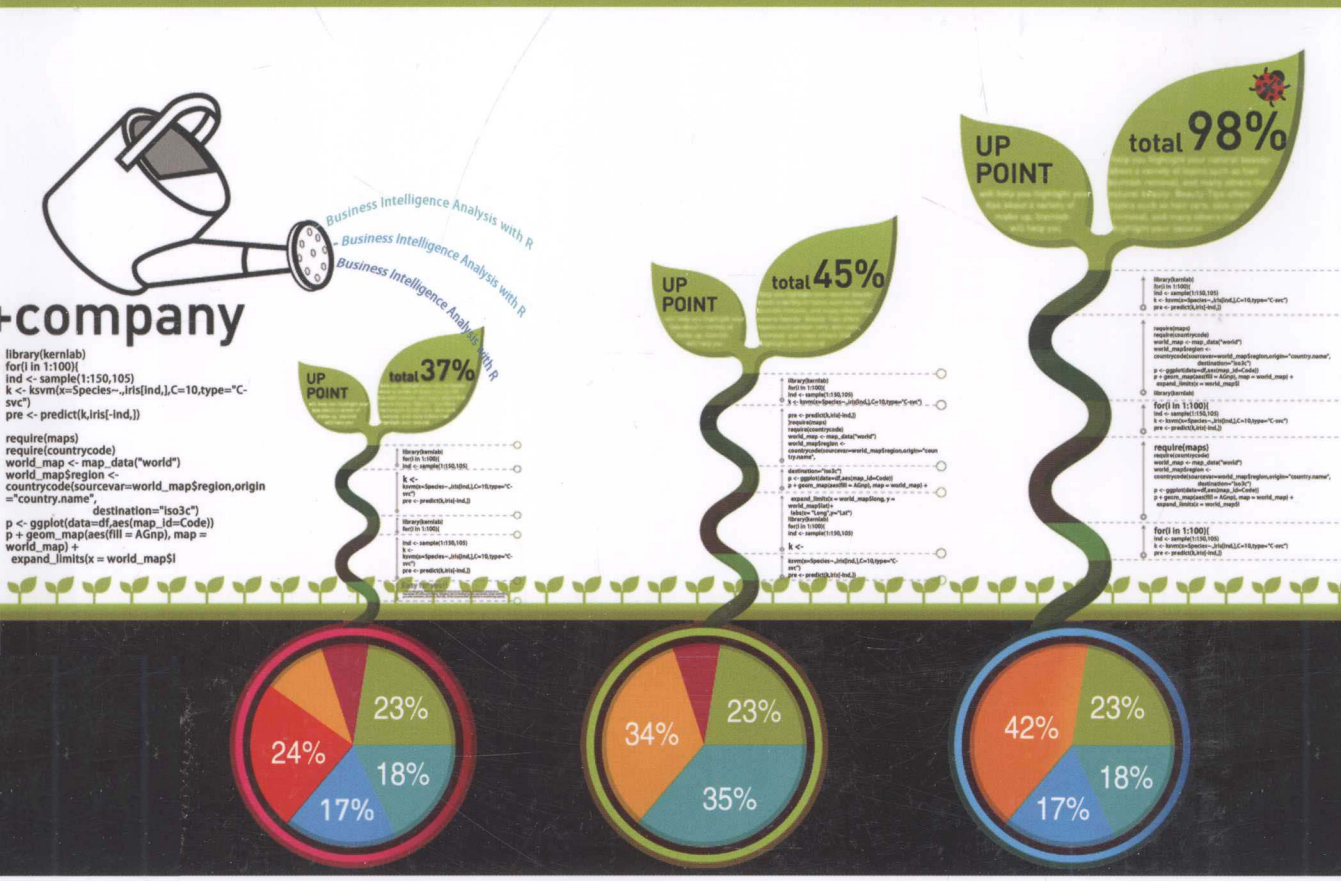
R语言与商业智能

韩　伟　毛俊杰 编著



電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

R语言与商业智能

韩 伟 毛俊杰 编著

電子工業出版社
Publishing House of Electronics Industry
北京 • BEIJING

内 容 简 介

本书主要介绍了R语言在商业智能中的应用，全书共分为两部分：第一部分介绍了R语言的基本使用，包括基本操作、数据结构、数据整理、基础和高级函数及其应用等；第二部分通过实例介绍了R语言在具体学科中的运用，涉及到金融、统计、经济、数据挖掘等多方面的案例。

本书偏向于实际应用而非技术或方法的研究，希望能起到抛砖引玉的作用，有效引领读者开始体验商业智能和R语言的魅力。

本书既可以作为财经类大学营销、物流、金融、经济、统计、工商管理、经济科学、国贸、计算机、信息管理等专业选修数据挖掘与商业智能的教材，也可用于社会培训机构作为R语言课程教材，同时也适用于没有编程经验的数据分析工作者和爱好者。

图书在版编目（CIP）数据

R 语言与商业智能/韩伟，毛俊杰编著. —北京：电子工业出版社，2014.1
ISBN 978-7-121-21706-7

Ⅰ.①R… Ⅱ.①韩… ②毛… Ⅲ.①程序语言—程序设计—应用—经济 Ⅳ.①TP312②F-39

中国版本图书馆 CIP 数据核字（2013）第 247252 号

策划编辑：张贵芹　张洪锐
责任编辑：张贵芹
文字编辑：张洪锐
印　　刷：三河市鑫金马印装有限公司
装　　订：三河市鑫金马印装有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　邮编　100036
开　　本：787×1 092　1/16　印张：12.5　字数：296 千字
印　　次：2014 年 1 月第 1 次印刷
定　　价：29.80 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

服务热线：（010）88258888。

前　言

在大数据时代来临的背景下，R 语言作为一种迅速兴起的数据分析工具软件，正越来越多地成为数据分析项目的实施基础。

在商业智能的教学过程中，如果缺少相应的实际训练，是很难掌握好理论的。而传统使用的软件主要有两种：一种是全封闭的，用户无法得知软件在运行中究竟完成了什么工作，从而理论和实际得不到紧密地结合，且定制性很难满足教学实践的需求；另外一种是可编程的，这类工具虽然可以很好地定制，能够通过编程来实现模型和计算，但是在使用时也会遇到很多问题，比如这些工具太过于底层，在使用这些软件的时候，会花费很多时间在底层的计算上。因此需要一个工具，它既有很好的定制性，能透明地了解其工作原理，并且对于底层的一些工作已经有了现成的实现。R 在很大程度上就能够满足这种需要，R 是一个具备计算和绘图环境的语言，在实践中需要使用的一些底层的数据计算、统计计算都已经在 R 软件或其自带的一些基础包中有了很好地实现，并且 R 是可以免费下载的开源软件。此外，R 和其他很多工具都能很好地结合，比如 Microsoft SQL Server 以及在本书中介绍的 MySQL 数据库。

本书正是基于此而编写，主要介绍如何在商业智能中使用 R 软件。本书依托金融、营销、统计学等学科基础，利用行业背景知识，讲解 R 语言及其在商业智能上的应用。本书内容主要分为两部分：第一部分介绍 R 语言的基本使用，包括基本操作、数据结构、数据整理、函数和绘图等内容。在每一章的开始处，对本章内容进行了简单的整体介绍，这有助于读者对本章内容有大体的了解；在结尾处，对每章节进行补充，所补充的是正文中使用到的但没有细讲的那部分内容。第二部分为案例分析，结合具体案例来深入介绍 R 在商业智能方面的运用，案例分析的章末对本章的知识点进行了总结，列出了重要知识点，有助于在案例分析以后对所用、所学的知识的积累。

本书既不是为了全面地介绍商业智能的各个方面或者商业智能使用的各种技术细节，也不是为了介绍 R 作为一门计算机语言的各种实现细节，而是从实际应用的角度向读者介绍如何在商业智能中使用 R 软件。因此在编写的时候，尽量避免写成 R 的使用手册，也极力避免过度纠缠商业智能技术的细节。希望本书能够抛砖引玉，引领读者开始体验商业智能和 R 的魅力。

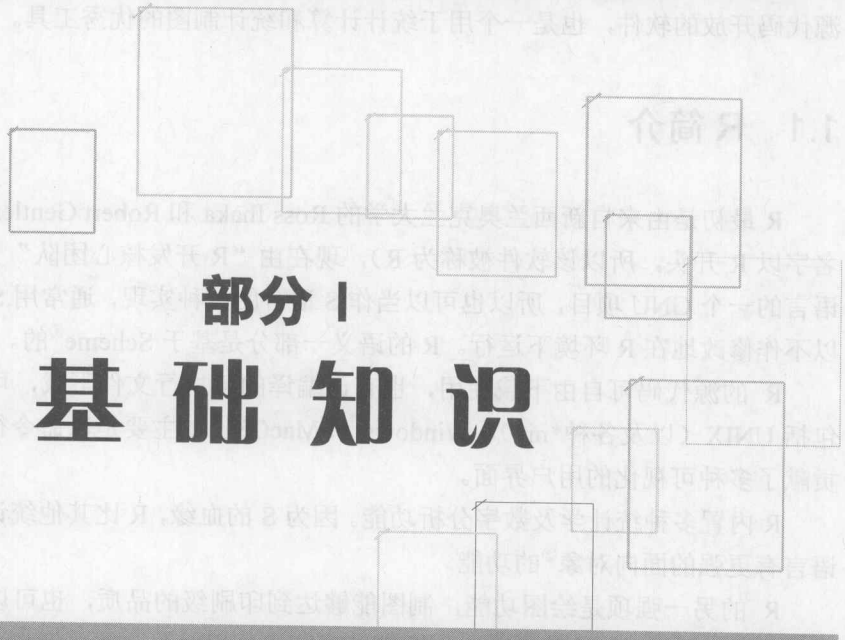
在本书的写作过程中，参考了大量的有关 R 的在线文档，这些文档大多会和本书使用到的加载包一起发布。此外，互联网上大量的关于 R 的博客也对本书的完成提供了思路和借鉴，在此表示感谢。同时对 R 核心小组和大量的 R 包的开发者表示感谢，正是因为他们的贡献才有了 R 的发展。

编著者

目　录

部分 I　基 础 知 识

部分 II 案 例



部分 I

基础知识

第1章 R 简 介

R 是用于统计分析、统计绘图的语言和操作环境，是属于 GNU 系统的一个自由、免费、源代码开放的软件，也是一个用于统计计算和统计制图的优秀工具。

1.1 R 简介

R 最初是由来自新西兰奥克兰大学的 Ross Ihaka 和 Robert Gentleman 开发（由于他们的名字以 R 开头，所以该软件被称为 R），现在由“R 开发核心团队”负责开发。R 是基于 S 语言的一个 GNU 项目，所以也可以当作 S 语言的一种实现，通常用 S 语言编写的代码都可以不作修改地在 R 环境下运行。R 的语义一部分是基于 Scheme[①]的。

R 的源代码可自由下载使用，也有已编译的可执行文件下载，可在多种平台下运行，包括 UNIX（以及各种*nix）、Windows 和 MacOS。R 主要是以命令行操作，但也有志愿者贡献了多种可视化的用户界面。

R 内置多种统计学及数字分析功能。因为 S 的血缘，R 比其他统计学或数学专用的编程语言有更强的面向对象[②]的功能。

R 的另一强项是绘图功能，制图能够达到印刷级的品质，也可以在图中方便地加入数学符号。虽然 R 主要用于统计分析或者开发统计相关的软件，但也可以用作包括矩阵计算在内的数值计算（需要加载其他包，其实 R 中使用了优秀的开源数学计算库）。其计算速度在很多时候可媲美 GNU Octave[③]，甚至商业软件 MATLAB。

R 的功能能够通过由用户撰写的包来增强。增加的功能包括但不限于统计技术、绘图功能、编程界面和数据输入/输出功能。这些软件包是由 R 语言、LaTeX、Java、C 语言和 Fortran 编写。下载的可执行文件会带有一批核心功能的加载包，根据 CRAN 纪录，有千种不同的软件包。在 CRAN 上对这些包进行了分类，具体的类别可以查看网站 http://cran.r-project.org/web/views/。

还有很重要的一点就是：R is free，在这里 free 不仅仅只是指免费，更指“自由”这个含义。R 是一个免费的自由软件，它有 UNIX、Linux、MacOS 和 Windows 版本，都是可以免费下载和使用的。在 http://www.r-project.org/网站可以下载到 R 的安装程序、各种加载包和文档。在 R 的安装程序中只包含了八个基础模块，其他模块可以在网站上获得。

---

① 函式编程语言，一种 Lisp 方言，更多信息参见 http://zh.wikipedia.org/wiki/Scheme
② 其实 R 也不是严格意义上的面向对象的语言
③ 模仿 MATLAB 设计的，语法和 MATLAB 极为相似，和 MATLAB 的关系类似于 R 和 S-Plus 的关系

1.2　R 的发展历史

R 自从 1993 年首次发布以来，就一直没有停止过前进的脚步，相对于其他商业软件来说，R 的发展是飞速的，抛开其快速增长的加载包不说，其软件本身也在不断地更新中。最近的一次重大改变是 R3.0.0（代号为 Masked Marvel）的发布，这次主要改变了其对长向量的支持，即对长度为 2^{31} 以上向量的全面支持（仅限 64-bit 系统，32-bit 系统将会报错）。

1.3　R 的功能

R 是一套完整的数据处理、计算和制图软件系统。其功能包括：数据存储和处理系统、数组运算工具（在向量、矩阵运算方面的功能尤其强大）、完整连贯的统计分析工具、优秀的统计制图功能、简便而强大的编程语言、可操纵数据的输入和输出，可实现分支、循环，用户可自定义功能。

与其说 R 是一种统计软件，还不如说 R 是一种数学计算的环境，因为 R 并不是仅仅提供若干统计程序，使用者只需指定数据来源和若干参数便可进行统计分析。R 的思想是：它可以提供一些集成的统计工具，但更大量的是它提供各种数学计算、统计计算的函数，从而使使用者能灵活机动地进行数据分析，甚至创造出符合需要的新的统计计算方法。

R 语言的语法表面上类似 C，但在语义上是函数设计语言（functional programming language）的变种，并且和 Lisp 以及 APL 有很强的兼容性。特别的是，它允许在“语言上计算”（computing on the language）。这使得它可以把表达式作为函数的输入参数，而这种做法对统计模拟和绘图非常有用。

1.4　CRAN 和 Bioconductor

CRAN 为 Comprehensive R Archive Network（R 综合典藏网）的简称。它除了收藏了 R 可执行文件、源代码和说明文件，也收录了用户撰写的各种软件包。现在全球有超过一百个 CRAN 镜像站。在中国大陆公开的镜像有五个，比较常用的有北京交通大学、厦门大学、中国科学技术大学三个镜像，选择合适的镜像可以加快下载的速度。更多的镜像可以查看 http://cran.r-project.org/mirrors.html。

另外一个和 CRAN 类似的网站是 Bioconductor，也是一个集合了很多 R 包的资源丰富的网站，只是其在生物学领域的使用更加广泛一点，上面的很多包在数据的交互可视化方面是很优秀的，更多的信息可以访问其网站 http://www.bioconductor.org/。Bioconductor 的镜像比较少，目前大陆公开的镜像只有中国科学技术大学一个。

1.5　R 的缺点

和任何事物一样，R 也不是完美的。R 现在主要的问题有三个，首先 R 是一种解释性语言，和编译性语言相比，速度显得略慢一点，但是随着硬件和 R 自身的发展，这个问题

已经慢慢消失了，而且如果能够熟练使用向量化运算，可以大大提高速度，即使使用 R 无法满足速度的要求，我们也可以使用 C 语言或者 Fortran 来编写底层代码供 R 调用。

