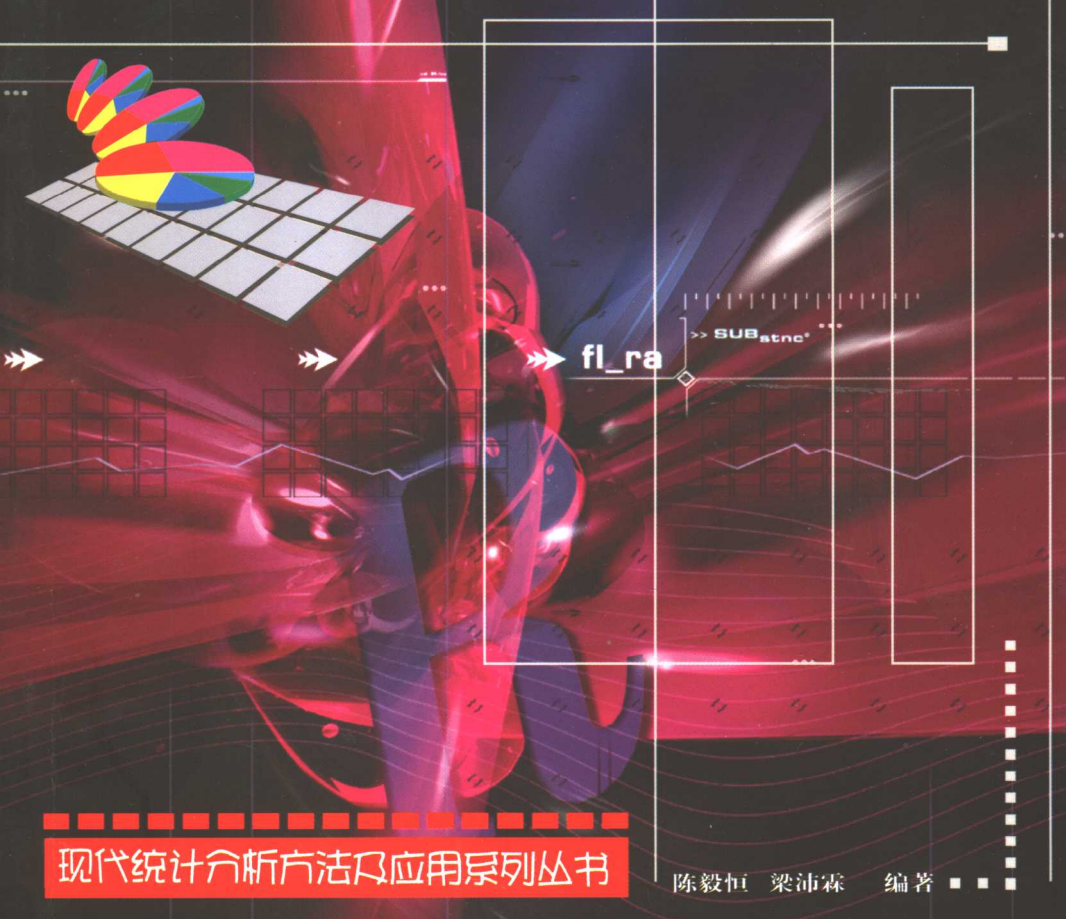


现代统计分析方法及应用系列丛书

陈毅恒 梁沛霖 编著

R 软件操作入门

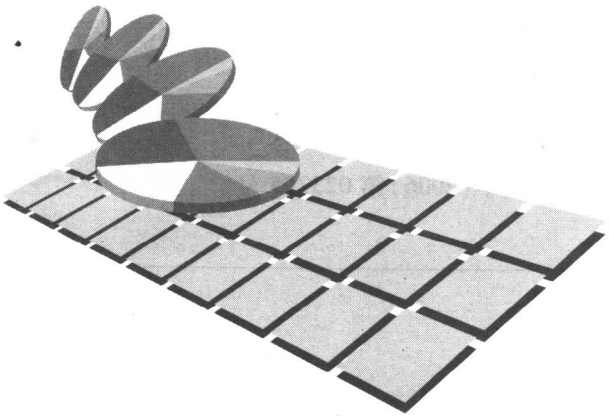
中国统计出版社
China Statistics Press

R 软件操作入门

现代统计分析方法及应用系列丛书

R 软件操作入门

陈毅恒　梁沛霖　编著



中国统计出版社
China Statistics Press

(京)新登字 041 号

图书在版编目(CIP)数据

R 软件操作入门/陈毅恒,梁沛霖编著.
—北京:中国统计出版社,2006.5
ISBN 7-5037-4865-6

Ⅰ.R…
Ⅱ.①陈… ②梁…
Ⅲ.统计分析—应用软件
Ⅳ.C819

中国版本图书馆 CIP 数据核字(2006)第 023205 号

作　　者/陈毅恒　梁沛霖
责任编辑/吕　军
封面设计/艺编广告
出版发行/中国统计出版社
通信地址/北京市西城区月坛南街 75 号　邮政编码/100826
办公地址/北京市丰台区西三环南路甲 6 号
电　　话/邮购(010)63376907　书店(010)68783172
印　　刷/河北天普润印刷厂
经　　销/新华书店
开　　本/880×1230mm　1/32
字　　数/90 千字
印　　张/4
印　　数/1-4000 册
版　　别/2006 年 6 月第 1 版
版　　次/2006 年 6 月第 1 次印刷
书　　号/ISBN 7-5037-4865-6/C·2100
定　　价/15.00 元

R

前言

随着国内经济迅速发展，统计分析在各个领域中扮演着举足轻重的角色。举例说，2008 年在北京举行的奥运会，有大量数据需要处理以作市场分析与预测。要分析数据，便要应用统计软件。因现时的计算机科技与日俱增，不同的计算机软件层出不穷，让人眼花撩乱。这本小册子只有一个目的，就是扼要地介绍一个免费而专业的统计计算机语言及软件平台——R。

早于廿年前，专业统计界已经广泛应用另一统计软件 S-PLUS。S-PLUS 为收费软件平台，其计算机语言称为 S，由于这软件十分昂贵，因此未能普及于教育界。直至 90 年代末，跟 S 语言极相近的 R 面世，而最重要是 R 的软件平台是免费的，这大大增加了它的普及性。虽然 R 为免费软件，但它拥有很多其它收费软件如 SAS 或 SPSS 没有的优点：

