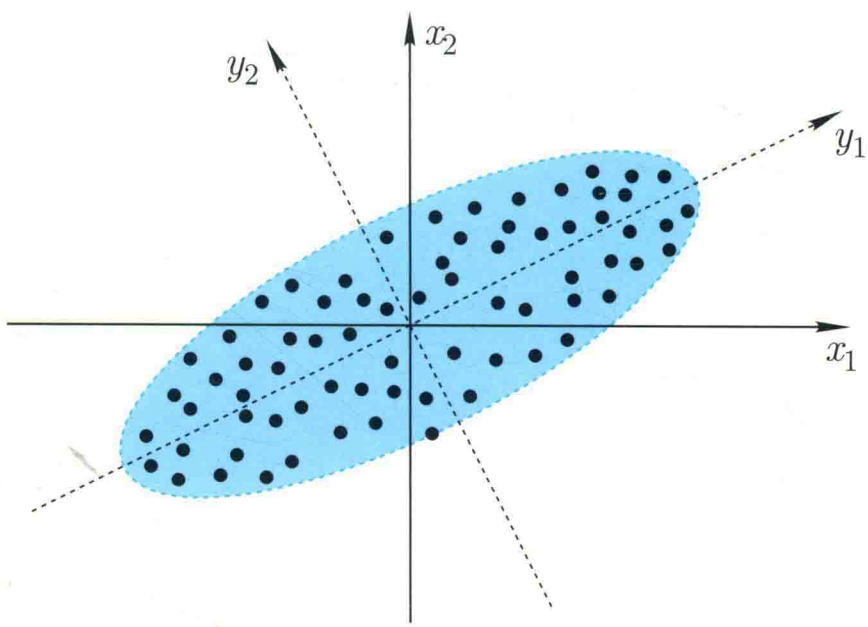
基于R应用的统计学丛书

# 多元统计分析

## ——基于R

*Multivariate Statistical Analysis with R*

主编 费宇　副主编 郭民之 陈贻娟




中国人民大学出版社

基于R应用的统计学丛书

Statistics Series with R

# 多元统计分析

## ——基于R

*Multivariate Statistical Analysis with R*

主 编 费 宇　副主编 郭民之 陈贻娟

藏书

中国人民大学出版社
·北京·

**图书在版编目（CIP）数据**

多元统计分析：基于R／费宇主编. —北京：中国人民大学出版社，2014.9
（基于R应用的统计学丛书）
ISBN 978-7-300-19952-8

Ⅰ.①多… Ⅱ.①费… Ⅲ.①多元分析－统计分析－高等学校－教材 Ⅳ.①O212.4

中国版本图书馆CIP数据核字（2014）第209791号

基于R应用的统计学丛书
**多元统计分析——基于R**
主　编　费　宇
副主编　郭民之　陈贻娟
Duoyuan Tongji Fenxi：Jiyu R

| | | | |
|---|---|---|---|
| **出版发行** | 中国人民大学出版社 | | |
| **社　　址** | 北京中关村大街31号 | **邮政编码** | 100080 |
| **电　　话** | 010－62511242（总编室） | 010－62511770（质管部） | |
| | 010－82501766（邮购部） | 010－62514148（门市部） | |
| | 010－62515195（发行公司） | 010－62515275（盗版举报） | |
| **网　　址** | http://www.crup.com.cn | | |
| | http://www.ttrnet.com（人大教研网） | | |
| **经　　销** | 新华书店 | | |
| **印　　刷** | 三河市汇鑫印务有限公司 | | |
| **规　　格** | 185 mm×260 mm　16开本 | **版　　次** | 2014年10月第1版 |
| **印　　张** | 9.75插页1 | **印　　次** | 2014年10月第1次印刷 |
| **字　　数** | 210 000 | **定　　价** | 28.00元 |