另外一个问题是，R 所有的计算都是在内存中进行的，也就是说，在处理数据的过程中，数据必须完整地装入内存当中，这在处理小型数据时是没有任何问题的，但是当遇到很大的数据的时候，问题就会变得很严重。其实，这个问题也得到了一定的解决，我们在本书后面的章节中会使用数据库来缓解这个问题。

最后一个问题是，由于 R 语言的自由，各种包的编写者来自不同的领域，所以在一定程度上是比较混乱的，没有统一的命名格式，参数格式不一，源代码和文档质量良莠不齐。当然，这个问题对于一般的用户来说并不是很严重，这是开源带来的问题之一。但是，也正是因为 R 的自由开源的性质，造成其问题会被迅速发现，然后提交给原作者，可以很快得到修改。自由软件开放性带来的利弊一直是争议不断的。

1.6 安装 R

以下简述 R for Windows 的安装和使用。在 http://www.r-project.org/下可以找到 R 的各个版本的安装程序和源代码。单击进入“Download R for Windows”，再单击“base”，下载 R-3.0.1-win.exe，便是 R For Windows 的安装程序。双击 R-3.0.1-win.exe，按照提示一步步安装即可。安装完成后，程序会创建 R 程序组，并在桌面上创建 R 主程序的快捷方式（也可以在安装过程中选择不要创建）。通过快捷方式运行 R，便可调出 R 的主窗口，如图 1.1 所示。

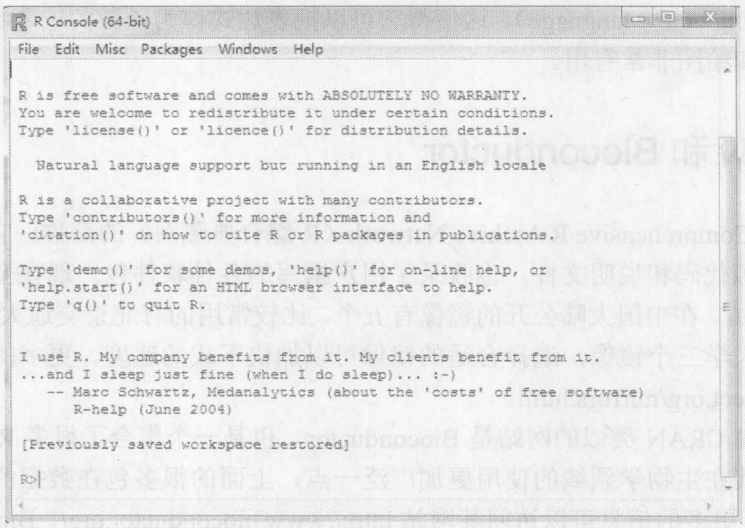


图 1.1 RGUI 启动界面

类似于许多以编程为主要工作方式的软件，R 的界面简单而朴素，只有不多的几个菜单和快捷按钮。快捷按钮下面的窗口便是命令输入窗口，它也是部分运算结果的输出窗口，有些运算结果则会输出在新建的窗口中。

主窗口上方的一些文字是刚运行 R 时出现的一些说明和指引，文字下的“：>”符号便是 R 的命令提示符（可以使用 options 函数来设置自己喜欢的提示符，比如截图中的提示符

就是使用 options(prompt="R>")修改后的结果），在其后可输出命令。在 R 朴素的界面下，是丰富而复杂的运算功能。

安装完成后，可以对 R 进行一系列的配置，这也正是 R 自由的地方，配置方法可以参考 https://github.com/yihui/r-ninja/blob/master/01-setup.md 上的内容。

1.7　R 的使用

1.7.1　第一次使用 R

在初次使用 R 的时候，面对的是一个等待输入的界面。R 的语句由函数和赋值语句组成，在 R 中可以使用等号“=”来赋值，也可以使用箭头“<-”来赋值，在实际使用中，这两种方法几乎没有区别。但是，一般在赋值的时候使用箭头，在传递参数的时候使用等号（必须使用等号，虽然有的时候用箭头也可以运行，但是很危险）。可以在 R 中输入以下的例子：

```
#  第一个例子
x  <-  1:10
x

##  [1] 1   2   3   4   5   6   7   8   9  10

y  =  1:10
y

##  [1] 1   2   3   4   5   6   7   8   9  10
```

在上面的例子中，使用等号和使用箭头的结果是一样的。我们不推荐使用等号赋值，当然这不是强制的。在这个例子中有一行注释，在 R 中使用井号“#”来注释。

1.7.2　获取帮助

在使用 R 的过程中，如果遇到了问题，在大部分情况下，阅读 R 自带的帮助文档就可以很好地解决。比如想生成 100 个正态分布的随机数，虽然知道要使用函数 rnorm，但是不知道如何使用，这时就需要获取帮助，获取帮助的方式有两种：

```
#  查看函数的帮助
help(rnorm)          #等价于?rnorm
#  模糊查询
help.secrch("rnor") #等价于??rnor
```

这样就可以查看 rnorm 函数的文档了。R 中还有其他一些和帮助相关的函数，可以使用 example("函数名")来查看一个函数的例子的运行结果，也可以使用 vignette()来列出包当中的小短文，这些小短文是很实用的，一般是 pdf 格式，比如使用 vignette("arules")就可以查看 arules 在实际中是如何使用的，这对于学习者是很有帮助的。

1.7.3　工作空间和工作目录

R 中一个重要的概念是工作空间，R 是在内存中运行的，所使用的数据和函数等都在内存中，这被称之为工作空间。可以使用 ls()来列出当前工作空间中的所有对象，使用 rm()

来删除工作空间当中的某一个对象，特别地，可以使用 rm(list= ls(all.names=T))来删除所有的对象，包括其中隐藏的对象。

另外一个概念是工作目录，工作目录是一个文件夹的路径，这个路径表示的是当前在哪一个文件夹下工作。在存取文件时，如果不指定路径的话，就会默认为这个文件夹。在 R 中，可以使用 getwd（即 get work directory）函数来获得当前目录，如果要改变当前工作目录，可以使用 setwd 函数来设置。

```
getwd()
##  [1]  "G:/RBook/1"
setwd("G:/")
```

1.8 包的安装和使用

R 的一个主要特点是其具有众多的加载包，这些包大部分是由某一领域的专业人士编写，大部分包都可以从 CRAN 和 Bioconductor 上找到。所谓的包就是一系列的函数集合，可能还带有数据和文档等内容。在 R 的基础包中只有很少的一部分包（相对来说，这部分包是比较稳定和高质量的），其他的包用户如果要使用的话，需要另外安装。安装的方法有两种，第一种是直接去 CRAN 的网站上下载，另一种是在 R 中使用命令来安装。第二种方法的好处是不需要考虑各种包之间的依赖关系[①]，而且不用去网站上寻找需要的包。在这里推荐第二种方法，下面演示如何安装使用一个包：

```
#  使用第二种方式安装包
install.packages("安装的包名",  dependencies  =  TRUE)
#  加载一个包
library("包名")
#  查看一个包的使用说明
library(help  =  "包名")
```

我们使用 install.packages 来安装一个包，但是，安装完成后，如果要使用，还需要使用 library 函数来加载。如果想查看一个包的使用说明，也可以使用 library 函数，说明中包含了这个包的简述、作者的信息以及所有的函数。

1.9 其他辅助工具

与 SAS 和 SPSS 等软件相比，R 的缺点在于没有很好的操作菜单，很多不习惯使用命令行的用户起初会觉得有困难。但是，R 自由的特性得到了很好的发挥，有用户贡献了 R 包，实现了很多功能的菜单化操作。下面介绍两种可以实现菜单化操作的包以及一个比较友好的编辑器。