- 它可以处理不同类型的数据，如数字或非数字的数据也可一并处理，应用层面增广，举例说，市场数据往往存在大量的非数字数据，例如性别、籍贯等等。假若要把非数字数据分开处理，便费时误事。而 R 却能一并处理，十分方便；
- R 为独立程序，用者可以自行编写所需的程序，不像一般软件，规范严谨，缺乏弹性；

- R 拥有高质素的图像工具，这对统计分析提供了很大的帮助。再者，R 的图像工具弹性很高，用者可加进自己的要求，绘出所需的图像。R 的另一特点为拥有七彩缤纷的绘图能力，增加统计的乐趣；
- 大部份常用与非常用的统计工具已被写成 R 程序，用者可按需要去加减，从而进行统计分析。再者，相关的 R 网页已存有大量写成的程序，读者只要在相关网页下载便能免费使用。而一般收费软件的平台却价钱昂贵，一般人难以支付；
- R 也可与其它计算机语言如 C ++ 及 FORTRAN 合并，很多先进的统计概念可以运算出来。它的应用层面除了统计，还遍布医疗、金融经济、商业等等。

基于上述种种优点，R 在近几年间快速普及起来，在欧美十分通行。现在北京大学和中国人民大学都已经开始运用 R 教授统计分析课程。虽然统计分析在中文为主的社会中已十分普及，但有关 R 的中文课本却寥寥可数。这是笔者写这本书的主要动机之一，希望能藉此推广 R 的应用。

全书共分九章，第一章为基本入门的简单指令；第二章介绍数据及对象的类型；第三章介绍各类统计分布及如何在 R 进行模拟实验；第四章介绍有关 R 程序编写；第五章是关于读取及整理数据。第六及第七章为关于 R 的统计模型；第六章主要是常用的线性统计模型如线性回归、方差分析及广义线性模型；而第七章是关于一些常用多元统计方法如聚类分析、分类树及人工神经网络等；第八章是有关 R 的造像功能；最后一章则关于最优化方法。

在编写过程中，笔者以一般读者为对象。除了第六章及第七章关于统计模型外，并没有假设读者具备深厚的统计或计算知识。本书其中一大特点是以实例来解释 R 的秘诀，读者只要按部就班，根据书中的实例一步步地练习，应很容易地掌握 R 的特征及要点。换句话说，假如读者能边读边练习，把例子重复，便能很有效地学会怎样利用 R 作统计分析与推断。

本书所用的数据，大部份都是 R 内置数据。可以依照书内的指

令，直接在 R 中读取。只有几个很小的数据档案不是内置的。读者可到以下网站下载或自行输入亦不会有困难。一般来说，假若读者能花上数小时于每章内，不用数星期便能读毕全书。应用 R 便能驾轻就熟，应付自如。当然，假若读者完成这书后有兴趣了解更深入的统计或 R 概念，读者大可参考其它书籍，而这书便成读者入门的良伴。

