Broadview®
www.broadview.com.cn

R数据分析
方法与案例详解

方匡南 朱建平 姜叶飞 编著

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

R数据分析
方法与案例详解

方匡南 朱建平 姜叶飞 编著

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

内容简介

本书是一本R语言和数据分析的入门教材，循序渐进、深入浅出，每个知识点尽量从实际的应用案例出发，以问题为导向，在解决问题中学习统计方法、R语言的基本使用以及编程技巧。

本书内容涵盖R数据结构、函数与优化、抽样模拟、统计分析、假设检验、回归分析、统计绘图和R包制作等内容。

本书的定位是为业界数据分析人员、经济管理类、医学的学生提供方法和程序上的参考，在写作过程中尽量删去比较理论的数学原理，这样能够帮助读者轻松上手学习。

图书在版编目（CIP）数据

R数据分析：方法与案例详解 / 方匡南，朱建平，姜叶飞编著. —北京：电子工业出版社，2015.2
ISBN 978-7-121-25290-7

Ⅰ. ①R… Ⅱ. ①方… ②朱… ③姜… Ⅲ. ①程序语言—程序设计—教材 Ⅳ. ①TP312

中国版本图书馆CIP数据核字（2014）第305474号

责任编辑：张月萍
印　　刷：三河市双峰印刷装订有限公司
装　　订：三河市双峰印刷装订有限公司
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：787×980　1/16　　印张：24.5　　字数：583千字
版　　次：2015年2月第1版
印　　次：2015年2月第1次印刷
定　　价：69.00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

R语言是由新西兰奥克兰大学的Ross Ihaka与Robert Gentleman一起开发的一个面向对象的编程语言，因两人的名字都是以R开头，所以命名为“R”。R是“GNU S”，一个免费开源、能够自由有效地用于统计计算和绘图的语言和环境，可以在UNIX、Windows和Mac OS系统中运行，它提供了广泛的统计分析和绘图技术，包括回归分析、时间序列、分类和聚类等方法。2009年，《纽约时报》发表了题为“Data Analysts Captivated by R's Power”的社评，集中讨论了R语言在数据分析领域的发展，并引发了SAS和R用户广泛而激烈的争论。文章认为，让R变得如此有用和如此快地广受欢迎的原因是统计学家、工程师和科学家们在不断精炼代码或编写各种特有而具体的包。而且现在R软件增添了很多高级算法、作图颜色和文本注释，以及为与数据库链接等提供挖掘技术。文中引用了几位科学家对R做的高度评价，如Google首席经济学家Hal Varian讲了一句很好的话：“R最优美的地方是它能够修改很多前人编写的包的代码做各种你所需要的事情，实际上你是站在巨人的肩膀上。”

2010年，美国统计协会（American Statistical Association）将第一届“统计计算及图形奖”授予R语言，用于表彰其在统计应用和统计研究方面广泛的影响。

笔者自接触R语言至今已近10年，记得最早听到R语言是2004年日本同志社文化大学情报研究所金明哲教授前来讲座，他给我们讲了一个关于利用文本挖掘方法破了人身意外保险诈骗杀人案。第一次听到数据挖掘竟然可以如此神奇的破案，便被深深地吸引住了，当初金教授提到其主要分析工具就是R语言。这是笔者第一次接触R语言。

其后，笔者的硕士导师王斌会教授要求我们学习R语言，坦诚地讲，R语言对于刚入门的人来讲确实有点困难，而且那个时候R语言的参考书少得可怜，当时市面上还没有R语言中文教材，用的人也很少。我参与了王斌会老师R语言中文教材的编写小组，一边学习R语言英文经典教材，一边整理学习心得和相关资料。一年之后，第一本R语言中文教材《R语言统计分析软件教程》面世了。通过编写教材，笔者从中体会到了学习R语言的乐趣，自从学会了R语言，笔者之后的数据分析主要使用的都是它。

后来，人大经济论坛邀请笔者为他们录制R语言视频课程，笔者利用闲暇时间整理讲义，先后录制了《R语言初级》《R语言高级》《R金融时间序列初级》和《R金融时间序列高级》等视频课程，反响很好，很多学员要求笔者写相应的系列教材，由于繁重的科研压力以及个人

的惰性，虽有此想法已久，但却迟迟未动笔。后来成为厦门大学教师，系里让教《计算机在统计中的应用》一课，笔者觉得讲R语言是最恰当不过了，因此边上课，边整理讲义，并不断地完善，教该课程已经4年，讲义也先后修改了4年。

再后来，人大经济论坛又邀请笔者为他们在北京、上海等地开设暑期和节假日现场公开课，前来听课的学生既有国内外著名高校的教师和研究生，又有医药公司的数据分析人员以及互联网企业的数据分析人员等，他们对笔者的讲义提了很多有用的建议。笔者在这些公开课的讲课中，不断地完善讲义，最终形成了此书。关于如何取一个恰当而又响亮的书名，确实是一件挺难的事，笔者想了好几个书名，但最终还是认为取名为《R数据分析——方法与案例详解》可能比较合适。

该书是一本R语言和数据分析的入门教材，内容循序渐进、深入浅出，每个知识点都尽量从实际的应用案例出发，以问题为导向，在解决问题中学习统计方法、R语言的基本使用以及编程技巧。本书的定位是为业界数据分析人员、经济管理类和医学类的学生提供方法和程序上的参考，在写作过程中尽量删去比较理论的数学原理，当然有些原理部分无法跳过，所以在学习时需要有一定高等数学和概率论的基础，但如果想真正掌握某个统计方法，那么学习其背后的原理还是非常有必要的。当然，如果对方法的原理确实不感兴趣，只是为了用R程序实现某种方法，或者分析某个有意义的数据，则可以跳过原理，只看案例和程序即可。

该书的姐妹书，会讲解更为高深的统计方法，涉及聚类分析、分类分析、关联规则、时间序列等问题，更注重在互联网、金融、企业营销、基因分析等领域的应用，目前暂定书名为《R数据挖掘——方法与案例详解》。

该书适合作为高校数据分析相关专业的教科书，也适合供医学、市场调查、金融以及互联网等企业的相关数据分析人员阅读。

本书第1章和第2章由方匡南、姜叶飞撰写，第3章和第4章由方匡南撰写，第5章由方匡南、易煌迪撰写，第6章到第9章由方匡南撰写，第10章到第15章由朱建平、方匡南撰写，第16章由方匡南撰写，第17章由方匡南、姜叶飞撰写。第18章由方匡南、张声威撰写。最后由方匡南、朱建平进行加工总纂、修改定稿。

王秉权参与了第4章到第12章的校对、修改等工作。欧阳汉参与了第7章到第9章的校对、修改。

感谢成都道然科技有限责任公司的专业意见和建议。感谢四川大学严珂玮为本书配的精美插画。感谢夫人张晶在我的写作过程中给予的支持和帮助。再次感谢为本书提供了直接或者间接帮助的各位朋友，没有你们的帮助，本书的出版没有这么顺利。

该书的数据可以从网址zhiliaobang.com或者kuangnanfang.com下载。

由于作者水平有限，书中难免有错误和不足之处，恳请读者批评指正！

方匡南

2014.8.1

目录

第1章

初识R语言

本章主要对R语言的基础内容做讲解，使读者可以尽快熟悉这款流行的数据分析工具。本章内容涵盖：R语言的优势、安装流程、扩展包、常用编辑器和工作空间。

1.1 什么是R语言

你现在最常用的统计软件是什么？SAS，MATLAB，还是Eviews？如果你是一个经常和数据分析打交道，需要运用或自己编写各种统计方法的人，但还没用上R，那么你已经脱离现在的主流了。

R语言是由新西兰奥克兰大学的Ross Ihaka与Robert Gentleman一起开发的一个面向对象的编程语言，因两人的名字都是以R开头，所以命名为“R”。R是“GNU S”，一个免费开源、能够自由有效地用于统计计算和绘图的语言和环境，可以在UNIX、Windows和Mac OS系统中运行，它提供了广泛的统计分析和绘图技术，包括回归分析、时间序列、分类和聚类等方法。R语言的前身是S语言，S语言是贝尔实验室（Bell Laboratories）的Rick Becker、John Chambers和Allan Wilks共同开发的，提供了一系列统计和图形显示工具，这个语言过去一度是数据分析领域里面的标准语言，但现在正逐步被R语言取代。

R是一套完整的数据处理、计算和制图软件系统，它是一套开源的数据分析解决方案，由一个庞大而活跃的全球性社区维护。其功能包括：数据存储和处理系统；数组运算工具（其向量、矩阵运算方面功能尤其强大）；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是一种统计分析与计算的环境，因为R不仅提供若干统计程序，而且还可进行统计分析，只需使用者指定数据库和若干参数即可。R的思想是：它可以提

供一些集成的统计工具，更重要的是，它还可以提供各种数学计算、统计计算的函数，从而令使用者（用户）能够灵活地进行数据分析，甚至创造出符合需要的新的统计计算方法。随着R功能的不断完善，也提供了如图1-1所示的许多高级功能。

图1-1 R能够提供的高级功能

1.2 为什么用R语言

2009年，《纽约时报》发表了题为“Data Analysts Captivated by R's Power”的社评，集中讨论了R语言在数据分析领域的发展，并引发了SAS和R用户广泛而激烈的争论。文章认为，让R变得如此有用和如此快地广受欢迎的原因是统计学家、工程师和科学家们在不断精炼代码或编写各种特有、具体的包。而且现在R软件增添了很多高级算法、作图颜色、文本注释以及与数据库等链接的挖掘技术。

KDnuggets网站每年都会做一些数据分析和数据挖掘软件使用的专题问卷调查。据KDnuggets网站2011年对570个数据挖掘和数据分析的工作者关于过去12个月数据挖掘和数据分析所使用的编程语言的调查显示（http://www.kdnuggets.com/2011/08/poll-languages-for-data-mining-analytics.html），R语言名列榜首，如图1-2所示，占接近半壁江山（45%），而紧随其后的SQL、Python和Java则在某一领域具有各自独到的优势，而SAS和MATLAB分别名列第5和第6，被R远远甩在后面。

除了R之外，现在还有许多人使用SPSS、SAS或MATLAB，但大都用的是盗版软件。免费是R流行开来的最大的一个因素。不过，R的最大优点是出色的可视化图形、丰富的统计方法及高效的更新速度。其主要特点如下。

① 高效的数据处理和保存机制。

② 完整的数组和矩阵操作运算符，以及完整的数据分析工具。

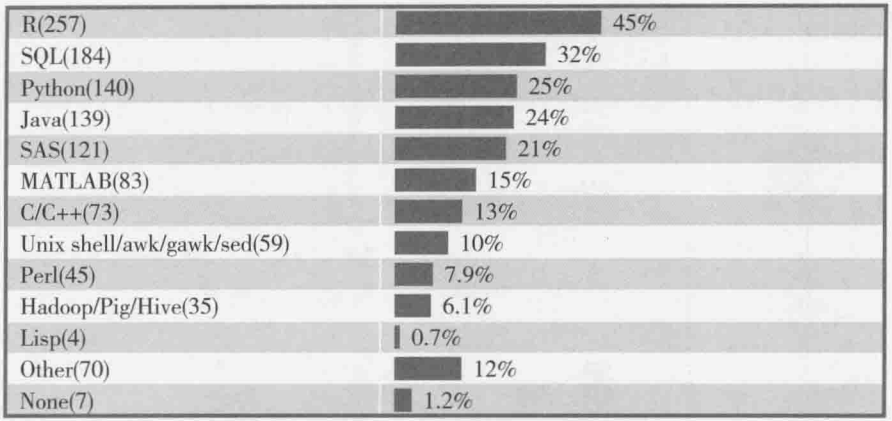

图1-2　数据挖掘和数据分析编程语言使用调查结果

③ 出色的图形统计功能。

④ 简单高效的建模工具。

⑤ 作为一个开源软件，R具有丰富的扩展包（packages），可以自由加载其他开发者提供的函数和数据包，直接利用可以节省很多重新编写算法的精力。

⑥ 提供了很多高级功能。除统计外，还可以使用R来给电脑关机、发微博、发校内状态、下五子棋，或是配合LaTeX撰写动态统计报告以及自动生成概率统计的试卷和答案，等等。

⑦ 几乎兼容全平台。除了支持OS X、Linux和Windows之外，甚至可以在iOS设备上编辑和运行R的程序，还可以在iPhone等移动设备上安装R程序。

⑧ 逐渐支持多国语言。作为一个开源软件，R在其主页上提供了可供大家添加自己国家语言的文件(http://developer.r-project.org/Translations.html)，其中，中文的翻译在https://github.com/r-cn/r-cntrans 上面得到了国内众多的R爱好者的支持。

⑨ 更新速度快。R几乎囊括了所有统计方法，当其他软件还不能完成一些最新的统计方法时，在R中几乎都可以完成。R的更新速度是以周来计算的。

R功能强大、内容丰富，各种帮助文档很多。目前R已有近5000个扩展包，这些包由独立贡献者编写，质量参差不齐，文档零散而且很难找到。不得不承认R的学习曲线比较陡峭。要掌握R的所有功能是非常困难的，因此必须有一本能带你从入门到精通的参考书，这样会让你事半功倍！本书总结作者10多年R的使用经验，深入浅出地讲解了R语言如何进行数据分析和作图。

1.3 安装R

R的官方网站是http://www.r-project.org，打开该网站，界面如图1-3所示。

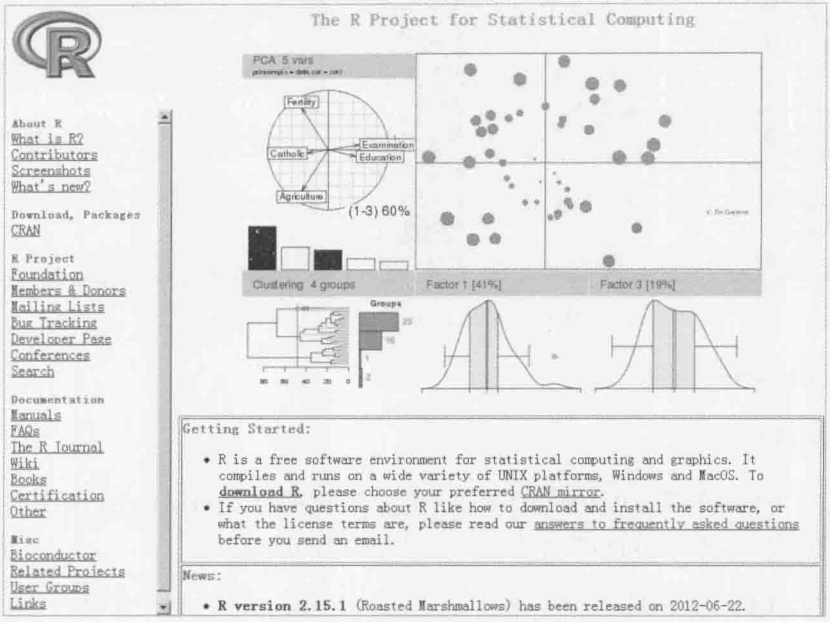

图1-3 R官方网站首页

R安装软件和安装包（package）的下载链接在首页左侧的CRAN。R除了自己的主程序外，还有5000多个用户贡献的扩展（附加）包，以及各种文档，这一大箩筐东西被复制到世界各地几十台服务器上供用户下载。

作为中国用户，一般都会选择一个国内的镜像。目前，中国大陆的镜像有如下4个。

http://ftp.ctex.org/mirrors/CRAN/	CTEX.ORG
http://mirror.bjtu.edu.cn/cran	Beijing Jiaotong University, Beijing
http://mirrors.ustc.edu.cn/CRAN/	University of Science and Technology of China
http://mirrors.xmu.edu.cn/CRAN/	Xiamen University

在R官方主页单击download R，选择对应的镜像后，在右侧下载和安装包栏目里会出现三种操作系统的R版本（Linux、OS X和Windows），选择相应的操作系统后，再选择base就会进入R的下载页面，在页面上会出现R的最新版本和安装说明等文件。

1.4 R扩展包

整个R系统主要是由一系列程序包（package）组成的。R包是函数、数据和预编译代码以

一种定义完善的格式组成的集合。这些包默认储存在library里。R安装好后，默认自带了base、datasets、 utils、stats、grDevices、graphics和methods包，除这些包外，CRAN上还有数千附加包，由R核心开发团队之外的用户自行提交，如此多的程序包基本上保证了R可以实现目前几乎所有的统计分析。扩展包需要下载安装并载入到会话中才能使用，接下来介绍如何安装和载入R扩展包。

1.4.1　R扩展包的安装与载入

R有多种安装方法，这里主要介绍其中两种。

① 在线安装。例如，需要安装“class”这个扩展包，则输入命令 install.packages("class")执行即可。也可以同时安装多个包，例如，需同时安装“class”和“cluster”两个包，则输入命令install.packages(c("class","cluster"))执行即可。

② 利用RStudio安装。RStudio在1.5节会详细介绍。在RStudio右下角栏目里单击“packages”，然后再单击 “install packages”，会出现如图1-4所示对话框，在packages对话框里输入需要安装的包名，如输入 class，就可以安装包。

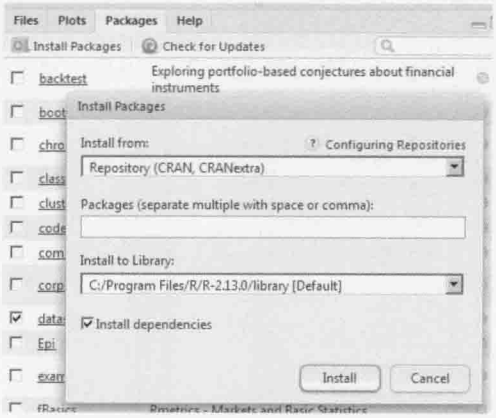

图1-4　用RStudio安装包

安装完R扩展包后，并不能马上使用，需要载入到当前会话中才能使用。R包的载入有两种方法。

① 在R Console中输入命令library()载入。如需要载入class包，则输入 library(class)即可。

② 利用RStudio载入。安装好后的包会出现在RStudio右下方的栏目里，即如图1-5所示对话框，单击左侧的方框，即可载入相应的包。

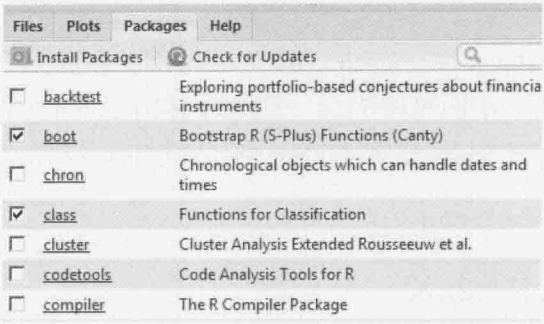

图1-5 RStudio载入包

一个包仅需安装一次即可，但是R的扩展包经常被更新，要想使用更新后的扩展包，则需要对已安装的包进行更新。下面介绍两种常用方法。

① 在R Console中输入命令update.packages()，则会出现需要更新的包和相关信息，并可以选择是否需要更新。如果选择“y”，则会在线更新。

```
> update.packages()
boot :
 Version 1.2-43 installed in C:/Program Files/R/R-2.13.0/library
 Version 1.3-2 available at http://mirrors.xmu.edu.cn/CRAN
Update (y/N/c)?
```

② 利用RStudio更新R包。单击“Check for Updates”按钮，会出现如图1-6所示对话框。在左侧单击需要更新的包，即可进行更新。

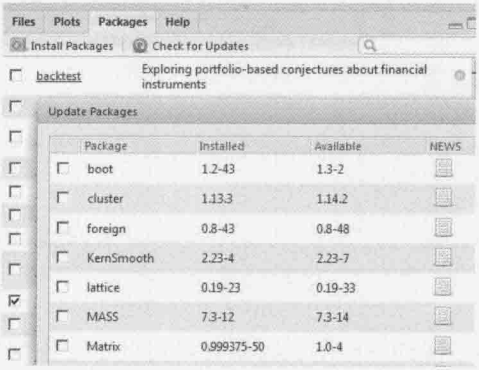

图1-6 利用RStudio更新R包

1.4.2 R包的使用

载入R包后，可以使用该扩展包的函数和数据。例如，我们载入“class”包后，想要查看该包里的函数，一种方法是在R Console里输入命令 help(package="class")，会返回一个包含了“class”包里所有函数和数据集名称的帮助文档，如图1-7所示；另一种方法是在RStudio右侧