**版权所有　侵权必究　　印装差错　负责调换**

# 作 者 简 介

费　宇　二级教授, 博士生导师, 统计学博士, 英国曼彻斯特大学博士后. 现任云南财经大学统计与数学学院常务副院长, 主要从事统计理论与方法、应用统计、数据挖掘和计量经济分析方面的研究. 在国内外学术期刊上发表论文 40 余篇, 在科学出版社出版学术专著 2 部, 在高等教育出版社和科学出版社出版教材 3 部, 获省部级以上奖 8 项, 先后访问过荷兰尼津洛德大学、英国基尔大学、曼彻斯特大学、美国本特利大学和香港大学. 现担任全国经济数学与管理数学学会常务理事、云南省统计学会理事、云南省应用统计学会副理事长、云南财经大学统计与数学学院统计调查与数据挖掘研究所所长. 2010 年获云南省 "青年学术技术带头人" 称号, 2012 年获云南省 "教学名师" 称号, 2012 年获云南省 "有突出贡献中青年专家" 称号. 2004 年获得英国皇家学会王宽诚奖学金 (KC Wong Fellowship, 当年获此奖全国仅 6 人), 主编的教材《应用数理统计——基本概念与方法》2010 年获第十届全国统计科研优秀成果二等奖, 2009 年获云南省高等教育教学成果一等奖, 2013 年获云南省高等教育教学成果二等奖.

# 前　言

多元统计分析是统计学应用性最强的一个分支, 在社会、经济、管理、生物、医学、体育和环境科学等很多领域应用广泛, 是数学、统计学、经济和管理类本科生和研究生的一门重要课程. 然而, 多元统计分析这门课不好教、不好学, 一个重要的原因就是多元统计分析的理论比较抽象, 涉及的计算复杂, 需要借助软件在计算机上实现.

目前关于多元统计分析的教材一般分为两种: 一种注重系统讲授多元统计理论, 比如张尧庭和方开泰教授编写的经典教材《多元统计分析引论》; 一种强调多元统计方法的应用, 结合统计软件讲解多元统计理论与方法, 比如何晓群教授编写的《多元统计分析》教材. 第一种教材比较适合统计类和数学类学生使用, 第二种教材比较适合经济和管理类学生使用.

本书属于第二种教材, 结合目前非常流行的 R 软件来讲解多元统计分析的基本理论和方法, 力求采用简洁明了的语言来阐述理论, 使用 R 软件来实现具体的计算分析, 试图帮助读者在最短的时间里领会多元统计分析的真谛所在.

本书的编写有以下特点: (1) 言简意赅, 为了节约篇幅, 省略了一些烦琐的理论证明和公式推导; (2) 强调应用, 采用生动具体的例子来讲解多元统计分析方法, 方便读者学习; (3) 与 R 密切结合, 采用 R 软件来实现多元统计的计算和分析, 并解读 R 软件的分析结果; (4) 使用方便, 本书所有例题、案例和习题的数据文件以及相应的 R 程序都放在中国人民大学出版社工商管理出版分社网站 www.rdjg.com.cn 上供读者下载使用. 读者也可以通过电子邮件向作者索取, 邮箱地址: 1350691353@qq.com (费宇).

全书共 10 章, 第 1, 2, 3, 4, 7 章由费宇编写, 第 5, 6, 10 章由郭民之编写, 第 8, 9 章由陈贻娟编写. 本书可作为经济学和管理学类专业的本科生和硕士研究生教材, 也可以作为统计工作者的参考书.

本书参阅了许多国内外教材和资料, 并引用了部分例题和习题, 在此向有关的作者表示衷心的感谢; 本书得到了云南省教育厅 “统计学” 省院省校教育合作咨询、共建省级重点学科项目的支持, 得到了云南省教育厅 “统计学人才培养模式创新实验区” 项目的支持, 还得到了云南财经大学三年提升计划 “统计学精品视频公开课” 项目的

试读结束：需要全本请在线购买： www.ertongbook.com

支持, 在此表示感谢; 本书的出版得到中国人民大学出版社的大力支持和帮助, 在此表示诚挚的谢意.

由于作者水平有限, 难免有不妥和谬误之处, 恳请同行专家及广大读者提出宝贵意见和建议.