在编写本书过程中，得到人民大学吴喜之教授提供宝贵建议，大大改进本书的内容及编排，在此作者衷心感谢吴教授的支持及协助。作者亦感谢中国香港特别行政区研究资助局基金赞助。当然，本书内之漏洞错误皆为作者之责任。

数据档案下载网站：http://www.sta.cuhk.edu.hk/books/Rdata/

目录

第一章

R 入门

这一章主要介绍 R 的基本使用方法，包括开启及关闭 R、输入数字、向量和一些简单指令及内置函数的应用。

1.1 下载及安装 R

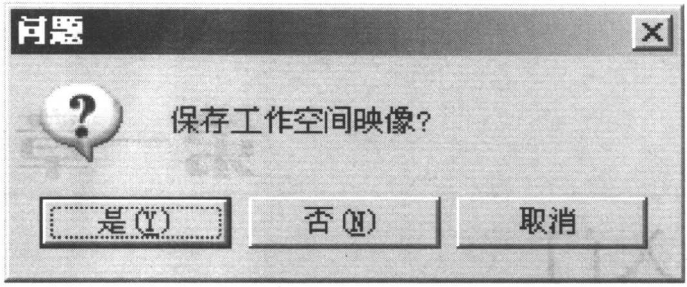
R 是免费软件，读者可以根据以下程序于网上下载：

1. 进入 http://www.r-project.org/；
2. 点击 CRAN，然后选择最近的镜子网站（mirror）；
3. 选择 Windows（95 and later），假设读者使用微软窗口；
4. 选择 base，然后点击 rw2011.exe 开始下载至硬盘（档案大约有 25 兆字节）。
5. 下载后双击 rw2001.exe 进行安装。

1.2 启动和退出 R 环境

启动和退出 R 程序的方法和其它微软窗口软件一样，在程序集里按出 R 选项便可进入 R 程序；点右上角的 框或输入 q() 就能退出。在退出 R 时会出现以下的选项框：

使用者可选择保存或不保存工作空间映像。如果选择保存，则这次在 R 出现的数据对象及自设函数等都会保存。（关于 R 数据对象及自设函数等本书将详细介绍）。下一次启动 R 时这些数据对象及自设函数等都会自动加载。如果选择不保存，则这些数据对象及



自设函数等都会消失。至于保存与否,则视使用者喜好而定。

启动 R 后,使用者可以看见以下文字:

```
R : Copyright 2005, The R Foundation for Statistical Computing
Version 2.2.0 (2005-10-06 r35749)
ISBN 3-900051-07-0

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for a HTML browser interface to help.
Type 'q()' to quit R.

>
```

在 R 窗口中出现“ > ”这符号,就是表示 R 等待使用者在这里输入指令。

1.3 输入指令

使用者只需在 > 后输入指令,然后按 enter 就可以执行指令。举例说,输入 2 + 3 后按 enter,R 就会显示答案:

```
> 2 + 3
[1] 5
```

四则运算的次序跟数学的法则相同，乘号、除号及指数符号（Exponentiation）分别以 * 、/ 及 ^ 来表示。

函数的格式可以分为“有自变量”和“无自变量”两类。如果函数 functionname 是没有自变量的，那么只需输入 functionname( ) 便可。但假若它是需要自变量 arg1，arg2 的话，那么使用者便要输入 functionname(arg1, arg2)。如果只输入 functionname（没有括号），R 便会把它的定义显示出来。使用者必须注意一点，在 R 中，大小写是有分别的，所以 functionname( ) 和 Functionname( ) 代表着两个不同的函数。

使用者可参考以下例子，平方根（square root）和绝对值（absolute value）都是需要输入一个自变量的：

```
> sqrt(9)
[1] 3

>abs( -5)
[1] 5
```

这里的 [1] 代表着所得答案的第一个元素。这种表示方式在表达一串数字所组成的向量时是很重要的。下列例子中第 36 个数值是 1000：

```
[1]    0.73   1.46   2.19   2.92   3.65   4.38   5.11   5.84    6.57
[10]   7.30   8.03   8.76   9.49  10.22  10.95  11.68  12.41   13.14
[19]  13.87  14.60  15.33  16.06  16.79  17.52  18.25  18.98   19.71
[28]  20.44  21.17  21.90  22.63  23.36  24.09  24.82  25.55 1000.00
[37]   1.04   1.30   1.56   1.82   2.08   2.34    2.6   2.86    3.12
[46]   3.38   3.64   3.90   4.16   4.42   4.68   4.94
```