1.9.1 Rcmdr 包：实现 R 的菜单化操作

可能大多数人刚开始学 R，对 R 的 Console 深有感触，它非常简洁，没有提供像 SPSS、STATA、SPLUS 等菜单式操作，因此很多人对 R 望而却步，不知从何下手。R 最传统的工作方式为命令行，很多人开始学习 R 就是从命令行开始的。

R 也完全可以支持“鼠标为主”的用户工作方式，对于初学者，我们推荐 Rcmdr 包，

① 所谓包的依赖关系是指使用 A 包的同时，A 包又使用了 B 包，所以在安装 A 包的时候需要安装上 B 包

该包主要使用了 R 中最基础的 tcl/tk 等包，只要安装了推荐的一些包，再加上 Rcmdr 包中的图像，就可以使用 R 中几乎所有的统计分析工具了。Rcmdr 让你的统计分析菜单化，但不傻瓜化，因为菜单操作同时也提供了 R 对应的命令。我们可以在单击菜单的同时，学习使用 R 中的命令行，这是很有帮助的，因为这样做有助于把命令行和它们的作用联系起来。

另外，R 和 Java、Gtk 等结合得非常好，也有人做了 R 的(D)COM。已经有一些 GUI 方式的界面出来，“鼠标化”的程度想必会越来越高。Rcmdr 是基于 R 最基础的包而来的，GUI 具有很好的扩展性。

可以按以下步骤安装 Rcmdr：

（1）安装 Rcmdr 包，输入：install.packages("Rcmdr")，回车，接着让其自动操作，选择一下镜像网站就可以了。

（2）输入：library(Rcmdr)，回车就可以运行，如果关闭后，需要再次运行，则需输入 Commander()。

运行后就会出现如图 1.2 所示的图形界面，在这个界面上可以实现几乎所有的统计分析方法。

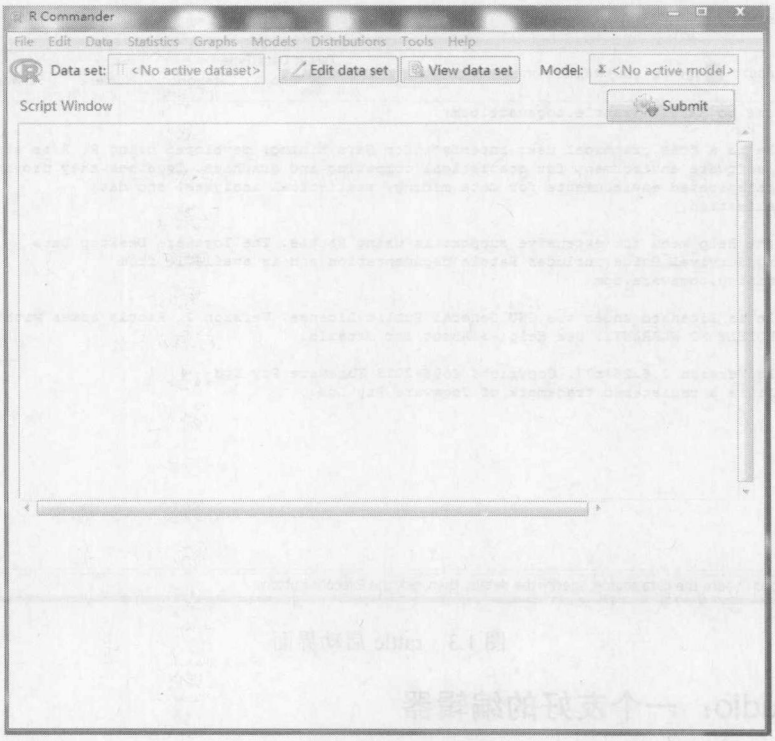


图 1.2　Rcmdr 启动界面

1.9.2　rattle：可视化数据挖掘工具

数据可视化旨在借助图形化手段，清晰有效地传达与沟通信息。但是，这并不意味着，数据可视化就一定因为要实现其功能用途，而令人感到枯燥乏味，或者是为了看上去绚丽多彩，而显得极端复杂。为了有效地传达思想概念，美学形式与功能需要齐头并进，通过