**费 宇**
于云南财经大学

# 目　录

**第 1 章　R 与多元统计分析简介** . . . . . **1**

1.1　R 简介 . . . . . 1

1.1.1　R 的特点 . . . . . 1

1.1.2　R 的安装与运行 . . . . . 2

1.1.3　R 的基本原理 . . . . . 3

1.1.4　R 的帮助 . . . . . 6

1.2　多元统计分析简介 . . . . . 6

1.2.1　多元统计分析的用途 . . . . . 6

1.2.2　多元统计分析的内容 . . . . . 7

习　题 . . . . . 8

**第 2 章　多元线性模型** . . . . . **10**

2.1　多元线性模型 . . . . . 10

2.1.1　模型定义 . . . . . 10

2.1.2　模型的参数估计和检验 . . . . . 12

2.2　变量选择 . . . . . 14

2.3　回归诊断 . . . . . 16

2.3.1　残差分析和异常点探测 . . . . . 16

2.3.2　回归诊断: 一般的方法 . . . . . 18

2.4　回归预测 . . . . . 20

习　题 . . . . . 21

**第 3 章 广义线性模型 · · · · · 28**
3.1 广义线性模型概述 · · · · · 28
3.2 Logistic 模型 · · · · · 29
3.3 对数线性模型 · · · · · 31
习 题 · · · · · 33

**第 4 章 聚类分析 · · · · · 38**
4.1 相似性的度量 · · · · · 38
4.2 系统聚类法 · · · · · 39
4.3 $k$ 均值聚类法 · · · · · 43
4.4 案例: 世界 146 个国家和地区人文发展情况的聚类分析 · · · · · 45
习 题 · · · · · 46

**第 5 章 判别分析 · · · · · 52**
5.1 距离判别 · · · · · 52
5.1.1 距离 · · · · · 52
5.1.2 两个总体的距离判别 · · · · · 53
5.1.3 多个总体的距离判别 · · · · · 56
5.2 Fisher 判别 · · · · · 56
5.2.1 两总体 Fisher 判别 · · · · · 56
5.2.2 多总体 Fisher 判别 · · · · · 58
5.3 Bayes 判别 · · · · · 61
5.3.1 两总体的 Bayes 判别 · · · · · 61
5.3.2 多总体的 Bayes 判别 · · · · · 63
5.4 案例分析与 R 实现 · · · · · 65
习 题 · · · · · 70
附 录 · · · · · 74

**第 6 章 主成分分析 · · · · · 77**
6.1 主成分分析的基本思想 · · · · · 77
6.2 总体主成分 · · · · · 77
6.2.1 主成分的含义 · · · · · 77
6.2.2 主成分的计算 · · · · · 79
6.2.3 主成分的主要性质 · · · · · 79
6.2.4 主成分个数的确定 · · · · · 80
6.3 样本主成分 · · · · · 80
6.3.1 样本主成分性质和计算 · · · · · 81
6.3.2 主成分分析的步骤和相关 R 函数 · · · · · 82

6.4 案例: 主成分综合分析 · · · · · 88
习 题 · · · · · 92

**第 7 章 因子分析 · · · · · 97**

7.1 正交因子模型 · · · · · 97
7.2 因子模型的估计 · · · · · 99
7.3 因子正交旋转 · · · · · 101
7.4 因子得分 · · · · · 101
习 题 · · · · · 107

**第 8 章 对应分析 · · · · · 108**

8.1 对应分析的基本思想 · · · · · 108
8.2 对应分析的原理 · · · · · 109
8.3 对应分析的计算步骤 · · · · · 110
8.4 案例: 对应分析在现金支出定位中的应用及 R 操作 · · · · · 111
习 题 · · · · · 114

**第 9 章 典型相关分析 · · · · · 117**

9.1 典型相关分析基本理论 · · · · · 117
9.2 案例: 我国科学研究与开发机构科研投入与产出的典型相关分析及 R 操作 · · · · · 120
习 题 · · · · · 124

**第 10 章 多维标度分析 · · · · · 129**

10.1 多维标度法的基本思想 · · · · · 129
10.2 古典多维标度法 · · · · · 129
10.2.1 多维标度法的几个基本概念 · · · · · 130
10.2.2 已知距离矩阵时 CMDS 解的计算 · · · · · 134
10.2.3 已知相似系数矩阵时 CMDS 解的计算 · · · · · 136
10.3 非度量多维标度法 · · · · · 138
10.4 案例分析与 R 实现 · · · · · 140
习 题 · · · · · 143

参考文献 · · · · · 147

# 第 1 章
# Chapter 1 R 与多元统计分析简介

多元统计分析是研究多个 (随机) 变量之间相互关系和规律的统计学分支. 在实际生活中, 受多个变量作用和影响的现象很多, 如果变量之间是相互独立或互不相关的, 我们可以把多个变量分开来进行研究, 一次分析一个变量, 即采用一元统计分析的方法进行分析, 但如果变量之间是相关的, 则采用一元统计方法就会丢失很多信息, 因为这种分析方法忽略了多个变量间的相关性. 多元统计分析就是把多个变量合在一起进行研究的统计学方法, 它在自然科学、经济学、管理学和社会科学等领域有广泛的应用.

R 是目前流行的一款统计软件系统, 本章将对 R 软件和多元统计分析做一个简要的介绍.

## 1.1 R 简介

### 1.1.1 R 的特点

R 是一款统计分析和作图软件系统, 它是美国贝尔实验室开发的 S 语言的一种实现或形式, 它与商业软件 S-PLUS 有很多相似之处, 二者都是基于 S 语言的软件系统, 但 R 是一个免费的软件系统, 最先是由新西兰奥克兰大学的 Robert Gentleman 和 Ross Ihaka 共同创立的, 现在由 R 开发核心小组 (R Develop Core Team) 维护.

作为一款优秀的统计分析软件系统, R 具有如下特点:

(1) 免费和开放. R 是一款由志愿者维护的完全免费的统计分析软件, 它的安装文件和程序包都可以从 CRAN (Comprehensive R Archive Network) 社区 (www.r-project.org) 下载, 作为教学使用非常方便, 国外很多大学的统计教学都使用这款软件; 而且 R 的源代码是公开的, 这样方便使用者了解 R 程序的计算方法, 并且可以对程序进行修改和扩展处理.

(2) 统计分析功能完善. R 内嵌了许多统计分析函数, 可以直接调用进行统计分析, R 的部分统计功能整合在 R 语言的低层, 但大多数功能是以各种程序包的形式提供的, 大约有 25 个 "标准" 程序包和 R 同时发布, 但更多的程序包可以通过 CRAN 社区下载安装, 而且程序包的更新比商业软件及时, 使用非常方便.

(3) 作图功能强大. R 内嵌的作图函数能在图形窗口输出漂亮美观的图形, 这些图

形可以保存为各种形式的文件 (比如 jpg, bmp, ps, pdf, emf, png, pictex, xfig 等), 方便使用.

(4) 可移植性强. R 程序可以很容易地移植到 S-PLUS 中, 同时, S-PLUS 的程序也可以方便地移植到 R 中使用. R 可以读入很多分析软件 (比如 SAS, SPSS, Excel, Stata 等) 的数据文件, 而 R 的数据文件可以保存为文本格式供其他统计软件使用, 这样 R 与其他统计软件就建立了一个良好的联系机制.

(5) 使用灵活. R 可以运行于 UNIX, Linux, Windows 和 Macinton 等操作系统中, R 的分析结果都存放在一个对象里, 用户可以有选择地显示感兴趣的结果, 这些结果可以直接用于进一步的分析.

### 1.1.2 R 的安装与运行

从 CRAN 社区下载最新的 R 安装程序, 就可以进行安装了, 通常默认的安装目录为 C:\Program Files\R\R-x.x.x, 安装完成后点击桌面上的 R-x.x.x 图标就可以启动 R 软件了. 在 RGui 命令窗口的命令提示符 ">" 后输入命令就可以完成相应的操作. 如果要退出 R 系统, 可以在命令行输入 q( ), 也可以点击 RGui 右上角的叉号退出. 退出时可以保存工作空间, 比如将工作空间保存在 "C:\Work\" 目录下, 名称为 "W.RData", 则以后可以通过命令 load("C:\\Work\\ W.RData") 来加载这个空间, 或者通过菜单 "文件" 下的 "载入工作空间" 加载.

R 软件的程序包的安装有三种方式:

(1) 菜单方式: 在联网情况下, 按照 "程序包 → 安装程序包 → 选择 CRAN Mirror 服务器 → 选择要安装的程序包" 的步骤进行在线安装.

(2) 命令方式: 在联网情况下, 在命令提示符后输入命令

```
>install.package(Rcmdr)
```

完成程序包 Rcmdr 的安装.

(3) 本地安装: 要安装本机上的程序包, 可以按 "程序包 → 从本地 zip 文件安装程序包" 的步骤选择本机上的程序包进行安装.

新安装的程序包 (除了 R 的标准程序包, 比如 base) 必须先载入才能使用, 可以采取如下方式载入:

(1) 菜单方式: 按照 "程序包 → 加载程序包 → 选择要加载的程序包" 的步骤进行加载.

(2) 命令方式: 在命令提示符后输入命令

```
>library(Rcmdr)
```

完成程序包的加载.

此外, 我们还可以通过 "程序包 → 更新程序包 ·····" 的步骤对程序包进行实时更新.

试读结束：需要全本请在线购买： www.ertongbook.com

### 1.1.3　R 的基本原理

R 是一种解释性语言, 它的语法非常简单, 比如求变量 $x$ 的方差的命令为 var(x), 而命令 lm(y∼x) 表示以 $y$ 为因变量, $x$ 为自变量拟合一个线性回归模型.

需要注意的是, 只有先给变量赋值才能进行相应的计算, 统计分析中最常见的变量是向量和矩阵, 下面介绍数值型向量和矩阵的建立方法. 为了说明方便, 每一个语句都给出一个注释语句, 井号 (#) 表示注释的开始, 即 # 后面的是注释语句.

#### 1. 数值型向量的建立

```
x1<-seq(2,6,by=1)   #生成序列 x1, x1=(2,3,4,5,6), 这里赋值符号 ''<-'' 也可以用
#等号 ''=''
x2<-c(8,10,12,16,21)   #生成一个 5 维向量 x2, x2=(8,10,12,16,21)
x3<-rep(2:4,2)   #生成序列 x3, x3=(2,3,4,2,3,4)
z.dat<-data.frame(x=x1,y=x2)   #生成数据框 (数据文件) z.dat, 具体形式如下
  x  y
1 2  8
2 3 10
3 4 12
4 5 16
5 6 21
cbind(x1,x2)   #将 x1 和 x2 按列合并得如下数据:
     x1 x2
[1,] 2  8
[2,] 3 10
[3,] 4 12
[4,] 5 16
[5,] 6 21
rbind(x1,x2)   #将 x1 和 x2 按行合并得如下数据:
   [,1] [,2] [,3] [,4] [,5]
x1    2    3    4    5    6
x2    8   10   12   16   21
```

#### 2. 矩阵的建立

```
A<-matrix(1,nr=2,nc=2)   #建立一个所有元素都为 1 的 2 阶方阵
B<-diag(3)   #生成一个 3 阶单位阵
x<-c(2,3,4)
D<-diag(x)   #生成一个对角元素是 (2,3,4) 的 3 阶方阵
x<-matrix(0,nr=2,nc=3)   #建立一个所有元素都为 0 的 2×3 阶矩阵
```

```
x1<-c(2,3,4)
x2<-c(1,2,5)
X<-rbind(x1,x2)   #将 X 的第 1 行赋值为 x1, 第 2 行赋值为 x2, 得到如下阶矩阵:
     [,1]   [,2]   [,3]
x1      2      3      4
x2      1      2      5
```

下面以一个例子来具体说明 R 的工作原理.

**【例 1.1】**(数据文件为 eg1.1.txt) 表 1—1 给出了我国 2007 年 31 个地区城镇居民年人均可支配收入和年人均消费性支出数据, 该数据文件是 txt 格式的文件, 请将数据读入 R 生成相应的 R 数据文件, 并建立年人均消费性支出 $y$ 关于年人均可支配收入 $x$ 的线性回归模型.

**表 1—1　　城镇居民年人均可支配收入和年人均消费性支出数据**　　单位: 元

| 地区 | 可支配收入 | 消费性支出 | 地区 | 可支配收入 | 消费性支出 |
|---|---|---|---|---|---|
| 北　京 | 21 988.71 | 15 330.44 | 湖　北 | 11 485.80 | 8 701.18 |
| 天　津 | 16 357.35 | 12 028.88 | 湖　南 | 12 293.54 | 8 990.72 |
| 河　北 | 11 690.47 | 8 234.97 | 广　东 | 17 699.30 | 14 336.87 |
| 山　西 | 11 564.95 | 8 101.84 | 广　西 | 12 200.44 | 8 151.26 |
| 内蒙古 | 12 377.84 | 9 281.46 | 海　南 | 10 996.87 | 8 292.89 |
| 辽　宁 | 12 300.39 | 9 429.73 | 重　庆 | 12 590.78 | 9 890.31 |
| 吉　林 | 11 285.52 | 8 560.30 | 四　川 | 11 098.28 | 8 691.99 |
| 黑龙江 | 10 245.28 | 7 519.28 | 贵　州 | 10 678.40 | 7 758.69 |
| 上　海 | 23 622.73 | 17 255.38 | 云　南 | 11 496.11 | 7 921.83 |
| 江　苏 | 16 378.01 | 10 715.15 | 西　藏 | 11 130.93 | 7 532.07 |
| 浙　江 | 20 573.82 | 14 091.19 | 陕　西 | 10 763.34 | 8 427.06 |
| 安　徽 | 11 473.58 | 8 531.90 | 甘　肃 | 10 012.34 | 7 875.78 |
| 福　建 | 15 506.05 | 11 055.13 | 青　海 | 10 276.06 | 7 512.39 |
| 江　西 | 11 451.69 | 7 810.73 | 宁　夏 | 10 859.33 | 7 817.28 |
| 山　东 | 14 264.70 | 9 666.61 | 新　疆 | 10 313.44 | 7 874.27 |
| 河　南 | 11 477.05 | 7 826.72 | | | |

**解**: 假定数据文件为 eg1.1.txt, 保存在 "C:\data" 子目录下, 我们先读入数据, 计算 $x$ 与 $y$ 的相关系数并绘制散点图, 具体程序如下:

```
setwd("C:/data")   #设定工作路径, R 中路径的斜线符号为 ''/'', 与 Windows 中的相
#应符号 ''\'' 不一样
dat=read.table("eg1.1.txt",header=T)   #从 eg1.1.txt 中读入数据, 记为 dat
#header=T 表示将 eg1.1.txt 文件的第 1 行作为表头行, 也可以写为 header=TRUE
#header=F 或 header=FALSE 则表示文件的第 1 行不作为表头行
cor(dat)   #计算 x 和 y 的相关系数
plot(y~x,data=dat)   #绘制 x 和 y 的散点图
```

运行结果为:

```
        x           y
x  1.0000000   0.9760254
y  0.9760254   1.0000000
```

在图形窗口可以得到 $x$ 和 $y$ 的散点图, 如图 1—1 所示.

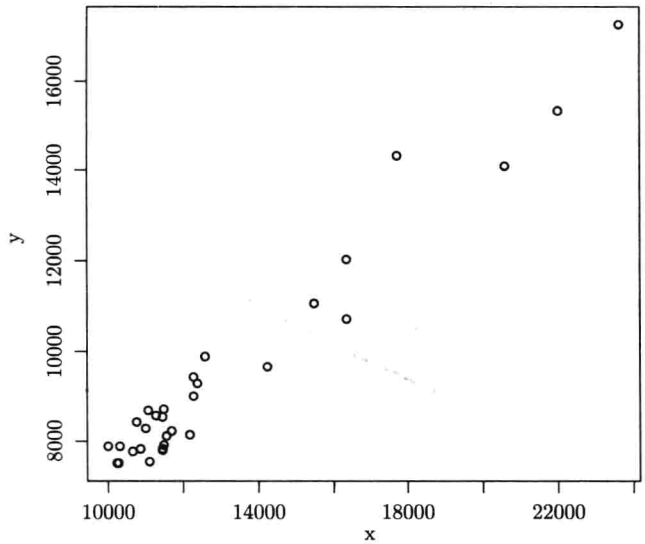



**图 1—1　年人均可支配收入 $x$ 和年人均消费性支出 $y$ 的散点图**

从图 1—1 可以看出年人均消费性支出 $y$ 与年人均可支配收入 $x$ 之间的线性关系非常明显, 二者的相关系数为 0.976, 因此可以建立年人均消费性支出 $y$ 关于年人均可支配收入 $x$ 的线性回归模型, 具体程序如下:

```
fitlm<-lm(y~x,data=dat)   #使用数据文件 dat 中的数据, 建立 y 关于 x 的回归方程,
#并将回归结果保存在 fitlm 中, 这里赋值符号 ''<-'' 也可以用等号 ''=''
summary(fitlm)   #显示 fitlm 的内容, 即输出回归分析的结果
```

运行结果为:

```
Call:
lm(formula = y ~ x, data = dat)
Residuals:
    Min      1Q  Median      3Q     Max
-1068.3  -417.3   -20.5   301.9  1639.1
Coefficients:
             Estimate   Std.Error  t value  Pr(> |t|)
(Intercept)  450.33408  388.90559    1.158    0.256
x              0.69197    0.02865   24.148   <2e-16 ***
---
Signif.codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'  0.1 ' ' 1
```

```
Residual standard error:  559.3 on 29 degrees of freedom
Multiple R-squared:  0.9526,         Adjusted R-squared:  0.951
F-statistic:  583.1 on 1 and 29 DF,  p-value:  < 2.2e-16
```

### 1.1.4 R 的帮助

R 是一种编程语言, 它的语法简单直观, 统计分析和绘制图形主要是通过 R 中的各种函数来实现的. R 中的程序包由大量的统计分析函数组成, 要编写程序进行统计计算和分析, 就必须理解各种统计分析函数的含义, 熟悉它们的使用方法, 初学者可以通过 R 的帮助系统获得相应的帮助.

比如, 要获得 R 的基本知识, 可以启动 R 软件, 在 RGui 的窗口中选择 "帮助" 菜单中的 "R FAQ" (R 的常见问题), 获得 R 的特点、安装、使用、界面和编程规则等基本知识; 也可以选择 "帮助" 菜单中的 "手册 (PDF 文件)" 提供的 8 本帮助手册: *An Introduction to R*, *R Reference Manual*, *R Data Import/Export*, *R Language Definition*, *Writing R Extensions*, *R Internals*, *R Installation and Administration*, *Sweave User*, 其中第一本 *An Introduction to R* 是最基本的手册. 通过命令 ">help.start()" 也可以获得类似的帮助.

如果要了解有关函数的含义和使用方法, 可以采用如下命令:

```
help(lm)   #获得名为 ''lm'' 的函数的帮助页面
?lm   #此命令与上面的命令效果一样
```

## 1.2 多元统计分析简介

### 1.2.1 多元统计分析的用途

多元统计分析是 20 世纪初发展起来的统计分析方法, 它是通过对多个随机变量观测数据的分析来研究多个随机变量之间的相互关系并揭示变量内在规律的分析方法. 多元统计分析方法可以应用于经济、管理、生物、医学、教育学、心理学、工业、农业等很多领域, 是一种常用的统计分析方法. 实践中, 多元统计分析方法通常用于解决以下四个方面的问题:

(1) 多变量的相关性分析. 分析多个变量之间的相关性是实践中非常常见的问题, 简单相关分析、偏相关分析和复相关分析是分析多个变量相关性的常用方法, 而典型相关分析可以用于分析两组变量的相依关系.

(2) 预测分析. 通过建立分析模型来预测和估计我们关心的变量, 这种分析通常采用多元回归分析来完成.

(3) 分类和组合. 根据事物 (个体) 的多个指标, 将事物按照相似程度来进行分类和组合, 或者根据个体的多个指标测量值, 将考察的个体具体划分到某个类别, 这样的

分类和组合问题可以通过聚类分析和判别分析来完成.

(4) 数据简化 (降维). 将多个变量的主要信息用很少的几个变量来表示, 降低数据的维度, 从而达到化简数据的目的. 主成分分析和因子分析就是常用的数据简化方法.

### 1.2.2　多元统计分析的内容

多元统计分析的主要内容包括: 多元回归分析、聚类分析、判别分析、主成分分析、因子分析、对应分析、典型相关分析和多维标度分析等.