使用者如输入不完整的指令时（例如：没有输入关括号），符号 + 便会出现，它表示 R 正等待使用者输入余下的指令，例如：

```
> sqrt(16
+
```

使用者只要输入“)”就可以完成指令：

```
> sqrt(16
+ )
[1] 4
```

1.4 R 原始码的使用(R Source code)

使用者除了可每次输入一个指令外，还可以选择把一组程序编写成指令文件，并在 R 里执行。一般来说，指令文件可以把多个指令放在一起，以便易于修改，如计算仿真模型等。但对于处理一些需要互动输入及探索性的数据分析，则用处不大。使用者可利用文本编辑器来编写指令文件，并把所有指令输入文件内。(注意：在档尾段必须留有空白行，否则 R 可能不能正常执行最后一行指令。)

使用者先用记事本在某数据夹（例如 C:\ test）建立及编写一个文字模式的 R 原始码档案（例如 myfile. r). 在 R 的选单中选取 file - > change dir. . . ,更改预设的目录为 C:\ test。然后输入指令：source(“myfile. r”,echo = T)。R 就会执行在 myfile. r 档案中的每句指令。自变量 echo = T 是要显示 R 执行的指令。

1.5 R 在线说明

R 的使用说明只提供个别功能的用法和定义，并无提供整体的功能指南。使用者最初会感到困难，但一旦熟悉了，便会发现这种方式能有效地说明功能的用法、所需自变量及例子等。例如，若想找出绝对值(absolute value)的应用方法，只需输入 help(abs) 即可。这个窗口提供了一个接口给使用者从不同的题目中找寻适合的功能使用说明。还有，在 R 的选单中选取 help，里面有很多说明档案。在此推荐 Html help 中的 An Introduction to R，特别适合初学人士。

1.6 常用功能

R 会把使用者所输入的全部指令储存，使用者可以利用上、下光标键来选取以前输入的指令。选取适当指令后，可用输入键重新执行此指令；亦可用左、右光标键来更改指令。应用光标键来更改及执行以前输入的指令是十分方便快捷的。还有，在 R 的选单下面有很多不同的小图像，包括加载 R 原始码、加载及储存影像文件、剪贴及复制和停止目前运算等。

1.7 变量与赋值

使用者可以用“=”或“<-”来将数值赋给一个变量。在旧版中只可使用“<-”，至于用哪个符号则视使用者喜好而定。任何英文字母、数字、“-”及“.”都可作为变量名称。但是第一个字必须是英文字母。除此之外，R 不容许变量名称中有空格或“-”，如 Water Depth 就必须改为 WaterDepth 或 Water. Depth，x-1 可改为 x_1。如果要给变量 x 赋予数值 9，只需输入：

```
> x = 9    或    > x <- 9
```

使用者可能已注意到赋值后是不会显示出它的数值的。如使用者想确定赋值是否成功，可以输入该变量名称看看。当 x 有其数值时，就可以使用它作运算：

```
> sqrt(x)
[1] 3
> y = (5 * ( x + 2 )) - 3
> y
[1] 52
```

这些运算不会影响 x 的数值。倘若想重新给 x 赋值，可参考以下例子：

```
> sqrt(x)
[1] 3
```

```
> x
[1] 9
> x = sqrt(x)
> x
[1] 3
```

另外,使用者也可赋予一个字符串,如:

```
> y = "hello"
> y
[1] hello
```

现在变量 x 和 y 都储存在 R 的工作空间(workspace)中,使用者可输入 objects() 来检查目标中对象的清单。要删除不需要的对象,如 x,可用 remove("x")或 rm(x)。

问题:怎样可以把 x 和 y 的数值对调呢?

1.8 向量

上述所用的例子都是标量的,但在统计学里,多数数据都是以一组来表达的。R 中,使用者能以向量形式来输入一组数字。举例来说,在一个重复实验中得出 10 个结果如下:

2, 4.6, 1, 3.7, 5.9, 4.0, 6.7, 2.8, 1.4, 3.1

使用者可以利用向量形式把这组数据储存在一个变量中:

```
> observations = c(2, 4.6, 1, 3.7, 5.9, 4.0, 6.7, 2.8, 1.4, 3.1)
```

在这里 observations 是一个包含了 10 个数值的向量,而 c()则指示 R 在括号中的数值是以向量形式输入。而再输入 observations 时,则会显示变量的内容:

```
> observations
[1] 2.0 4.6 1.0 3.7 5.9 4.0 6.7 2.8 1.4 3.1
```

向量的运算与标量的运算是相同的。例如使用者可将 observations 的单位由英吋转换成厘米: