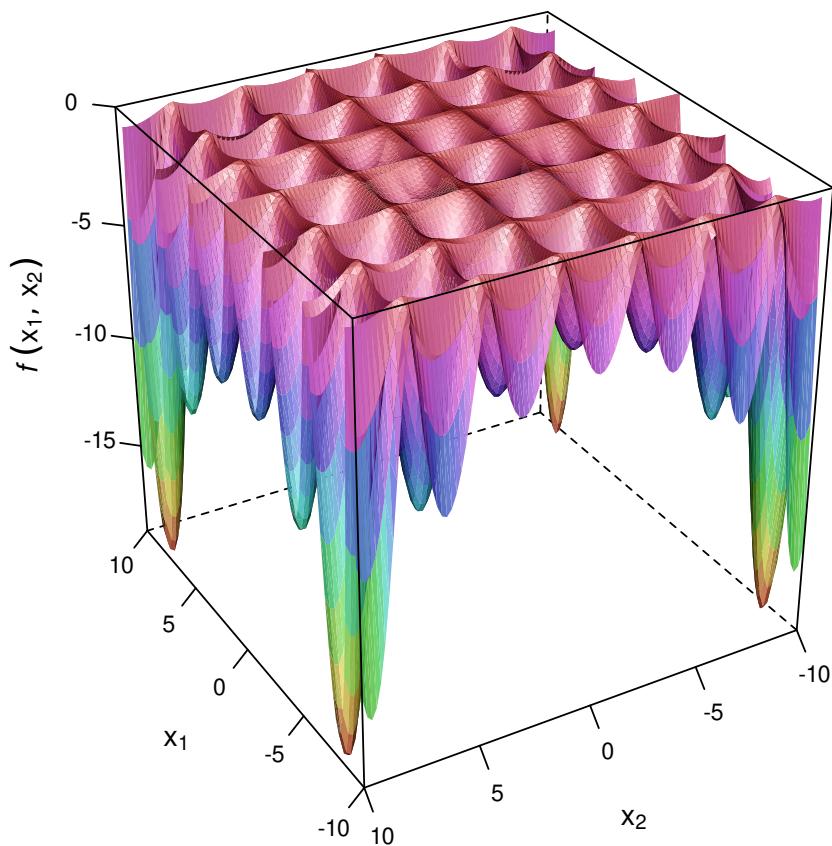


图 33.10: Hölder 函数

```

df <- expand.grid(
 x = seq(-4, 4, length = 101),
 y = seq(-4, 4, length = 101)
)
df$fnxy = apply(df, 1, fn)

wireframe(
 data = df, fnxy ~ x * y,
 shade = TRUE, drape = FALSE,
 xlab = expression(x[1]),
 ylab = expression(x[2]),
 zlab = list(expression(italic(f) ~ group("(", list(x[1], x[2]), ")"))), rot = 90),
 scales = list(arrows = FALSE, col = "black"),
 par.settings = list(axis.line = list(col = "transparent")),
 screen = list(z = -60, x = -70, y = 0)
)

```

### 33.4.2.8 超级复杂函数

有如下复杂的目标函数

$$\begin{aligned} \min_x \quad & \cos(x_1) \cos(x_2) - \sum_{i=1}^5 \left( (-1)^i \cdot i \cdot 2 \cdot \exp \left( -500 \cdot ((x_1 - i \cdot 2)^2 + (x_2 - i \cdot 2)^2) \right) \right) \\ s.t. \quad & -50 \leq x_1, x_2 \leq 50 \end{aligned}$$

```

subfun <- function(i, m) {
 (-1)^i * i * 2 * exp(-500 * ((m[1] - i * 2)^2 + (m[2] - i * 2)^2))
}

fn <- function(x) {
 cos(x[1]) * cos(x[2]) -
 sum(mapply(FUN = subfun, i = 1:5, MoreArgs = list(m = x)))
}

```

目标函数的图像见图 33.12，搜索区域  $[-50, 50] \times [-50, 50]$  内几乎没有变化的梯度，给寻优过程带来很大困难。

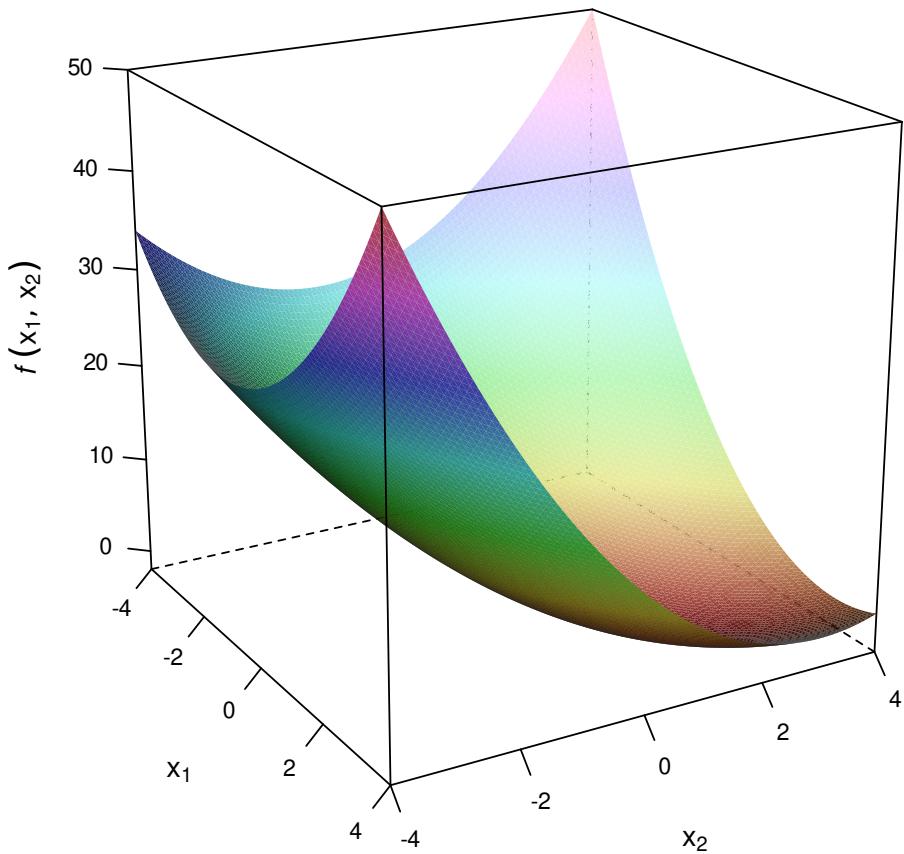


图 33.11: Trid 函数

```

df <- expand.grid(
 x = seq(-50, 50, length.out = 101),
 y = seq(-50, 50, length.out = 101)
)

df$fnxy = apply(df, 1, fn)

wireframe(
 data = df, fnxy ~ x * y,
 shade = TRUE, drape = FALSE,
 xlab = expression(x[1]),
 ylab = expression(x[2]),
 zlab = list(expression(italic(f) ~ group("(", list(x[1], x[2]), ")"))), rot = 90),
 scales = list(arrows = FALSE, col = "black"),
 par.settings = list(axis.line = list(col = "transparent")),
 screen = list(z = 120, x = -65, y = 0)
)

```

将区域  $[0, 12] \times [0, 12]$  的图像绘制出来，不难发现，有不少局部陷阱。

```

df <- expand.grid(
 x = seq(0, 12, length.out = 201),
 y = seq(0, 12, length.out = 201)
)

df$fnxy = apply(df, 1, fn)

wireframe(
 data = df, fnxy ~ x * y,
 shade = TRUE, drape = FALSE,
 xlab = expression(x[1]),
 ylab = expression(x[2]),
 zlab = list(expression(italic(f) ~ group("(", list(x[1], x[2]), ")"))), rot = 90),
 scales = list(arrows = FALSE, col = "black"),
 par.settings = list(axis.line = list(col = "transparent")),
 screen = list(z = 120, x = -65, y = 0)
)

```

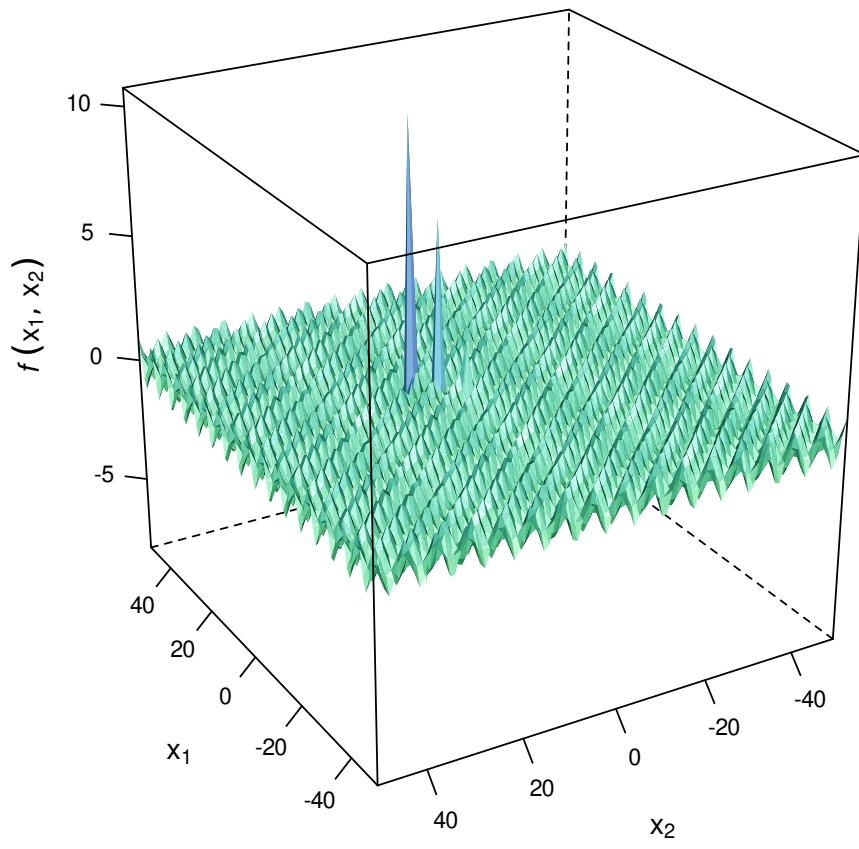


图 33.12: 函数图像

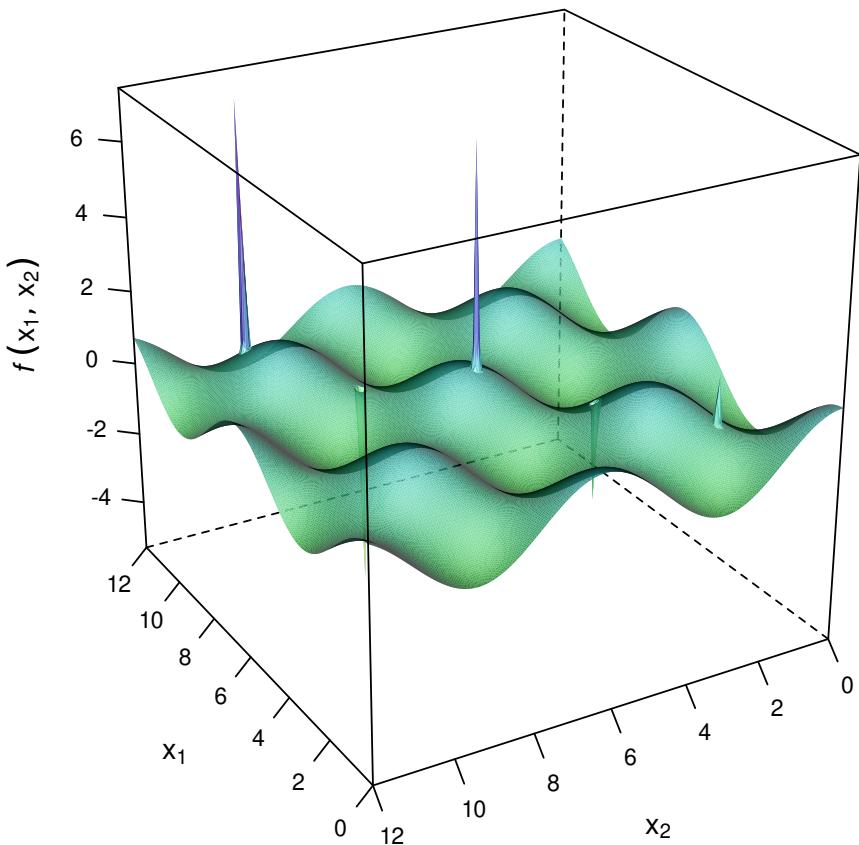


图 33.13: 局部放大函数图像



最优解在  $(7.999982, 7.999982)$  取得，目标函数值为 -7.978832。

```
fn(x = c(7.999982, 7.999982))
```

```
[1] -7.978832
```

面对如此复杂的函数，调用全局优化器

```
op <- OP(
 objective = F_objective(fn, n = 2L),
 bounds = V_bound(ld = -50, ud = 50, nobj = 2L)
)
nlp <- ROI_solve(op, solver = "nloptr.directL")
nlp$solution
```

```
[1] 22.22222 0.00000
```

```
nlp$objval
```

```
[1] -0.9734211
```

实际上，还是陷入局部最优解。

SETS:

P/1..5/;

Endsets

```
Min=@cos(x1) * @cos(x2) - @Sum(P(j): (-1)^j * j * 2 * @exp(-500 * ((x1 - j * 2)^2 + (x2
@Bnd(-50, x1, 50);
@Bnd(-50, x2, 50);
```

Lingo 18.0 启用全局优化求解器后，在  $(x_1 = 7.999982, x_2 = 7.999982)$  取得最小值 -7.978832。而默认未启用全局优化求解器的情况下，在  $(x_1 = 18.84956, x_2 = -40.84070)$  取得局部极小值 -1.000000。

### 33.4.3 多元非线性约束优化

R 自带的函数 `nlminb()` 可求解无约束、箱式约束优化问题，`constrOptim()` 还可求解线性不等式约束优化，其中包括带线性约束的二次规划。`optim()` 提供一大类优化算法，且包含随机优化算法—模拟退火算法，可求解无约束、箱式约束优化问题。

### 33.4.3.1 普通箱式约束

有如下箱式约束优化问题，目标函数和香蕉函数有些相似。

$$\begin{aligned} \min_x \quad & (x_1 - 1)^2 + 4 \sum_{i=1}^{n-1} (x_{i+1} - x_i^2)^2 \\ \text{s.t.} \quad & 2 \leq x_1, x_2, \dots, x_n \leq 4 \end{aligned}$$

```
fn <- function(x) {
 n <- length(x)
 sum(c(1, rep(4, n - 1)) * (x - c(1, x[-n])^2)^2)
}
```

$n$  维目标函数是非线性的，给定初值  $(3, 3, \dots, 3)$ ，下面求解 25 维的箱式约束，

```
nlminb(start = rep(3, 25), objective = fn, lower = rep(2, 25), upper = rep(4, 25))

$par
[1] 2.000000 2.000000 2.000000 2.000000 2.000000 2.000000 2.000000 2.000000
[9] 2.000000 2.000000 2.000000 2.000000 2.000000 2.000000 2.000000 2.000000
[17] 2.000000 2.000000 2.000000 2.000000 2.000000 2.000000 2.000000 2.109093
[25] 4.000000
##
$objective
[1] 368.1059
##
$convergence
[1] 0
##
$iterations
[1] 6
##
$evaluations
function gradient
10 177
##
$message
[1] "relative convergence (4)"
```

`nlminb()` 出于历史兼容性的原因尚且存在，最优解的第 24 个分量没有在可行域

的边界上。使用 `constrOptim()` 函数求解，默认求极小，需将箱式或线性不等式约束写成矩阵形式，即  $Ax \geq b$  的形式，参数 `ui` 是  $k \times n$  的约束矩阵  $A$ ，`ci` 是右侧  $k$  维约束向量  $b$ 。以上面的优化问题为例，将箱式约束  $2 \leq x_1, x_2 \leq 4$  转化为矩阵形式，约束矩阵和向量分别为：

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -1 & 0 \\ 0 & -1 \end{bmatrix}, \quad b = (2, 2, -4, -4)$$

```
constrOptim(
 theta = rep(3, 25), # 初始值
 f = fn, # 目标函数
 method = "Nelder-Mead", # 没有提供梯度，则必须用 Nelder-Mead 方法
 ui = rbind(diag(rep(1, 25)), diag(rep(-1, 25))),
 ci = c(rep(2, 25), rep(-4, 25))
)

$par
[1] 2.006142 2.002260 2.003971 2.003967 2.004143 2.004255 2.001178 2.002990
[9] 2.003883 2.006029 2.017345 2.009236 2.000949 2.007793 2.025831 2.007896
[17] 2.004514 2.004381 2.008771 2.015695 2.005803 2.009127 2.017988 2.257782
[25] 3.999846

$value
[1] 378.4208

$counts
function gradient
12048 NA

$convergence
[1] 1

$message
NULL
##
```



```
$outer.iterations
[1] 25
##
$barrier.value
[1] -0.003278963
```

从求解的结果来看，`convergence = 1` 意味着迭代次数到达默认的极限 `maxit = 500`，结合 `nlminb()` 函数的求解结果来看，实际上还没有收敛。如果没有提供梯度，则必须用 Nelder-Mead 方法，下面增加迭代次数到 1000。

```
constrOptim(
 theta = rep(3, 25), # 初始值
 f = fn, # 目标函数
 method = "Nelder-Mead",
 control = list(maxit = 1000),
 ui = rbind(diag(rep(1, 25)), diag(rep(-1, 25))),
 ci = c(rep(2, 25), rep(-4, 25))
)

$par
[1] 2.000081 2.000142 2.001919 2.000584 2.000007 2.000003 2.001097 2.001600
[9] 2.000207 2.000042 2.000250 2.000295 2.000580 2.002165 2.000453 2.000932
[17] 2.000456 2.000363 2.000418 2.000474 2.009483 2.001156 2.003173 2.241046
[25] 3.990754
##
$value
[1] 370.8601
##
$counts
function gradient
18036 NA
##
$convergence
[1] 1
##
$message
NULL
##
$
```

```
$outer.iterations
[1] 19

$barrier.value
[1] -0.003366467
```

还是没有收敛，可见 Nelder-Mead 方法在这个优化问题上收敛速度比较慢。下面考虑调用基于梯度的优化算法 — BFGS 方法。

```
输入 n 维向量, 输出 n 维向量
gr <- function(x) {
 n <- length(x)
 c(2 * (x[1] - 2), rep(0, n - 1))
 + 8 * c(0, x[-1] - x[-n]^2)
 - 16 * c(x[-n], 0) * c(x[-1] - x[-n]^2, 0)
}

constrOptim(
 theta = rep(3, 25), # 初始值
 f = fn, # 目标函数
 grad = gr,
 method = "BFGS",
 control = list(maxit = 1000),
 ui = rbind(diag(rep(1, 25)), diag(rep(-1, 25))),
 ci = c(rep(2, 25), rep(-4, 25))
)

$par
[1] 2.000000 2.000000 2.000000 2.000000 2.000000 2.000000 2.000000 2.000000
[9] 2.000000 2.000000 2.000000 2.000000 2.000000 2.000000 2.000000 2.000000
[17] 2.000000 2.000000 2.000000 2.000000 2.000000 2.000000 2.000000 2.000001
[25] 3.000000

$value
[1] 373

$counts
function gradient
```

云  
湘  
黄  
◎

```
3721 464
##
$convergence
[1] 0
##
$message
NULL
##
$outer.iterations
[1] 3
##
$barrier.value
[1] -0.003327104
```

相比于 Nelder-Mead 方法，目标值 373 更大，可见已陷入局部最优解，下面通过 ROI 包，分别调用求解器 L-BFGS 和 directL，发现前者同样陷入局部最优解，而后者可以获得与 nlmminb() 函数一致的结果。

```
调用改进的 BFGS 算法
op <- OP(
 objective = F_objective(fn, n = 25L, G = gr),
 bounds = V_bound(ld = 2, ud = 4, nobj = 25L)
)
nlp <- ROI_solve(op, solver = "nloptr.lbfgs", start = rep(3, 25))
nlp$objval

[1] 373
nlp$solution
```

```
[1] 2 3
```

```
调全局优化算法
nlp <- ROI_solve(op, solver = "nloptr.directL")
nlp$objval

[1] 368.1061
nlp$solution
```

```
[1] 2.000000 2.000000 2.000000 2.000000 2.000000 2.000000 2.000000
```

下面再与函数 `optim()` 提供的 L-BFGS-B 算法比较

```
optim(
 par = rep(3, 25), fn = fn, gr = NULL, method = "L-BFGS-B",
 lower = rep(2, 25), upper = rep(4, 25)
)
```

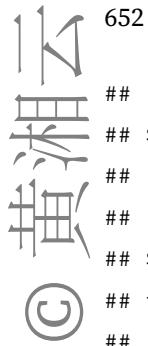
```

$par
[1] 2.000000 2.000000 2.000000 2.000000 2.000000 2.000000 2.000000 2.000000
[9] 2.000000 2.000000 2.000000 2.000000 2.000000 2.000000 2.000000 2.000000
[17] 2.000000 2.000000 2.000000 2.000000 2.000000 2.000000 2.000000 2.109093
[25] 4.000000
##
$value
[1] 368.1059
##
$counts
function gradient
6 6
##
$convergence
[1] 0
##
$message
[1] "CONVERGENCE: REL_REDUCTION_OF_F <= FACTR*EPSMCH"

```

值得注意的是，当提供梯度信息的时候，虽然求解速度提升了，但是最优解变差了。

```
optim(
 par = rep(3, 25), fn = fn, gr = gr, method = "L-BFGS-B",
 lower = rep(2, 25), upper = rep(4, 25)
)
```



```

$value
[1] 373

$counts
function gradient
2 2

$convergence
[1] 0

$message
[1] "CONVERGENCE: NORM OF PROJECTED GRADIENT <= PGTOL"
```

### 33.4.3.2 非线性严格不等式约束

第一个例子，目标函数是非线性的，约束条件也是非线性的，非线性不等式约束不包含等号。

$$\begin{aligned} \min_x \quad & (x_1 + 3x_2 + x_3)^2 + 4(x_1 - x_2)^2 \\ \text{s.t. } & \begin{cases} x_1 + x_2 + x_3 = 1 \\ 6x_2 + 4x_3 - x_1^3 > 3 \\ x_1, x_2, x_3 > 0 \end{cases} \end{aligned}$$

```
目标函数
fn <- function(x) (x[1] + 3 * x[2] + x[3])^2 + 4 * (x[1] - x[2])^2
目标函数的梯度
gr <- function(x) {
 c(
 2 * (x[1] + 3 * x[2] + x[3]) + 8 * (x[1] - x[2]), # 对 x[1] 求偏导
 6 * (x[1] + 3 * x[2] + x[3]) - 8 * (x[1] - x[2]), # 对 x[2] 求偏导
 2 * (x[1] + 3 * x[2] + x[3]) # 对 x[3] 求偏导
)
}
等式约束
heq <- function(x) {
```



```
x[1] + x[2] + x[3] - 1
}
等式约束的雅可比矩阵
这里只有一个等式约束，所以雅可比矩阵行数为 1
heq.jac <- function(x) {
 matrix(c(1, 1, 1), ncol = 3, byrow = TRUE)
}
不等式约束
要求必须是严格不等式，不能带等号，方向是 x > 0
hin <- function(x) {
 c(6 * x[2] + 4 * x[3] - x[1]^3 - 3, x[1], x[2], x[3])
}
不等式约束的雅可比矩阵
其实是有 4 个不等式约束，3 个目标变量约束，雅可比矩阵行数是 4
hin.jac <- function(x) {
 matrix(c(
 -3 * x[1]^2, 6, 4,
 1, 0, 0,
 0, 1, 0,
 0, 0, 1
), ncol = 3, byrow = TRUE)
}
```

调用 **alabama** 包的求解器

```
set.seed(12)
初始值
p0 <- runif(3)
求目标函数的极小值
ans <- alabama::constrOptim.nl(
 par = p0,
 # 目标函数
 fn = fn,
 gr = gr,
 # 等式约束
 heq = heq,
 heq.jac = heq.jac,
```

黄湘云

```
不等式约束
hin = hin,
hin.jac = hin.jac,
不显示迭代过程
control.outer = list(trace = FALSE)
)
ans
```

```
$par
[1] 7.390292e-04 4.497160e-12 9.992610e-01
##
$value
[1] 1.000002
##
$counts
function gradient
1230 163
##
$convergence
[1] 0
##
$message
NULL
##
$hessian
[,1] [,2] [,3]
[1,] 120517098 120517087 120517091
[2,] 120517087 120517115 120517095
[3,] 120517091 120517095 120517091
##
$outer.iterations
[1] 13
##
$lambda
[1] 4.481599
##
```



```
$sigma
[1] 120517089

$barrier.value
[1] 0.003472071

$K
[1] 4.269112e-08
```

ans 是 `constrOptim.nl()` 返回的一个 list, `convergence = 0` 表示迭代成功收敛, `value` 表示目标函数在迭代终止时的取值, `par` 表示满足约束条件, 成功收敛的情况下, 目标函数的参数值, `counts` 表示迭代过程中目标函数及其梯度计算的次数。

# 不提供梯度函数, 照样可以求解

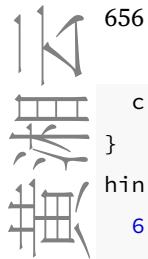
```
ans <- alabama:::constrOptim.nl(par = p0, fn = fn, heq = heq, hin = hin)
```

注意

等式和不等式约束的雅可比矩阵必须以 `matrix` 数据类型存储, 而不能以 `vector` 类型存储。要注意和后面 ROI 包的调用形式区别。

实际上, 可以用 ROI 调用 alabama 求解器的方式, 这种方式可以简化目标函数梯度和约束条件的表示

```
目标函数
fn <- function(x) (x[1] + 3 * x[2] + x[3])^2 + 4 * (x[1] - x[2])^2
目标函数的梯度
gr <- function(x) {
 c(
 2 * (x[1] + 3 * x[2] + x[3]) + 8 * (x[1] - x[2]),
 6 * (x[1] + 3 * x[2] + x[3]) - 8 * (x[1] - x[2]),
 2 * (x[1] + 3 * x[2] + x[3])
)
}
heq <- function(x) {
 x[1] + x[2] + x[3]
}
heq.jac <- function(x) {
```



```
c(1, 1, 1)
}

hin <- function(x) {
 6 * x[2] + 4 * x[3] - x[1]^3
}

hin.jac <- function(x) {
 c(-3 * x[1]^2, 6, 4)
}
```

通过 ROI 调用 alabama 求解器

```
set.seed(2020)
初始值
p0 <- runif(3)
定义目标规划
op <- OP(
 objective = F_objective(F = fn, n = 3L, G = gr), # 4 个目标变量
 constraints = F_constraint(
 F = list(heq = heq, hin = hin),
 dir = c("==", ">"),
 rhs = c(1, 3),
 # 等式和不等式约束的雅可比
 J = list(heq.jac = heq.jac, hin.jac = hin.jac)
),
 bounds = V_bound(lb = 0, ub = Inf, nobj = 3L),
 maximum = FALSE # 求最小
)
nlp <- ROI_solve(op, solver = "alabama", start = p0)
nlp$solution
```

```
[1] 1.674812e-06 9.994336e-08 9.999982e-01
nlp$objval
```

```
[1] 1
```



### 33.4.3.3 非线性和箱式约束

与上面的例子不同，下面这个例子的不等式约束包含等号，还有箱式约束，优化问题来源于[Ipopt 官网](#)，提供的初始值为  $x_0 = (1, 5, 5, 1)$ ，最优解为  $x_* = (1.00000000, 4.74299963, 3.82114998, 1.37940829)$ 。优化问题的具体内容如下：

$$\begin{aligned} \min_x \quad & x_1 x_4 (x_1 + x_2 + x_3) + x_3 \\ s.t. \quad & \begin{cases} x_1^2 + x_2^2 + x_3^2 + x_4^2 = 40 \\ x_1 x_2 x_3 x_4 \geq 25 \\ 1 \leq x_1, x_2, x_3, x_4 \leq 5 \end{cases} \end{aligned}$$

考虑用 ROI 调 nloptr 实现，看结果是否和例子一致，nloptr 支持不等式约束包含等号，支持箱式约束。

```
一个 4 维的目标函数
fn <- function(x) {
 x[1] * x[4] * (x[1] + x[2] + x[3]) + x[3]
}

目标函数的梯度
gr <- function(x) {
 c(
 x[4] * (2 * x[1] + x[2] + x[3]), x[1] * x[4],
 x[1] * x[4] + 1, x[1] * (x[1] + x[2] + x[3])
)
}

等式约束
heq <- function(x) {
 sum(x^2)
}

等式约束的雅可比
heq.jac <- function(x) {
 2 * c(x[1], x[2], x[3], x[4])
}

不等式约束
hin <- function(x) {
 prod(x)
}
```



```
不等式约束的雅可比
hin.jac <- function(x) {
 c(prod(x[-1]), prod(x[-2]), prod(x[-3]), prod(x[-4]))
}

定义目标规划
op <- OP(
 objective = F_objective(F = fn, n = 4L, G = gr), # 4 个目标变量
 constraints = F_constraint(
 F = list(heq = heq, hin = hin),
 dir = c("==", ">="),
 rhs = c(40, 25),
 # 等式和不等式约束的雅可比
 J = list(heq.jac = heq.jac, hin.jac = hin.jac)
),
 bounds = V_bound(ld = 1, ud = 5, nobj = 4L),
 maximum = FALSE # 求最小
)
```

```
目标函数初始值
fn(c(1, 5, 5, 1))
```

```
[1] 16
目标函数最优值
fn(c(1.00000000, 4.74299963, 3.82114998, 1.37940829))
```

```
[1] 17.01402
```

求解一般的非线性约束问题，求解器 `nloptr.mma` / `nloptr.cobyla` 仅支持非线性不等式约束，不支持等式约束，而 `nlminb` 只支持等式约束，因此，下面分别调用 `nloptr.auglag`、`nloptr.slsqp` 和 `nloptr.isres` 来求解上述优化问题。

```
nlp <- ROI_solve(op, solver = "nloptr.auglag", start = c(1, 5, 5, 1))
nlp$solution
```

```
[1] 1.000000 4.743025 3.821117 1.379413
nlp$objval
```

```
[1] 17.01402
```



```
nlp <- ROI_solve(op, solver = "nloptr.slsqp", start = c(1, 5, 5, 1))
nlp$solution

[1] 1.000000 4.742996 3.821155 1.379408

nlp$objval

[1] 17.01402

nlp <- ROI_solve(op, solver = "nloptr.isres", start = c(1, 5, 5, 1))
nlp$solution

[1] 1.043554 4.919033 3.573490 1.395034

nlp$objval

[1] 17.45605
```

可以看出，nloptr 提供的优化能力可以覆盖Ipopt 求解器，推荐使用 nloptr.slsqp 求解器。

#### 33.4.3.4 非线性混合整数约束

$$\begin{aligned} \max_x \quad & 1.5(x_1 - \sin(x_1 - x_2))^2 + 0.5x_2^2 + x_3^2 - x_1x_2 - 2x_1 + x_2x_3 \\ s.t. \quad & \begin{cases} -20 < x_1 < 20 \\ -20 < x_2 < 20 \\ -10 < x_3 < 10 \\ x_1, x_2 \in \mathbb{R}, \quad x_3 \in \mathbb{Z} \end{cases} \end{aligned}$$

```
fn <- function(x) {
 1.5 * (x[1] - sin(x[1] - x[2]))^2 + 0.5 * x[2]^2 + x[3]^2
 -x[1] * x[2] - 2 * x[1] + x[2] * x[3]
}

gr <- function(x) {
 c(
 3 * (x[1] - sin(x[1] - x[2])) * (1 - cos(x[1] - x[2])) - x[2] - 2,
 3 * (x[1] - sin(x[1] - x[2])) * cos(x[1] - x[2]) - x[2] - x[1] + x[3],
 2 * x[3] + x[2]
)
}
```



目前 ROI 还解不了

```
初始值
p0 <- c(2.1, 5.1, 5)
定义目标规划
op <- OP(
 objective = F_objective(F = fn, n = 3L, G = gr), # 3 个目标变量
 types = c("C", "C", "I"), # 目标变量的类型
 bounds = V_bound(lb = c(-20, -20, -10), ub = c(20, 20, 10), nobj = 3L),
 maximum = FALSE # 求最小
)
nlp <- ROI_solve(op, solver = "auto", start = p0)
nlp$solution
```

目标函数在 (4.49712, 9.147501, -4) 取得最小值 -86.72165

```
fn(x = c(4.49712, 9.147501, -4))
```

```
[1] -86.72165
```