直观地传达关键的方面与特征，从而实现对相当稀疏而又复杂的数据集的深入洞察。然而，设计人员往往并不能很好地把握设计与功能之间的平衡，从而创造出华而不实的数据可视化形式，无法达到传达与沟通信息的目的。

R 语言有众多的绘图工具包，例如 ggplot2、lattice 等。而在动态绘图方面，则可以利用 rggobi 与 ggobi 软件进行协同工作。对懒得敲命令的读者来说，还可以利用 rattle[①]工具包的图形界面进行数据挖掘和可视化工作。在 R 程序内运行如下命令：install.packages ('rattle', dep=T)，即可安装 rattle。

与 Rcmdr 的安装和使用方式很像，安装加载完成后，输入 rattle()后，就会出现如图 1.3 所示的界面。

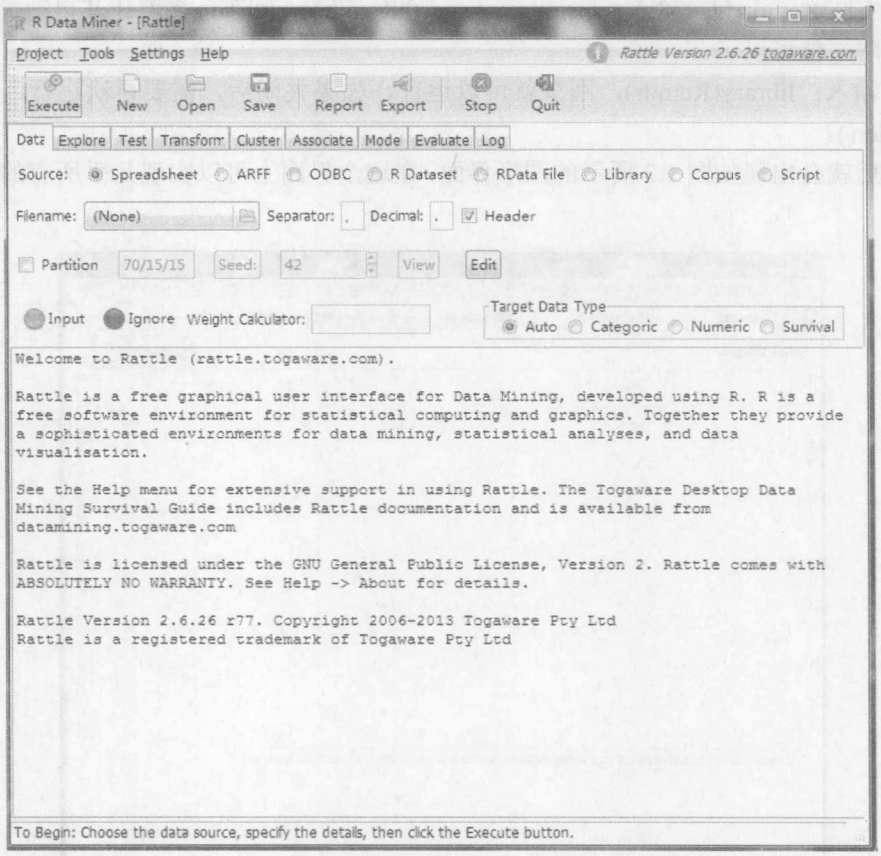


图 1.3 rattle 启动界面

1.9.3 Rstudio：一个友好的编辑器

R 自身带的编辑器很不好用，因此可以寻找很多的替代方案，比如可以选择 Emacs 和 Vim 来替代。我们这里推荐 Rstudio，它是专门用于 R 语言环境的 IDE。Rstudio 可以从其官网 http://www.rstudio.com/上免费下载安装。Rstudio 分为桌面版本和服务器版本，一般情况下，下载安装桌面版本即可。安装完成的启动界面如图 1.4 所示。

① rattle 是 R Analytic Tool To Learn Easily 的首字母缩写，有趣的是，这个单词自身的意思是“震颤、翻滚”，在这里可以表示翻滚数据的意思