#### 1. 多元回归分析

多元回归分析是研究一个因变量 (随机变量) 随多个自变量 (通常假定为非随机变量) 的变化而变化的情况, 通过建立多元回归模型 (线性模型和广义线性模型等) 来分析二者之间的依赖关系. 普通线性模型适合因变量是连续型变量的情况, 如果因变量是离散型变量, 则要采用广义线性模型来处理. 第 2 章将介绍多元线性模型, 第 3 章将讨论广义线性模型.

#### 2. 聚类分析

聚类分析是根据聚类对象 (若干个个体的集合) 的多个变量 (指标) 的测量值, 按照某种标准把这些个体分成若干类. 它是研究如何做到 "物以类聚" 的一种统计分析方法, 聚类方法分为系统聚类法和分解聚类法两种, 系统聚类法是将类由多变少的聚类方法, 而分解聚类法则是将类由少变多的聚类方法, 第 4 章将介绍两种常用的聚类方法: 系统聚类法和 $k$ 均值聚类法.

#### 3. 判别分析

判别分析是在已知分类的前提下, 将给定的新样品按照某种分类规则判入某个类中, 它是研究如何将个体 "归类" 的一种统计分析方法. 常用的判别分析方法主要有距离判别法、Fisher 判别法和 Bayes 判别法三种. 距离判别法和 Fisher 判别法属于确定性判别法; Bayes 判别法属于概率判别法, 判别以个体归属某类的概率最大或错判总平均损失最小为标准. 第 5 章将介绍距离判别法、Fisher 判别法和 Bayes 判别法.

#### 4. 主成分分析

主成分分析是一种降维分析方法, 即将多个存在相关关系的变量转化为少数几个综合变量 (即主成分) 的统计分析方法, 每个主成分都是原始变量的线性组合, 这些主成分保留了原始变量的大部分信息, 从而可以简化数据, 揭示变量之间的内在联系. 第 6 章将介绍主成分分析方法.

### 5. 因子分析

因子分析最早起源于 Karl Pearson 和 Chales Spearman 等人关于智力的定义和测量工作, 因子分析的基本目的是用少数几个随机变量 (称为因子) 去描述多个随机变量之间的协方差关系, 从这点上看, 因子分析与主成分分析有相似之处, 但因子分析中的因子是不可观测的, 也不必是相互正交的变量. 因子分析可以视为主成分分析的一种推广, 它的基本思想是: 根据相关性大小把变量分组, 使得组内的变量相关性较强, 但不同组的变量相关性较弱, 则每组变量可以代表一个基本结构, 称为因子, 它反映已经观测到的相关性. 第 7 章将讨论因子分析方法.

### 6. 对应分析

对应分析是在因子分析的基础上发展起来的, 因子分析分为针对变量的 R 型因子分析和针对样品的 Q 型因子分析, 对应分析把 R 型因子分析和 Q 型因子分析有机地结合起来, 同时把变量和样品反映到有相同坐标轴 (因子轴) 的一张图上来说明变量与样品之间的对应关系. 第 8 章将介绍对应分析方法.

### 7. 典型相关分析

典型相关分析是一般相关分析的推广, 是用于研究两组随机变量之间的相互依赖关系的一种统计分析方法. 它利用主成分分析的思想来讨论两组变量的相关性问题, 把两组变量的相关性研究转化为少数几对变量之间的相关性研究, 而这少数几对变量之间是不相关的, 这样能比较清楚地反映两组变量之间的相互关系. 第 9 章将讨论典型相关分析.

### 8. 多维标度分析

多维标度分析是以空间分布的形式表现对象之间相似性或亲疏关系的一种多元分析方法. 给定 $n$ 个个体, 它们是由多个变量反映的个体, 我们知道这 $n$ 个个体之间的某种距离 (比如欧氏距离) 或某种相似性, 我们从这种距离或相似性出发, 在低维的欧氏空间中把 $n$ 个个体的图形绘制出来, 反映这些个体之间的结构关系, 这就是多维标度分析. 第 10 章将讨论多维标度分析.

## 习 题

1.1 R 软件与 SPSS 软件有何区别? R 的主要特点是什么?

1.2 到 R 的网站 (www.r-project.org) 上下载并安装最新版的 R 软件, 运行 1.3 节有关数值型向量和矩阵的建立语句以及例 1.1 的程序.

试读结束: 需要全本请在线购买: www.ertongbook.com