### 33.4.3.5 含复杂目标函数

下面这个目标函数比较复杂，约束条件也是非线性的

$$\begin{aligned} \max_x \quad & \frac{(\sin(2\pi x_1))^3 \sin(2\pi x_2)}{x_1^3(x_1+x_2)} \\ s.t. \quad & \begin{cases} x_1^2 - x_2 + 1 \leq 0 \\ 1 - x_1 + (x_2 - 4)^2 \geq 0 \\ 0 \leq x_1, x_2 \leq 10 \end{cases} \end{aligned}$$

```
目标函数
fn <- function(x) (sin(2*pi*x[1]))^3 * sin(2*pi*x[2])/(x[1]^3*(x[1] + x[2]))
目标函数的梯度
gr <- function(x) {
 numDeriv::grad(fn, c(x[1], x[2]))
}

hin <- function(x) {
 c(
```



```
x[1]^2 - x[2] + 1,
1 - x[1] + (x[2] - 4)^2
)
}

hin.jac <- function(x) {
 matrix(c(
 2 * x[1], -1,
 -1, 2 * x[2]
),
 ncol = 2, byrow = TRUE
)
}

初始值
p0 <- c(2, 5)

定义目标规划
op <- OP(
 objective = F_objective(F = fn, n = 2L, G = gr), # 2 个目标变量
 constraints = F_constraint(
 F = list(hin = hin),
 dir = c("<=", "<="),
 rhs = c(0, 0),
 # 不等式约束的雅可比
 J = list(hin.jac = hin.jac)
),
 bounds = V_bound(ld = 0, ud = 10, nobj = 2L),
 maximum = TRUE # 求最大
)
nlp <- ROI_solve(op, solver = "nloptr.isres", start = p0)
nlp$solution
```

```
[1] 1.227973 4.245372
nlp$objval
```

```
[1] 0.09582504
```



下面再给一个来自 [Octave 优化文档](#) 的示例，该优化问题包含多个非线性的等式约束。

$$\begin{aligned} & \min_x \quad e^{\prod_{i=1}^5 x_i} - \frac{1}{2}(x_1^3 + x_2^3 + 1)^2 \\ & s.t. \quad \begin{cases} \sum_{i=1}^5 x_i^2 - 10 = 0 \\ x_2 x_3 - 5x_4 x_5 = 0 \\ x_1^3 + x_2^3 + 1 = 0 \end{cases} \end{aligned}$$

```
一个 5 维的目标函数
fn <- function(x) {
 exp(prod(x)) - 0.5 * (x[1]^3 + x[2]^3 + 1)^2
}

目标函数的梯度
gr <- function(x) {
 c(
 exp(prod(x))*prod(x[-1]) - 3*(x[1]^3 + x[2]^3 + 1)*x[1]^2,
 exp(prod(x))*prod(x[-2]) - 3*(x[1]^3 + x[2]^3 + 1)*x[2]^2,
 exp(prod(x))*prod(x[-3]),
 exp(prod(x))*prod(x[-4]),
 exp(prod(x))*prod(x[-5])
)
}

等式约束
heq <- function(x) {
 c(
 sum(x^2) - 10,
 x[2] * x[3] - 5 * x[4] * x[5],
 x[1]^3 + x[2]^3 + 1
)
}

等式约束的雅可比
heq.jac <- function(x) {
 matrix(c(2 * x[1], 2 * x[2], 2 * x[3], 2 * x[4], 2 * x[5],
 0, x[3], x[2], -5 * x[5], -5 * x[4],
 3 * x[1]^2, 3 * x[2]^2, 0, 0, 0),
 ncol = 5, byrow = TRUE)
```

```

)
}

定义目标规划
op <- OP(
 objective = F_objective(F = fn, n = 5L, G = gr), # 5 个目标变量
 constraints = F_constraint(
 F = list(heq = heq,
 dir = "==" ,
 rhs = 0,
 # 等式的雅可比
 J = list(heq.jac = heq.jac)
),
 bounds = V_bound(lb = -Inf, ub = Inf, nobj = 5L),
 maximum = FALSE # 求最小
)
)

```

调用 SQP (序列二次规划) 求解器

```
nlp <- ROI_solve(op, solver = "nloptr.slsqp", start = c(-1.8, 1.7, 1.9, -0.8, -0.8))
nlp$solution
```

```
[1] -1.7171435 1.5957096 1.8272458 -0.7636431 -0.7636431
```

计算结果和 Octave 的示例一致。

### 33.4.3.6 含复杂约束条件

$$\begin{aligned} \min_x \quad & \exp(\sin(50 \cdot x)) + \sin(60 \cdot \exp(y)) + \sin(70 \cdot \sin(x)) \\ & + \sin(\sin(80 \cdot y)) - \sin(10 \cdot (x + y)) + \frac{(x^2 + y^2)^{\sin(y)}}{4} \\ s.t. \quad & \left\{ \begin{array}{l} x - ((\cos(y))^x - x)^y = 0 \\ -50 \leq x_1, x_2 \leq 50 \end{array} \right. \end{aligned}$$

Lingo 代码如下：

```
Min = @exp(@sin(50 * x)) + @sin(60 * @exp(y)) + @sin(70 * @sin(x))
 + @sin(@sin(80 * y)) - @sin(10 * (x + y)) + (x^2 + y^2)^@sin(y) / 4;

x - ((@cos(y))^x - x)^y = 0;
```



@bnd(-50, x, 50);  
@bnd(-50, y, 50);

启用全局优化求解器，求解 14 分钟，在 (0.08256372, 24.56510) 取得极小值 -2.863497。不启用全局优化器就没法解，Lingo 会报错，找不到最优解，勉强找到一个可行解 (0.06082750, 44.12793)，目标值为 -1.29816。

```
fn <- function(x) {
 exp(sin(50 * x[1])) + sin(60 * exp(x[2])) +
 sin(70 * sin(x[1])) + sin(sin(80 * x[2])) -
 sin(10 * (x[1] + x[2])) + (x[1]^2 + x[2]^2)^(sin(x[2])) / 4
}

gr <- function(x){
 numDeriv::grad(fn, c(x[1], x[2]))
}

heq <- function(x){
 x[1] - (cos(x[2]))^x[1] - x[1])^x[2]
}

heq.jac <- function(x){
 numDeriv::grad(heq, c(x[1], x[2]))
}

fn(x = c(0.06082750, 44.12793))

[1] -1.29816
fn(x = c(1, 0))

[1] 1.966877
heq(x = c(0.06082750, 44.12793))

[1] 1.923673e-08
heq(x = c(1, 0))

[1] 0

定义目标规划
op <- OP(
 objective = F_objective(F = fn, n = 2L, G = gr), # 2 个目标变量
```



```
constraints = F_constraint(
 F = list(heq = heq),
 dir = "==" ,
 rhs = 0,
 J = list(heq.jac = heq.jac)
),
bounds = V_bound(lb = -50, ub = 50, nobj = 2L),
maximum = FALSE # 求最小
)
```

nloptr.auglag 无法求解此优化问题

```
nlp <- ROI_solve(op, solver = "nloptr.auglag", start = c(1, 0))
nlp$solution
```

调 nloptr.isres 求解器，每次执行都会得到不同的局部最优解

```
nlp <- ROI_solve(op, solver = "nloptr.isres", start = c(1, 0))
nlp$solution
```

```
[1] 28.86779 18.44372
```

```
nlp$objval
```

```
[1] -3.320204
```

比如下面三组

```
fn(x = c(40.95941, 41.52914))
```

```
[1] -1.025926
```

```
heq(x = c(40.95941, 41.52914))
```

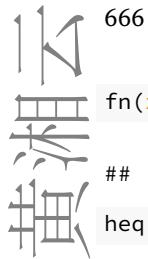
```
[1] NaN
```

```
fn(x = c(-21.88091, 28.96994))
```

```
[1] -1.467513
```

```
heq(x = c(-21.88091, 28.96994))
```

```
[1] NaN
```



```
fn(x = c(-49.921967437, 4.8499336803))
[1] -3.466596
heq(x = c(-49.921967437, 4.8499336803))
[1] -8.515447e+208
```

## 33.5 非线性方程

### 33.5.1 一元非线性方程

牛顿-拉弗森方法

```
library(rootSolve)
```

### 33.5.2 非线性方程组

```
library(BB)
```

二项混合泊松分布的参数最大似然估计

```
poissmix.loglik <- function(p, y) {
 # Log-likelihood for a binary Poisson mixture distribution
 i <- 0:(length(y) - 1)

 loglik <- y * log(p[1] * exp(-p[2]) * p[2]^i / exp(lgamma(i + 1)) +
 (1 - p[1]) * exp(-p[3]) * p[3]^i / exp(lgamma(i + 1)))

 sum(loglik)
}
Data from Hasselblad (JASA 1969)
介绍应用场景
poissmix.dat <- data.frame(death = 0:9,
 freq = c(162, 267, 271, 185, 111, 61, 27, 8, 3, 1))
lo <- c(0, 0, 0) # lower limits for parameters
hi <- c(1, Inf, Inf) # upper limits for parameters
```

黄湘云

```
p0 <- runif(3, c(0.2, 1, 1), c(0.8, 5, 8))
a randomly generated vector of length 3
y <- c(162, 267, 271, 185, 111, 61, 27, 8, 3, 1)

ans1 <- spg(
 par = p0, fn = poissmix.loglik, y = y, lower = lo, upper = hi,
 control = list(maximize = TRUE, trace = FALSE)
)
ans2 <- BBoptim(
 par = p0, fn = poissmix.loglik, y = y,
 lower = lo, upper = hi, control = list(maximize = TRUE)
)

iter: 0 f-value: -2136.431 pgrad: 236.9752
iter: 10 f-value: -1995.89 pgrad: 2.961353
iter: 20 f-value: -2041.139 pgrad: 2.57697
iter: 30 f-value: -1989.974 pgrad: 0.4742151
iter: 40 f-value: -1989.949 pgrad: 0.2614752
iter: 50 f-value: -1989.946 pgrad: 0.01959506
iter: 60 f-value: -1989.946 pgrad: 0.002494289
Successful convergence.

ans2

$par
[1] 0.3598829 1.2560906 2.6634011
##
$value
[1] -1989.946
##
$gradient
[1] 0.0001000444
##
$fn.reduction
[1] -146.4848
##
$iter
[1] 68
```

云  
湘  
黄  
◎

```

$feval
[1] 170

$convergence
[1] 0

$message
[1] "Successful convergence"

$cpar
method M
2 50
```

计算最大似然处的黑塞矩阵以及参数的标准差

```
hess <- numDeriv:::hessian(x = ans2$par, func = poissmix.loglik, y = y)
Note that we have to supplied data vector 'y'
hess

[,1] [,2] [,3]
[1,] -907.1105 270.2287 341.2543
[2,] 270.2287 -113.4794 -61.6819
[3,] 341.2543 -61.6819 -192.7822
se <- sqrt(diag(solve(-hess)))
se
```

```
[1] 0.1946836 0.3500308 0.2504769
```

从不同初始值出发尝试寻找全局最大值，实际找的是一系列局部最大值

```
3 randomly generated starting values
p0 <- matrix(runif(30, c(0.2, 1, 1), c(0.8, 8, 8)), 10, 3, byrow = TRUE)
ans <- multiStart(
 par = p0, fn = poissmix.loglik, action = "optimize",
 y = y, lower = lo, upper = hi, control = list(maximize = TRUE)
)

Parameter set : 1 ...
iter: 0 f-value: -2076.377 pgrad: 266.5811
```

```
iter: 10 f-value: -1991.788 pgrad: 3.394882
iter: 20 f-value: -1990.932 pgrad: 8.266675
iter: 30 f-value: -1989.958 pgrad: 0.2441652
iter: 40 f-value: -1989.946 pgrad: 0.001411991
Successful convergence.

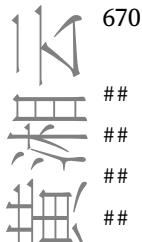
Parameter set : 2 ...
iter: 0 f-value: -3999.343 pgrad: 6.350898
iter: 10 f-value: -2015.457 pgrad: 2.400803
Successful convergence.

Parameter set : 3 ...
iter: 0 f-value: -2526.385 pgrad: 3.959104
iter: 10 f-value: -1997.785 pgrad: 4.651176
iter: 20 f-value: -2041.124 pgrad: 130.6335
iter: 30 f-value: -1989.979 pgrad: 0.4133676
iter: 40 f-value: -1989.953 pgrad: 0.2001525
iter: 50 f-value: -1989.946 pgrad: 0.02953584
Successful convergence.

Parameter set : 4 ...
iter: 0 f-value: -4036.966 pgrad: 7.725057
iter: 10 f-value: -1993.146 pgrad: 3.356279
iter: 20 f-value: -1992.445 pgrad: 3.162911
iter: 30 f-value: -1999.964 pgrad: 3.124857
iter: 40 f-value: -1990.201 pgrad: 0.9762675
iter: 50 f-value: -1989.962 pgrad: 0.3950169
iter: 60 f-value: -1989.946 pgrad: 0.0507498
iter: 70 f-value: -1989.946 pgrad: 0.0001978151
Successful convergence.

Parameter set : 5 ...
iter: 0 f-value: -2048.809 pgrad: 2.862445
iter: 10 f-value: -1992.344 pgrad: 2.68979
iter: 20 f-value: -1990.604 pgrad: 7.2791
iter: 30 f-value: -1989.978 pgrad: 0.3772993
iter: 40 f-value: -1989.946 pgrad: 0.004172307
iter: 50 f-value: -1989.946 pgrad: 0.004260983
Successful convergence.

Parameter set : 6 ...
```



```
iter: 0 f-value: -4777.283 pgrad: 7.596832
iter: 10 f-value: -1991.838 pgrad: 11.02078
iter: 20 f-value: -1990.272 pgrad: 0.5307333
iter: 30 f-value: -1989.963 pgrad: 2.230793
iter: 40 f-value: -1989.946 pgrad: 0.008421921
iter: 50 f-value: -1989.946 pgrad: 0.0001841727
Successful convergence.

Parameter set : 7 ...
iter: 0 f-value: -2019.928 pgrad: 3.485709
iter: 10 f-value: -1990.626 pgrad: 1.833378
iter: 20 f-value: -1989.999 pgrad: 1.098717
iter: 30 f-value: -1989.947 pgrad: 0.3092782
iter: 40 f-value: -1989.946 pgrad: 0.007039489
Successful convergence.

Parameter set : 8 ...
iter: 0 f-value: -2764.625 pgrad: 4.891128
iter: 10 f-value: -2001.398 pgrad: 2.273737e-06
Successful convergence.

Parameter set : 9 ...
iter: 0 f-value: -2167.165 pgrad: 195.5499
iter: 10 f-value: -2001.54 pgrad: 2.194864
iter: 20 f-value: -2000.825 pgrad: 0.6559458
iter: 30 f-value: -1992.777 pgrad: 7.064828
iter: 40 f-value: -1991.747 pgrad: 3.357115
iter: 50 f-value: -1989.983 pgrad: 2.772795
iter: 60 f-value: -1989.946 pgrad: 0.03392643
iter: 70 f-value: -1989.946 pgrad: 0.0003728928
Successful convergence.

Parameter set : 10 ...
iter: 0 f-value: -2100.94 pgrad: 317.5313
iter: 10 f-value: -1991.327 pgrad: 2.7843
iter: 20 f-value: -1990.415 pgrad: 1.435174
iter: 30 f-value: -1990.046 pgrad: 3.248585
iter: 40 f-value: -1989.946 pgrad: 0.06813025
iter: 50 f-value: -1989.946 pgrad: 0.001450644
Successful convergence.
```



```
selecting only converged solutions
pmat <- round(cbind(ans$fvalue[ans$conv], ans$par[ans$conv,]), 4)
dimnames(pmat) <- list(NULL, c("fvalue", "parameter 1", "parameter 2", "parameter 3"))
pmat[!duplicated(pmat),]

fvalue parameter 1 parameter 2 parameter 3
[1,] -1989.946 0.6401 2.6634 1.2561
[2,] -1997.263 0.4922 2.4559 1.8567
[3,] -1989.946 0.3599 1.2561 2.6634
[4,] -2000.039 0.7931 2.0681 2.4778
[5,] -1989.946 0.3599 1.2560 2.6634
```

用一个具体的参数估计问题，求极大似然点，混合正态分布隐函数方程组求解非线性方程组 [Varadhan and Gilbert, 2009]

## 33.6 多目标规划

多目标规划的基本想法是将多目标问题转化为单目标问题，常见方法有理想点法、线性加权法、非劣解集法、极大极小法。理想点法是先在给定约束条件下分别求解单个目标的最优值，构造新的单目标函数。线性加权法是给每个目标函数赋予权重系数，各个权重系数之和等于 1。非劣解集法是先求解其中一个单目标函数的最优值，然后将其设为等式约束，将其最优值从最小值开始递增，然后求解另一个目标函数的最小值。极大极小法是采用标准的简面体爬山法和通用全局优化法求解多目标优化问题。

R 环境中，**GParato** 主要用来求解多目标规划问题。[试验设计和过程优化与 R 语言的约束优化](#) 章节，[优化和解方程](#)

$$\begin{aligned} \min_x & \left\{ \begin{array}{l} f_1(x) = 0.5x_1 + 0.6x_2 + 0.7 \exp\left(\frac{x_1+x_3}{10}\right) \\ f_2(x) = (x_1 - 2x_2)^2 + (2x_2 - 3x_3)^2 + (5x_3 - x_1)^2 \end{array} \right. \\ \text{s.t. } & x_1 \in [10, 80], x_2 \in [20, 90], x_3 \in [15, 100] \end{aligned}$$

```
library(DiceKriging)
library(emoa)
library(GPareto)
library(DiceDesign)
```



```
library(Ternary)
TernaryPlot(
 atip = "Top", btip = "Bottom", ctip = "Right",
 axis.col = "red", col = rgb(0.8, 0.8, 0.8)
)
HorizontalGrid(grid.lines = 2, grid.col = "blue", grid.lty = 1)
```

## 33.7 经典优化问题

旅行商问题、背包问题、指派问题、选址问题、网络流量问题

规划快递员送餐的路线：从快递员出发地到各个取餐地，再到顾客家里，如何规划路线使得每个顾客下单到拿到餐的时间间隔小于 50 分钟，完成送餐，快递员的总时间最少？

## 33.8 回归与优化

简单线性回归

是否能给大家提供一些思路？

Lasso [Tibshirani, 1996]

Least Angle Regression [Efron et al., 2004]

为了解决 Lasso 的有偏估计问题，自适应 Lasso、松弛 Lasso SCAD (Smoothly Clipped Absolute Deviation)[Kim et al., 2008] MCP (Minimax Concave Penalty)[Zhang, 2010]

由于缺少高效的求解算法，Lasso 在高维小样本特征选择研究中没有广泛流行，最小角回归 (Least Angle Regression, LAR) 算法 [Efron et al., 2004] 的出现有力促进了 Lasso 在高维小样本数据中的应用

**bestsubset** 最优子集回归

经典的普通最小二乘、广义最小二乘、岭回归、逐步回归、Lasso 回归、最优子集回归都可转化为优化问题，一般形式如下



$$\underbrace{\hat{\theta}_{\lambda_n}}_{\text{待估参数}} \in \arg \min_{\theta \in \Omega} \left\{ \underbrace{\mathcal{L}(\theta; Z_1^n)}_{\text{损失函数}} + \lambda_n \underbrace{\mathcal{R}(\theta)}_{\text{正则化项}} \right\}.$$

下面尝试以 nloptr 包的优化器来展示求解过程，并与 Base R、glmnet 和 MASS 实现的回归模型比较。

$$\arg \min_{\beta, \lambda} \frac{1}{2} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_1$$

其中， $X \in \mathbb{R}^{m \times n}$ ， $y \in \mathbb{R}^m$ ， $\beta \in \mathbb{R}^n$ ， $0 < \lambda \in \mathbb{R}$

$$y = X\beta + \epsilon$$

$$\hat{\beta} = (X^\top X)^{-1} X^\top y, \quad \hat{Y} = X(X^\top X)^{-1} X^\top y$$

```
set.seed(123)
n <- 200
p <- 50
x <- matrix(rnorm(n * p), n)
y <- rnorm(n)
lm(y ~ x + 0)
y 的估计
教科书版
fit_base = function(x, y) {
 x %*% solve(t(x) %*% x) %*% t(x) %*% y
}
先向量计算，然后矩阵计算
fit_vector = function(x, y) {
 x %*% (solve(t(x) %*% x) %*% (t(x) %*% y))
}
$X'X$ 是对称的，防止求逆
fit_inv = function(x, y) {
 x %*% solve(crossprod(x), crossprod(x, y))
}
```

QR 分解  $X_{n \times p} = Q_{n \times p} R_{p \times p}$ ,  $n > p$ ,  $Q^\top Q = I$ ,  $R$  是上三角矩阵,  
 $\hat{Y} = X(X^\top X)^{-1} X^\top y = QQ^\top y$



```
fit_qr <- function(x, y) {
 decomp <- qr(x)
 qr.qy(decomp, qr.qty(decomp, y))
}
lm.fit(x, y)
```

若  $A = X^\top X$  是正定矩阵，则  $A = LL^\top$ ， $L$  是下三角矩阵

```
fit_chol <- function(x, y) {
 decomp <- chol(crossprod(x))
 lxy <- backsolve(decomp, crossprod(x, y), transpose = TRUE)
 b <- backsolve(decomp, lxy)
 x %*% b
}
```

```
Using C/C++
system.time(RcppEigen::fastLmPure(x, y, method = 1)) ## QR
system.time(RcppEigen::fastLmPure(x, y, method = 2)) ## Cholesky
system.time(RcppArmadillo::fastLmPure(x, y, method = 1)) ## QR
system.time(RcppArmadillo::fastLmPure(x, y, method = 2)) ## Cholesky
```

## 33.9 对数似然

随机变量  $X$  服从参数为  $\lambda > 0$  的指数分布，密度函数  $p(x)$  为

$$p(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

其中， $\lambda > 0$ ，下面给定一系列模拟样本观察值  $x_1, x_2, \dots, x_n$ ，估计参数  $\lambda$ 。对数似然函数  $\ell(\lambda) = \log \prod_{i=1}^n f(x_i) = n \log \lambda - \lambda \sum_{i=1}^n x_i$ 。解此方程即可得到  $\lambda$  的极大似然估计  $\lambda_{mle} = \frac{1}{\bar{X}}$ ，极大值  $\ell(\lambda_{mle}) = -n(1 + \log \bar{X})$ 。

根据上述样本，计算样本均值  $(\mu - 1.5 * \sigma / \sqrt{n}, \mu + 1.5 * \sigma / \sqrt{n})$  和方差  $(0.8\sigma, 1.5\sigma)$ 。已知正态分布  $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$  的对数似然形式  $\ell(\mu, \sigma^2) = \log \prod_{i=1}^n f(x_i) = \sum_{i=1}^n \log f(x_i)$ 。正态分布的密度函数的对数可用 `dnorm(..., log = TRUE)` 计算。

生成服从指数分布的样本，计算样本的均值和方差，依据均值和方差构造区间，然后将区间网格化，在此网格上绘制正态分布的对数似然函数。绕那么大一个



圈子，其实就是绘制正态分布的对数似然函数。

```
set.seed(2021)
n <- 20 # 随机数的个数
x <- rexp(n, rate = 5) # 服从指数分布的随机数
m <- 40 # 网格数

mu <- seq(
 mean(x) - 1.5 * sd(x) / sqrt(n),
 mean(x) + 1.5 * sd(x) / sqrt(n),
 length.out = m
)
sigma <- seq(0.8 * sd(x), 1.5 * sd(x), length.out = m)
df <- expand.grid(x = mu, y = sigma)
正态分布的对数似然
loglik <- function(b, x0) -sum(dnorm(x0, b[1], b[2]), log = TRUE)

df$fnxy = apply(df, 1, loglik, x0 = x)

wireframe(
 data = df, fnxy ~ x * y,
 shade = TRUE, drape = FALSE,
 xlab = expression(mu),
 ylab = expression(sigma),
 zlab = list(expression(-loglik(mu, sigma)), rot = 90),
 scales = list(arrows = FALSE, col = "black"),
 par.settings = list(axis.line = list(col = "transparent")),
 screen = list(z = 120, x = -70, y = 0)
)
```

## 33.10 微分方程

### ode45 求解偏微分方程

**pracma** 实现了 `ode23`, `ode23s`, `ode45` 等几个自适应的 Runge-Kutta 求解器, **de-Solve** 包求解 ODE (常微分方程), DAE (微分代数方程), DDE (延迟微分方程,

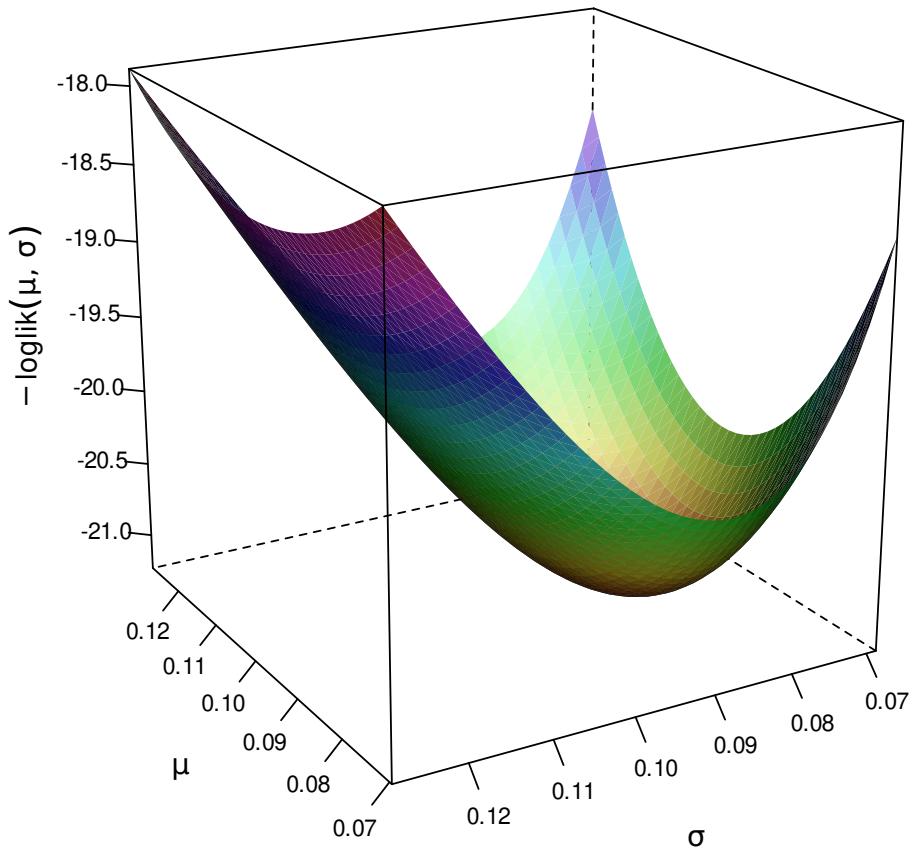


图 33.14: 正态分布参数的负对数似然函数



包含刚性和非刚性方程) 和 PDE (偏微分方程), **bvpSolve** 包求解 DAE/ODE 方程的边值问题。**ReacTran** [Soetaert and Meysman, 2012] 可将偏微分方程转为常微分方程组, 解决反应运输问题, 在笛卡尔、极坐标、圆柱形和球形网格上离散偏微分方程。**sundials** 提供一系列非线性方程、常微分方程、微分代数方程求解器, Satyaprakash Nayak 开发了相应的 **sundialr** 包。

### 33.10.1 常微分方程

[洛伦兹系统](#)是一个常微分方程组, 系统参数的默认值为 ( $\sigma = 10$ ,  $\rho = 28$ ,  $\beta = 8/3$ ), 初值为  $(-13, -14, 47)$ 。

$$\begin{cases} \frac{\partial x}{\partial t} = \sigma(y - x) \\ \frac{\partial y}{\partial t} = x(\rho - z) - y \\ \frac{\partial z}{\partial t} = xy - \beta z \end{cases}$$

```
library(deSolve)
参数
pars <- c(a = -8 / 3, b = -10, c = 28)
初值
state <- c(X = 1, Y = 1, Z = 1)
时间间隔
times <- seq(0, 100, by = 0.01)
定义方程组
lorenz_fun <- function(t, state, parameters) {
 with(as.list(c(state, parameters)), {
 dX <- a * X + Y * Z
 dY <- b * (Y - Z)
 dZ <- -X * Y + c * Y - Z
 list(c(dX, dY, dZ))
 })
}
out <- ode(
 y = state, times = times,
 func = lorenz_fun, parms = pars
)
```

调用 **scatterplot3d** 绘制三维曲线图, 如图33.15 所示

```
library(scatterplot3d)

scatterplot3d(
 x = out[, "X"], y = out[, "Y"], z = out[, "Z"],
 col.axis = "black", type = "l", color = "gray",
 xlab = expression(x), ylab = expression(y), zlab = expression(z),
 col.grid = "gray", main = "Lorenz"
)
```

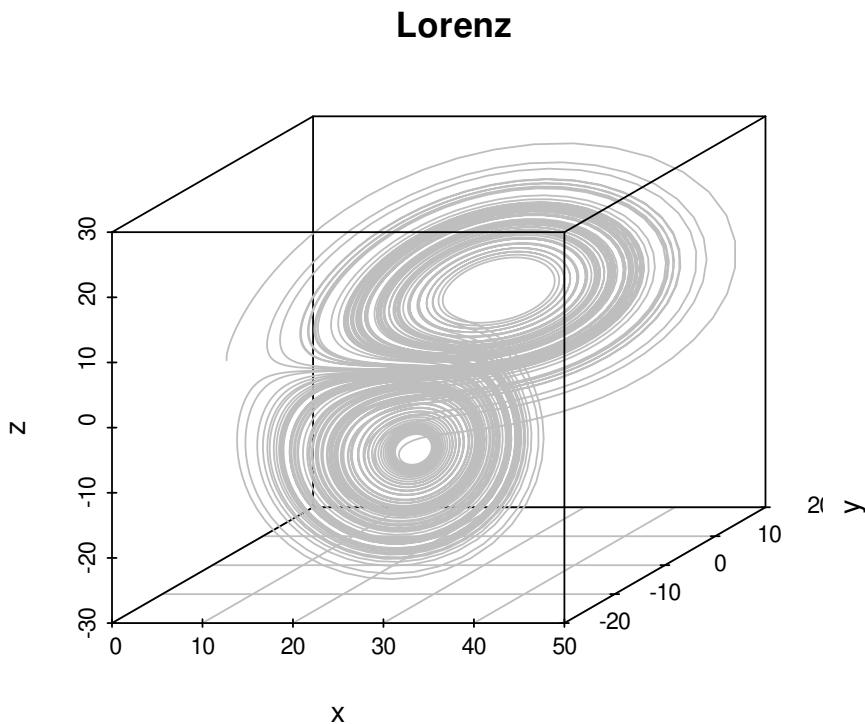


图 33.15: 洛伦兹曲线

### 33.10.2 偏微分方程

ReacTran 的几个关键函数介绍



## 一维热传导方程

$$\left\{ \begin{array}{l} \frac{\partial y}{\partial t} = D \frac{\partial^2 y}{\partial x^2} \end{array} \right.$$

参数  $D = 0.01$ , 边界条件  $y_{t,x=0} = 0, y_{t,x=1} = 1$ , 初始条件  $y_{t=0,x} = \sin(\pi x)$ 。

```
library(shape)
persp(volcano,
 theta = 30, phi = 20,
 r = 50, d = 0.1, expand = 0.5, ltheta = 90, lphi = 180,
 shade = 0.1, ticktype = "detailed", nticks = 5, box = TRUE,
 col = drapecol(volcano, col = terrain.colors(100)),
 xlab = "X", ylab = "Y", zlab = "Z", border = "transparent",
 main = "Topographic Information \n on Auckland's Maunga Whau Volcano"
)

library(ReacTran)

N <- 100
xgrid <- setup.grid.1D(x.up = 0, x.down = 1, N = N)
x <- xgrid$x.mid
D.coeff <- 0.01

Diffusion <- function(t, Y, parms) {
 tran <- tran.1D(
 C = Y, C.up = 0, C.down = 1,
 D = D.coeff, dx = xgrid
)
 list(
 dY = tran$dC,
 flux.up = tran$flux.up,
 flux.down = tran$flux.down
)
}
yini <- sin(pi * x)
times <- seq(from = 0, to = 5, by = 0.01)
out <- ode.1D(
 y = yini, times = times, func = Diffusion,
```

### Topographic Information on Auckland's Maunga Whau Volcano

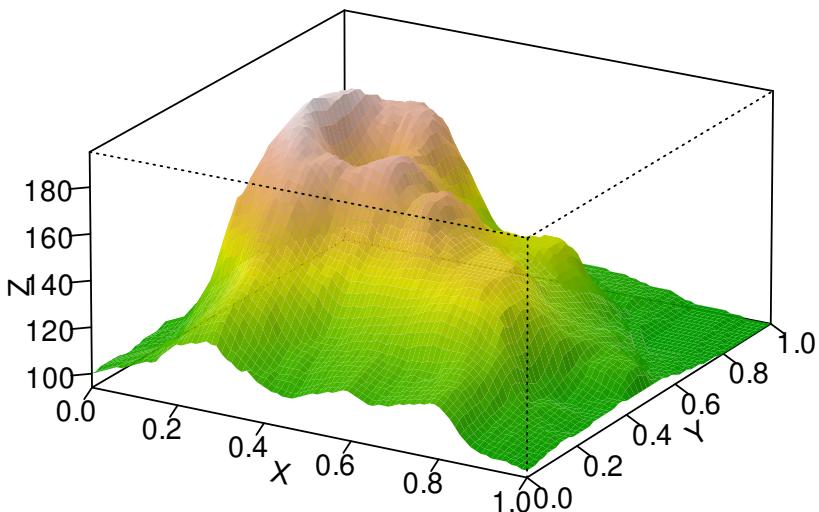


图 33.16: Auckland Maunga Whau 火山地形图  $10m \times 10m$ 。火山的实况地形图  
[https://en.wikipedia.org/wiki/Maungawhau\\_-\\_Mount\\_Eden](https://en.wikipedia.org/wiki/Maungawhau_-_Mount_Eden)。

```
parms = NULL, dimens = N
)

image(out,
 grid = xgrid$x.mid, xlab = "times",
 ylab = "Distance", main = "PDE", add.contour = TRUE
)
```

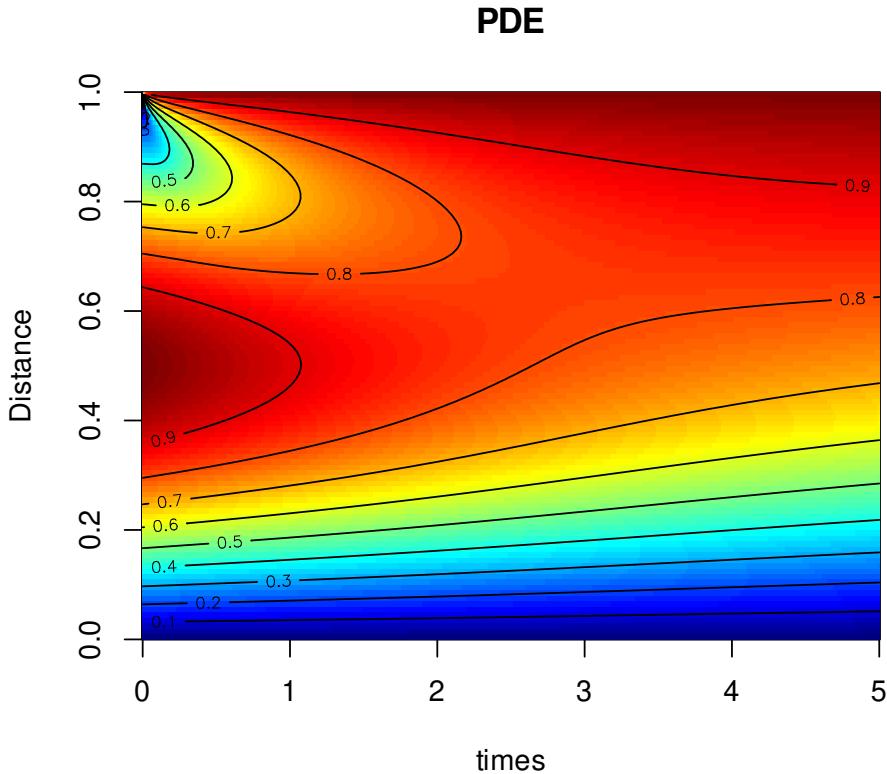


图 33.17: 一维热传导方程的数值解热力图

二维拉普拉斯方程

$$\left\{ \begin{array}{l} \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \end{array} \right.$$

边界条件

数值  
方法  
插值  
拟合

(C)

它有解析解

$$\begin{cases} u_{x=0,y} = u_{x=1,y} = 0 \\ \frac{\partial u_{x,y=0}}{\partial y} = 0 \\ \frac{\partial u_{x,y=1}}{\partial y} = \pi \sinh(\pi) \sin(\pi x) \end{cases}$$

$$u(x, y) = \sin(\pi x) \cosh(\pi y)$$

其中  $x \in [0, 1], y \in [0, 1]$

```
fn <- function(x, y) {
 sin(pi * x) * cosh(pi * y)
}

x <- seq(0, 1, length.out = 101)
y <- seq(0, 1, length.out = 101)
z <- outer(x, y, fn)

image(z, col = terrain.colors(20))
contour(z, method = "flat", add = TRUE, lty = 1)
```

```
persp(z,
 theta = 30, phi = 20,
 r = 50, d = 0.1, expand = 0.5, ltheta = 90, lphi = 180,
 shade = 0.1, ticktype = "detailed", nticks = 5, box = TRUE,
 col = drapecol(z, col = terrain.colors(20)),
 border = "transparent",
 xlab = "X", ylab = "Y", zlab = "Z",
 main = "")
```

求解 PDE

```
dx <- 0.2
xgrid <- setup.grid.1D(-100, 100, dx.1 = dx)
x <- xgrid$x.mid
N <- xgrid$N

uini <- exp(-0.05 * x^2)
vini <- rep(0, N)
```

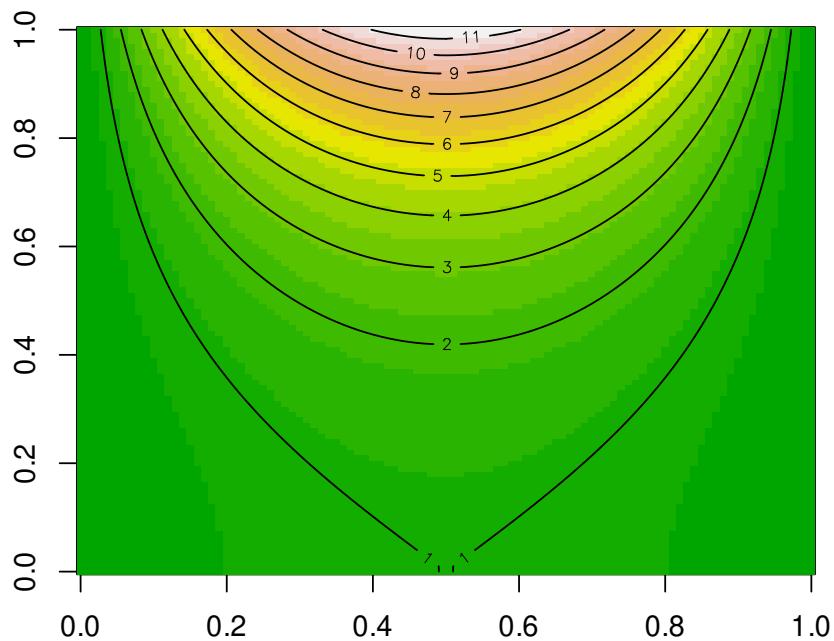


图 33.18: 解析解的二维图像

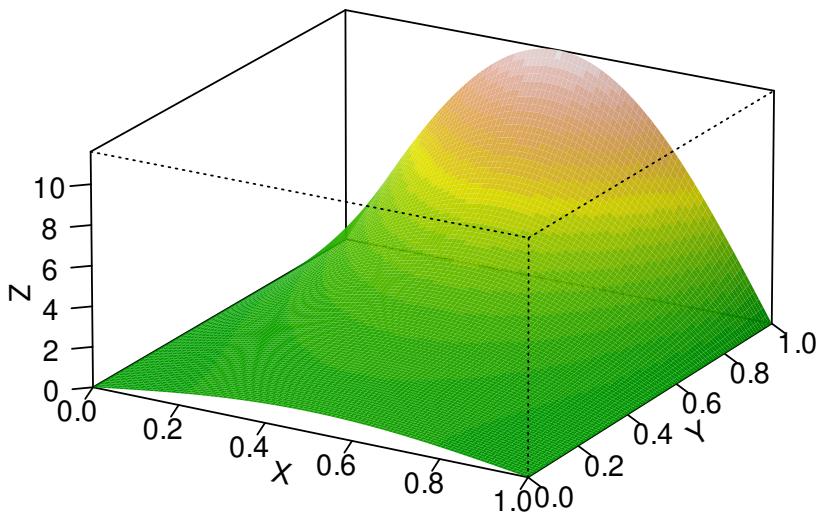


图 33.19: 解析解的三维透视图像



```
yini <- c(uini, vini)
times <- seq(from = 0, to = 50, by = 1)

wave <- function(t, y, parms) {
 u1 <- y[1:N]
 u2 <- y[-(1:N)]
 du1 <- u2
 du2 <- tran.1D(C = u1, C.up = 0, C.down = 0, D = 1, dx = xgrid)$dC
 return(list(c(du1, du2)))
}

out <- ode.1D(
 func = wave, y = yini, times = times, parms = NULL,
 nspec = 2, method = "ode45", dimens = N, names = c("u", "v")
)
```

### 33.10.3 延迟微分方程

```
library(PBSddesolve) # DAE 延迟微分方程
```

**PBSddesolve** [Couture-Beil et al., 2019] **PBSmodelling PBSmapping**

**nlmeODE** 通过微分方程整合用于混合效应模型的 **odesolve** 和 **nlme** 包。

### 33.10.4 随机微分方程

**Sim.DiffProc**

```
library(Sim.DiffProc)
```

种群 ODE 建模，

**nlmixr** 借助 **RxODE** 求解基于常微分方程的非线性混合效应模型



## 33.11 运行环境

```
sessionInfo()

R version 4.1.0 (2021-05-18)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 20.04.2 LTS
##
Matrix products: default
BLAS: /usr/lib/x86_64-linux-gnublas/libblas.so.3.9.0
LAPACK: /usr/lib/x86_64-linux-gnulapack/liblapack.so.3.9.0
##
locale:
[1] LC_CTYPE=en_US.UTF-8 LC_NUMERIC=C
[3] LC_TIME=en_US.UTF-8 LC_COLLATE=en_US.UTF-8
[5] LC_MONETARY=en_US.UTF-8 LC_MESSAGES=en_US.UTF-8
[7] LC_PAPER=en_US.UTF-8 LC_NAME=C
[9] LC_ADDRESS=C LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
attached base packages:
[1] stats graphics grDevices utils datasets methods base
##
other attached packages:
[1] quadprog_1.5-8 kableExtra_1.3.4
[3] tibble_3.1.3 Sim.DiffProc_4.8
[5] nlmeODE_1.1 nlme_3.1-152
[7] PBSddesolve_1.12.6 ReacTran_1.4.3.1
[9] shape_1.4.6 scatterplot3d_0.3-41
[11] deSolve_1.28 BB_2019.10-1
[13] rootSolve_1.8.2.2 kernlab_0.9-29
[15] lattice_0.20-44 ROI.plugin.quadprog_1.0-0
[17] ROI.plugin.lpsolve_1.0-1 ROI.plugin.nloptr_1.0-0
[19] ROI.plugin.alabama_1.0-0 ROI_1.0-0
[21] lpSolve_5.6.15
```



```
loaded via a namespace (and not attached):
[1] svglite_2.0.0 digest_0.6.27 utf8_1.2.2
[4] slam_0.1-48 R6_2.5.0 alabama_2015.3-1
[7] evaluate_0.14 httr_1.4.2 pillar_1.6.2
[10] rlang_0.4.11 rstudioapi_0.13 nloptr_1.2.2.2
[13] rmarkdown_2.9 webshot_0.5.2 stringr_1.4.0
[16] munsell_0.5.0 compiler_4.1.0 numDeriv_2016.8-1.1
[19] Deriv_4.1.3 xfun_0.24 pkgconfig_2.0.3
[22] systemfonts_1.0.2 htmltools_0.5.1.1 bookdown_0.22
[25] viridisLite_0.4.0 fansi_0.5.0 crayon_1.4.1
[28] MASS_7.3-54 grid_4.1.0 lifecycle_1.0.0
[31] registry_0.5-1 magrittr_2.0.1 scales_1.1.1
[34] stringi_1.7.3 xml2_1.3.2 ellipsis_0.3.2
[37] vctrs_0.3.8 lpSolveAPI_5.5.2.0-17.7 tools_4.1.0
[40] glue_1.4.2 parallel_4.1.0 yaml_2.2.1
[43] colorspace_2.0-2 rvest_1.0.1 knitr_1.33
```

## 附录 A 命令行操作

Bash 文件查找、查看（内容、大小）、移动（重命名）、删除、创建、修改权限

Linux 命令行工具是非常强大的，命令行中的数据科学 <https://www.datascienceatthecommandline.com/>，Linux 命令行 <https://github.com/jaywcjlove/linux-command>

`optparse`、`docopt`、`littler` 包提供了很多便捷的命令行工具，`sys`、`fs` 在 R 中运行操作系统命令

如表A.1所示，总结了 R 和 Shell 命令的等价表示，下面以 `list.files()` 和 `ls` 为例，介绍其等价的内容

表 A.1: R 和 Shell 命令的等价表示<sup>1</sup>

	R	Shell
查看文件	<code>list.files()</code>	<code>ls</code>
查看目录	<code>list.dirs()</code>	<code>dir</code>
目录层次	<code>fs::dir_tree()</code>	<code>tree</code>

### A.1 查看文件

`ls`/`mkdir`/`mv`/`du`

查看文件

```
```bash
ls -a
```

¹CentOS 系统默认没有安装 `tree` 软件，需要先安装才能使用此命令 `sudo dnf install -y tree`

...

列出目录下所有文件

```
```bash
ls -1
````
```

一行显示一个文件或文件夹

```
```bash
ls -l
````
```

按从 aA-zZ 的顺序列出所有文件以及所属权限

```
```bash
ls -rl
````
```

相比于 `ls -l` 文件是逆序排列

```
```bash
ls -lh
````
```

列出文件或文件夹（不包含子文件夹）的大小

```
```bash
ls -ld
````
```

列出当前目录本身，而不是其所包含的内容



A.2 创建文件夹

```
```bash
mkdir images
```

```

创建文件用 `touch` 如 `touch .Rprofile`

```
```bash
删除文件夹及子文件夹, 递归删除
rm -rf images/
删除文件
rm .Rprofile
```

```

A.3 移动文件

在当前目录下

```
```bash
移动文件夹 images 下的所有文件到 figures 文件夹下
mv images/* figures/
images 文件夹移动到 figures 文件夹下
mv images/ figures/
移动特定的文件
mv images/*.png figures/
```

```

同一目录下有两个文件 `R-3.5.1.tar.gz` 未下载完整 和 `R-3.5.1.tar.gz.1` 完全下载

```
```bash
删除 R-3.5.1.tar.gz
rm R-3.5.1.tar.gz
重命名 R-3.5.1.tar.gz.1
mv R-3.5.1.tar.gz.1 R-3.5.1.tar.gz
```

```

...

A.4 查看文件大小

当前目录下各文件夹的大小，`-h` 表示人类（相对于机器来说）可读的方式显示，如 Kb、Mb。

```
```bash
```

```
du -h -d 1 ./
```

```
...
```

```
```bash
```

```
# 对当前目录下的文件/夹 按大小排序
```

```
du -sh * | sort -nr
```

```
...
```

A.5 终端模拟器

oh-my-zsh 是 Z Shell 扩展，开发在 Github 上 <https://github.com/ohmyzsh/ohmyzsh>。

zsh 相比于 bash，在语法高亮、自动补全等方面有优势

```
sudo dnf install -y zsh
```

```
sh -c "$(curl -fsSL https://raw.github.com/ohmyzsh/ohmyzsh/master/tools/install.sh)"
```

RStudio 集成的终端支持 Zsh，操作路径 Tools -> Global Options -> Terminal，见图 A.1

A.6 压缩和解压缩

最常见的压缩文件格式有 .tar、.tar.gz、.tar.bz2、.zip 和 .rar，分别对应于 Tar <https://www.gnu.org/software/tar/>、Gzip <https://www.gzip.org/>、Bzip2 <https://www.bzip.org/>、UnZip/Zip <http://www.info-zip.org> 和 WinRAR <https://www.rarlab.com/>。Tar 提供了基本的打包和解包工具，Gzip 和 Bzip2 在 Tar 打包的基础上提供了压缩功能，UnZip/Zip 是兼容 Windows 原生压缩/解压

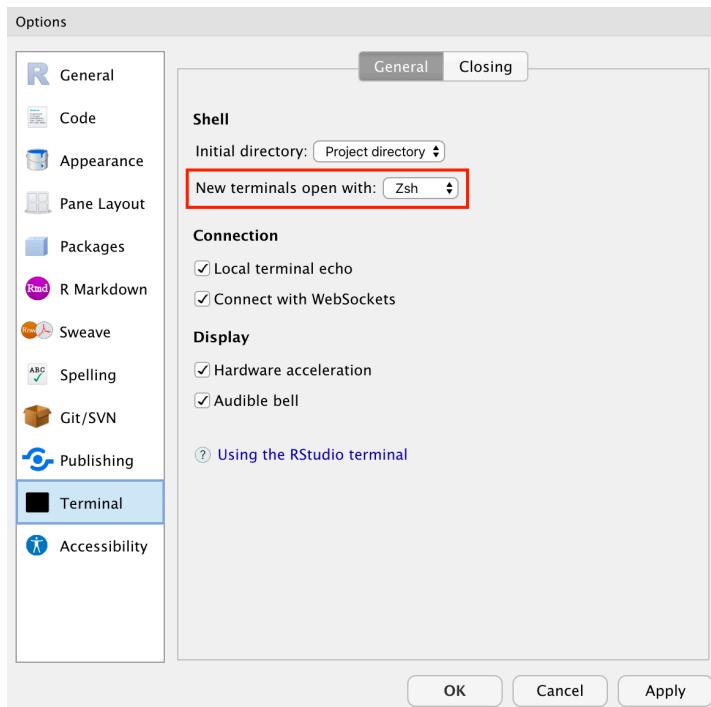


图 A.1: RStudio IDE 集成的 Zsh 终端模拟器

缩功能的程序，WinRAR 是广泛流行于 Windows 系统的压缩/解压缩收费软件，除了 WinRAR，其它都是免费甚至开源软件。下面以 .tar.gz 和.tar.bz2 两种格式的压缩文件为例，介绍文件压缩和解压缩的操作，其它文件格式的操作类似²。WinRAR <https://www.rarlab.com/> 是收费的压缩和解压缩工具，也支持 Linux 和 macOS 系统，鉴于它是收费软件，这里就不多展开介绍了，详情请见官网。

```
sudo dnf install -y tar gzip zip unz sudo dnf install -y bzip2
# 将目录 ~/tmp 压缩成文件 filename.tu # 将目录 ~/tmp 压缩成文件 filename.tar.bz2
tar -czf **.tar.gz ~/tmp           tar -cjf filename.tar.bz2 ~/tmp
# 将文件 filename.tar.gz 解压到目录 . # 将文件 filename.tar.bz2 解压到目录 ~/tmp
tar -xzf **.tar.gz -C ~/tmp       tar -xjf filename.tar.bz2 -C ~/tmp
解压不带 tar 的.gz 文件，比如 tex.eps.gz 解压后变成 tex.eps
gzip filename.gz -d ~/tmp
```

A.7 从源码安装 R

从源码编译 R 的需求大概有以下几点：

1. 爱折腾的极客：玩配置，学习 make 相关工具和 Linux 世界的依赖
2. 追求性能：如 LFS 支持 和 Intel MKL 加速
3. 环境限制：CentOS 或者红帽系统，自带的 R 版本比较落后

```
./configure --prefix=/opt/R/R-devel \
--enable-R-shlib --enable-byte-compiled-packages \
--enable-BLAS-shlib --enable-memory-profiling --with-blas="-lopenblas"

R is now configured for x86_64-pc-linux-gnu

Source directory: .
Installation directory: /opt/R/R-devel

C compiler: gcc -g -O2
Fortran fixed-form compiler: gfortran -fno-optimize-sibling-calls -g -O2

Default C++ compiler: g++ -std=gnu++11 -g -O2
C++14 compiler: g++ -std=gnu++14 -g -O2
```

²zip 格式的文件需要额外安装 zip 和 unzip 两款软件实现压缩和解压缩。



```
C++17 compiler:           g++ -std=gnu++17 -g -O2
C++20 compiler:           g++ -std=gnu++2a -g -O2
Fortran free-form compiler: gfortran -fno-optimize-sibling-calls -g -O2
Obj-C compiler:

Interfaces supported:      X11, tcltk
External libraries:         pcre2, readline, BLAS(OpenBLAS), curl
Additional capabilities:   PNG, JPEG, TIFF, NLS, cairo, ICU
Options enabled:           shared R library, shared BLAS, R profiling, memory profiling

Capabilities skipped:
Options not enabled:

Recommended packages:      yes
```

配置好了以后，可以编译安装了

```
make && sudo make install
```

flexiblas 支持多种 BLAS 库自由切换

A.8 安装软件

本书在后续章节中陆续用到新的 R 包，其安装过程不会在正文中呈现，下面以在 CentOS 8 上安装 **sf** 包为例介绍。首先需要安装一些系统依赖，具体安装哪些依赖参见 **sf** 包开发站点 <https://github.com/r-spatial/sf>。

```
sudo dnf config-manager --set-disabled PowerTools # openblas-devel
sudo dnf install -y sqlite-devel gdal-devel \
proj-devel geos-devel udunits2-devel
```

然后，在 R 命令行窗口中，执行安装命令：

```
install.packages('sf')
```

至此，安装完成。如遇本地未安装的新 R 包，可从其官方文档中找寻安装方式。如果你完全不知道自己应该安装哪些，考虑把下面的依赖都安装上

```
sudo dnf install -y \
# magick
```

```
ImageMagick-c++-devel \
# pdf tools
poppler-cpp-devel \
# gifski
cargo
```

软件包管理器架构图，各个命令分别担负什么样的功能，每个命令学习的一般路径是什么，而不是详细介绍每个命令、每个参数的使用，只需给出一个命令的完整使用即可，其余给出一个查询命令帮助手册

```
dnf copr
dnf config-manager
```

A.9 安装 R 包

Iñaki Ucar 开发的 [cran2copr](#) 项目实现在 Fedora 上安装预编译好的二进制 R 包，项目的类似 Debian 平台上的 [cran2deb](#)

1. [devtools](#) 是开发 R 包的常用工具，同时具有很重的依赖，请看

```
tools::package_dependencies('devtools', recursive = TRUE)

## $devtools
## [1] "usethis"      "callr"        "cli"          "desc"         "ellipsis"
## [6] "fs"           "httr"         "lifecycle"    "memoise"      "pkgbuild"
## [11] "pkgload"       "rcmdcheck"   "remotes"     "rlang"        "roxygen2"
## [16] "rstudioapi"   "rversions"    "sessioninfo" "stats"        "testthat"
## [21] "tools"         "utils"        "withr"        "processx"    "R6"
## [26] "glue"          "crayon"       "rprojroot"   "methods"     "curl"
## [31] "jsonlite"      "mime"         "openssl"     "cachem"      "prettyunits"
## [36] "digest"        "xopen"        "brew"        "commonmark"  "knitr"
## [41] "purrr"         "Rcpp"         "stringi"     "stringr"     "xml2"
## [46] "brio"          "evaluate"    "magrittr"    "praise"      "ps"
## [51] "waldo"         "clipr"        "gert"        "gh"          "rappdirs"
## [56] "whisker"       "yaml"        "graphics"    "grDevices"   "fastmap"
## [61] "askpass"       "credentials" "sys"        "zip"        "gitcreds"
## [66] "ini"           "highr"       "markdown"   "xfun"        "diffobj"
## [71] "fansi"         "rematch2"   "tibble"     "pillar"     "pkgconfig"
```



```
## [76] "vctrs"          "utf8"
```

其中，依赖关系见表 A.2

表 A.2: devtools 的系统依赖

	curl	git2r	openssl
Ubuntu	libcurl-dev ³	libgit2-dev	libssl-dev
CentOS	libcurl-devel	libgit2-devel	openssl-devel

1. `sf` 是处理空间数据的常用工具

```
tools::package_dependencies('sf', recursive = TRUE)
```

```
## $sf
## [1] "methods"    "classInt"    "DBI"        "graphics"   "grDevices"
## [6] "grid"        "magrittr"    "Rcpp"       "s2"         "stats"
## [11] "tools"       "units"       "utils"      "e1071"     "class"
## [16] "KernSmooth" "wk"         "MASS"      "proxy"     "cpp11"
```

其主要的系统依赖分别是 GEOS 3.5.1, GDAL 2.2.2, PROJ 4.9.2

```
sudo add-apt-repository -y ppa:ubuntugis/ubuntugis-unstable
sudo apt-get update
sudo apt-get install -y libudunits2-dev libgdal-dev libgeos-dev libproj-dev
```

这样也同时解决了 `udunits2`、`rgdal` 和 `rgeos` 等 3 个 R 包的系统依赖，其中 `udunits2` 使用如下命令安装

```
install.packages(' udunits2' , configure.args = '--with-udunits2-include=/usr/include/udunits2')
```

2. 图形设备支持 cairo png jpeg tiff

```
sudo apt-get install -y libcairo2-dev libjpeg-dev libpng-dev libtiff-dev
```

3. 图像处理 `imager` 和 `magick`

```
sudo yum install fftw-devel # CentOS
sudo apt-get install libfftw3-dev # Ubuntu
```

在 Ubuntu 系统上安装最新的 `libmagick++-dev` 库

³`libcurl-dev` 是一个虚包 virtual package，由 `libcurl4-openssl-dev` 或 `libcurl4-nss-dev` 或 `libcurl4-gnutls-dev` 实际提供，选择其中一个安装即可。



```
sudo add-apt-repository -y ppa:opencpu/imagemagick  
sudo apt-get update  
sudo apt-get install -y libmagick++-dev
```

在 CentOS 系统上

```
sudo yum install -y ImageMagick-c++-devel
```

然后安装 R 包 `install.packages(c('imager', 'magick'))`

4. `rgl` 是绘制真三维图形的重量级 R 包

```
sudo apt-get install libcgal-dev libglu1-mesa-dev libx11-dev # Ubuntu  
sudo yum install mesa-libGLU mesa-libGLU-devel # CentOS
```

然后安装 R 包

```
install.packages('rgl')
```

在 Ubuntu 系统上还可以这样安装

```
sudo add-apt-repository ppa:marutter/rrutter3.5  
sudo apt-get update  
sudo apt-get install r-cran-rgl
```

5. `rJava` 是 Java 语言和 R 语言之间实现通信交流的桥梁

```
sudo apt-get install -y default-jdk  
sudo R CMD javareconf
```

然后安装 `rJava` 包 `install.packages('rJava')`

6. `igraph` 是网络数据分析的必备 R 包，为了发挥其最大性能，需要安装三个系统依赖

```
sudo apt-get install -y libgmp-dev libxml2-dev libglpk-dev
```

然后安装 R 包

```
install.packages('igraph')
```

7. `gpuR` 是基于 GPU 进行矩阵计算的扩展包，依赖 `RcppEigen` 确保安装 OpenCL 和 `RViennaCL` 或者安装 Nvidia 驱动和 CUDA，使用 `gpuRcuda` 和 `gputools` 扩展包，下面安装指导来自其 Wiki



```
# Install OpenCL headers
sudo apt-get install opencl-headers opencv-dev

# Install NVIDIA Drivers and CUDA
sudo add-apt-repository -y ppa:xorg-edgers/ppa
sudo apt-get update
sudo apt-get install nvidia-346 nvidia-settings
```

8. **nloptr** 是 **NLOpt** 的 R 语言接口，首先安装 NLOpt 程序库 `sudo apt-get install libnlopt-dev` 然后安装 R 包 `install.packages('nloptr')`，**nloptr** 被 700+ R 包依赖，如 **lme4**, **spaMM**, **glmmTMB**, **rstanarm** 等。

9. **Rmpfr**

```
sudo apt-get install libmpfr-dev

install.packages('Rmpfr')
```

10. **geojson**

```
sudo yum install jq-devel protobuf-devel

install.packages(c('geojson', 'geojsonio', 'jqr', 'protolite'))
```

11. **lgcp**

```
sudo yum install bwidget

install.packages(c('rpanel', 'lgcp'))
```

12. **ijtiff**

```
sudo yum install jbigkit-devel

install.packages('ijtiff')
```

13. **webshot** 包用于截图

```
sudo apt install phantomjs

install.packages('webshot')
```

14. **gifski** 包合成 GIF 动图

```
sudo apt-get install cargo  
install.packages('gifski')
```

A.10 软件包管理器

A.10.1 dnf

1. 清理升级后的 CentOS 8 系统内核

查找系统安装的内核

```
rpm -qa | sort | grep kernel
```

```
kernel-4.18.0-147.8.1.el8_1.x86_64  
kernel-4.18.0-193.6.3.el8_2.x86_64  
kernel-core-4.18.0-147.8.1.el8_1.x86_64  
kernel-core-4.18.0-193.6.3.el8_2.x86_64  
kernel-headers-4.18.0-193.6.3.el8_2.x86_64  
kernel-modules-4.18.0-147.8.1.el8_1.x86_64  
kernel-modules-4.18.0-193.6.3.el8_2.x86_64  
kernel-tools-4.18.0-193.6.3.el8_2.x86_64  
kernel-tools-libs-4.18.0-193.6.3.el8_2.x86_64
```

仅保留一个版本的内核，其它旧的内核都删除掉

```
sudo dnf remove $(dnf repoquery --installonly --latest-limit=1 -q)
```

模块依赖问题

问题 1: conflicting requests

- nothing provides module(perl:5.26) needed by module perl-DBD-MySQL:4.046:80100201

问题 2: conflicting requests

- nothing provides module(perl:5.26) needed by module perl-DBI:1.641:80100201911132

问题 3: conflicting requests

- nothing provides module(perl:5.26) needed by module perl-YAML:1.24:80100201911140

依赖关系解决。

```
=====
```

软件包

架构

版本

仓库



事务概要

移除 3 软件包

将会释放空间: 78 M

确定吗? [y/N]: y

运行事务检查

事务检查成功。

运行事务测试

事务测试成功。

运行事务

准备中 :

1/1

删除 : kernel-4.18.0-147.8.1.el8_1.x86_64

运行脚本: kernel-4.18.0-147.8.1.el8_1.x86_64

删除 : kernel-modules-4.18.0-147.8.1.el8_1.x86_64

运行脚本: kernel-modules-4.18.0-147.8.1.el8_1.x86_64

运行脚本: kernel-core-4.18.0-147.8.1.el8_1.x86_64

删除 : kernel-core-4.18.0-147.8.1.el8_1.x86_64

运行脚本: kernel-core-4.18.0-147.8.1.el8_1.x86_64

验证 : kernel-4.18.0-147.8.1.el8_1.x86_64

验证 : kernel-core-4.18.0-147.8.1.el8_1.x86_64

验证 : kernel-modules-4.18.0-147.8.1.el8_1.x86_64

已移除:

kernel-4.18.0-147.8.1.el8_1.x86_64

kernel-core-4.18.0-147.8.1.el8_1.x86_64

kernel-modules-4.18.0-147.8.1.el8_1.x86_64

完毕!

[解决上述模块依赖问题的办法](#) 是重置三个 Perl 模块



```
sudo dnf module reset perl-DBD-MySQL perl-YAML perl-DBI
```

依赖关系解决。

软件包	架构	版本	仓库
-----	----	----	----

重置模块：

```
perl-DBD-MySQL
```

```
perl-DBI
```

```
perl-YAML
```

事务概要

确定吗? [y/N]: y

完毕!

A.10.2 apt

添加或删除 PPA (Personal Package Archive)，比如在 Ubuntu 20.04 及之前的版本上安装新版 Inkscape

```
sudo add-apt-repository ppa:inkscape.dev/stable  
sudo add-apt-repository --remove ppa:inkscape.dev/stable
```

```
sudo apt-get install build-essential # 修复依赖问题  
sudo apt update # 更新资源列表  
sudo apt-get upgrade # 更新软件包  
sudo apt-get autoclean # 删除已卸的软件的备份  
sudo apt-get clean # 删除已装或已卸的软件的备份  
sudo apt-get autoremove --purge * # 推荐卸载软件的方式  
apt-get list --upgradable # 列出可升级的包
```

找到并删除旧的内核

```
dpkg --list | grep linux-image  
sudo apt-get purge linux-image-3.19.0-[18,20,21,25]  
sudo update-grub2
```



```
# 搜索
apt-cache search octave | grep octave
# 查询
apt show octave
# 安装
sudo apt install octave

sudo apt-get install lsb-core
lsb_release -a

adduser cloud2016 # 添加用户
passwd cloud2016 # 用户密码设为 cloud
whereis sudoers # 查找文件位置
chmod -v u+w /etc/sudoers # 给文件 sudoers 添加写权限
vim /etc/sudoers # 添加 cloud2016 管理员权限
chmod -v u-w /etc/sudoers # 收回权限
```

安装确认 openssh-server 服务