单击“class”，则会返回如图1-8所示的帮助文档，如需进一步了解该函数的使用方法，则继续单击相应的函数，就会返回该函数的详细帮助文档。

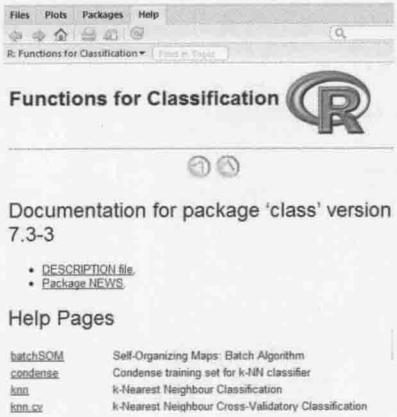

图1-7　class帮助文档

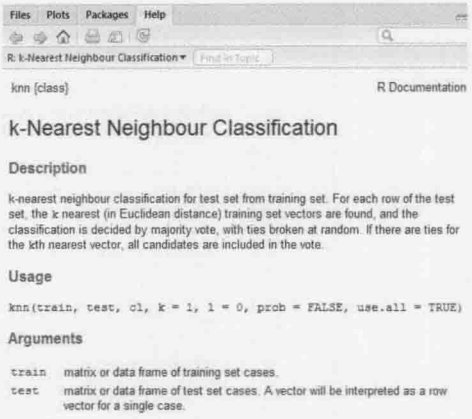

图1-8　knn函数帮助文档

1.5　R编辑器

Windows版本下载安装完成后，在桌面会出现图标。双击该快捷图标，打开R的操作平台（R Console），如图1-9所示。虽然可以直接在此操作平台上输入命令，但是R本身默认的IDE不是很友好，甚至可以用粗糙来形容。通常情况下，建议不要直接在操作平台上输入命令，而是使用R的编辑软件。

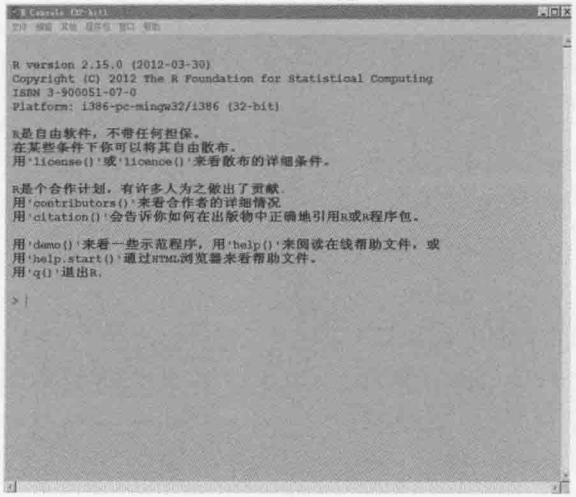

图 1-9　R的操作平台

作为一个开源软件，R不乏各种优秀的IDE。本书主要使用其中一款优秀的跨平台开源IDE——RStudio。RStudio将常用的窗口都整合在一起，使开发者无须在命令行和绘图窗口之间跳

来跳去，更便于控制。不过，要想使用RStudio，除了需要安装R开发环境外，还需要到http://www.rstudio.org/download/desktop/ 上下载适合你电脑的RStudio版本并安装。

打开RStudio，如图1-10所示，左上角是脚本编辑窗口，左下角是命令窗口，右上角是工作空间和历史记录窗口，右下角是作图、帮助和包管理窗口。

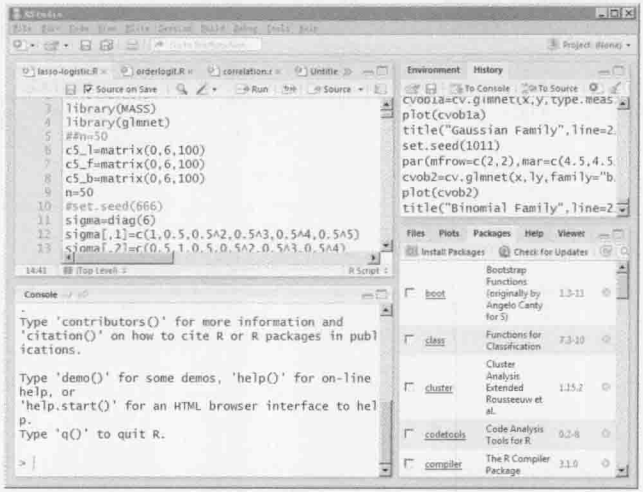

图1-10 RStudio用户界面

如果需要编写R语言代码，则可以用这里提供的两种途径运行。

① 在左下角的命令窗口输入并回车运行。

② 使用左上角的脚本编辑窗口编写好代码之后，根据你的需要选择运行代码。

图1-11 运行R脚本程序

如果需要使用脚本编辑代码，则需要在RStudio中执行File→New File→RScript命令（快捷键Ctrl+Shift+N）来新建一个脚本编辑窗口。写好代码之后，可以通过单击如图1-11所示的Run按钮来运行程序。如果直接单击Run按钮，则只是运行当前行代码。如果先用鼠标选好要运行的代码，然后再单击Run按钮，则会运行所选代码。

在使用脚本编辑窗口进行编辑时，RStudio还提供了一些高级功能。