```
sudo apt install openssh-server
sudo /etc/init.d/ssh start
ps -aux | grep ssh
```

附录 B 其它软件

I think, therefore I R.

— William B. King¹

B.1 文本编辑器

代码文件也是纯文本，RStudio 集成了编辑器，支持语法高亮。Windows 系统上优秀的代码编辑器有 Notepad++ 非常轻量。Markdown 文本编辑器我们推荐 Typora 编辑器，它是跨平台的，下面以 Ubuntu 环境为例，介绍安装和使用过程：

```
# or run:  
# sudo apt-key adv --keyserver keyserver.ubuntu.com --recv-keys BA300B7755AFCFAE  
wget -qO - https://typora.io/linux/public-key.asc | sudo apt-key add -  
  
# add Typora's repository  
sudo add-apt-repository 'deb https://typora.io/linux ./'  
sudo apt-get update  
  
# install typora  
sudo apt-get install typora
```

设置中文环境，并且将主题风格样式配置为 Vue，见图B.1（右），Vue 主题可从 Typora 官网下载 <https://theme.typora.io/theme/Vue/>。

1. Atom 编辑器 <https://atom.io/>

¹<https://www2.coastal.edu/kingw/statistics/R-tutorials/>

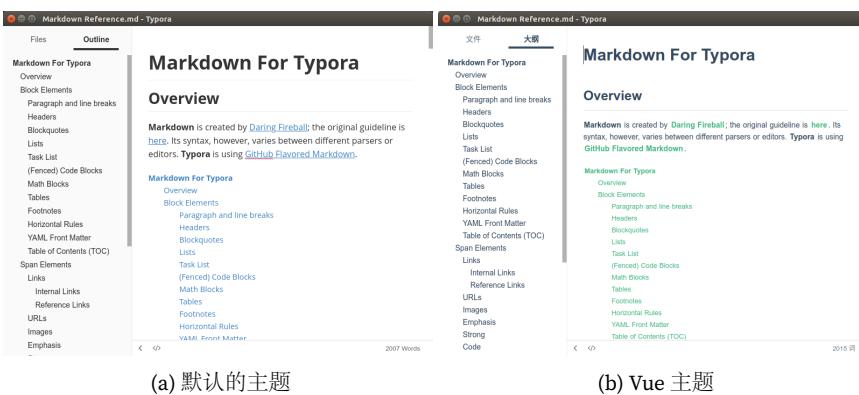


图 B.1: Typora 主题

```
sudo add-apt-repository ppa:webupd8team/atom
sudo apt-get update
sudo apt-get install atom
```

1. Code 编辑器微软出品 <https://code.visualstudio.com/>
2. Notepad++ 开源的 Windows 平台上的编辑器 <https://notepad-plus-plus.org/>
3. VI & VIM 开源的跨平台编辑器
4. Atom 和 Code 有商业公司支持的开源免费的跨平台的编辑器
5. VI/VIM 和 Emacs 是跨平台的编辑器
6. Markdown 编辑器 + blogdown 记笔记
7. Typora Markdown 编辑器，支持自定义 CSS 样式

B.2 代码编辑器

VS Code, Sublime Text 和 Atom

B.3 集成开发环境

RStudio 公司的愿景，介绍 RStudio 开发环境提供的效率提升工具或功能

B.3.1 RStudio 桌面版

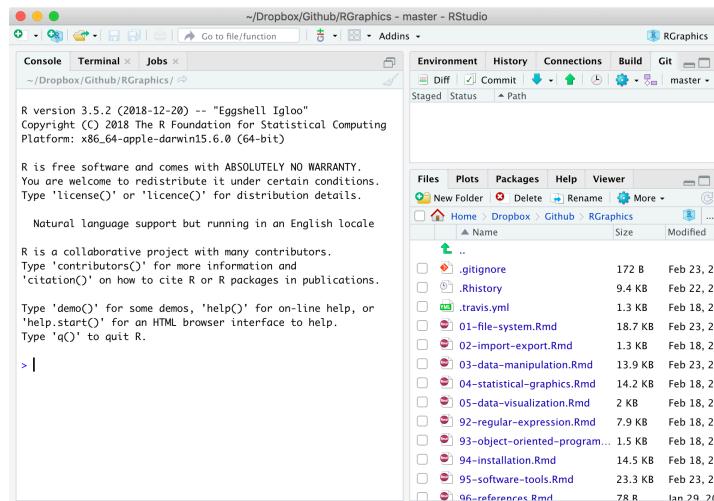


图 B.2: 开源桌面版 RStudio 集成开发环境

```
# mongolite
sudo dnf install -y openssl-devel cyrus-sasl-devel
# sodium
sudo dnf install -y libodium-devel
# rJava
R CMD javareconf

# https://github.com/s-u/rJava
# shinytest::installDependencies()
db <- rstudioapi::getRStudioPackageDependencies()

invisible(lapply(db$name, function(pkg) {
  if (system.file(package = pkg) == "") {
    install.packages(pkg)
  }
}))
```



rsthemes 主题

B.3.2 RStudio 服务器版

RStudio Server 开源服务器版可以放在虚拟机里或者容器里，RStudio 桌面版装在服务器上，服务器为 Ubuntu/CentOS/Windows 系统，然后本地是 Windows 系统，可以通过远程桌面连接服务器，使用 RStudio；



图 B.3: 虚拟机里的 RStudio

服务器上启动 Docker，运行 RStudio 镜像，本地通过桌面浏览器，如谷歌浏览器登陆连接。

1. 下载 RStudio IDE

我们从 RStudio 官网[下载](https://www.rstudio.com/products/rstudio/download/)开源桌面或服务器版本，服务器版本的使用介绍见[文档](#)，最常见的就是设置端口

```
wget https://download2.rstudio.org/rstudio-server-1.1.456-amd64.deb
sudo apt-get install gdebi
sudo gdebi rstudio-server-1.1.456-amd64.deb
```

2. 设置端口

在文件 `/etc/rstudio/rserver.conf` 下，设置

```
www-port=8181
```

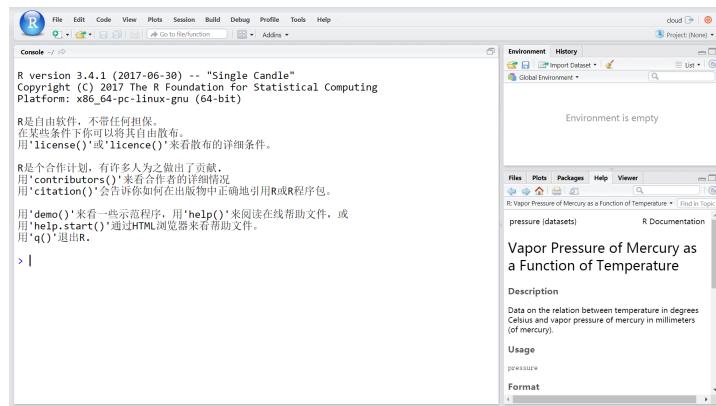


图 B.4: 容器里的 RStudio

注意：修改 `rserver.conf` 文件后需要重启才会生效

```
sudo rstudio-server stop
sudo rstudio-server start
```

接着获取机器的 IP 地址，如 192.168.141.3

```
ip addr

1: lo: <LOOPBACK,UP,LOWER_UP> mtu 65536 qdisc noqueue state UNKNOWN group default
    link/loopback 00:00:00:00:00:00 brd 00:00:00:00:00:00
    inet 127.0.0.1/8 scope host lo
        valid_lft forever preferred_lft forever
    inet6 ::1/128 scope host
        valid_lft forever preferred_lft forever

2: enp0s3: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc fq_codel state UP group default
    link/ether 08:00:27:59:c0:fb brd ff:ff:ff:ff:ff:ff
    inet 10.0.2.15/24 brd 10.0.2.255 scope global dynamic enp0s3
        valid_lft 83652sec preferred_lft 83652sec
    inet6 fe80::a00:27ff:fe59:c0fb/64 scope link
        valid_lft forever preferred_lft forever

3: enp0s8: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc fq_codel state UP group default
    link/ether 08:00:27:09:33:0d brd ff:ff:ff:ff:ff:ff
    inet 192.168.141.3/24 brd 192.168.141.255 scope global dynamic enp0s8
        valid_lft 547sec preferred_lft 547sec
    inet6 fe80::a00:27ff:fe09:330d/64 scope link
```



```
valid_lft forever preferred_lft forever
```

然后,就可以从本地浏览器登陆 RStudio 服务器版本,如<http://192.168.141.3:8181/>

提示

rstudio-server 已经收录在 Fedora 33+ 仓库中了, 详情见 <https://cran.r-project.org/bin/linux/fedora/>

授权问题 ERROR system error 13 (Permission denied) How to Disable SELinux Temporarily or Permanently

B.3.3 Shiny 服务器版

shiny 开源服务器版

B.3.4 Eclipse + StatET

Eclipse 配合 StatET 插件 <http://www.walware.de/goto/statet> 提供 R 语言的集成开发环境 <https://projects.eclipse.org/projects/science.statet>

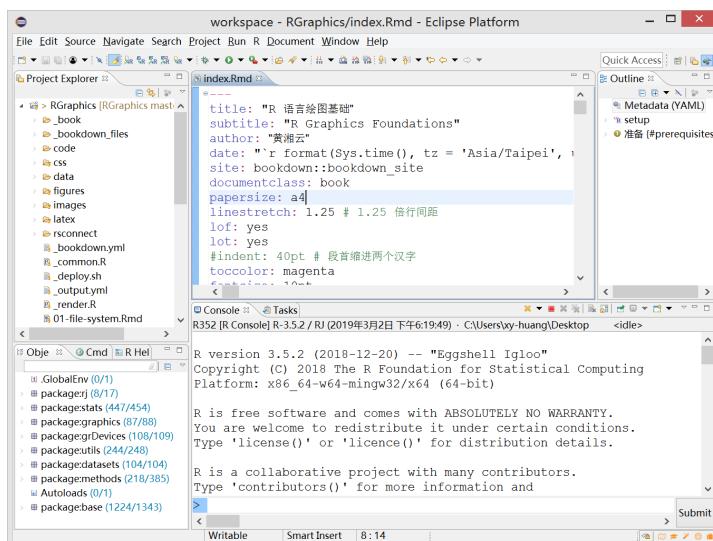


图 B.5: 基于 Eclipse 的 R 集成开发环境 StatET

StatET 基于 Eclipse 首次建立索引很慢，估计半小时到一个小时，添加新的 R 包后，每次启动 StatET 也会建立索引缓存，此外，Eclipse 开发环境占用内存比较多，配置 StatET 的过程如下

B.3.5 Emacs + ESS

Emacs 配合 ESS 插件 <https://ess.r-project.org/>

B.3.6 Nvim-R

Nvim-R 是一个基于 Vim 的集成开发环境 <https://github.com/jalvesaq/Nvim-R>

B.4 Git 版本控制

Git 操作

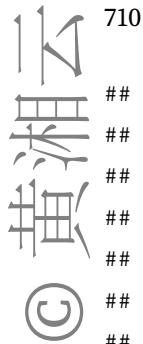
MacOS 上用 Homebrew 安装 [git-delta](#)

```
brew install git-delta
```

[gitdown](#)

只考虑 Ubuntu 18.04 环境下的三剑客 Git & Github & Gitlab

```
summary(git2r::repository())  
  
## Local:    devel /home/runner/work/masr/masr  
## Remote:   devel @ origin (https://github.com/XiangyunHuang/masr)  
## Head:     [4a4bc4f] 2021-07-19: update pandoc  
##  
## Branches:      1  
## Tags:         0  
## Commits:      5  
## Contributors: 1  
## Stashes:      0  
## Ignored files: 14  
## Untracked files: 80  
## Unstaged files: 0
```



```
## Staged files:      0
##
## Latest commits:
## [4a4bc4f] 2021-07-19: update pandoc
## [ec490cb] 2021-07-19: use https in references
## [d00acdd] 2021-07-18: build full pdf
## [9bc7292] 2021-07-18: fix WARNING when build pdf
## [a49be96] 2021-07-18: fix error when build pdf
```

仓库 [masr](#) 哪些人给我点赞加星了

```
library(gh)
my_repos <- gh("GET /repos/:owner/:repo/stargazers", owner = "xiangyunhuang", repo = "masr")
vapply(my_repos, "[[", "", "login")
```

[1]	"ddd1007"	"boltomli"	"JackieMium"	"AXGL"	"fyemath"
[6]	"rogerclarkgc"	"swsoyee"	"joegaotao"	"YTLogos"	"Accelerator086"
[11]	"yimingfish"	"gaospecial"	"shenxiangzhuang"	"shuaiwang88"	"LusiXie"
[16]	"llxlr"	"TingjieGuo"	"oiazt"	"XiaogangHe"	"xwydq"
[21]	"guohongwang1"	"yinandong"	"algony-tony"	"XiangyunHuang"	"perlatex"
[26]	"talegari"	"hao-shefer"	"zhouyisu"	"tsitong"	"liuyadong"



图 B.6: Git 代码版本管理

Jeroen Ooms 开发的 [gert](#) 包实现在 R 环境中操作 Git，我们可以从幻灯片 – [Gert: A minimal git client for R](#) 学习重点内容。

```
library(gert)
library(magrittr)
git_log(max = 10) %>%
  subset(subset = grepl("Yihui Xie", x = author), select = c("author", "message"))
```

提供了 `git_rm()`、`git_status()`、`git_add()` 和 `git_commit()` 等函数，其中包含



`git_reset()` 高级的 Git 操作。此外，还有 `git_branch_*`() 系列分支操作函数

B.4.1 安装配置

Ubuntu 16.04.5 默认安装的 Git 版本是 2.7.4，下面安装最新版本 Git 和配置自己的 GitHub 账户

1. 根据官网安装指导 <https://git-scm.com/download/linux>, 在 Ubuntu 14.04.5 和 Ubuntu 16.04.5 安装最新版 GIT

```
sudo add-apt-repository -y ppa:git-core/ppa  
sudo apt update && sudo apt install git
```

- ## 2. 配置账户

```
git config --global user.name "你的名字"
git config --global user.email "你的邮件地址"
touch .git-credentials
# 记住密码
echo "https://username:password@github.com" >> .git-credentials
git config --global credential.helper store
```

以 Fedora 为例 [安装 tig](#)，首先安装必要的依赖，然后从官网下载源码，编译安装，之后切到任意本地 Git 仓库下，输入 `tig` 就可以看到如图 B.7 所示的样子了。

```
sudo yum install readline-devel ncurses-devel asciidoc docbook-utils xsltproc
```

`tig` 主要用于查看 git 提交的历史日志。

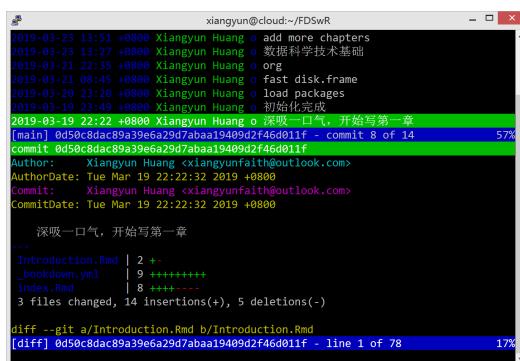


图 B.7: Git 日志查看器



B.4.2 追踪文件

```
git add .
```

提交新文件 (new) 和被修改 (modified) 文件，不包括被删除 (deleted) 文件

```
git add -u
```

提交被修改 (modified) 和被删除 (deleted) 文件，不包括新文件 (new)，`git add --update` 的缩写

```
git add -A
```

提交所有变化，`git add --all` 的缩写

```
git init  
git remote add origin https://github.com/XiangyunHuang/masr.git  
git add -A  
git commit -m "添加提交说明"  
git push -u origin master
```

往远程的空的 Github 仓库添加本地文件

B.4.3 合并上流

```
git clone --depth=5 https://github.com/XiangyunHuang/cosx.org.git  
git submodule update --init --recursive
```

查看远程分支

```
cd cosx.org  
git remote -v
```

```
origin https://github.com/XiangyunHuang/cosx.org.git (fetch)  
origin https://github.com/XiangyunHuang/cosx.org.git (push)
```

```
# 添加上流分支
```

```
git remote add upstream https://github.com/cosname/cosx.org.git  
# 查看远程分支  
git remote -v
```

```
origin https://github.com/XiangyunHuang/cosx.org.git (fetch)
```



```
origin https://github.com/XiangyunHuang/cosx.org.git (push)
upstream      https://github.com/cosname/cosx.org.git (fetch)
upstream      https://github.com/cosname/cosx.org.git (push)

# 获取上游 commit 并且合并到我的 master 分支
git fetch upstream
git merge upstream/master master
git push origin master
```

B.4.4 大文件支持

```
sudo apt install git-lfs
git lfs install
git lfs track "*.psd"
git add .gitattributes
git commit -m "track *.psd files using Git LFS"
git push origin master
```

这玩意迟早需要你购买存储空间，慎用

B.4.5 新建分支

```
git checkout -b stan      # 新建 stan 分支
git branch -v             # 查看本地分支 stan 前有个星号标记
git pull --rebase git@github.com:XiangyunHuang/cosx.org.git master
# 同步到远程分支 stan
git push --set-upstream origin stan
git push origin master:stan

git add .
git commit -m "balabala"
git push --set-upstream origin stan
```

本地新建仓库推送至远程分支



```
git remote add origin https://github.com/XiangyunHuang/notesdown.git  
git add .  
git commit -m "init cos-art"  
# 此时远程仓库 notesdown 还没有 cos-art 分支  
git push origin master:cos-art
```

位于 [Github Git Community Book 中译本](#)

B.4.6 创建 Github Pages 站点

基于 GitHub Pages 创建站点用于存放图片和数据

1. 在 Github 上创建一个空的仓库，命名为 uploads，没有 readme.md 和 LICENSE
2. 在本地创建目录 uploads
3. 切换到 uploads 目录下

```
git init  
git checkout -b gh-pages  
git remote add origin https://github.com/XiangyunHuang/uploads.git
```

添加图片或者数据，并且 git add 和 commit 后

```
git push --set-upstream origin gh-pages
```

这样仓库 uploads 只包含 gh-pages 分支，数据地址即为以日期为分割线

https://xiangyunhuang.github.io/uploads/data/eqList2018_05_18.xls

B.4.7 博客主题

初始化博客网站

```
git subtree add --squash --prefix=themes/hugo-lithium \  
git@github.com:yihui/hugo-lithium.git master
```

在 Github 创建新的空仓库，本地创建空的目录 xiangyun

```
cd xiangyun  
git init  
git remote add origin https://github.com/XiangyunHuang/xiangyun.git
```



```
git add .gitignore  
git commit -m 'upload'  
git push --set-upstream origin master
```

`git subtree` 将另外一个仓库收缩为当前仓库的一个目录，且只产生一条提交记录

```
# 子库分支  
git subtree add --squash --prefix=themes/hugo-xmag \  
-m "add hugo-xmag" git@github.com:yihui/hugo-xmag.git master  
# 或者子库分支  
git subtree add --squash --prefix=themes/hugo-xmag \  
-m "add hugo-xmag" https://github.com/yihui/hugo-xmag.git master  
  
# 移除 git subtree 添加的 hugo 主题  
git filter-branch --index-filter 'git rm --cached --ignore-unmatch -rf themes/hugo'
```

B.4.8 修改远程仓库的位置

有时候我们将自己的仓库转移给别人/组织，或者我们将远程仓库的名字改变了，这时候需要修改远程仓库的位置。比如最近我将博客仓库从 <https://github.com/XiangyunHuang/xiangyun> 转移到 <https://github.com/rbind/xiangyun>

转移前

```
git remote -v
```

```
origin https://github.com/XiangyunHuang/xiangyun.git (fetch)  
origin https://github.com/XiangyunHuang/xiangyun.git (push)
```

转移命令

```
git remote set-url origin https://github.com/rbind/xiangyun.git
```

转移后

```
git remote -v
```

```
origin https://github.com/rbind/xiangyun.git (fetch)  
origin https://github.com/rbind/xiangyun.git (push)
```



B.4.9 统计代码仓库的提交量

比如统计之都的主站仓库，提交量最大的 20 个人

```
git shortlog -sn | head -n 20
```

提交量	用户名
153	Dawei Lang
106	Yihui Xie
89	Beilei Bian
46	王佳
42	雷博文
39	Ryan Feng Lin
35	Xiangyun Huang
32	fanchao
32	闫晗
30	Lin Feng
28	Jiaao Yu
25	fyears
24	Yixuan Qiu
24	Miao YU
22	Yuxuan Li
22	qinwf
20	Alice敏
19	yanshi
18	Shuyi.Yang
13	黄湘云

B.4.10 账户共存

本节介绍如何使 Gitlab/Github 账户共存在一台机器上

如何生成 SSH 密钥见 Github 文档 – [使用 SSH 连接到 GitHub](#)。有了密钥之后只需在目录 `~/.ssh` 下创建一个配置文件 `config`

生成 SSH Key

```
ssh-keygen -t rsa -f ~/.ssh/id_rsa_github -C "name1@xxx1.com"  
ssh-keygen -t rsa -f ~/.ssh/id_rsa_gitlab -C "name2@xxx2.com"
```

将 GitHub/GitLab 公钥分别上传至服务器，然后创建配置文件



```
touch ~/.ssh/config
```

配置文件内容如下

```
#  
# Github  
#  
Host github.com // 个人的代码仓库服务器地址  
HostName github.com  
User XiangyunHuang  
IdentityFile ~/.ssh/id_rsa_github  
  
#  
# company  
#  
Host xx.xx.xx.xx //  
IdentityFile ~/.ssh/id_rsa_gitlab
```

配置成功，你会看到

```
ssh -T git@xx.xx.xx.xx
```

```
Welcome to GitLab, xiangyunhuang!
```

和

```
ssh -T git@github.com
```

```
Hi XiangyunHuang! You've successfully authenticated, but GitHub does not provide shel
```

B.4.11 回车换行

CR (Carriage Return) 表示回车，LF (Line Feed) 表示换行，Windows 下用回车加换行表示下一行，UNIX/Linux 采用换行符 (LF) 表示下一行，MAC OS 则采用回车符 (CR) 表示下一行

```
git config --global core.autocrlf false
```



B.4.12 子模块

- 添加子模块到目录 `templates/` 下

```
git submodule add git://github.com/jgm/pandoc-templates.git templates
```

- 移除子模块

<https://stackoverflow.com/questions/1260748/how-do-i-remove-a-submodule/>

B.4.13 克隆项目

```
git clone --depth=10 --branch=master --recursive \
git@github.com:XiangyunHuang/pandoc4everything.git
```

B.4.14 创建 PR

```
git pull --rebase git@github.com:yihui/xaringan.git master
# then force push to your master branch
```

参考 <https://github.com/yihui/xaringan/pull/107>

I don't recommend you to use your master branch for pull requests, because all commits will be squashed before merging, e.g. c2c2055 Then you will have some trouble with syncing your master branch with the master branch here (your choices are (1) delete your repo and fork again; or (2) force push; either option is not good). For pull requests, I recommend that you always use different branches for different pull requests.

B.4.15 修改 PR

之前一直有一个思想在阻止自己，就是别人的 `repo` 我是不能修改的，但是在这里，我拥有修改原始仓的权限，那么别人的复制品衍生的分支，我也有修改权限

```
git fetch origin refs/pull/771/head:patch-2
# 771 是 PR 对应的编号
```



```
git checkout patch-2

# 你的修改

git add -u
git commit -m "描述你的修改"

git remote add LalZzy https://github.com/LalZzy/cosx.org.git

git push --set-upstream LalZzy patch-2
```

整理自统计之都论坛的讨论 <https://d.cosx.org/d/420363>

1. GitHub/Git 小抄英文版 <https://www.runoob.com/manual/github-git-cheat-sheet.pdf>
2. GitHub/Git 小抄中文版 https://github.github.com/training-kit/downloads_zh_CN/github-git-cheat-sheet/
3. Github 秘籍 <https://github.com/tiimgreen/github-cheat-sheet/blob/master/README.zh-cn.md>
4. Git 简明指南 <https://rogerdudler.github.io/git-guide/index.zh.html>
5. Git 奇技淫巧 <https://github.com/521xueweihan/git-tips>
6. Git 官方书籍 <https://git-scm.com/book/zh/v2>
7. Git 时代的 VIM 不完全使用教程 <http://beiyuu.com/git-vim-tutorial>
8. 最佳搭档: 利用 SSH 及其配置文件节省你的生命 <https://liam.page/2017/09/12/rescue-your-life-with-SSH-config-file/>

B.5 Pandoc 文档处理

Pandoc 是一个万能文档转化器, 安装 pandoc, 下载网址 <https://github.com/jgm/pandoc/releases/latest>

```
sudo apt-get install gdebi-core
wget https://github.com/jgm/pandoc/releases/download/2.9.2/pandoc-2.9.2-1-amd64.deb
sudo chmod +x pandoc-2.9.2-1-amd64.deb
sudo gdebi pandoc-2.9.2-1-amd64.deb
```

rmarkdown 包裹了 Pandoc 工具, 使用 `rmarkdown::render()` 函数即可将 R Mark-



down 文档转化为 HTML、LaTeX 和 Markdown 等格式。

B.6 Calibre 书籍管理



Calibre 是一款电子书转化和管理软件，首先安装 calibre

```
sudo -v && wget -nv -O- https://download.calibre-ebook.com/linux-installer.sh | sudo sh
```

calibre 可以将 epub 格式电子书文档转化为 mobi 格式，bookdown 已经给这个工具穿上了一件马甲，用户只需调用 `bookdown::calibre()` 函数即可实现电子书格式的转换。

B.7 ImageMagick 图像处理

图像的各种操作，包括合成、转换、旋转等等

首先安装 ImageMagick 软件包中的 convert 程序

```
asy -f jpg test.asy
```

指定分辨率

```
convert -geometry 1000x3000 -density 300 -units PixelsPerInch example.eps example.png
```

这样不改变图像的像素数，只是给出一个每个像素应该显示多大的提示。

```
convert -quality 100 -density 300x300 filename.pdf filename.png
```

高质量大图，给定像素，转化 eps 格式图片，需要先安装 Ghostscript

```
convert -geometry 1000x3000 example.eps example.png
```

多页的 PDF 文件转化为多张 PNG 图片

```
convert -quality 100 -density 300x300 input.pdf output.png
```

将多页 PDF 文件合成为 GIF 动图

```
convert -delay 60 -density 300x300 -background white -alpha remove \
-dispose previous pdf-mobile.pdf -layers coalesce pdf-mobile.gif
```

B.8 OptiPNG 图片优化

OptiPNG 是一个非常好的图片压缩、优化工具

现在，我们设置 chunk 选项 optipng 为非空 (non-NULL) 的值，例如，'' 去激活这个 hook （益辉称之为钩子，这里勾的是 optipng 这个图片优化工具）

```
knitr::knit_hooks$set(optipng = knitr::hook_optipng)

library(ggplot2)
ggplot(data = iris, aes(x = Sepal.Length, y = Sepal.Width)) +
  geom_point()
```

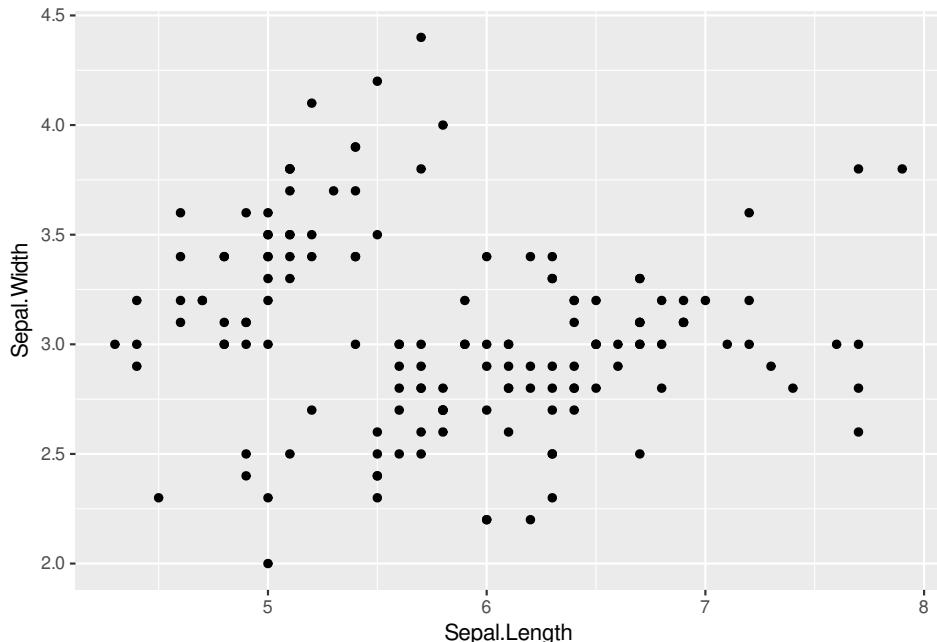


图 B.8: 没有优化

```
library(ggplot2)
ggplot(data = iris, aes(x = Sepal.Length, y = Sepal.Width)) +
  geom_point()
```

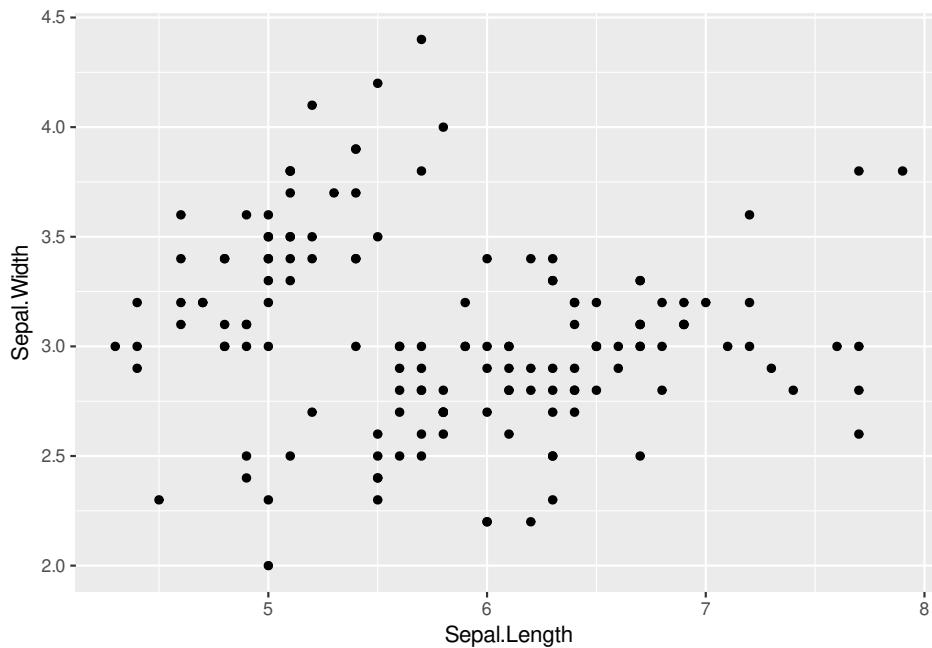


图 B.9: 优化

```
optipng -o5 filename.png
```

B.9 PDFCrop 裁剪边空

PDFCrop 可将 PDF 图片中留白的部分裁去，再也不用纠结 par 了

B.10 PhantomJS 网页截图

Winston Chang 开发了 webshot 包网页截图，它依赖 PhantomJS，所以首先需要安装

```
install.packages("webshot")
webshot::install_phantomjs()
```

以截取网页 <https://www.r-project.org/> 为例，



```
library(webshot)
webshot("https://www.r-project.org/", "r.png")
webshot("https://www.r-project.org/", "r.pdf") # Can also output to PDF
```

还可以截取 R Markdown 文档内容，注意是先编译 R Markdown 文档为 HTML 文档，然后截取网页

```
rmdshot(system.file("examples/knitr-minimal.Rmd", package = "knitr"), file = "screen.png")
```

裁剪出特定大小的图片，需要额外的系统依赖 GraphicsMagick (recommended) or ImageMagick installed

```
# Can specify pixel dimensions for resize()
```

```
webshot("https://www.r-project.org/", "r-small.png") %>%
  resize("400x") %>%
  shrink()
```

```
** Processing: r-small.png
400x442 pixels, 4x8 bits/pixel, RGB+alpha
Reducing image to 3x8 bits/pixel, RGB
Input IDAT size = 70570 bytes
Input file size = 70867 bytes
```

Trying:

```
zc = 9  zm = 8  zs = 0  f = 0           IDAT size = 59441
zc = 9  zm = 8  zs = 1  f = 0
zc = 1  zm = 8  zs = 2  f = 0
zc = 9  zm = 8  zs = 3  f = 0
zc = 9  zm = 8  zs = 0  f = 5
zc = 9  zm = 8  zs = 1  f = 5
zc = 1  zm = 8  zs = 2  f = 5
zc = 9  zm = 8  zs = 3  f = 5
```

Selecting parameters:

```
zc = 9  zm = 8  zs = 0  f = 0           IDAT size = 59441
```

```
Output IDAT size = 59441 bytes (11129 bytes decrease)
```

```
Output file size = 59714 bytes (11153 bytes = 15.74% decrease)
```



B.11 Inkscape 矢量绘图

Inkscape 是一款开源、免费、跨平台的矢量绘图软件。是替代 Adobe Illustrator (简称 AI) 最佳工具，没有之一

Ubuntu 20.04 及之前版本

```
sudo add-apt-repository ppa:inkscape.dev/stable  
sudo apt update  
sudo apt install inkscape
```

PDF 图片格式转化为 SVG 格式

```
inkscape -l output-filename.svg input-filename.pdf
```

SVG 转 PDF 格式

```
inkscape -f input-filename.svg -A output-filename.pdf
```

Jeroen Ooms 开发的 `rsvg` 包支持将 SVG 格式图片导出为 PNG、PDF、PS 等格式。使用它可以批量将 SVG 格式文件转化为其它格式文件，比如 PDF (`rsvg::rsvg_pdf`)，PS (`rsvg::rsvg_ps`) 和 PNG (`rsvg::rsvg_png`)

```
svg_paths = list.files(path = "images", pattern = "*.svg", full.names = T)  
for (svg in svg_paths) {  
  rsvg::rsvg_pdf(svg, file = gsub(pattern = "\\.svg", replacement= "\\.pdf", svg))  
}
```

B.12 QPDF PDF 文件操作

Jeroen Ooms 开发的另一个 `qpdf` 包将 C++ 库 `qpdf` 搬运到 R 环境中，用于 PDF 文件的拆分 `pdf_split()`，组合 `pdf_combine()`，加密（传递 `password` 参数值即可加密），提取 `pdf_subset()` 和压缩 `pdf_compress()` 等。下面以组合为例，就是将多个 PDF 文件合成一个 PDF 文件。

```
library(qpdf)  
pdf_paths = list.files(path = "images", pattern = "*.pdf", full.names = T)  
pdf_combine(input = pdf_paths, output = "images/all.pdf", password = "")
```

PDF 操作：价值数百美元的开源替代方案，参考 Adobe Acrobat 的功能

B.13 UML 标准建模图

UML (Unified Modeling Language) 表示统一建模语言

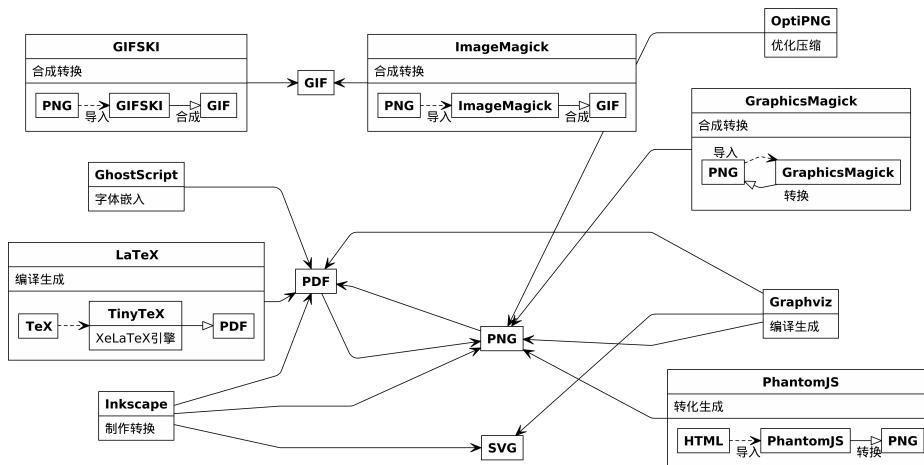


图 B.10: 图片制作、合成、优化、转换等常用工具

Javier Luraschi 将 UML 绘图库 `nomnoml` 引入 R 社区，开发了 `nomnoml` 包，相比于 `DiagrammeR` 包，它显得非常轻量，网站 <https://www.nomnoml.com/> 还可以在线编辑、预览、下载 UML 图。`webshot` 包可以将网页截图并插入 PDF 文档中。其它制作图形的工具见 B.10。

`nomnoml` 调 `webshot` 包对网页截图生成 PNG 格式的图片，其中 `webshot` 调 `phantomjs` 软件。`nomnoml` 制作 R Markdown 生态图，导出为 PNG 格式

安装 `nomnoml`

```
install.packages("nomnoml")
```

安装 `PhantomJS`

```
brew install --cask phantomjs
```

```
nomnoml::nomnoml("
#stroke: #26A63A
#.box: dashed visual=ellipse
#direction: down
```

[<box>HTML] -> [网页三剑客]

```
[<box>JavaScript] -> [网页三剑客]
[<box>CSS]          -> [<table>网页三剑客|htmlwidgets|htmltools||sass|bslib||thematic|jqu
[设计布局|bs4Dash|flexdashboard|shinydashboard] -> [<actor>开发应用|R Shiny]
[设计交互|waiter|shinyFeedback|shinyToastify] -> [<actor>开发应用|R Shiny]
[权限代理|shinyproxy|shinyauthr|shinymanager] -> [<actor>开发应用|R Shiny]

[网页三剑客] -> [<actor>开发应用|R Shiny]
[网页三剑客] -> [<actor>开发应用|R Shiny]
[网页三剑客] -> [<actor>开发应用|R Shiny]

[开发应用] <- [<table>处理数据|Base R|SQL||data.table|dplyr||tidyR|purrr]
[开发应用] <- [<table>制作表格|DT|gt||reactable|formatable||kableExtra|sparkline]
[开发应用] <- [<table>制作图形|ggplot2|plotly||echarts4r|leaflet||dygraphs|visNetwork]
", png = "shiny-app.png")
```

B.14 Graphviz 流程图

Graphviz 官网 <http://www.graphviz.org/>，常用于绘制流程图，广泛用于 tensorflow 和 mxnet 的模型描述中

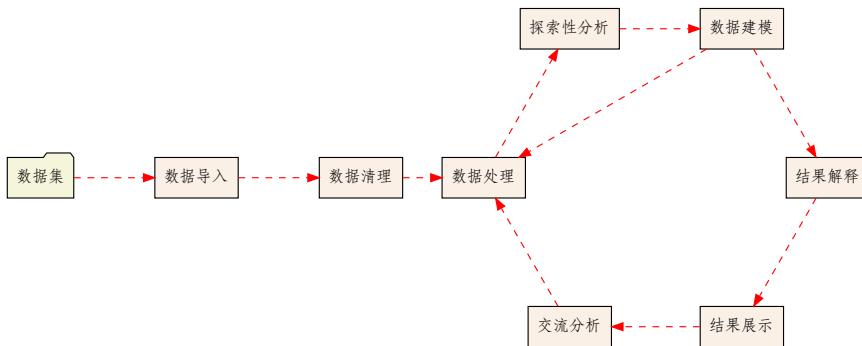


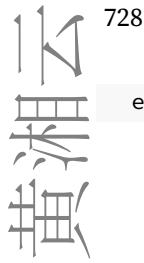
图 B.11: 数据分析流程图

[DiagrammeR](#) 包将 Graphviz 引入 R 语言



```
library(DiagrammeR)
library(DiagrammeRsvg)
library(magrittr)
library(rsvg)

graph <-
  "graph {
    rankdir=LR; // Left to Right, instead of Top to Bottom
    a -- { b c d };
    b -- { c e };
    c -- { e f };
    d -- { f g };
    e -- h;
    f -- { h i j g };
    g -- k;
    h -- { o l };
    i -- { l m j };
    j -- { m n k };
    k -- { n r };
    l -- { o m };
    m -- { o p n };
    n -- { q r };
    o -- { s p };
    p -- { s t q };
    q -- { t r };
    r -- t;
    s -- z;
    t -- z;
  }
"
# 导出图形
grViz(graph) %>%
  export_svg %>% charToRaw %>% rsvg_pdf("graph.pdf")
grViz(graph) %>%
  export_svg %>% charToRaw %>% rsvg_png("graph.png")
grViz(graph) %>%
```



```
export_svg %>% charToRaw %>% rsvg_svg("graph.svg")
```

B.15 LaTeX 排版工具



另外值得一提的是 TikZ 和 PGF (Portable Graphic Format) 宏包，支持强大的绘图功能，图形质量达到出版级别，详细的使用说明见宏包手册 <https://pgf-tikz.github.io/pgf/pgfmanual.pdf>。

B.15.1 TinyTeX 发行版

```
library(tinytex)
# 升级 TinyTeX 发行版
upgrade_tinytex <- function(repos = NULL) {
  # 此处还要考虑用户输错的情况和选择离用户最近(快)的站点
  if(is.null(repos)) repos = "https://mirrors.tuna.tsinghua.edu.cn/CTAN/"

  file_ext <- if (.Platform$OS.type == "windows") ".exe" else ".sh"
  tlmgr_url <- paste(repos, "/systems/texlive/tlnet/update-tlmgr-latest", file_ext, sep = "")
  file_name <- paste0("update-tlmgr-latest", file_ext)
  download.file(url = tlmgr_url, destfile = file_name,
                mode = if (.Platform$OS.type == "windows") "wb" else "w")

  # window下 命令行窗口下 如何执行 exe 文件
  if(.Platform$OS.type == "windows"){
    shell.exec(file = file_name)
    file.remove("update-tlmgr-latest.exe")
  }
  else{
    system("sudo sh update-tlmgr-latest.sh -- --upgrade")

    file.remove("update-tlmgr-latest.sh")
  }

  # 类似地 Linux 下执行 sh
```



```
# 升级完了 删除 update-tlmgr-latest.exe  
}
```

Winston Chang 整理了一份 LaTeX 常用命令速查小抄 <https://wch.github.io/latexsheet/latexsheet.pdf>

B.15.2 安装和更新

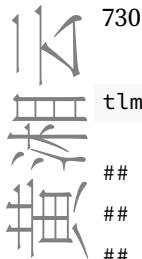
tlmgr (TeXLive Manager) 是 LaTeX 包管理器

```
# 就近选择 CTAN 镜像站点  
tlmgr option repository https://mirrors.tuna.tsinghua.edu.cn/CTAN/systems/texlive  
tlmgr option repository http://mirror.ctan.org/systems/texlive/tlnet  
# 可更新的 TeX 包列表  
tlmgr update --list  
# 更新所有已经安装的 TeX 包  
tlmgr update --all  
# 更新 tlmgr 管理器本身  
tlmgr update --self  
# 安装  
tlmgr install ctex fandol  
# 列出套装  
tlmgr list schemes  
tlmgr list collections  
# 列出已经安装的 TeX 包  
tlmgr list --only-installed  
# 安装 GPG 公钥 (只限 Win/Mac)  
tlmgr --repository http://www.preining.info/tlgpg/ install tlgpg
```

B.15.3 查询和搜索

```
tlmgr search *what*
```

参数 *what* 是正则表达式



```
tlmgr search --file tikz.sty  
## langsci:  
## texmf-dist/tex/xelatex/langsci/langsci-tikz.sty  
## pgf:  
## texmf-dist/tex/latex/pgf/frontendlayer/tikz.sty
```

等价于

```
tinytex::tlmgr_search('tikz.sty')
```

这样，我们就可以知道要使用 `\usepackage{tikz}` 就得先安装 **pgf** 包，此外，管道命令也是支持的

```
tlmgr search --file font | grep math
```

查询 CTAN 仓库列表

```
tlmgr repository list
```

一般地，只显示已安装的 LaTeX 宏包的名字及大小

```
tlmgr info --list --only-installed --data name,size
```

更多命令详见[tlmgr 管理器手册](#)

B.15.4 TikZ 绘图工具

TikZ 绘制书籍封面 <https://latexdraw.com/how-to-create-a-beautiful-cover-page-in-latex-using-tikz/>

TikZ 绘制知识清单，书籍章节结构等 <https://www.latexstudio.net/index/lists/barsearch/author/1680.html>

更多例子参考 <https://github.com/FriendlyUser/LatexDiagrams>

B.16 Octave 科学计算

```
%% fig1  
tx = ty = linspace (-8, 8, 41)';  
[xx, yy] = meshgrid (tx, ty);
```



```
r = sqrt (xx .^ 2 + yy .^ 2) + eps;
tz = sin (r) ./ r;
mesh (tx, ty, tz);
xlabel ("tx");
ylabel ("ty");
zlabel ("tz");
title ("3-D Sombrero plot");

% fig2
x = 0:0.01:3;
hf = figure ();
plot (x, erf (x));
hold on;
plot (x, x, "r");
axis ([0, 3, 0, 1]);
text (0.65, 0.6175, ['$\leftarrow x = {2 \over \sqrt{\pi}} \int_0^x e^{-t^2} dt = 0.6175$']);
xlabel ("x");
ylabel ("erf (x)");
title ("erf (x) with text annotation");
set (hf, "visible", "off");
print (hf, "plot15_7.pdf", "-dpdflatexstandalone");
set (hf, "visible", "on");
system ("pdflatex plot15_7");
open ("plot15_7.pdf");

%% fig3
clf ();
surf (peaks);
peaks(50)
print -dpswrite -PPS_printer

%% images/peaks-inc
hf = figure (1);
```



```
surf (peaks);
print (hf, "peaks.pdf", "-dpdflatexstandalone");

%% windows
hf = figure (1);
peaks(10);
print (hf, "peaks.pdf", "-dpdf");
print (hf, "peaks.eps", "-color", "-deps");

print (hf, "peaks.svg", "-color", "-dsvg");

%% windows
hf = figure (1);
peaks(50);
print (hf, "peaks-more.eps", "-color", "-deps");

print (hf, "peaks-more.svg", "-color", "-dsvg");
```

B.17 Python 环境配置

首先创建一个 Python 虚拟环境，环境隔离可以减少对系统的侵入，方便迭代更新和项目管理。创建一个虚拟环境，步骤非常简单，下面以 CentOS 8 为例：

1. 安装虚拟模块 virtualenv

```
sudo dnf install -y virtualenv
```

2. 准备 Python 虚拟环境存放位置

```
sudo mkdir -p /opt/.virtualenvs/r-tensorflow
```

3. 给虚拟环境必要的访问权限

```
sudo chown -R $(whoami):$(whoami) /opt/.virtualenvs/r-tensorflow
```

4. 初始化虚拟环境



```
virtualenv -p /usr/bin/python3 /opt/.virtualenvs/r-tensorflow
```

5. 激活虚拟环境，安装必要的模块

```
source /opt/.virtualenvs/r-tensorflow/bin/activate  
pip install numpy
```

一般来讲，系统自带的 pip 版本较低，可以考虑升级 pip 版本。

```
pip install -i https://pypi.tuna.tsinghua.edu.cn/simple pip -U
```

根据项目配置文件 requirements.txt 安装多个 Python 模块，每个 Python 项目都应该有这么个文件来描述项目需要的依赖环境，包含 Python 模块及其版本号。

```
pip install -i https://pypi.tuna.tsinghua.edu.cn/simple -r requirements.txt
```

指定 Python 模块的镜像地址，加快下载速度，特别是对于国内的环境，加速镜像站点非常有意义，特别是遇到大型的 Python 模块，比如 tensorflow 框架

```
pip install -i https://pypi.tuna.tsinghua.edu.cn/simple tensorflow
```

conda 创建 Python 3.8 虚拟环境，并命名为 tensorflow

```
conda create -n tensorflow python=3.8
```

激活 tensorflow 环境

```
conda activate tensorflow
```

B.18 Python 基础绘图

Python 的 matplotlib 模块支持保存的图片格式有 eps, pdf, pgf, png, ps, raw, rgba, svg, svgz，不支持 cairo_pdf 绘图设备，所以这里使用 pdf 设备，但是这样会导致图形没有字体嵌入，从而不符合出版要求。一个解决办法是在后期嵌入字体，图形默认使用数学字体 STIX 和英文字体 DejaVu Sans，所以需要预先安装这些字体。

```
# CentOS 8  
sudo dnf install -y dejavu-fonts-common dejavu-sans-fonts \  
dejavu-serif-fonts dejavu-sans-mono-fonts
```

借助 grDevices 包提供的 embedFonts() 函数，它支持 postscript 和 pdf 图形设备，嵌入字体借助了 Ghostscript 以及 PDF 阅读器 MuPDF



注意

Windows 系统下需要手动指定 Ghostscript 安装路径，特别地，如果你想增加可选字体范围，需要指定相应字体搜索路径，而 Linux/MacOS 平台下不需要关心 Ghostscript 的安装路径问题，

```
Sys.setenv(R_GSCMD = "C:/Program Files/gs/gs9.26/bin/gswin64c.exe")
embedFonts(
  file = "cm.pdf", outfile = "cm-embed.pdf",
  fontpaths = system.file("fonts", package = "fontcm")
)
embedFonts(file = "cm.pdf", outfile = "cm-embed.pdf")
```

另一个解决办法是使用 LaTeX 渲染图片中的文字，这就需要额外安装一些 LaTeX 宏包，此时默认执行渲染的 LaTeX 引擎是 PDFLaTeX。

```
tlmgr install type1cm cm-super dvipng psnfss ucs ncntrsbk helvetica
```

每年 4 月是 TeX Live 的升级月，升级指导见 <https://www.tug.org/texlive/upgrade.html>，升级之后，需要更新所有 LaTeX 宏包。

```
tlmgr update --self --all
```

如图 B.12 所示，我们采用第二个方法，它可以支持更好的数学公式显示，更多详情见 <https://matplotlib.org/tutorials/text/mathtext.html>。

```
## [<matplotlib.lines.Line2D object at 0x7fcf5febba90>]
## Text(0.5, 0, 'Coord $x$')
## Text(0, 0.5, 'Coord $y$')
```

提示

如果你的系统是 Windows/MacOS 可以添加 GPG 验证以增加安全性，最简单的方式就是：

```
tlmgr --repository http://www.preining.info/tlgpg/ install tlgpg
```

二维函数 $f(x, y) = 20 + x^2 + y^2 - 10 * \cos(2 * \pi * x) - 10 * \cos(2 * \pi * y)$ 最小值 0
最大值 80

```
from math import cos, pi
import numpy as np
from mpl_toolkits.mplot3d import Axes3D
```

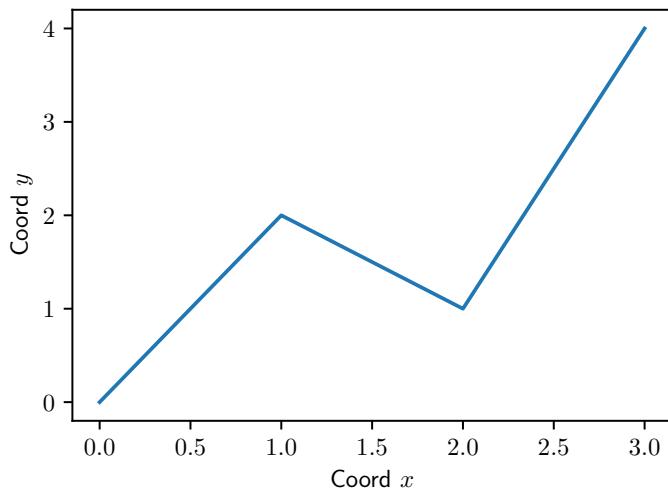


图 B.12: matplotlib 示例

```
import matplotlib.pyplot as plt
from matplotlib import cm

from matplotlib import rcParams
rcParams.update({'font.size': 18, 'text.usetex': True}) # 其它可配置选项见 rcParams
plt.switch_backend('agg')

xDomain = np.arange(-5.12, 5.12, .08)
yDomain = np.arange(-5.12, 5.12, .08)

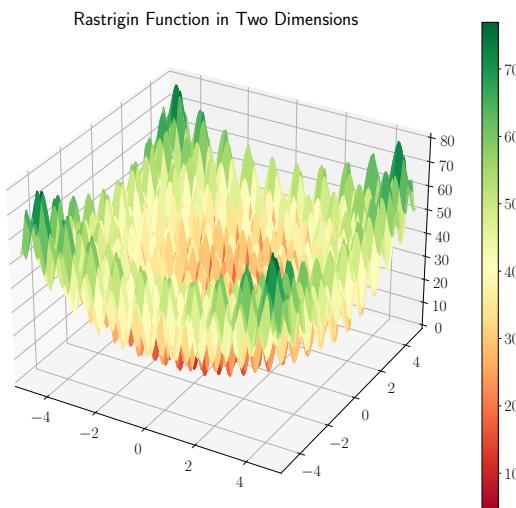
X, Y = np.meshgrid(xDomain, yDomain)
z = [20 + x**2 + y**2 - (10*(cos(2*pi*x) + cos(2*pi*y))) for x in xDomain for y in yDomain]
Z = np.array(z).reshape(128,128)

fig = plt.figure(figsize = (12,10))
ax = fig.gca(projection='3d')
## <string>:1: MatplotlibDeprecationWarning: Calling gca() with keyword arguments
surf = ax.plot_surface(X, Y, Z, cmap=cm.RdYlGn, linewidth=1, antialiased=False)

ax.set_xlim(-5.12, 5.12)
```



```
## (-5.12, 5.12)
ax.set_ylim(-5.12, 5.12)
## (-5.12, 5.12)
ax.set_zlim(0, 80)
## (0.0, 80.0)
fig.colorbar(surf, aspect=30)
## <matplotlib.colorbar.Colorbar object at 0x7fcf5c8f70a0>
plt.title(r'Rastrigin Function in Two Dimensions')
## Text(0.5, 0.92, 'Rastrigin Function in Two Dimensions')
plt.show()
```



B.19 Python 基础操作

- 张量操作 numpy <https://numpy.org/> 向量、矩阵操作
- 科学计算 scipy <https://scipy.org/> 统计、优化和方程
- 数据操作 pandas <https://pandas.pydata.org/> 面向数据分析
- 数据可视化 matplotlib <https://matplotlib.org/> 静态图形
- 交互可视化 bokeh <https://bokeh.org/>
- 机器学习 scikit-learn <https://scikit-learn.org/> 面向机器学习



- 深度学习 [tensorflow https://tensorflow.org/](https://tensorflow.org/) 面向深度学习

A Python implementation of global optimization with gaussian processes.
[Bayesian Optimization](#)

用 numpy 实现一个统计类的算法，比如线性回归、稳健的线性回归、广义线性回归，数据集用 Python 内置的

```
import numpy as np
np.zeros(3) # vector

## array([0., 0., 0.])

np.ones(3) # vector

## array([1., 1., 1.])

np.diag([1,1,1]) # identity matrix
# np.multiply()

## array([[1, 0, 0],
##        [0, 1, 0],
##        [0, 0, 1]])

np.cumsum([1,1,1])

## array([1, 2, 3])
```

Python 模块 scikit-learn [[Pedregosa et al., 2011](#)] 内置的数据集 iris 为例 <https://scikit-learn.org/stable/datasets/index.html>

导入正则表达式库，

```
import re
m = re.search('(?<=abc)def', 'abcdef')
m.group(0) # 必须调用 print 函数打印结果

## 'def'
print(m.group(0))

## def
import sys
print(sys.path)

## ['', '/usr/bin', '/usr/lib/python38.zip', '/usr/lib/python3.8', '/usr/lib/python
```



字符串基本操作，如拆分

```
dir(str)
```

```
## ['__add__', '__class__', '__contains__', '__delattr__', '__dir__', '__doc__', '__eq__',  
print(dir(str.split))
```

④
import re

```
print(dir(re.split))
```

```
## ['__annotations__', '__call__', '__class__', '__closure__', '__code__', '__defaults__',  
import sys  
# 模块存放路径  
print(sys.path)  
# 已安装的模块  
sys.modules.keys()
```

```
dict_keys(['sys', 'builtins', '_frozen_importlib', '_imp', '_warnings', '_frozen_importlib  
'_io', 'marshal', 'posix', '_thread', '_weakref', 'time', 'zipimport', '_codecs', 'code  
'encodings.alises', 'encodings.cp437', 'encodings', 'encodings.utf_8', '_signal', '_m  
'encodings.latin_1', '_abc', 'abc', 'io', '_stat', 'stat', '_collections_abc', 'generic  
'posixpath', 'os.path', 'os', '_sitebuiltins', 'site', 'readline', 'atexit', 'rlcomplet
```

```
pip3 install virtualenv
```

```
virtualenv -p python3 <desired-path>
```

```
source <desired-path>/bin/activate
```

```
source /opt/virtualenv/tensorflow/bin/activate
```

- LaTeX 专家黄晨成写的译文 [Matplotlib 教程](#)
- 周沫凡 制作的莫烦 Python 系列视频教程之 [Matplotlib 数据可视化神器](#)
- 陈治兵维护的在线 [Matplotlib 中文文档](#)
- Sebastian Raschka 和 Vahid Mirjalili 合著的 [Python Machine Learning \(3rd Edition\) \[Raschka and Mirjalili, 2017\]](#)

编译书籍使用的 Python 3 模块有

```
pip3 list --format=columns
```



Package	Version
absl-py	0.13.0
astunparse	1.6.3
cachetools	4.2.2
certifi	2021.5.30
charset-normalizer	2.0.4
cycler	0.10.0
flatbuffers	1.12
gast	0.4.0
google-auth	1.34.0
google-auth-oauthlib	0.4.5
google-pasta	0.2.0
graphviz	0.8.4
grpcio	1.34.1
h5py	3.1.0
idna	3.2
joblib	1.0.1
kaleido	0.2.1
keras-nightly	2.5.0.dev2021032900
Keras-Preprocessing	1.1.2
kiwisolver	1.3.1
Markdown	3.3.4
matplotlib	3.4.2
mpmath	1.2.1
mxnet	1.8.0.post0
numpy	1.21.1
oauthlib	3.1.1
opt-einsum	3.3.0
pandas	1.3.1
patsy	0.5.1
Pillow	8.3.1
pip	20.0.2
pkg-resources	0.0.0
plotly	5.1.0
protobuf	3.17.3
pyasn1	0.4.8



Package	Version
pyasn1-modules	0.2.8
pyparsing	2.4.7
python-dateutil	2.8.2
pytz	2021.1
requests	2.26.0
requests-oauthlib	1.3.0
rsa	4.7.2
scikit-learn	0.24.2
scipy	1.7.1
setuptools	44.0.0
six	1.15.0
statsmodels	0.12.2
sympy	1.8
tenacity	8.0.1
tensorboard	2.6.0
tensorboard-data-server	0.6.1
tensorboard-plugin-wit	1.8.0
tensorflow	2.5.0
tensorflow-estimator	2.5.0
termcolor	1.1.0
threadpoolctl	2.2.0
typing-extensions	3.7.4.3
urllib3	1.26.6
Werkzeug	2.0.1
wheel	0.36.2
wrapt	1.12.1

```
# 安装 Python 虚拟环境管理器 virtualenv
sudo dnf install -y python3-pip python3-virtualenv
# 创建虚拟环境
virtualenv -p /usr/bin/python3 $RETICULATE_PYTHON_ENV
# 激活虚拟环境
source $RETICULATE_PYTHON_ENV/bin/activate
# 将虚拟环境位置写入配置文件
```



```
echo "export RETICULATE_PYTHON_ENV=$HOME/.virtualenvs/r-tensorflow" >> ~/.bashrc
source ~/.bashrc
# 安装 numpy matplotlib 等模块
pip install -r requirements.txt
# 导出模块版本信息
pip freeze >> requirements.txt

import os
os.listdir('.git')
```

['FETCH_HEAD', 'branches', 'HEAD', 'index', 'config', 'info', 'refs', 'shallow', '']

多个代码块共享同一个 Python 进程

```
os.path
```

<module 'posixpath' from '/usr/lib/python3.8 posixpath.py'>

matplotlib 绘图，支持交叉引用²，如图 B.13 所示

```
import matplotlib.pyplot as plt
from matplotlib import rcParams
# 其它可配置选项见 rcParams.keys()
rcParams.update({'font.size': 10, 'text.usetex': True})
# rcParams.update({'font.family':      ['sans-serif'],
#                  'font.monospace': ['DejaVu Sans Mono'],
#                  'font.sans-serif': ['DejaVu Sans'],
#                  'font.serif':     ['DejaVu Serif']})
plt.switch_backend('agg')
plt.plot([0, 2, 1, 4])
```

[<matplotlib.lines.Line2D object at 0x7fcf42bd9f70>]

```
plt.xlabel(r'Coord $x$')
```

```
## Text(0.5, 0, 'Coord $x$')
```

²早些时候，在 R Markdown 中设置 `python.reticulate = TRUE` 调用 `reticulate` 包，带来的副作用是不支持交叉引用的 <https://d.cosx.org/d/420680-python-reticulate-true>。RStudio 1.2 已经很好地集成了 `reticulate`，对 Python 的支持更加到位了 <https://blog.rstudio.com/2018/10/09/rstudio-1-2-preview-reticulated-python/>。截至本文写作时间 2021 年 08 月 07 日使用 `reticulate` 版本 1.20，本文没有对之前的版本进行测试。

```
plt.ylabel(r'Coord $y$')  
## Text(0, 0.5, 'Coord $y$')  
plt.tight_layout()  
plt.show()
```

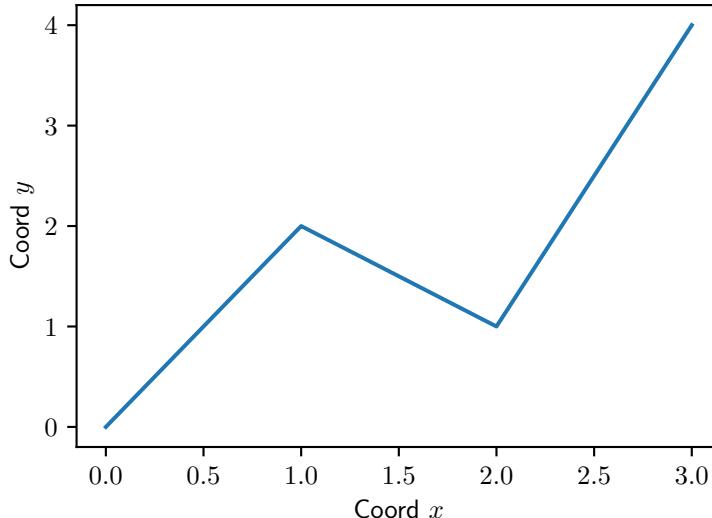


图 B.13: matplotlib 复制示例

有了 `reticulate` 包，我们可以把任意想要导入到 R 环境中的 Python 模块导进来，实现 R 与 Python 的数据交换和函数调用³

```
os <- reticulate::import("os") # 导入 Python 模块  
x <- os$listdir(".git") # 调用 os.listdir() 函数  
x # 得到 python 中的向量 vector 或数组 array  
  
## [1] "FETCH_HEAD"   "branches"      "HEAD"          "index"        "config"  
## [6] "info"         "refs"          "shallow"       "objects"      "description"  
## [11] "logs"         "hooks"  
  
# https://docs.bokeh.org/en/latest/docs/user\_guide/quickstart.html#userguide-quickstart  
from bokeh.plotting import figure, output_file, show  
# 准备一些数据
```

³朱俊辉的帖子 – 在 R 中使用 gluon <https://d.cosx.org/d/419785-r-gluon>



```
x = [1, 2, 3, 4, 5]
y = [6, 7, 2, 4, 5]
# 将动态图形以静态 HTML 文件的方式保存
output_file("lines.html")
# 创建一个简单的图形, 设置标题、x,y 轴标签
p = figure(title="simple line example", x_axis_label='x', y_axis_label='y')
# 添加一条折线, 设置图例, 线宽
p.line(x, y, legend_label="Temp.", line_width=2)
# 显示结果
show(p)
```

将静态图形嵌入到 R Markdown 中

```
htmltools:::includeHTML("lines.html")
```

R 和 Python 之间的交互, Python 负责数据处理和建模, R 负责绘图, 有些复杂的机器学习模型及其相关数据操作需要在 Python 中完成, 数据集清理至数据框的形式后导入到 R 中, 画各种静态或者动态图, 这时候需要加载 reticulate 包, 只是设置 `python.reticulate = TRUE` 还不够

提示

R Markdown 文档 [Xie et al., 2018] 中的 Python 代码块是由 knitr 包 [Xie, 2015] 负责调度处理的, 展示 Matplotlib 绘图的结果使用了 reticulate 包 [Ushey et al., 2021] 提供的 Python 引擎而不是 knitr 自带的。

在 `knitr::opts_chunk` 中设置 `python.reticulate = TRUE` 意味着所有的 Python 代码块共享一个 Python Session, 而 `python.reticulate = FALSE` 意味着使用 knitr 提供的 Python 引擎, 所有的 Python 代码块独立运行。

pandas 读取数据, 整理后由 reticulate 包传递给 R 环境中的 `data.frame` 对象, 加载 ggplot2 绘图

```
library(ggplot2)
theme_set(theme_minimal())
library(patchwork)
p1 <- ggplot(py$iris2, aes(x = Sepal.Length, y = Sepal.Width)) +
  geom_point(aes(color = Species)) +
  labs(title = "Call iris from Python")
p2 <- ggplot(iris, aes(x = Sepal.Length, y = Sepal.Width)) +
  geom_point(aes(color = Species)) +
```



```
labs(title = "Call iris from R")  
p1 + p2
```

以 NumPy 为例

```
import numpy as np  
a = np.arange(15).reshape(3, 5)  
a
```

```
## array([[ 0,  1,  2,  3,  4],  
##          [ 5,  6,  7,  8,  9],  
##          [10, 11, 12, 13, 14]])
```

```
a.shape
```

```
## (3, 5)
```

```
a.ndim
```

```
## 2
```

```
a.dtype.name
```

```
## 'int64'
```

```
a.itemsize
```

```
## 8
```

```
a.size
```

```
## 15
```

```
type(a)
```

```
## <class 'numpy.ndarray'>
```

```
b = np.array([6, 7, 8])  
b
```

```
## array([6, 7, 8])
```

```
type(b)
```

```
## <class 'numpy.ndarray'>
```



```
a.transpose() @ b  
## array([115, 136, 157, 178, 199])
```

Python 里面的点号 . 对应于 R 里面的 \$

```
library(reticulate)  
np <- import("numpy", convert=FALSE) # 导入 Python 模块  
a <- np$arange(0, 15)$reshape(3L, 5L)  
a
```

```
## [[ 0.  1.  2.  3.  4.]  
##   [ 5.  6.  7.  8.  9.]  
##   [10. 11. 12. 13. 14.]]
```

```
a$shape
```

```
## (3, 5)
```

```
a$ndim
```

```
## 2
```

```
a$dtype$name
```

```
## float64
```

```
a$itemsize
```

```
## 8
```

```
a$size
```

```
## 15
```

```
a$ctypes
```

```
## <numpy.core._internal._ctypes>
```

```
a$dtype # data type 数据类型
```

```
## float64
```

```
a$astype
```

```
## <built-in method astype of numpy.ndarray>
```

```
builtins <- import_builtins() # Python 内建的函数，不需要导入第三方模块
builtins$type(a)
```



```
## <class 'numpy.ndarray'>
```



基本线性代数运算

```
a$transpose() # 转置
```

```
## [[ 0.  5. 10.]
##   [ 1.  6. 11.]
##   [ 2.  7. 12.]
##   [ 3.  8. 13.]
##   [ 4.  9. 14.]]
```

```
a$trace() # 迹
```

```
## 18.0
```

```
np$eye(2L) # 单位矩阵
```

```
## [[1. 0.]
##   [0. 1.]]
```

```
a$diagonal() # 对角
```

```
## [ 0.  6. 12.]
```

两个矩阵的乘法

```
b <- np$array(c(6, 7, 8, 9, 10))$reshape(5L, 1L)
```

```
b
```

```
## [[ 6.]
##   [ 7.]
##   [ 8.]
##   [ 9.]
##   [10.]]
```

```
b$shape
```

```
## (5, 1)
```



```
np$multiply(b$transpose(), a) # b 乘以 a  
## [[ 0.   7.  16.  27.  40.]  
##  [ 30.  42.  56.  72.  90.]  
##  [ 60.  77.  96. 117. 140.]]
```

Python 对象转化为 R 对象

```
py_to_r(b)
```

```
##      [,1]  
## [1,]    6  
## [2,]    7  
## [3,]    8  
## [4,]    9  
## [5,]   10
```

B.20 VBox 虚拟机

B.20.1 从命令行启动虚拟机

当前我的虚拟机里安装了两个系统 Fedora 29 和 CentOS 8.2

```
VBoxManage list vms
```

```
"Fedora 29" {d316fe8d-c053-4941-8a45-a59fd476898d}  
"CentOS 8.2" {f1613f26-ea65-4f02-9cb6-6a79a758a60e}
```

以无图形化界面的方式启动虚拟机 CentOS 8.2

```
VBoxManage startvm "CentOS 8.2" --type headless  
# 或者  
VBoxHeadless --startvm "CentOS 8.2"
```

其它常用的命令还有

```
VBoxManage list runningvms # 列出运行中的虚拟机  
VBoxManage controlvm "CentOS 8.2" acpipowerbutton # 关闭虚拟机, 等价于点击系统关闭  
VBoxManage controlvm "CentOS 8.2" poweroff # 关闭虚拟机, 等价于直接关闭电源, 非正常关  
VBoxManage controlvm "CentOS 8.2" pause # 暂停虚拟机的运行
```



```
VBoxManage controlvm "CentOS 8.2" resume # 恢复暂停的虚拟机
VBoxManage controlvm "CentOS 8.2" savestate # 保存当前虚拟机的运行状态
```

更多细节解释见 [VBox 官方文档](#)



B.21 Docker 虚拟环境

`docker` 创建云实例 `rstudio` DigitalOcean, `docker` 支持的驱动类型 <https://docs.docker.com/machine/drivers/>。Rocker 项目组提供的 shiny 容器 <https://github.com/rocker-org/shiny> 和构建过程 <https://hub.docker.com/r/rocker/shiny/dockerfile>

主机 80 端口映射给 shiny 容器 3838 端口

```
docker run --user shiny -d -p 80:3838 \
-v /srv/shinyapps/:/srv/shiny-server/ \
-v /srv/shinylog/:/var/log/shiny-server/ \
rocker/shiny
```

shiny 服务器默认支持从 80 端口访问 <http://localhost:80>, shiny 应用放在目录 `/srv/shinyapps/appdir`, 访问 Shiny 应用的位置 <http://localhost/appdir/>, 使用 `boot2docker` 则访问 <http://192.168.59.103:80/appdir/>

Docker 相比虚拟机占用资源少, 拉起来就可以用, 虚拟机还需要各种环境配置, 很多与 R 有关的项目现在都提供 Docker 镜像, 大大方便了开发人员和数据分析师。当然 `docker` 的环境隔离性, 对主机系统侵入小, 即使挂了, 再拉起来也就是了, 安全性和可靠性高。

基于 [The Rocker Project](#) 快速构建数据分析环境, [Rocker 项目](#) 站在 [Debian](#) 和 [R](#) 的肩膀上, 在 [Docker](#) 内配置众多数据分析和开发的工具, 免去用户手动配置的复杂性。此事非有心者不能为之, 因为需费时费力找寻依赖库, 编译 R 包, 还要尽可能地给 Docker 镜像减负, 以便部署。如果想抢先试水的赶快去 Rocker 项目主页。

- 由 Dirk Eddelbuettel 等人担纲的 Rocker 项目, [项目主页](#) 和 [Docker 镜像](#)
- Wei-Chen Chen 等人的大数据编程项目 Programming with Big Data in R, [项目主页](#) 和 [Docker 镜像](#)
- 非常详细的 `docker` 笔记



- Dockerfile 最佳实践 https://docs.docker.com/develop/develop-images/dockerfile_best-practices/
- build 构建 <https://docs.docker.com/engine/reference/builder/#usage>

其它容器相关项目有 [Singularity](#) 和 [Kubernetes](#) 容器集群管理，更多参见高策的博客 <https://gaocegege.com>

本节介绍与本书配套的 VBox 镜像和 Docker 容器镜像，方便读者直接运行书籍原稿中的例子，尽量不限于软件环境配置的苦海中，因为对于大多数初学者来说，软件配置是一件不小的麻烦事。

本书依赖的 R 包和配置环境比较复杂，所以将整个运行环境打包成 Docker 镜像，方便读者重现，构建镜像的 Dockerfile 文件随同书籍源文件一起托管在 Github 上，方便读者研究。本地编译书籍只需三步走，先将存放在 Github 上的书籍项目克隆到本地，如果本地环境中没有 Git，你需要从它的官网 <https://git-scm.com/> 下载安装适配本地系统的 Git 软件。

```
git clone https://github.com/XiangyunHuang/masr.git
```

然后在 Git Bash 的模拟终端器中，启动虚拟机，拉取准备好的镜像文件。为了方便读者重现本书的内容，特将书籍的编译环境打包成 Docker 镜像。在启动镜像前需要确保本地已经安装 Docker 软件 <https://www.docker.com/products/docker-desktop>，安装过程请看官网教程。

```
docker-machine.exe start default  
docker pull xiangyunhuang/masr
```

最后 cd 进入书籍项目所在目录，运行如下命令编译书籍

```
docker run --rm -u docker -v "${PWD}://home/docker/workspace" \  
xiangyunhuang/masr make gitbook
```

编译成功后，可以在目录 _book/ 下看到生成的文件，点击文件 index.html 选择谷歌浏览器打开，不要使用 IE 浏览器，推荐使用谷歌浏览器获取最佳阅读体验，尽情地阅读吧！

如果你想了解编译书籍的环境和过程，我推荐你阅读随书籍源文件一起的 Dockerfile 文件，Docker Hub 是根据此文件构建的镜像，打包成功后，大约占用空间 2 Gb，本书在 RStudio IDE 下用 R Markdown [Xie et al., 2018] 编辑的，编译本书获得电子版还需要一些 R 包和软件。Pandoc <https://pandoc.org/> 软件是系统 Fedora 30 仓库自带的，版本是 2.2.1，较新的 RStudio IDE 捆绑的 Pandoc 软件一般会高于此版本。如果你打算在本地系统上编译书籍，RStudio IDE 捆绑的 Pandoc 软件



版本已经足够，当然你也可以在 <https://github.com/jgm/pandoc/releases/latest> 下载安装最新版本，此外，你还需参考书籍随附的 Dockerfile 文件配置 C++ 代码编译环境，安装所需的 R 包，并确保本地安装的版本不低于镜像内的版本。

镜像中已安装的 R 包列表可运行如下命令查看。

```
docker run --rm xiangyunhuang/masr \
  Rscript -e 'xfun::session_info(.packages(TRUE))'
```

Docker & Docker Machine & Docker Swarm

1. 容器与镜像的操作

```
docker --version
# Docker version 18.03.0-ce, build 0520e24302
```

查看容器

```
docker ps -a
```

删除容器 docker rm 容器 ID，删除前要确认已经停止该容器的运行

```
docker rm 6f932357e6ce
```

查看镜像

```
docker images
```

删除镜像

```
docker rmi 镜像 ID
```

```
docker rmi 811281c54b23
```

2. 拉取镜像

```
docker pull rocker/verse:latest
```

3. 运行容器

```
docker run --name verse -d -p 8282:8080 -e ROOT=TRUE \
-e USER=rstudio -e PASSWORD=cloud rocker/verse
```

将主机端口 8282 映射给虚拟机/容器的 8080 端口，RStudio Server 默认使用的端口是 8787，因此改为 8080 需要修改 /etc/rstudio/rserver.conf 文件，添加

```
www-port=8080
```



然后重启 RStudio Server，之后可以在浏览器中登陆，登陆网址为 <http://ip-addr:8080>，其中 ip-addr 可在容器中运行如下一行命令获得

```
ip addr
```

更多关于服务器版本的 RStudio 介绍，请参考 <https://docs.rstudio.com/ide/server-pro/access-and-security.html>

Docker Machine

基本命令

- 查看 docker machine 版本信息

```
docker-machine --version
# docker-machine.exe version 0.14.0, build 89b8332
```

- 列出创建的虚拟机

```
# 启动前
docker-machine ls
# NAME      ACTIVE     DRIVER      STATE      URL      SWARM      DOCKER      ERRORS
# default    -          virtualbox   Stopped
# 启动后
docker-machine ls
# NAME      ACTIVE     DRIVER      STATE      URL      SWARM
# default    *          virtualbox   Running   tcp://192.168.99.100:2376      SWARM
```

- 查看虚拟机 default 的 ip

```
docker-machine ip default
# 192.168.99.100
```

- 启动虚拟机

```
docker-machine start default
# Starting "default"...
# (default) Check network to re-create if needed...
# (default) Windows might ask for the permission to configure a dhcp server.
# (default) Waiting for an IP...
# Machine "default" was started.
# Waiting for SSH to be available...
# Detecting the provisioner...
```



Started machines may have new IP addresses. You may need to re-run the `docker-mn` command.

- 进入 Docker 环境

```
docker-machine ssh default
```

- 查看容器

```
docker ps -a
```

#	CONTAINER ID	IMAGE	COMMAND	CREATED	S
#	69e6929d269e	rocker/verse	"/init"	3 weeks ago	E

- 启动/停止容器

```
docker start verse  
# verse  
docker stop verse  
# verse
```

- 查看虚拟机 default 的环境

```
docker-machine env default  
# export DOCKER_TLS_VERIFY="1"  
# export DOCKER_HOST="tcp://192.168.99.100:2376"  
# export DOCKER_CERT_PATH="D:\Docker\machines\default"
```

```
# export DOCKER_MACHINE_NAME="default"
# export COMPOSE_CONVERT_WINDOWS_PATHS="true"
# # Run this command to configure your shell:
## eval $("C:\Program Files\Docker Toolbox\docker-machine.exe" env default)
```

- 关闭虚拟机 default

```
docker-machine stop default
# Stopping "default"...
# Machine "default" was stopped.
```

更多详情见帮助文档 <https://docs.docker.com/machine/get-started>

B.22 安装的 R 包

警告

本小节仅用于展示目前书籍写作过程中安装的 R 包依赖，不会出现在最终的书稿中

```
sessionInfo(sort(.packages(T)))

## R version 4.1.0 (2021-05-18)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 20.04.2 LTS
##
## Matrix products: default
## BLAS:    /usr/lib/x86_64-linux-gnublas/libblas.so.3.9.0
## LAPACK:  /usr/lib/x86_64-linux-gnulapack/liblapack.so.3.9.0
##
## locale:
## [1] LC_CTYPE=en_US.UTF-8          LC_NUMERIC=C
## [3] LC_TIME=en_US.UTF-8          LC_COLLATE=en_US.UTF-8
## [5] LC_MONETARY=en_US.UTF-8       LC_MESSAGES=en_US.UTF-8
## [7] LC_PAPER=en_US.UTF-8          LC_NAME=C
## [9] LC_ADDRESS=C                  LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8   LC_IDENTIFICATION=C
##
```

attached base packages:

[1] base compiler datasets graphics grDevices grid methods
[8] parallel splines stats stats4 tcltk tools utils

other attached packages:

[1] ABCoptim_0.15.0 abind_1.4-5
[3] agricardat_1.18 alabama_2015.3-1
[5] arrow_5.0.0 arules_1.6-8
[7] askpass_1.1 assertive.base_0.0-9
[9] assertive.properties_0.0-4 assertive.types_0.0-3
[11] assertthat_0.2.1 autoplotly_0.1.4
[13] backports_1.2.1 base64enc_0.1-3
[15] bayesplot_1.8.1 BB_2019.10-1
[17] bbmle_1.0.24 bdsmatrix_1.3-4
[19] beanplot_1.2 beeswarm_0.4.0
[21] BH_1.75.0-0 BiocGenerics_0.38.0
[23] BiocManager_1.30.16 BiocVersion_3.13.1
[25] bit_4.0.4 bit64_4.0.5
[27] bitops_1.0-7 blob_1.2.2
[29] bookdown_0.22 boot_1.3-28
[31] brew_1.0-6 bridgesampling_1.1-2
[33] brio_1.1.2 brms_2.15.0
[35] Broddingnag_1.2-6 broom_0.7.9
[37] broom.mixed_0.2.7 bslib_0.2.5.1
[39] cachem_1.0.5 Cairo_1.5-12.2
[41] callr_3.7.0 car_3.0-11
[43] carData_3.0-4 cellranger_1.1.0
[45] checkmate_2.0.0 circlize_0.4.13
[47] class_7.3-19 classInt_0.4-3
[49] cli_3.0.1 clipr_0.7.1
[51] clue_0.3-59 cluster_2.1.2
[53] cmdstanr_0.4.0 coda_0.19-4
[55] codetools_0.2-18 colorspace_2.0-2
[57] colourpicker_1.1.0 commonmark_1.7
[59] ComplexHeatmap_2.8.0 config_0.3.1
[61] conquer_1.0.2 corrplot_0.90

```
## [63] countrycode_1.3.0          cowplot_1.1.1
## [65] cpp11_0.3.1                 crayon_1.4.1
## [67] credentials_1.3.1           crosstalk_1.1.1
## [69] curl_4.3.2                  data.table_1.14.0
## [71] DBI_1.1.1                   dbplyr_2.1.1
## [73] dendextend_1.15.1           Deriv_4.1.3
## [75] desc_1.3.0                  deSolve_1.28
## [77] devtools_2.4.2              DiagrammeR_1.0.6.1
## [79] diffobj_0.3.4               digest_0.6.27
## [81] distributional_0.2.2        doParallel_1.0.16
## [83] downlit_0.2.1               downloader_0.4
## [85] dplyr_1.0.7                 DT_0.18
## [87] dtplyr_1.1.0                dygraphs_1.1.1.6
## [89] e1071_1.7-8                 echarts4r_0.4.1
## [91] egg_0.4.5                   ellipsis_0.3.2
## [93] emo_0.0.0.9000             equatiomatic_0.2.0
## [95] evaluate_0.14               extrafont_0.17
## [97] extrafontdb_1.0              fansi_0.5.0
## [99] farver_2.1.0                fastmap_1.1.0
## [101] filehash_2.4-2             flexdashboard_0.5.2
## [103] fontcm_1.1                  forcats_0.5.1
## [105] foreach_1.5.1              foreign_0.8-81
## [107] forge_0.2.0                 formatR_1.11
## [109] formattable_0.2.1          Formula_1.2-4
## [111] fresh_0.2.0                fs_1.5.0
## [113] future_1.21.0              GA_3.2.1
## [115] gamm4_0.2-6                 gapminder_0.3.0
## [117] gargle_1.2.0                gclus_1.3.2
## [119] gdtools_0.2.3              generics_0.1.0
## [121] geoR_1.8-1                 gert_1.3.1
## [123] GetoptLong_1.0.5            ggalluvial_0.12.3
## [125] gganimate_1.0.7            ggbeeswarm_0.6.0
## [127] ggbump_0.1.99999           ggdendro_0.1.22
## [129] ggfittext_0.9.1            ggfortify_0.4.12
## [131] ggmosaic_0.3.3             ggnormalviolin_0.1.2
## [133] ggplot2_3.3.5              ggpubr_0.4.0
```



```

## [135] ggrepel_0.9.1           ggridges_0.5.3
## [137] ggsci_2.9                ggsignif_0.6.2
## [139] ggstream_0.1.0            gh_1.3.0
## [141] gifski_1.4.3-1           git2r_0.28.0
## [143] gitcreds_0.1.1            glmmTMB_1.1.2
## [145] glmnet_4.1-2              GlobalOptions_0.1.2
## [147] globals_0.14.0             glue_1.4.2
## [149] googledrive_2.0.0          googlesheets4_1.0.0
## [151] graph_1.70.0              gridBase_0.4-7
## [153] gridExtra_2.3              gt_0.3.0
## [155] gtable_0.3.0              gtools_3.9.2
## [157] haven_2.4.3               heatmaply_1.2.1
## [159] hexbin_1.28.2              highcharter_0.8.2
## [161] highr_0.9                 Hmisc_4.5-0
## [163] hms_1.1.0                 hrbrthemes_0.8.0
## [165] htmlTable_2.2.1            htmltools_0.5.1.1
## [167] htmlwidgets_1.5.3          httpuv_1.6.1
## [169] httr_1.4.2                ids_1.0.1
## [171] igraph_1.2.6              influenceR_0.1.0
## [173] ini_0.3.1                INLA_21.02.23
## [175] inline_0.3.19             IRanges_2.26.0
## [177] isoband_0.2.5             iterators_1.0.13
## [179] jpeg_0.1-9                jquerylib_0.1.4
## [181] jsonlite_1.7.2            kableExtra_1.3.4
## [183] Kendall_2.2               kernlab_0.9-29
## [185] KernSmooth_2.23-20         knitr_1.33
## [187] labeling_0.4.2             later_1.2.0
## [189] lattice_0.20-44            latticeExtra_0.6-29
## [191] lazyeval_0.2.2             leaflet_2.0.4.1
## [193] leaflet.extras_1.0.0        leaflet.providers_1.9.0
## [195] leafletCN_0.2.1            lifecycle_1.0.0
## [197] lightgbm_3.2.1             listenv_0.8.0
## [199] lme4_1.1-27.1              loo_2.4.1
## [201] lpSolve_5.6.15             lpSolveAPI_5.5.2.0-17.7
## [203] lubridate_1.7.10            magick_2.7.2
## [205] magrittr_2.0.1              mapdata_2.3.0

```

云
湘
黄
©

```
## [207] mapproj_1.2.7           maps_3.3.0
## [209] maptools_1.1-1            markdown_1.1
## [211] MASS_7.3-54                Matrix_1.3-4
## [213] MatrixModels_0.5-0        matrixStats_0.60.0
## [215] maxLik_1.5-2              mcmc_0.9-7
## [217] memoise_2.0.0             mgcv_1.8-36
## [219] mime_0.11                 miniUI_0.1.1.1
## [221] minqa_1.2.4              miscTools_0.6-26
## [223] modelr_0.1.8             munsell_0.5.0
## [225] mvtnorm_1.1-2            networkD3_0.4
## [227] nleqslv_3.3.2            nlme_3.1-152
## [229] nlmeODE_1.1              nloptr_1.2.2.2
## [231] NMOF_2.4-1                nnet_7.3-16
## [233] nomnoml_0.2.3            numDeriv_2016.8-1.1
## [235] odbc_1.3.2                openssl_1.4.4
## [237] openxlsx_4.2.4            optimx_2021-6.12
## [239] packrat_0.6.0              palmerpenguins_0.1.0
## [241] parallelly_1.27.0          patchwork_1.1.1
## [243] pbkrtest_0.5.1            PBSddesolve_1.12.6
## [245] pdftools_3.0.1            pdist_1.2
## [247] pillar_1.6.2              pkgbuild_1.2.0
## [249] pkgconfig_2.0.3            pkgload_1.2.1
## [251] logr_0.2.0                 plotly_4.9.4.1
## [253] plyr_1.8.6                png_0.1-7
## [255] polynom_1.4-0              posterior_1.0.1
## [257] praise_1.0.0               prettydoc_0.4.1
## [259] prettyunits_1.1.1           PrevMap_1.5.3
## [261] processx_3.5.2             productplots_0.1.1
## [263] progress_1.2.2              projpred_2.0.2
## [265] promises_1.2.0.1            proxy_0.4-26
## [267] ps_1.6.0                   pso_1.0.3
## [269] pspearman_0.3-0            purrr_0.3.4
## [271] pwr_1.3-0                  qap_0.1-1
## [273] qpdf_1.1                   quadprog_1.5-8
## [275] quantmod_0.4.18            quantreg_5.86
## [277] r2d3_0.2.5                 R6_2.5.0
```



```
## [279] RandomFields_3.3.8           RandomFieldsUtils_0.5.3
## [281] randomForest_4.6-14            rappdirs_0.3.3
## [283] raster_3.4-13                 rasterly_0.2.0
## [285] rasterVis_0.50.3              rcmdcheck_1.3.3
## [287] RColorBrewer_1.1-2             Rcpp_1.0.7
## [289] RcppArmadillo_0.10.6.0.0       RcppEigen_0.3.3.9.1
## [291] RcppParallel_5.1.4              reactable_0.2.3
## [293] reactR_0.4.4                  ReacTran_1.4.3.1
## [295] readr_2.0.0                   readxl_1.3.1
## [297] registry_0.5-1                rematch_1.0.1
## [299] rematch2_2.1.2                remotes_2.4.0
## [301] renv_0.14.0                  reprex_2.0.0
## [303] reshape2_1.4.4                reticulate_1.20
## [305] rgdal_1.5-23                 rgeos_0.5-5
## [307] RgoogleMaps_1.4.5.3          Rgraphviz_2.36.0
## [309] rio_0.5.27                  rJava_1.0-4
## [311] rjson_0.2.20                 rlang_0.4.11
## [313] rlist_0.4.6.1                rmarkdown_2.9
## [315] rngtools_1.5                 ROI_1.0-0
## [317] ROI.plugin.alabama_1.0-0     ROI.plugin.lpsolve_1.0-1
## [319] ROI.plugin.nloptr_1.0-0      ROI.plugin.quadprog_1.0-0
## [321] rootSolve_1.8.2.2            roxygen2_7.1.1
## [323] rpart_4.1-15                 rprojroot_2.0.2
## [325] rsconnect_0.8.18              RSQLite_2.2.7
## [327] rstan_2.26.2                rstantools_2.1.1
## [329] rstatix_0.7.0                rstudioapi_0.13
## [331] Rttf2pt1_1.3.9               rversions_2.1.1
## [333] rvest_1.0.1                 s2_1.0.6
## [335] S4Vectors_0.30.0             sandwich_3.0-1
## [337] sass_0.4.0                  scales_1.1.1
## [339] scatterplot3d_0.3-41         selectr_0.4-2
## [341] seriation_1.3.0              servr_0.22
## [343] sessioninfo_1.1.1            sf_1.0-2
## [345] sfarrow_0.4.0                shades_1.4.0
## [347] shape_1.4.6                 shiny_1.6.0
## [349] shinydashboard_0.7.1          shinydashboardPlus_2.0.2
```

```
## [351] shinyjs_2.0.0          shinystan_2.5.0
## [353] shinythemes_1.2.0       shinyWidgets_0.6.0
## [355] showtext_0.9-3          showtextdb_3.0
## [357] Sim.DiffProc_4.8         slam_0.1-48
## [359] sm_2.2-5.6              sourcetools_0.1.7
## [361] sp_1.4-5                sparkline_2.0
## [363] sparklyr_1.7.1         SparseM_1.81
## [365] spatial_7.3-14          spDataLarge_0.5.4
## [367] splancs_2.01-42         splines2_0.4.3
## [369] StanHeaders_2.26.2      stringi_1.7.3
## [371] stringr_1.4.0            SuppDists_1.1-9.5
## [373] survival_3.2-11         svglite_2.0.0
## [375] symengine_0.1.5          sys_3.4
## [377] sysfonts_0.8.4           systemfonts_1.0.2
## [379] tensorA_0.36.2          tensorflow_2.5.0
## [381] terra_1.3-4              testthat_3.0.4
## [383] tfruns_1.5.0             threejs_0.3.3
## [385] tibble_3.1.3             tidyverse_1.3.1
## [387] tidyselect_1.1.1          timeline_0.9
## [389] tikzDevice_0.12.3.1      tint_0.1.3
## [391] timelineS_0.1.1          TMB_1.7.20
## [393] tinytex_0.32              treemap_2.4-2
## [395] transformr_0.1.3         truncnorm_1.0-8
## [397] treemapify_2.5.5         TTR_0.24.2
## [399] TSP_1.1-10              tzdb_0.1.2
## [401] tweenr_1.0.2             usethis_2.0.1
## [403] units_0.7-2              uuid_0.1-4
## [405] utf8_1.2.2               vctrs_0.3.8
## [407] V8_3.4.2                 viper_0.4.5
## [409] vioplot_0.3.7            viridisLite_0.4.0
## [411] viridis_0.6.1             vistime_1.2.1
## [413] visNetwork_2.0.9          waiter_0.2.3
## [415] vroom_1.5.3              webshot_0.5.2
## [417] waldo_0.2.5              withr_2.4.2
## [419] whisker_0.4               xaringan_0.22
```



```

## [423] xaringantheme_0.4.0      xfun_0.24
## [425] xgboost_1.4.1.1          xkcd_0.0.6
## [427] XML_3.99-0.6             xml2_1.3.2
## [429] xopen_1.0.0                xtable_1.8-4
## [431] xts_0.12.1               yaml_2.2.1
## [433] zip_2.2.0                 zoo_1.8-9

library(magrittr)
pdb <- tools::CRAN_package_db()
pkg <- subset(desc::desc_get_deps(), subset = type == "Imports", select = "package", dr
pkg <- tools::package_dependencies(packages = pkg, db = pdb, recursive = FALSE) %>% # 是
  unlist() %>%
  as.vector() %>%
  c(., pkg) %>%
  unique() %>%
  sort()

pkg_quote <- c(
  "Armadillo", "Rcpp", "R", "Stan", "DataTables", "Dygraphs", "ggplot2",
  "Grobs", "Geospatial", "Eigen", "Sundown", "plog", "TeX Live", "Tidyverse",
  "LaTeX", "ADMB", "matplotlib", "Yihui Xie", "With", "Highcharts",
  "kable", "plotly.js", "Python", "Formattable"
)
# 单引号
pkg-regexp <- paste("'"(, paste(pkg_quote, collapse = "|"), ")'", sep = "")
# R 包列表
subset(pdb,
  subset = !duplicated(pdb$Package) & Package %in% pkg,
  select = c("Package", "Version", "Title")
) %>%
  transform(.,
    Title = gsub("\\\\n", " ", Title),
    Package = paste("**", Package, "**", sep = ""))
) %>%
  transform(., Title = gsub(pkg-regexp, "\\\\$1", Title)) %>%
  transform(., Title = gsub('"(Grid)"', "\\\\$1", Title)) %>%
knitr::kable(.,
```



```
caption = "依赖的 R 包", format = "pandoc",
booktabs = TRUE, row.names = FALSE
)
```

表 B.2: 依赖的 R 包

Package	Version	Title
abind	1.4-5	Combine Multidimensional Arrays
agridat	1.18	Agricultural Datasets
alabama	2015.3-1	Constrained Nonlinear Optimization
arrow	4.0.1	Integration to ‘Apache’ ‘Arrow’
arules	1.6-8	Mining Association Rules and Frequent Itemsets
assertive.types	0.0-3	Assertions to Check Types of Variables
assertthat	0.2.1	Easy Pre and Post Assertions
autoplotty	0.1.4	Automatic Generation of Interactive Visualizations for S
backports	1.2.1	Reimplementations of Functions Introduced Since R-3.0.
base64enc	0.1-3	Tools for base64 encoding
bayesplot	1.8.1	Plotting for Bayesian Models
beanplot	1.2	Visualization via Beanplots (like Boxplot/Stripchart/Violin
beeswarm	0.4.0	The Bee Swarm Plot, an Alternative to Stripchart
BH	1.75.0-0	Boost C++ Header Files
BiocManager	1.30.16	Access the Bioconductor Project Package Repository
bit64	4.0.5	A S3 Class for Vectors of 64bit Integers
bitops	1.0-7	Bitwise Operations
blob	1.2.1	A Simple S3 Class for Representing Vectors of Binary Data
bookdown	0.22	Authoring Books and Technical Documents with R Markdown
boot	1.3-28	Bootstrap Functions (Originally by Angelo Canty for S)
bridgesampling	1.1-2	Bridge Sampling for Marginal Likelihoods and Bayes Factor
brms	2.15.0	Bayesian Regression Models using Stan
broom	0.7.8	Convert Statistical Objects into Tidy Tibbles
broom.mixed	0.2.7	Tidying Methods for Mixed Models
bslib	0.2.5.1	Custom ‘Bootstrap’ ‘Sass’ Themes for ‘shiny’ and ‘rmarkd
cachem	1.0.5	Cache R Objects with Automatic Pruning
callr	3.7.0	Call R from R
checkmate	2.0.0	Fast and Versatile Argument Checks
classInt	0.4-3	Choose Univariate Class Intervals



Package	Version	Title
cli	3.0.1	Helpers for Developing Command Line Interfaces
coda	0.19-4	Output Analysis and Diagnostics for MCMC
colorspace	2.0-2	A Toolbox for Manipulating and Assessing Colors and Palettes
commonmark	1.7	High Performance CommonMark and Github Markdown Rend
config	0.3.1	Manage Environment Specific Configuration Values
conquer	1.0.2	Convolution-Type Smoothed Quantile Regression
corrplot	0.90	Visualization of a Correlation Matrix
countrycode	1.3.0	Convert Country Names and Country Codes
cowplot	1.1.1	Streamlined Plot Theme and Plot Annotations for ggplot2
cpp11	0.3.1	A C++11 Interface for R's C Interface
crayon	1.4.1	Colored Terminal Output
crosstalk	1.1.1	Inter-Widget Interactivity for HTML Widgets
curl	4.3.2	A Modern and Flexible Web Client for R
data.table	1.14.0	Extension of <code>data.frame</code>
DBI	1.1.1	R Database Interface
dbplyr	2.1.1	A 'dplyr' Back End for Databases
dendextend	1.15.1	Extending 'dendrogram' Functionality in R
Deriv	4.1.3	Symbolic Differentiation
desc	1.3.0	Manipulate DESCRIPTION Files
deSolve	1.28	Solvers for Initial Value Problems of Differential Equations ('O
devtools	2.4.2	Tools to Make Developing R Packages Easier
DiagrammeR	1.0.6.1	Graph/Network Visualization
digest	0.6.27	Create Compact Hash Digests of R Objects
downloader	0.4	Download Files over HTTP and HTTPS
dplyr	1.0.7	A Grammar of Data Manipulation
DT	0.18	A Wrapper of the JavaScript Library DataTables
dtplyr	1.1.0	Data Table Back-End for 'dplyr'
echarts4r	0.4.1	Create Interactive Graphs with 'Echarts JavaScript' Version 5
egg	0.4.5	Extensions for ggplot2: Custom Geom, Custom Themes, Plot A
ellipsis	0.3.2	Tools for Working with ...
equatiomatic	0.2.0	Transform Models into LaTeX Equations
evaluate	0.14	Parsing and Evaluation Tools that Provide More Details than th
extrafont	0.17	Tools for using fonts
extrafontdb	1.0	Package for holding the database for the extrafont package
fastmap	1.1.0	Fast Data Structures



Package	Version	Title
filehash	2.4-2	Simple Key-Value Database
fontcm	1.1	Computer Modern font for use with extrafont package
forcats	0.5.1	Tools for Working with Categorical Variables (Factors)
foreach	1.5.1	Provides Foreach Looping Construct
forge	0.2.0	Casting Values into Shape
formatR	1.11	Format R Code Automatically
fs	1.5.0	Cross-Platform File System Operations Based on 'libuv'
future	1.21.0	Unified Parallel and Distributed Processing in R for Every
gapminder	0.3.0	Data from Gapminder
gdtools	0.2.3	Utilities for Graphical Rendering
generics	0.1.0	Common S3 Generics not Provided by Base R Methods R
geoR	1.8-1	Analysis of Geostatistical Data
ggalluvial	0.12.3	Alluvial Plots in ggplot2
gganimate	1.0.7	A Grammar of Animated Graphics
ggbeeswarm	0.6.0	Categorical Scatter (Violin Point) Plots
ggbump	0.1.0	Bump Chart and Sigmoid Curves
ggfittext	0.9.1	Fit Text Inside a Box in ggplot2
ggfortify	0.4.12	Data Visualization Tools for Statistical Analysis Results
ggmosaic	0.3.3	Mosaic Plots in the ggplot2 Framework
ggnormalviolin	0.1.2	A ggplot2 Extension to Make Normal Violin Plots
ggplot2	3.3.5	Create Elegant Data Visualisations Using the Grammar o
ggpubr	0.4.0	ggplot2 Based Publication Ready Plots
ggrepel	0.9.1	Automatically Position Non-Overlapping Text Labels with
ggridges	0.5.3	Ridgeline Plots in ggplot2
ggsci	2.9	Scientific Journal and Sci-Fi Themed Color Palettes for g
ggsignif	0.6.2	Significance Brackets for ggplot2
ggstream	0.1.0	Create Streamplots in ggplot2
gifski	1.4.3-1	Highest Quality GIF Encoder
git2r	0.28.0	Provides Access to Git Repositories
glmnet	4.1-2	Lasso and Elastic-Net Regularized Generalized Linear M
globals	0.14.0	Identify Global Objects in R Expressions
glue	1.4.2	Interpreted String Literals
googledrive	2.0.0	An Interface to Google Drive
googlesheets4	0.3.0	Access Google Sheets using the Sheets API V4
gridBase	0.4-7	Integration of base and grid graphics



Package	Version	Title
gridExtra	2.3	Miscellaneous Functions for Grid Graphics
gt	0.3.0	Easily Create Presentation-Ready Display Tables
gttable	0.3.0	Arrange Grobs in Tables
haven	2.4.1	Import and Export ‘SPSS’, ‘Stata’ and ‘SAS’ Files
heatmaps	1.2.1	Interactive Cluster Heat Maps Using ‘plotly’ and ggplot2
hexbin	1.28.2	Hexagonal Binning Routines
highcharter	0.8.2	A Wrapper for the Highcharts Library
highr	0.9	Syntax Highlighting for R Source Code
Hmisc	4.5-0	Harrell Miscellaneous
hms	1.1.0	Pretty Time of Day
hrbrthemes	0.8.0	Additional Themes, Theme Components and Utilities for ggplot2
htmltools	0.5.1.1	Tools for HTML
htmlwidgets	1.5.3	HTML Widgets for R
httpuv	1.6.1	HTTP and WebSocket Server Library
httr	1.4.2	Tools for Working with URLs and HTTP
igraph	1.2.6	Network Analysis and Visualization
influenceR	0.1.0	Software Tools to Quantify Structural Importance of Nodes in a Network
inline	0.3.19	Functions to Inline C, C++, Fortran Function Calls from R
isoband	0.2.5	Generate Isolines and Isobands from Regularly Spaced Elevation Data
jsonlite	1.7.2	A Simple and Robust JSON Parser and Generator for R
kableExtra	1.3.4	Construct Complex Table with kable and Pipe Syntax
Kendall	2.2	Kendall rank correlation and Mann-Kendall trend test
knitr	1.33	A General-Purpose Package for Dynamic Report Generation in R
later	1.2.0	Utilities for Scheduling Functions to Execute Later with Event Loops
lattice	0.20-44	Trellis Graphics for R
latticeExtra	0.6-29	Extra Graphical Utilities Based on Lattice
lazyeval	0.2.2	Lazy (Non-Standard) Evaluation
leaflet	2.0.4.1	Create Interactive Web Maps with the JavaScript ‘Leaflet’ Library
leaflet.providers	1.9.0	Leaflet Providers
lifecycle	1.0.0	Manage the Life Cycle of your Package Functions
lightgbm	3.2.1	Light Gradient Boosting Machine
lme4	1.1-27.1	Linear Mixed-Effects Models using Eigen and S4
loo	2.4.1	Efficient Leave-One-Out Cross-Validation and WAIC for Bayesian Models
lpSolve	5.6.15	Interface to ‘Lp_solve’ v. 5.5 to Solve Linear/Integer Programs
lpSolveAPI	5.5.2.0-17.7	R Interface to ‘lp_solve’ Version 5.5.2.0



Package	Version	Title
lubridate	1.7.10	Make Dealing with Dates a Little Easier
magick	2.7.2	Advanced Graphics and Image-Processing in R
magrittr	2.0.1	A Forward-Pipe Operator for R
mapdata	2.3.0	Extra Map Databases
mapproj	1.2.7	Map Projections
maps	3.3.0	Draw Geographical Maps
markdown	1.1	Render Markdown with the C Library Sundown
MASS	7.3-54	Support Functions and Datasets for Venables and Ripley
Matrix	1.3-4	Sparse and Dense Matrix Classes and Methods
MatrixModels	0.5-0	Modelling with Sparse and Dense Matrices
matrixStats	0.59.0	Functions that Apply to Rows and Columns of Matrices (and Arrays)
maxLik	1.4-8	Maximum Likelihood Estimation and Related Tools
mcmc	0.9-7	Markov Chain Monte Carlo
memoise	2.0.0	Memoisation of Functions
mgcv	1.8-36	Mixed GAM Computation Vehicle with Automatic Smoothing Parameter Selection
mime	0.11	Map Filenames to MIME Types
minqa	1.2.4	Derivative-free optimization algorithms by quadratic approximation
modelr	0.1.8	Modelling Functions that Work with the Pipe
mvtnorm	1.1-2	Multivariate Normal and t Distributions
networkD3	0.4	D3 JavaScript Network Graphs from R
nleqslv	3.3.2	Solve Systems of Nonlinear Equations
nlme	3.1-152	Linear and Nonlinear Mixed Effects Models
nloptr	1.2.2.2	R Interface to NLOpt
nomnoml	0.2.3	Sassy ‘UML’ Diagrams
numDeriv	2016.8-1.1	Accurate Numerical Derivatives
odbc	1.3.2	Connect to ODBC Compatible Databases (using the DBI Interface)
openssl	1.4.4	Toolkit for Encryption, Signatures and Certificates Based on OpenSSL
palmerpenguins	0.1.0	Palmer Archipelago (Antarctica) Penguin Data
patchwork	1.1.1	The Composer of Plots
pdftools	3.0.1	Text Extraction, Rendering and Converting of PDF Documents
pdist	1.2	Partitioned Distance Function
pillar	1.6.1	Coloured Formatting for Columns
pkgbuild	1.2.0	Find Tools Needed to Build R Packages
pkgconfig	2.0.3	Private Configuration for R Packages
pkgload	1.2.1	Simulate Package Installation and Attach



Package	Version	Title
plogr	0.2.0	The plog C++ Logging Library
plotly	4.9.4.1	Create Interactive Web Graphics via plotly.js
plyr	1.8.6	Tools for Splitting, Applying and Combining Data
png	0.1-7	Read and write PNG images
polynom	1.4-0	A Collection of Functions to Implement a Class for Univariate Polynomial Regression
prettydoc	0.4.1	Creating Pretty Documents from R Markdown
PrevMap	1.5.3	Geostatistical Modelling of Spatially Referenced Prevalence Data
processx	3.5.2	Execute and Control System Processes
productplots	0.1.1	Product Plots for R
progress	1.2.2	Terminal Progress Bars
projpred	2.0.2	Projection Predictive Feature Selection
promises	1.2.0.1	Abstractions for Promise-Based Asynchronous Programming
pspearman	0.3-0	Spearman's rank correlation test
purrr	0.3.4	Functional Programming Tools
pwr	1.3-0	Basic Functions for Power Analysis
qpdf	1.1	Split, Combine and Compress PDF Files
quadprog	1.5-8	Functions to Solve Quadratic Programming Problems
quantmod	0.4.18	Quantitative Financial Modelling Framework
quantreg	5.86	Quantile Regression
r2d3	0.2.5	Interface to 'D3' Visualizations
R6	2.5.0	Encapsulated Classes with Reference Semantics
RandomFields	3.3.8	Simulation and Analysis of Random Fields
rappdirs	0.3.3	Application Directories: Determine Where to Save Data, Cache Files
raster	3.4-13	Geographic Data Analysis and Modeling
rasterly	0.2.0	Easily and Rapidly Generate Raster Image Data with Support for Parallel Processing
rasterVis	0.50.2	Visualization Methods for Raster Data
rcmdcheck	1.3.3	Run 'R CMD check' from R and Capture Results
RColorBrewer	1.1-2	ColorBrewer Palettes
Rcpp	1.0.7	Seamless R and C++ Integration
RcppEigen	0.3.3.9.1	Rcpp Integration for the Eigen Templated Linear Algebra Library
RcppParallel	5.1.4	Parallel Programming Tools for Rcpp
reactable	0.2.3	Interactive Data Tables Based on 'React Table'
reactR	0.4.4	React Helpers
readr	1.4.0	Read Rectangular Text Data
readxl	1.3.1	Read Excel Files



Package	Version	Title
registry	0.5-1	Infrastructure for R Package Registries
remotes	2.4.0	R Package Installation from Remote Repositories, Including CRAN
reprex	2.0.0	Prepare Reproducible Example Code via the Clipboard
reshape2	1.4.4	Flexibly Reshape Data: A Reboot of the Reshape Package
reticulate	1.20	Interface to Python
rgdal	1.5-23	Bindings for the Geospatial Data Abstraction Library
rjson	0.2.20	JSON for R
rlang	0.4.11	Functions for Base Types and Core R and Tidyverse Features
rlist	0.4.6.1	A Toolbox for Non-Tabular Data Manipulation
rmarkdown	2.9	Dynamic Documents for R
ROI	1.0-0	R Optimization Infrastructure
ROI.plugin.alabama	1.0-0	‘alabama’ Plug-in for the R Optimization Infrastructure
ROI.plugin.lpsolve	1.0-1	‘lp_solve’ Plugin for the R Optimization Infrastructure
ROI.plugin.nloptr	1.0-0	‘nloptr’ Plug-in for the R Optimization Infrastructure
ROI.plugin.quadprog	1.0-0	‘quadprog’ Plug-in for the R Optimization Infrastructure
rootSolve	1.8.2.2	Nonlinear Root Finding, Equilibrium and Steady-State Analysis
roxygen2	7.1.1	In-Line Documentation for R
rprojroot	2.0.2	Finding Files in Project Subdirectories
RSQLite	2.2.7	‘SQLite’ Interface for R
rstan	2.21.2	R Interface to Stan
rstantools	2.1.1	Tools for Developing R Packages Interfacing with Stan
rstatix	0.7.0	Pipe-Friendly Framework for Basic Statistical Tests
rstudioapi	0.13	Safely Access the RStudio API
Rttf2pt1	1.3.8	‘ttf2pt1’ Program
rversions	2.1.1	Query R Versions, Including ‘r-release’ and ‘r-oldrel’
rvest	1.0.0	Easily Harvest (Scrape) Web Pages
s2	1.0.6	Spherical Geometry Operators Using the S2 Geometry Library
sass	0.4.0	Syntactically Awesome Style Sheets (“Sass”)
scales	1.1.1	Scale Functions for Visualization
scatterplot3d	0.3-41	3D Scatter Plot
seriation	1.3.0	Infrastructure for Ordering Objects Using Seriation
sessioninfo	1.1.1	R Session Information
sf	1.0-1	Simple Features for R
shape	1.4.6	Functions for Plotting Graphical Shapes, Colors
shiny	1.6.0	Web Application Framework for R



Package	Version	Title
shinystan	2.5.0	Interactive Visual and Numerical Diagnostics and Posterior Analysis for Stan
showtext	0.9-2	Using Fonts More Easily in R Graphs
showtextdb	3.0	Font Files for the ‘showtext’ Package
slam	0.1-48	Sparse Lightweight Arrays and Matrices
sm	2.2-5.6	Smoothing Methods for Nonparametric Regression and Density Estimation
sourcetools	0.1.7	Tools for Reading, Tokenizing and Parsing R Code
sp	1.4-5	Classes and Methods for Spatial Data
sparkline	2.0	‘jQuery’ Sparkline ‘htmlwidget’
sparklyr	1.7.1	R Interface to Apache Spark
SparseM	1.81	Sparse Linear Algebra
splancs	2.01-42	Spatial and Space-Time Point Pattern Analysis
StanHeaders	2.21.0-7	C++ Header Files for Stan
stringi	1.7.3	Character String Processing Facilities
stringr	1.4.0	Simple, Consistent Wrappers for Common String Operations
SuppDists	1.1-9.5	Supplementary Distributions
survival	3.2-11	Survival Analysis
svglite	2.0.0	An ‘SVG’ Graphics Device
symengine	0.1.5	Interface to the ‘SymEngine’ Library
sysfonts	0.8.3	Loading Fonts into R
tensorflow	2.5.0	R Interface to ‘TensorFlow’
terra	1.3-4	Spatial Data Analysis
testthat	3.0.4	Unit Testing for R
tfruns	1.5.0	Training Run Tools for ‘TensorFlow’
tibble	3.1.2	Simple Data Frames
tidyverse	1.1.3	Tidy Messy Data
tidyselect	1.1.1	Select from a Set of Strings
tidyverse	1.3.1	Easily Install and Load the Tidyverse
tikzDevice	0.12.3.1	R Graphics Output in LaTeX Format
timeline	0.9	Timelines for a Grammar of Graphics
timelineS	0.1.1	Timeline and Time Duration-Related Tools
tint	0.1.3	‘tint’ is not ‘Tufte’
tinytex	0.32	Helper Functions to Install and Maintain TeX Live, and Compile PDFs
transformr	0.1.3	Polygon and Path Transformations
treemap	2.4-2	Treemap Visualization
treemapify	2.5.5	Draw Treemaps in ggplot2



Package	Version	Title
truncnorm	1.0-8	Truncated Normal Distribution
TSP	1.1-10	Traveling Salesperson Problem (TSP)
tweenr	1.0.2	Interpolate Data for Smooth Animations
units	0.7-2	Measurement Units for R Vectors
usethis	2.0.1	Automate Package and Project Setup
uuid	0.1-4	Tools for Generating and Handling of UUIDs
V8	3.4.2	Embedded JavaScript and WebAssembly Engine for R
vctrs	0.3.8	Vector Helpers
vioplot	0.3.6	Violin Plot
vipor	0.4.5	Plot Categorical Data Using Quasirandom Noise and Densities
viridis	0.6.1	Colorblind-Friendly Color Maps for R
viridisLite	0.4.0	Colorblind-Friendly Color Maps (Lite Version)
visNetwork	2.0.9	Network Visualization using ‘vis.js’ Library
vistime	1.2.1	Pretty Timelines in R
webshot	0.5.2	Take Screenshots of Web Pages
withr	2.4.2	Run Code With Temporarily Modified Global State
xfun	0.24	Supporting Functions for Packages Maintained by Yihui Xie
xgboost	1.4.1.1	Extreme Gradient Boosting
xkcd	0.0.6	Plotting ggplot2 Graphics in an XKCD Style
xml2	1.3.2	Parse XML
xtable	1.8-4	Export Tables to LaTeX or HTML
xts	0.12.1	eXtensible Time Series
yaml	2.2.1	Methods to Convert R Data to YAML and Back
zoo	1.8-9	S3 Infrastructure for Regular and Irregular Time Series

提示

本书意欲覆盖的内容

```
inla_pdb <- data.frame(  
  Package = "INLA",  
  Title = paste(  
    "Full Bayesian Analysis of Latent Gaussian Models",  
    "using Integrated Nested Laplace Approximations"  
)  
)
```



```
pkgs <- c(  
  "ggplot2", "cowplot", "patchwork", "rgl", "MASS", "nlme", "mgcv",  
  "lme4", "gee", "gam", "gamm4", "cgam", "cglm", "pscl",  
  "GLMMadaptive", "gee4", "geor", "LaplacesDemon", "glmnet",  
  "betareg", "quantreg", "agridat", "moments", "R2BayesX",  
  "geoRglm", "spaMM", "spBayes", "CARBayes", "PrevMap",  
  "FRK", "lgcp", "HSAR", "spNNGP", "MuMin", "BANOVA",  
  "rpql", "QGglmm", "glmmsr", "glmmboot", "glmm",  
  "glmmML", "glmmEP", "r2glmm", "hglm", "glmmLasso",  
  "blme", "MCMCglmm", "MCMCpack", "glmmTMB", "geepack",  
  "glmmfields", "rstan", "rstanarm", "brms", "greta",  
  "BayesX", "Boom", "nimble", "rjags", "R2OpenBUGS",  
  "R2BayesX", "BoomSpikeSlab", "inlabru", "INLABMA",  
  "lmtest", "VGAM", "plotly", "leaflet", "LatticeKrig"  
)  
pdb <- tools::CRAN_package_db()  
book_pdb <- subset(pdb,  
  subset = !duplicated(pdb$Package) & Package %in% pkgs,  
  select = c("Package", "Title")  
)  
book_pdb <- rbind.data.frame(book_pdb, inla_pdb)  
book_pdb>Title <- gsub("\\\\n", " ", book_pdb>Title)  
book_pdb>Title <- gsub("(Armadillo|BayesX|Eigen|ggplot2|lme4|mgcv|Stan|Leaflet|plotly.",  
book_pdb$Package <- paste("**", book_pdb$Package, "**", sep = "")  
knitr::kable(book_pdb,  
  caption = "本书使用的 R 包", format = "pandoc",  
  booktabs = TRUE, row.names = FALSE  
)
```

表 B.3: 本书使用的 R 包

Package	Title
agridat	Agricultural Datasets
BANOVA	Hierarchical Bayesian ANOVA Models
BayesX	R Utilities Accompanying the Software Package BayesX
betareg	Beta Regression



Package	Title
blme	Bayesian Linear Mixed-Effects Models
Boom	Bayesian Object Oriented Modeling
BoomSpikeSlab	MCMC for Spike and Slab Regression
brms	Bayesian Regression Models using Stan
CARBayes	Spatial Generalised Linear Mixed Models for Areal Unit Data
cgam	Constrained Generalized Additive Model
cglm	Fits Conditional Generalized Linear Models
cowplot	Streamlined Plot Theme and Plot Annotations for ggplot2
FRK	Fixed Rank Kriging
gam	Generalized Additive Models
gamm4	Generalized Additive Mixed Models using mgcv and lme4
gee	Generalized Estimation Equation Solver
geepack	Generalized Estimating Equation Package
geoR	Analysis of Geostatistical Data
ggplot2	Create Elegant Data Visualisations Using the Grammar of Graphics
glmm	Generalized Linear Mixed Models via Monte Carlo Likelihood Approximation
GLMMadaptive	Generalized Linear Mixed Models using Adaptive Gaussian Quadrature
glmmboot	Bootstrap Resampling for Mixed Effects and Plain Models
glmmEP	Generalized Linear Mixed Model Analysis via Expectation Propagation
glmmfields	Generalized Linear Mixed Models with Robust Random Fields for Spatiotemporal Data
glmmLasso	Variable Selection for Generalized Linear Mixed Models by L1-Penalized Estimation
glmmML	Generalized Linear Models with Clustering
glmmssr	Fit a Generalized Linear Mixed Model
glmnet	Lasso and Elastic-Net Regularized Generalized Linear Models
greta	Simple and Scalable Statistical Modelling in R
hglm	Hierarchical Generalized Linear Models
HSAR	Hierarchical Spatial Autoregressive Model
INLABMA	Bayesian Model Averaging with INLA
inlabru	Bayesian Latent Gaussian Modelling using INLA and Extensions
LaplacesDemon	Complete Environment for Bayesian Inference
LatticeKrig	Multi-Resolution Kriging Based on Markov Random Fields
leaflet	Create Interactive Web Maps with the JavaScript Leaflet Library
lgcp	Log-Gaussian Cox Process
lme4	Linear Mixed-Effects Models using Eigen and S4
lmtest	Testing Linear Regression Models



Package	Title
MASS	Support Functions and Datasets for Venables and Ripley's MASS
MCMCglmm	MCMC Generalised Linear Mixed Models
MCMCpack	Markov Chain Monte Carlo (MCMC) Package
mgcv	Mixed GAM Computation Vehicle with Automatic Smoothness Estimation
moments	Moments, cumulants, skewness, kurtosis and related tests
MuMin	Multi-Model Inference
nimble	MCMC, Particle Filtering, and Programmable Hierarchical Modeling
nlme	Linear and Nonlinear Mixed Effects Models
patchwork	The Composer of Plots
plotly	Create Interactive Web Graphics via plotly.js
PrevMap	Geostatistical Modelling of Spatially Referenced Prevalence Data
pscl	Political Science Computational Laboratory
QGglmm	Estimate Quantitative Genetics Parameters from Generalised Linear Mixed Model
quantreg	Quantile Regression
R2BayesX	Estimate Structured Additive Regression Models with BayesX
r2glmm	Computes R Squared for Mixed (Multilevel) Models
R2OpenBUGS	Running OpenBUGS from R
rgl	3D Visualization Using OpenGL
rjags	Bayesian Graphical Models using MCMC
rpql	Regularized PQL for Joint Selection in GLMMs
rstan	R Interface to Stan
rstanarm	Bayesian Applied Regression Modeling via Stan
spaMM	Mixed-Effect Models, Particularly Spatial Models
spBayes	Univariate and Multivariate Spatial-Temporal Modeling
spNNGP	Spatial Regression Models for Large Datasets using Nearest Neighbor Gaussian Process
VGAM	Vector Generalized Linear and Additive Models
INLA	Full Bayesian Analysis of Latent Gaussian Models using Integrated Nested Laplace Approximation

附录 C 混合编程

R 语言 [Ihaka and Gentleman, 1996] 是一个统计计算和绘图的环境，以下各个节不介绍具体 R 包函数用法和参数设置，重点在历史发展趋势脉络，详细介绍去见《现代统计图形》的相应章节。R 语言的目标在于统计计算和绘图，设计优势在数据结构、图形语法、动态文档和交互图形

C.1 函数源码

`funflow` 包可以将函数调用的过程以流程图的方式呈现，代码结构一目了然，快速理清源代码

```
remotes::install_github('moodymudskipper/funflow')
funflow::view_flow('median.default')

methods(predict)

## [1] predict.ar*                  predict.Arima*
## [3] predict.arima0*              predict.glm
## [5] predict.HoltWinters*        predict.lm
## [7] predict.loess*               predict.mlm*
## [9] predict.nls*                 predict.poly*
## [11] predict.ppr*                predict.prcomp*
## [13] predict.princomp*           predict.smooth.spline*
## [15] predict.smooth.spline.fit* predict.StructTS*
## see '?methods' for accessing help and source code
```

stats 包里找不到这个函数



```
ls("package:stats", all.names = TRUE, pattern = "predict.poly")
## character(0)
predict.poly
## Error in eval(expr, envir, enclos): object 'predict.poly' not found
```

可见函数 predict.poly() 默认没有导出

```
stats:::predict.poly
```

```
## function (object, newdata, ...)
## {
##   if (missing(newdata))
##     object
##   else if (is.null(attr(object, "coefs")))
##     poly(newdata, degree = max(attr(object, "degree")), raw = TRUE,
##          simple = TRUE)
##   else poly(newdata, degree = max(attr(object, "degree")),
##            coefs = attr(object, "coefs"), simple = TRUE)
## }
## <bytecode: 0x55a97d07f158>
## <environment: namespace:stats>
```

或者

```
getAnywhere(predict.poly)

## A single object matching 'predict.poly' was found
## It was found in the following places
##   registered S3 method for predict from namespace stats
##   namespace:stats
## with value
##
## function (object, newdata, ...)
## {
##   if (missing(newdata))
##     object
##   else if (is.null(attr(object, "coefs")))
##     poly(newdata, degree = max(attr(object, "degree")), raw = TRUE,
```



```
##           simple = TRUE)
##   else poly(newdata, degree = max(attr(object, "degree")),
##           coefs = attr(object, "coefs"), simple = TRUE)
## }
## <bytecode: 0x55a97d07f158>
## <environment: namespace:stats>
getAnywhere("predict.poly")$where

## [1] "registered S3 method for predict from namespace stats"
## [2] "namespace:stats"
```

函数参数个数

```
names(formals(read.table))
```

```
## [1] "file"          "header"        "sep"          "quote"
## [5] "dec"           "numerals"       "row.names"     "col.names"
## [9] "as.is"          "na.strings"     "colClasses"    "nrows"
## [13] "skip"           "check.names"    "fill"          "strip.white"
## [17] "blank.lines.skip" "comment.char"  "allowEscapes"  "flush"
## [21] "stringsAsFactors" "fileEncoding"  "encoding"      "text"
## [25] "skipNul"
```

C.2 命名约定

R 语言当前的命名状态 https://journal.r-project.org/archive/2012-2/RJournal_2012-2_Baaaath.pdf 和 <https://essentials.togaware.com/StyleO.pdf>

R 与不同的编程语言如何交互

C.3 R 与 JavaScripts

```
library(htmlwidgets)
```

C.4 R 与 Python

R 包 knitr 和 reticulate 支持 R Markdown 文档中嵌入 Python 代码块，reticulate 包还支持 Python 和 R 之间的数据对象通信交流。

```
library(reticulate)
```

如图 C.1 所示，在 R Markdown 中执行 Python 绘图代码，并且将图形插入文档。

```
import matplotlib.pyplot as plt
plt.switch_backend('agg')

plt.plot([0, 2, 1, 4])
## [<matplotlib.lines.Line2D object at 0x7ff0b9103460>]
plt.show()
```

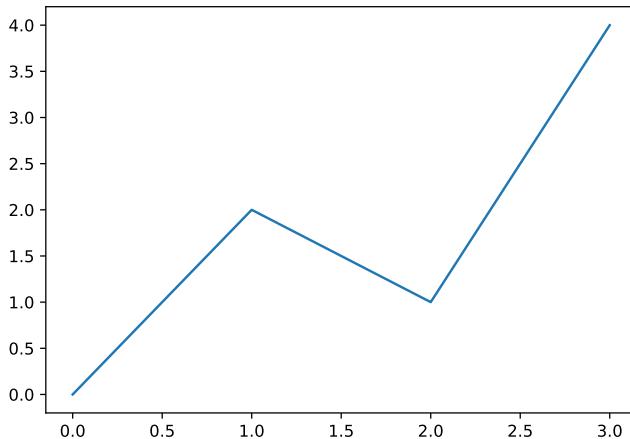


图 C.1: Python 图形

C.5 R 与 C

knitr 支持在 R Markdown 中嵌入 C 语言代码

```
void useC(int *i){
    i[0] = 11;
```



```
}
```

```
## make[1]: Entering directory '/home/runner/work/masr/masr'
## gcc -I"/opt/R/4.1.0/lib/R/include" -DNDEBUG -I/usr/local/include -fpic -g -O2
## gcc -shared -L/opt/R/4.1.0/lib/R/lib -L/usr/local/lib -o c6035490a9e65.so c603549
## make[1]: Leaving directory '/home/runner/work/masr/masr'
a <- rep(2,10)
out <- .C("useC", b = as.integer(a))
out
```

```
## $b
## [1] 11 2 2 2 2 2 2 2 2 2
```

```
out$b
```

```
## [1] 11 2 2 2 2 2 2 2 2 2
```

一步一步地命令行操作

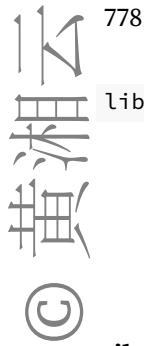
```
R CMD SHLIB useC1.c
```

```
dyn.load("useC1.dll")
a <- rep(2,10)
out <- .C("useC", b = as.integer(a))
out$b
```

C.6 R 与 C++

Dirk Eddelbuettel 是 Rcpp 的核心开发者。

- Dirk Eddelbuettel celebRtion 2020, Copenhagen, Denmark [Introduction to Rcpp: from simple examples to machine learning](#)
- Online Tutorial for useR! 2020 [Seamless R and C++ Introduction with Rcpp](#) 视频 <https://vimeo.com/438283959>
- James Balamuta [unofficial rcpp api documentation](#) <https://github.com/coatless/rcpp-api>
- Rcpp for everyone https://github.com/teuder/rcpp4everyone_en
- 课程 [Foundations of Data Science](#)



```
library(Rcpp)
```

C.7 R 与 LaTeX

tikzDevice 包将 LaTeX 公式和绘图系统 [TikZ](#) 引入 R 语言生态，贡献在于提供更加漂亮的公式输出，对图形进行后期布局排版加工，达到设计师出品的质量水平。图 C.2 展示了复杂的 TeX 生态系统，R 语言只是取其精华，使用 TikZ 绘制。

```
\begin{tikzpicture}
  \path [
    mindmap,
    text = white,
    level 1 concept/.append style =
      {font=\Large\bfseries\sffamily, sibling angle=90, level distance=125},
    level 2 concept/.append style =
      {font=\normalsize\bfseries\sffamily},
    level 3 concept/.append style =
      {font=\small\bfseries\sffamily},
    tex/.style = {concept, ball color=blue,
      font=\Huge\bfseries},
    engines/.style = {concept, ball color=green!50!black},
    formats/.style = {concept, ball color=purple!50!black},
    systems/.style = {concept, ball color=red!90!black},
    editors/.style = {concept, ball color=orange!90!black}
  ]
  node [tex] {\TeX} [clockwise from=0]
    child[concept color=green!50!black, nodes={engines}] {
      node {Engines} [clockwise from=90]
        child { node {\TeX} }
        child { node {pdf\TeX} }
        child { node {Xe\TeX} }
        child { node {Lua\TeX} }}
      child [concept color=purple, nodes={formats}] {
        node {Formats} [clockwise from=300]
          child { node {\LaTeX} }
```

```
    child { node {Con\TeX t} }}
```

```
child [concept color=red, nodes={systems}] {
```

```
    node {Systems} [clockwise from=210]
```

```
    child { node {\TeX Live} [clockwise from=300]}
```

```
        child { node {Mac \TeX} }}
```

```
        child { node {MiK\TeX} [clockwise from=60]}
```

```
            child { node {Pro \TeX t} }}}}
```

```
child [concept color=orange, nodes={editors}] {
```

```
    node {Editors} [clockwise from=180]
```

```
    child { node {WinEdt} }
```

```
    child { node {\TeX works} }
```

```
    child { node {\TeX studio} }
```

```
    child { node {\TeX maker} }};
```

```
\end{tikzpicture}
```

C.8 运行环境

```
sessionInfo()
```

```
## R version 4.1.0 (2021-05-18)
```

```
## Platform: x86_64-pc-linux-gnu (64-bit)
```

```
## Running under: Ubuntu 20.04.2 LTS
```

```
##
```

```
## Matrix products: default
```

```
## BLAS: /usr/lib/x86_64-linux-gnublas/libblas.so.3.9.0
```

```
## LAPACK: /usr/lib/x86_64-linux-gnulapack/liblapack.so.3.9.0
```

```
##
```

```
## locale:
```

```
## [1] LC_CTYPE=en_US.UTF-8          LC_NUMERIC=C
```

```
## [3] LC_TIME=en_US.UTF-8          LC_COLLATE=en_US.UTF-8
```

```
## [5] LC_MONETARY=en_US.UTF-8       LC_MESSAGES=en_US.UTF-8
```

```
## [7] LC_PAPER=en_US.UTF-8         LC_NAME=C
```

```
## [9] LC_ADDRESS=C                 LC_TELEPHONE=C
```

```
## [11] LC_MEASUREMENT=en_US.UTF-8   LC_IDENTIFICATION=C
```

```
##
```

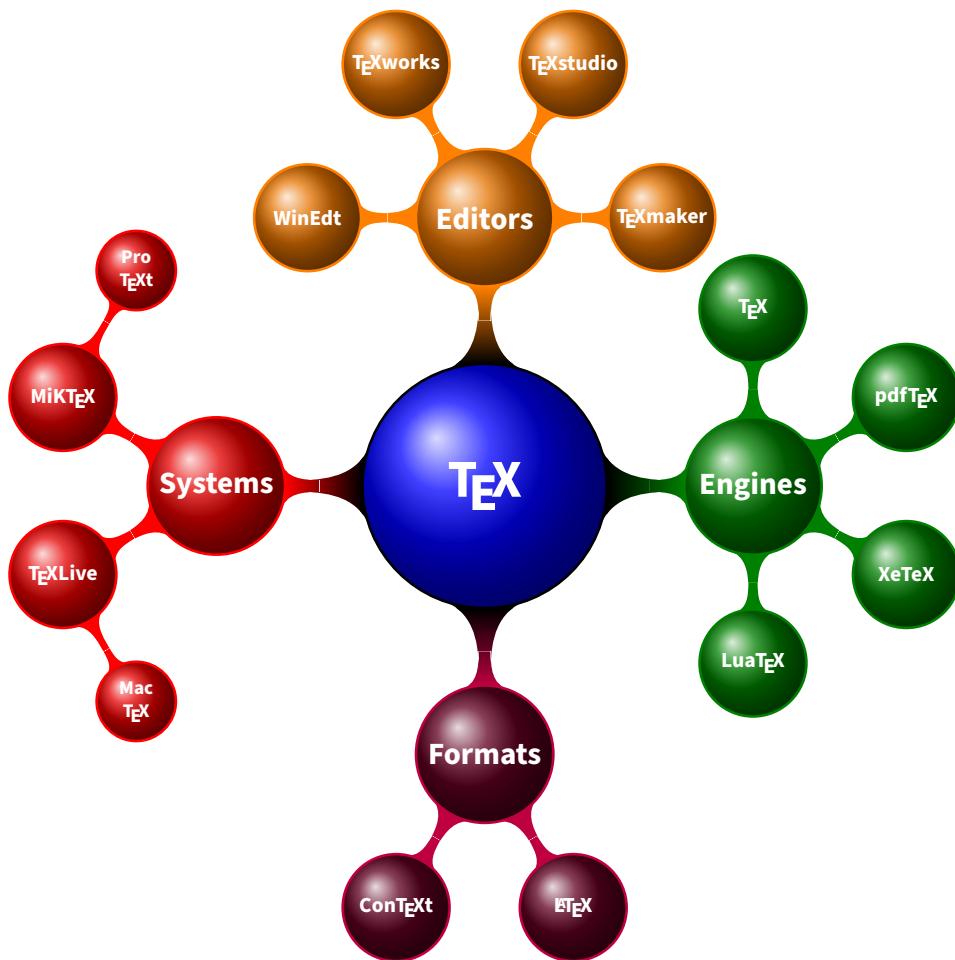


图 C.2: TeX 系统



```
## attached base packages:
## [1] stats      graphics   grDevices utils     datasets   methods    base
##
## other attached packages:
## [1] Rcpp_1.0.7       reticulate_1.20   htmlwidgets_1.5.3 shiny_1.6.0
## [5] magrittr_2.0.1
##
## loaded via a namespace (and not attached):
## [1] knitr_1.33        lattice_0.20-44   xtable_1.8-4       R6_2.5.0
## [5] rlang_0.4.11      fastmap_1.1.0    stringr_1.4.0     tools_4.1.0
## [9] grid_4.1.0        xfun_0.24       png_0.1-7        htmltools_0.5.1.1
## [13] ellipsis_0.3.2   yaml_2.2.1     digest_0.6.27    lifecycle_1.0.0
## [17] bookdown_0.22    Matrix_1.3-4    later_1.2.0      promises_1.2.0.1
## [21] evaluate_0.14    mime_0.11      rmarkdown_2.9    stringi_1.7.3
## [25] compiler_4.1.0   jsonlite_1.7.2   httpuv_1.6.1
```

附录 D 面向对象编程

进入这一章的读者都是对编程感兴趣的读者，希望在工程能力上有所提升的读者。那么最重要的是：

Code should be written to minimize the time it would take for someone else to understand it.

— The Art of Readable Code, Boswell, D. / Foucher, T.

代码可读性，代码复用性，代码维护性，代码扩展性，代码简洁性，代码高效性，代码容错性，我们共勉吧！如果读者已投身商业公司，应当以完成任务为第一，这自不必说！

D.1 环境

```
environment(fun = NULL)
environment(fun) <- value

is.environment(x)

.GlobalEnv
globalenv()
.BaseNamespaceEnv

emptyenv()
baseenv()

new.env(hash = TRUE, parent = parent.frame(), size = 29L)
```



```
parent.env(env)
parent.env(env) <- value

environmentName(env)

env.profile(env)
```

D.2 引用

```
get(x, pos = -1, envir = as.environment(pos), mode = "any",
     inherits = TRUE)

mget(x, envir = as.environment(-1), mode = "any", ifnotfound,
      inherits = FALSE)

dynGet(x, ifnotfound = , minframe = 1L, inherits = FALSE)
```

get Return the Value of a Named Object

exists Is an Object Defined?

```
exists(x, where = -1, envir = , frame, mode = "any",
       inherits = TRUE)

get0(x, envir = pos.to.env(-1L), mode = "any", inherits = TRUE,
      ifnotfound = NULL)
```

D.3 调用栈

Functions to Access the Function Call Stack

```
sys.call(which = 0)
sys.frame(which = 0)
sys.nframe()
```



```
sys.function(which = 0)
sys.parent(n = 1)

sys.calls()
sys.frames()
sys.parents()
sys.on.exit()
sys.status()
parent.frame(n = 1)

sys.source Parse and Evaluate Expressions from a File
```

D.4 闭包

An illustration of lexical scoping.

```
demo(scoping)
```

D.5 递归

Using recursion for adaptive integration

```
demo(recursion)
```

斐波那契数列

```
# 递归 Recall
fibonacci <- function(n) {
  if (n <= 2) {
    if (n >= 0) 1 else 0
  } else {
    Recall(n - 1) + Recall(n - 2)
  }
}
fibonacci(10) # 55
```



```
## [1] 55
```

D.6 异常

异常捕获和处理

```
demo(error.catching)
```

D.7 对象

判断对象类型

```
demo(is.things)
```

D.8 泛型

I'd like to prefix all these solutions with 'Here's how to do it, but don't actually do it you crazy fool'. It's on a par with redefining pi, or redefining '+'. And then redefining '<-'. These techniques have their proper place, and that would be in the currently non-existent obfuscated R contest. No, the R-ish (iRish?) way is to index vectors from 1. That's what the R gods intended!

— Barry Rowlingson ¹

如果要让下标从 0 开始的话，我们需要在现有的向量类型 `vector` 上定义新的向量类型 `vector0`，在其上并且实现索引运算 `[` 和赋值修改元素的运算 `[<-`

```
# https://stat.ethz.ch/pipermail/r-help/2004-March/048682.html
as.vector0 <- function(x) structure(x, class = "vector0") # 创建一种新的数据结构
as.vector.vector0 <- function(x) unclass(x)
"[.vector0" <- function(x, i) as.vector0(as.vector.vector0(x)[i + 1]) # 索引操作
"[<-vector0" <- function(x, i, value) { # 赋值操作
  x <- as.vector.vector0(x)
  x[i + 1] <- value}
```

¹<https://stat.ethz.ch/pipermail/r-help/2004-March/048688.html>



表 D.1: 泛型函数

A	B	C
plot.acf	plot.HoltWinters	plot.profile.nls
plot.ACF	plot.intervals.lmList	plot.ranef.lme
plot.augPred	plot.isoreg	plot.ranef.lmList
plot.compareFits	plot.jam	plot.raster
plot.data.frame	plot.lm	plot.shingle
plot.decomposed.ts	plot.lme	plot.simulate.lme
plot.default	plot.lmList	plot.spec
plot.dendrogram	plot.medpolish	plot.spline
plot.density	plot.mlm	plot.stepfun
plot.ecdf	plot.nffGroupedData	plot.stl
plot.factor	plot.nfnGroupedData	plot.table
plot.formula	plot.nls	plot.trellis
plot.function	plot.nmGroupedData	plot.ts
plot.gam	plot.pdMat	plot.tskernel
plot.gls	plot.ppr	plot.TukeyHSD
plot.hclust	plot.prcomp	plot.Variogram
plot.histogram	plot.princomp	plot.xyVector

```
as.vector0(x)
}

print.vector0 <- function(x) print(as.vector.vector0(x)) # 实现 print 方法
```

举个例子看看

```
1:10 # 是一个内置的现有向量类型 vector
```

```
## [1] 1 2 3 4 5 6 7 8 9 10

x <- as.vector0(1:10) # 转化为新建的 vector0 类型
x[0:4] <- 100 * x[0:4] # 对 x 的元素替换修改
x
```

```
## [1] 100 200 300 400 500 6 7 8 9 10
```

第三方 R 包大大扩展了 Base R 函数 `plot()` 的功能，比如 **mgcv**，**nlme** 包和 **lattice** 包等，表 D.1 列出当前环境下，`plot()` 绘图方法。



D.9 除虫

[Debugging with RStudio](#)

D.10 性能

D.11 质量

Github Action 提供的测试环境支持单元测试 testthat、静态代码检查 lintr、覆盖测试 covr、集成测试 Travis-CI、集成部署 Netlify 等一系列代码检查，还有额外的辅助工具，见 [Github Action 工具合集](#)，相关学习材料见快速参考手册 <https://github.com/github/actions-cheat-sheet> PDF 版本，以创建 R 包为例，展示工程开发的流程 <https://mdneuzaerling.com/post/data-science-workflows/>

标准计算和非标准计算 Standard and non-standard evaluation in R <https://www.brodieg.com/2020/05/05/on-nse/>

索引

bookdown, 18

Octave, 14

Pandoc, 18

Python, 14

信仰区间, 1

区间估计, 2

统计分布, 2

统计功效, 1

置信区间, 1



参考文献

- JJ Allaire, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, Winston Chang, and Richard Iannone. *rmarkdown: Dynamic Documents for R*, 2021. URL <https://CRAN.R-project.org/package=rmarkdown>. R package version 2.9.
- Douglas M. Bates and Donald G. Watts. *Nonlinear Regression Analysis and Its Applications*. John Wiley & Sons, New York, NY, 1988. URL <https://doi.org/10.1002/9780470316757.app2>.
- Geoffrey Beall. The transformation of data from entomological field experiments so that the analysis of variance becomes applicable. *Biometrika*, 32(3/4):243–262, 1942. doi: 10.2307/2332128. URL <https://www.jstor.org/stable/2332128>.
- Paul Berger and Robert Maurer. *Experimental Design*. Duxbury, 1st edition, 2002. ISBN 0-534-35822-5.
- Paul Berger, Robert Maurer, and Giovana B. Celli. *Experimental Design*. Springer International Publishing, New York, NY, 2nd edition, 2018. doi: 10.1007/978-3-319-64583-4. ISBN 978-3-319-64582-7.
- Mickaël Binois and Victor Picheny. GPareto: An R package for gaussian-process-based multi-objective optimization and analysis. *Journal of Statistical Software*, 89(8):1–30, 2019. doi: 10.18637/jss.v089.i08.
- Colin R. Blyth and David W. Hutchinson. Table of neyman-shortest unbiased confidence intervals for the binomial parameter. *Biometrika*, 47(3/4):381–391, 1960. URL <https://www.jstor.org/stable/2333308>.
- Hans W. Borchers. *pracma: Practical Numerical Math Functions*, 2021. URL <https://CRAN.R-project.org/package=pracma>. R package version 2.3.3.
- George E. P. Box, J. Stuart Hunter, and William G. Hunter. *Statistics for Experi-*



menters: *Design, Innovation, and Discovery*. John Wiley & Sons, Inc, Hoboken, New Jersey, 2nd edition, 2005. ISBN 978-0471-71813-0.

Stephen Boyd and Lieven Vandenberghe. *Introduction to Applied Linear Algebra: Vectors, Matrices, and Least Squares*. Cambridge University Press, New York, NY, 2018. URL <https://web.stanford.edu/~boyd/vmls/vmls.pdf>.



Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Journal of the American Statistical Association*, 16(3):199–231, 12 2001. doi: 10.1214/ss/1009213726.

David R. Brillinger. *Time Series: Data Analysis and Theory*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2001. ISBN 0-89871-501-6.

Lawrence D. Brown, T. Tony Cai, and Anirban DasGupta. Interval estimation for a binomial proportion. *Statistical Science*, (2):101–133, 2001. URL <https://projecteuclid.org/euclid.ss/1009213286>.

John M. Chambers. S, R, and Data Science. *The R Journal*, 12(1):462–476, 2020. doi: 10.32614/RJ-2020-028. URL <https://doi.org/10.32614/RJ-2020-028>.

Winston Chang, Alexej Kryukov, and Paul Murrell. *fontcm: Computer Modern font for use with extrafont package*, 2014. URL <https://github.com/wch/fontcm>. R package version 1.1.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 12 1934. doi: 10.1093/biomet/26.4.404.

Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, Hillsdale, NJ, 2nd edition, 1988. URL <https://www.utstat.toronto.edu/~brunner/oldclass/378f16/readings/CohenPower.pdf>. ISBN 0-8058-0283-5.

Jacob Cohen. The earth is round ($p < .05$). *American Psychologist*, 49(12):997–1003, 1994. doi: 10.1037/0003-066x.49.12.997.

Alex Couture-Beil, Jon T. Schnute, Rowan Haigh, Simon N. Wood, and Benjamin J. Cairns. *PBSddesolve: Solver for Delay Differential Equations*, 2019. URL <https://CRAN.R-project.org/package=PBSddesolve>. R package version 1.12.6.

Peter B Denton, Stephen J Parke, Terence Tao, and Xining Zhang. Eigenvectors from eigenvalues. 2019. URL <https://arxiv.org/pdf/1908.03795.pdf>.



- Annette J. Dobson. *An Introduction to Statistical Modelling*. Chapman and Hall/CRC, London, 1st edition, 1983. doi: 10.1007/978-1-4899-3174-0. ISBN 978-0412248603.
- Annette J. Dobson and Adrian G. Barnett. *An Introduction to Generalized Linear Models*. Chapman and Hall/CRC, Boca Raton, Florida, fourth edition, 2018. URL <https://www.crcpress.com/p/book/9781138741515>. ISBN 978-1138741515.
- Morris L. Eaton. *Chapter 8: The Wishart Distribution*, volume 53 of *Lecture Notes – Monograph Series*, pages 302–333. Institute of Mathematical Statistics, Beachwood, Ohio, USA, 2007. doi: 10.1214/lnms/1196285114. URL <https://doi.org/10.1214/lnms/1196285114>.
- Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004. doi: 10.1214/009053604000000067.
- T. W. Epps and Lawrence B. Pulley. A test for normality based on the empirical characteristic function. *Biometrika*, 70(3):723–726, 1983. doi: 10.2307/2336512.
- Mark A. Espeland and Siu L. Hui. A general approach to analyzing epidemiologic data that contain misclassification errors. *Biometrics*, 43(4):1001–1012, 1987. URL <https://www.jstor.org/stable/2531553>.
- John Fox and Sanford Weisberg. *An R Companion to Applied Regression*. Sage, Thousand Oaks CA, third edition, 2019. URL <https://socialsciences.mcmaster.ca/jfox/Books/Companion/>.
- Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001. URL <https://projecteuclid.org/euclid-aos/1013203451>.
- Mike Fritz and Paul D. Berger. *Improving the User Experience through Practical Data Analytics: Gain Meaningful Insight and Increase Your Bottom Line*. Morgan Kaufmann, 1st edition, 2015. ISBN 978-0128006351.
- Anqi Fu, Balasubramanian Narasimhan, and Stephen Boyd. CVXR: An R package for disciplined convex optimization. *Journal of Statistical Software*, 94(14):1–34, 2020. doi: 10.18637/jss.v094.i14.
- Alan E. Gelfand, Susan E. Hills, Amy Racine-Poon, and Adrian F. M. Smith. Illustration of bayesian inference in normal data models using gibbs sampling.



Journal of the American Statistical Association, 85(412):972–985, 1990. doi: 10.1080/01621459.1990.10474968. URL <https://www.tandfonline.com/doi/abs/10.1080/01621459.1990.10474968>.

Charles J. Geyer and Glen D. Meeden. Fuzzy and randomized confidence intervals and p-values. *Statistical Science*, 20(4):358–366, 11 2005. URL <https://www.jstor.org/stable/20061193>.

Manfred Gilli, Dietmar Maringer, and Enrico Schumann. *Numerical Methods and Optimization in Finance*. Elsevier/Academic Press, Waltham, MA, USA, second edition, 2019. URL <http://www.enricoschumann.net/NMOF/>. ISBN 978-0128150658.

Zuguang Gu, Roland Eils, and Matthias Schlesner. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, 2016.

Asad Hasan, Zhiyu Wang, and Alireza S. Mahani. Fast estimation of multinomial logit models: R package mnlogit. *Journal of Statistical Software*, 75(3):1–24, 2016. doi: 10.18637/jss.v075.i03.

C. C. Heyde, E. Seneta, P. Crépel, S. E. Fienberg, and J. Gani. *Statisticians of the Centuries*. Springer-Verlag, New York, NY, 2001. doi: 10.1007/978-1-4613-0179-0. ISBN 978-1-4613-0179-0.

David C. Hoaglin and Roy E. Welsch. The hat matrix in regression and anova. *The American Statistician*, 32(1):17–22, 1978. URL <https://www.jstor.org/stable/2683469>.

Charles C. Holt. Forecasting seasonals and trends by exponentially weighted moving averages. *International Journal of Forecasting*, 20(1):5–10, 2004. doi: 10.1016/j.ijforecast.2003.09.015.

David W. Hosmer and Stanley Lemeshow. *Applied Logistic Regression*. John Wiley & Sons, New York, NY, second edition, 2000. ISBN 0-471-72214-6.

Torsten Hothorn, Kurt Hornik, Mark A. van de Wiel, and Achim Zeileis. Implementing a class of permutation tests: The coin package. *Journal of Statistical Software*, 28(8):1–23, 2008. doi: 10.18637/jss.v028.i08.

P. L. HSU. Contribution to the theory of "student's" t -test as applied to the problem of two samples. *Statistical Research Memoirs*, 2:1–24, 1938.



- P. L. HSU. *Collected Papers*. Springer-Verlag, New York, NY, 1983. ISBN 978-1-49-392241-3.
- Ross Ihaka and Robert Gentleman. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314, 1996.
- H. Ishwaran and J.S. Rao. Spike and slab variable selection: frequentist and bayesian strategies. *Ann. Statist.*, 33(2):730–773, 2005. URL <http://arXiv.org/abs/math/0505633v1>.
- Norman L. Johnson and Samuel Kotz. *Leading Personalities in Statistical Sciences: From the Seventeenth Century to the Present*. John Wiley & Sons, New York, NY, 1997. ISBN 0-471-16381-3.
- Robert I. Kabacoff. *R in Action: Data Analysis and Graphics with R*. Manning Publications Co., Shelter Island, NY, 2nd edition, 2015. URL <https://github.com/kabacoff/RiA2>. ISBN 978-1617291388.
- Peter Kampstra. beanplot: A boxplot alternative for visual comparison of distributions. *Journal of Statistical Software*, 28(1):1–9, 2008. URL <http://www.jstatsoft.org/v28/c01/>.
- Peter Kasprzak, Lachlan Mitchell, Olena Kravchuk, and Andy Timmins. Six Years of Shiny in Research - Collaborative Development of Web Tools in R. *The R Journal*, 12(2):155–162, 2021. doi: 10.32614/RJ-2021-004. URL <https://doi.org/10.32614/RJ-2021-004>.
- Seock-Ho Kim and Allan S. Cohen. On the behrens-fisher problem: A review. *Journal of Educational and Behavioral Statistics*, 23(4):356–377, 1998. doi: 10.2307/1165281. URL <https://www.jstor.org/stable/1165281>.
- Yongdai Kim, Hosik Choi, and Hee-Seok Oh. Smoothly clipped absolute deviation on high dimensions. *Journal of the American Statistical Association*, 103(484):1665–1673, 2008. doi: 10.1198/016214508000001066.
- Christian Kleiber and Achim Zeileis. *Applied Econometrics with R*. Springer-Verlag, New York, 2008. URL <https://CRAN.R-project.org/package=AER>. ISBN 978-0-387-77316-2.
- Ron Kohavi, Diane Tang, and Ya Xu. *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing*. Cambridge University Press, Cambridge, United Kingdom, 2020. URL <https://experimentguide.com/>. ISBN 9781108724265.



Michael H. Kutner, Christopher J. Nachtsheim, John Neter, and William Li. *Applied Linear Statistical Models*. McGraw-Hill/Irwin, New York, NY, fifth edition, 2005. ISBN 0-07-238688-6.

Alexandra Kuznetsova, Per B. Brockhoff, and Rune H. B. Christensen. lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13):1–26, 2017. doi: 10.18637/jss.v082.i13.

Pierre Lafaye de Micheaux and Viet Anh Tran. PoweR: A reproducible research tool to ease monte carlo power simulation studies for goodness-of-fit tests in R. *Journal of Statistical Software*, 69(3):1–42, 2016. doi: 10.18637/jss.v069.i03.

John Lawson. *Design and Analysis of Experiments with R*. Chapman and Hall/CRC, Boca Raton, Florida, 1st edition, 2014. URL <http://www.mvstat.net/mvksa/mvksa.pdf>. ISBN 978-1498728485.

Lawrence M. Leemis. Relationships among common univariate distributions. *The American Statistician*, 40(2):143–146, 1986. URL <https://www.jstor.org/stable/2684876>.

Daniel Lüdecke, Philip Waggoner, and Dominique Makowski. insight: A unified interface to access information from model objects in r. *Journal of Open Source Software*, 4(38):1412, 2019. doi: 10.21105/joss.01412.

Dominique Makowski, Mattan Ben-Shachar, and Daniel Lüdecke. bayestestR: Describing effects and their uncertainty, existence and significance within the bayesian framework. *Journal of Open Source Software*, 4(40):1541, 2019. doi: 10.21105/joss.01541.

Ricardo A. Maronna, R. Douglas Martin, and Victor J. Yohai. *Robust Statistics, Theory and Methods*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd, 2006. ISBN 0-470-01092-4.

Peter McCullagh and John Nelder. *Generalized Linear Models*. Chapman and Hall/CRC, London, second edition, 1989. URL <https://www.crcpress.com/p/book/9780412317606>.

A.I. McLeod. *Kendall: Kendall rank correlation and Mann-Kendall trend test*, 2011. URL <http://www.stats.uwo.ca/faculty/aim>. R package version 2.2.

Aaron Meurer, Christopher P. Smith, Mateusz Paprocki, Ondřej Čertík, Sergey B. Kirpichev, Matthew Rocklin, Amit Kumar, Sergiu Ivanov, Jason K. Moore, Sar-



taj Singh, Thilina Rathnayake, Sean Vig, Brian E. Granger, Richard P. Muller, Francesco Bonazzi, Harsh Gupta, Shivam Vats, Fredrik Johansson, Fabian Pedregosa, Matthew J. Curry, Andy R. Terrel, Štěpán Roučka, Ashutosh Saboo, Isuru Fernando, Sumith Kulal, Robert Cimrman, and Anthony Scopatz. SymPy: symbolic computing in python. *PeerJ Computer Science*, 3:e103, January 2017. ISSN 2376-5992. doi: 10.7717/peerj-cs.103. URL <https://doi.org/10.7717/peerj-cs.103>.

Björn-Helge Mevik and Ron Wehrens. The pls package: Principal component and partial least squares regression in r. *Journal of Statistical Software*, 18(2):1–23, 2007. doi: 10.18637/jss.v018.i02. URL <https://www.jstatsoft.org/v018/i02>.

John C. Nash. On best practice optimization methods in r. *Journal of Statistical Software*, 60(2):1–14, 2014. doi: 10.18637/jss.v060.i02. URL <https://www.jstatsoft.org/v060/i02>.

Erich Neuwirth. *RColorBrewer: ColorBrewer Palettes*, 2014. URL <https://CRAN.R-project.org/package=RColorBrewer>. R package version 1.1-2.

Robert G. Newcombe. Interval estimation for the difference between independent proportions: comparison of eleven methods. *Statistics in Medicine*, 17(8): 873–890, 1998. doi: 10.1002/(SICI)1097-0258(19980430)17:8<873::AID-SIM779>3.0.CO;2-I.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Xiaoying Pu and Matthew Kay. A probabilistic grammar of graphics. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, page 1–13, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080. doi: 10.1145/3313831.3376466. URL <https://doi.org/10.1145/3313831.3376466>.

Yixuan Qiu. showtext: Using system fonts in R graphics. *The R Journal*, 7(1):99–108, jun 2015. doi: 10.32614/RJ-2015-008.

Sebastian Raschka and Vahid Mirjalili. *Python Machine Learning*. Packt Publishing, Birmingham, UK, 2nd edition, 2017. ISBN 978-1787125933.



Brian D. Ripley. Statistical methods need software: A view of statistical computing, 9 2002. URL <https://www.stats.ox.ac.uk/~ripley/RSS2002.pdf>.

Petr Savicky. *pspearman: Spearman's rank correlation test*, 2014. URL <https://CRAN.R-project.org/package=pspearman>. R package version 0.3-0.

Luca Scrucca. GA: A package for genetic algorithms in R. *Journal of Statistical Software*, 53(4):1–37, 2013. URL <https://www.jstatsoft.org/v53/i04/>.

Luca Scrucca. On some extensions to GA package: hybrid optimisation, parallelisation and islands evolution. *The R Journal*, 9(1):187–206, 2017. URL <https://journal.r-project.org/archive/2017/RJ-2017-008/>.

Karline Soetaert and Filip Meysman. Reactive transport in aquatic ecosystems: Rapid model prototyping in the open source software R. *Environmental Modelling & Software*, 32:49–60, 2012.

Stan Development Team. *Bayesian Statistics Using Stan*. 2019. URL <https://github.com/stan-dev/stan-book>.

Statisticat and LLC. *LaplaceDemon: Complete Environment for Bayesian Inference*, 2021. URL <https://www.bayesian-inference.com/>. R package version 16.1.6.

Reto Stauffer, Georg J. Mayr, Markus Dabernig, and Achim Zeileis. Somewhere over the rainbow: How to make effective use of colors in meteorological visualizations. *Bulletin of the American Meteorological Society*, 96(2):203–216, 2009. doi: 10.1175/BAMS-D-13-00155.1.

”Student”. The probable error of a mean. *Biometrika*, 6:1–25, 1908.

Yu-Sung Su and Masanao Yajima. *R2jags: Using R to Run JAGS*, 2020. URL <https://CRAN.R-project.org/package=R2jags>. R package version 0.6-1.

Yuan Tang. autoplotly: An r package for automatic generation of interactive visualizations for statistical results. *Journal of Open Source Software*, 3, 2018. URL <https://doi.org/10.21105/joss.00657>.

Yuan Tang, Masaaki Horikoshi, and Wenxuan Li. ggfortify: Unified interface to visualize statistical results of popular r packages. *The R Journal*, 8(2):474–485, 2016. doi: 10.32614/RJ-2016-060. URL <https://journal.r-project.org/archive/2016/RJ-2016-060/RJ-2016-060.pdf>.

Terry M. Therneau and Patricia M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. Springer, New York, 2000. ISBN 0-387-98784-3.



- Peter F. Thall and Stephen C. Vail. Some covariance models for longitudinal count data with overdispersion. *Biometrics*, 46(3):657–671, 1990. URL <https://www.jstor.org/stable/2532086>.
- Stefan Theußl, Florian Schwendinger, and Kurt Hornik. ROI: An extensible R optimization infrastructure. *Journal of Statistical Software*, 94(15):1–64, 2020. doi: 10.18637/jss.v094.i15.
- Robert Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. URL <http://www.jstor.org/stable/2346178>.
- Emilio Torres-Manzanera. *xkcd: Plotting ggplot2 Graphics in an XKCD Style*, 2018. R package version 0.0.6.
- Kenneth E. Train. *Discrete Choice Methods with Simulation*. Cambridge University Press, New York, NY, second edition, 2009. ISBN 9780511805271.
- Michail Tsagris and Manos Papadakis. Taking r to its limits: 70+ tips. *PeerJ Preprints*, 6:e26605v1, 2018. ISSN 2167-9843. doi: 10.7287/peerj.preprints.26605v1. URL <https://doi.org/10.7287/peerj.preprints.26605v1>.
- Berwin A. Turlach. *quadprog: Functions to Solve Quadratic Programming Problems.*, 2019. URL <https://CRAN.R-project.org/package=quadprog>. R package version 1.5-8.
- Kevin Ushey, JJ Allaire, and Yuan Tang. *reticulate: Interface to Python*, 2021. URL <https://github.com/rstudio/reticulate>. R package version 1.20.
- Ravi Varadhan and Paul Gilbert. BB: An R package for solving a large system of nonlinear equations and for optimizing a high-dimensional nonlinear objective function. *Journal of Statistical Software*, 32(4):1–26, 2009. URL <https://www.jstatsoft.org/v32/i04/>.
- W. N. Venables and Brian D. Ripley. *Modern Applied Statistics with S*. Springer-Verlag, New York, NY, fourth edition, 2002. URL <http://www.stats.ox.ac.uk/pub/MASS4>. ISBN 0-387-95457-0.
- Bob Wheeler. *SuppDists: Supplementary Distributions*, 2020. URL <https://CRAN.R-project.org/package=SuppDists>. R package version 1.1-9.5.
- Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, New York, 2nd edition, 2016. URL <https://ggplot2-book.org/>. ISBN 978-3319242774.



Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019. doi: 10.21105/joss.01686.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 6 1927. doi: 10.1080/01621459.1927.10502953.

Peter R. Winters. Forecasting sales by exponentially weighted moving averages. *Management Science*, 6(3):324–342, 1960. doi: 10.1287/mnsc.6.3.324.

Yihui Xie. animation: An R package for creating animations and demonstrating statistical methods. *Journal of Statistical Software*, 53(1):1–27, 2013. URL <http://www.jstatsoft.org/v53/i01/>.

Yihui Xie. *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition, 2015. URL <https://yihui.org/knitr/>. ISBN 978-1498716963.

Yihui Xie. *bookdown: Authoring Books and Technical Documents with R Markdown*. Chapman and Hall/CRC, Boca Raton, Florida, 2016. URL <https://github.com/rstudio/bookdown>. ISBN 978-1138700109.

Yihui Xie. TinyTeX: A lightweight, cross-platform, and easy-to-maintain latex distribution based on TeX Live. *TUGboat*, (1):30–32, 2019. URL <https://tug.org/TUGboat/Contents/contents40-1.html>.

Yihui Xie, J.J. Allaire, and Garrett Grolemund. *R Markdown: The Definitive Guide*. Chapman and Hall/CRC, Boca Raton, Florida, 2018. URL <https://bookdown.org/yihui/rmarkdown>. ISBN 9781138359338.

Derek S. Young. *Handbook of Regression Methods*. Chapman and Hall/CRC, Boca Raton, FL, 2017.

Jelmer Ypma. *R Interface to NLOpt*, 2020. URL <https://github.com/jyypma/nloptr>. R package version 1.2.2.2.



- Achim Zeileis and Torsten Hothorn. Diagnostic checking in regression relationships. *R News*, 2(3):7–10, 2002. URL <https://CRAN.R-project.org/doc/Rnews/>.
- Achim Zeileis, Kurt Hornik, and Paul Murrell. Escaping RGBland: Selecting colors for statistical graphics. *Computational Statistics & Data Analysis*, 53(9):3259–3270, 2009. doi: 10.1016/j.csda.2008.11.033.
- Achim Zeileis, Jason C. Fisher, Kurt Hornik, Ross Ihaka, Claire D. McWhite, Paul Murrell, Reto Stauffer, and Claus O. Wilke. colorspace: A toolbox for manipulating and assessing colors and palettes. arXiv 1903.06490, arXiv.org E-Print Archive, March 2019. URL <http://arxiv.org/abs/1903.06490>.
- Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894 – 942, 2010. doi: 10.1214/09-AOS729.
- 宋泽熙. 两个二项总体成功概率的比较. 中国校外教育（理论）, z1:81, 2011. doi: 10.3969/j.issn.1004-8502.B.2011.z1.0919.
- 茆诗松, 周纪芗, and 陈颖. 试验设计. 中国统计出版社, 北京, 1st edition, 2004. ISBN 7-5037-4316-6.
- 茆诗松, 程依明, and 潘晓龙. 高等数理统计. 高等教育出版社, 北京, 2nd edition, 2006. ISBN 978-7-04-019321-3.
- 陈希孺. 数理统计引论. 科学出版社, 北京, 1981.
- 韦博成. 《红楼梦》前 80 回与后 40 回某些文风差异的统计分析（两个独立二项总体等价性检验的一个应用）. 应用概率统计, 25(4):441–448, 2009. doi: 10.3969/j.issn.1001-4268.2009.04.012.