预防医学教学参考系列

R软件入门与基础

张志杰 编著

开源统计学

复旦大学出版社

R软件入门与基础

张志杰 编著

復旦大學出版社

图书在版编目(CIP)数据

R 软件入门与基础/张志杰编著. —上海:复旦大学出版社,2015.4
预防医学教学参考系列
ISBN 978-7-309-11207-8

Ⅰ. R… Ⅱ. 张… Ⅲ. 统计分析-应用软件-医学院校-教学参考资料 Ⅳ. C819

中国版本图书馆 CIP 数据核字(2015)第 010963 号

R 软件入门与基础
张志杰 编著
责任编辑/傅淑娟

复旦大学出版社有限公司出版发行
上海市国权路 579 号 邮编:200433
网址:fupnet@ fudanpress. com http://www. fudanpress. com
门市零售:86-21-65642857 团体订购:86-21-65118853
外埠邮购:86-21-65109143
上海浦东北联印刷厂

开本 787 × 1092 1/16 印张 20.25 字数 468 千
2015 年 4 月第 1 版第 1 次印刷

ISBN 978-7-309-11207-8/C · 294
定价: 49.00 元

内容提要

R 软件是一款开源、免费的统计软件，与其他软件相比，在图形绘制、统计分析上具有极高的优势，在国外的大学、科研机构及商业机构等很多单位与个人均在使用，应用领域覆盖了社会科学和自然科学的各个领域。本书基于 R3.1.1，从最基本也是最重要的数据管理工作入手，在工作流程框架的指导下基于数据类型以问题——解决方法的思路进行详细介绍。全书分为15章，内容包括R软件概述与简介、数据对象、属性及基本运算、数据集的建立与保存、单数据库数值型数据的编辑与整理、日期型数据的编辑与整理、分类型数据的编辑与整理、字符型数据的编辑与整理、多数据库的编辑与整理、数据的初步描述和分类汇总、图形绘制、R 语言的流程控制与函数编写、R 软件程序包的设计与开发、数据管理的程序包利器、R 与其他编程语言的相互调用及自动化统计分析报告的生成技术，其特点是内容充实、技能突出、实用性强，一切以应用为出发点，帮助大家迅速有效地掌握知识与技能。

本书适用于与数据有关专业的师生、科研人员及业务工作者（如经济学、医学、统计学、会计学、地理学等相关领域），具有很好的实用价值。

前言

随着计算机网络的发展，有效的网速越来越快，手机、Pad 等便携式设备已经被越来越多的人用于收发 email、社交等日常工作中，计算机技术和网络技术的发展融合催生了引领未来信息技术变革的云计算(cloud computing)这一新兴产物。随之而来的是大数据(big data)概念的持续火爆，几乎所有的领域都在寻找来自大数据的灵感，而 Google 公司利用其开发的谷歌流感趋势基于用户搜索数据预测流感的模型居然比美国疾病预防控制中心(CDC)的传统监测数据能平均提前 1～2 周预报流感暴发高峰，这对传统公共卫生的疾病监测技术提出了挑战。国内短时间内对此作出反应，如百度建立了百度疾病预测系统。一时间各种基于云计算、大数据的医学概念开始实践，移动医疗、智慧医疗等新词汇出现在公众视野中，人们对此感到新鲜不已。然而，静下心来思考一下，其实它的核心之一涉及了数据收集与数据分析等与统计学密切相关的内容，因此要让这些先进技术真正地发挥作用，统计学的作用是不可忽视的。在过去的几十年中，SAS，SPSS，STATA 以及 S-PLUS 等商业软件一直主导着这个市场，随着 2008 年 S-PLUS 被 TIBCO 收购、2009 年 7 月 SPSS 被 IBM 收购等事件的发生，商业统计软件的前景堪忧，似乎正在陷入一次“经济危机”。就像人类每一次全球经济危机后都将催生一次技术变革或产业革命、诞生某种新兴事物一样，“统计软件界”也正经历着这样一场变革。

过去，国内由于各种复杂原因使得盗版的商业统计软件泛滥，人们不需要花费昂贵的费用便可以使用，使得人们不太关注非商业软件的问题；但随着我国加入 WTO，版权问题日益重要，盗版现象在触及知识产权和相关方利益的同时，也伤害了我国的国际形象，相信在不久的将来盗版问题将彻底解决。本书介绍一款免费的非商业统计软件，确切地讲是笔者倡导的开源软件(open source software)：它不仅仅免费，而且可以获取统

计程序的源代码。这样既能像使用传统的商业统计软件那样完成工作，更重要的是可以了解所用方法的具体实现过程，彻底了解统计技术的“黑箱(blackbox)”，对于提高统计技能大有益处。基于此，笔者提出统计学的新概念：开源统计学(Open Source STATistics，OSSTAT)，定义为“基于开源软件学习统计学，将软件实践与统计理论更加有效地紧密整合，通过实践—理论—实践的反向循环学习模式，不仅学习统计学理论，更重要的是能够掌握统计学理论实现过程的黑箱机制，彻底掌握统计学，并有能力结合各自专业的特点，进一步发展新的统计学方法与技术”。笔者的远景是建立开源统计学的生态圈，向读者展示其生态链上各个环节的开源软件，并结合理论知识讲解，为开源统计学的发展贡献绵薄之力。

大家都知道，从项目的角度来讲，统计学将涉及研究设计、数据收集、数据整理、统计分析、结果解释以及报告撰写的内容，其中数据整理是最耗费时间的部分，因此作为开源统计学系列教程的第一本书，其定位重点就在这里。本书将以R3.1.1软件为基础，首先介绍R软件以及对象、属性及基本运算，然后在数据管理的基本工作流程框架指导下以“提出问题＋解决方法”的思路依次详细介绍，数据集的建立与保存、单数据库不同数据类型的编辑与整理(数值型、日期型、字符型以及特殊的分类型)，以及多数据库的编辑与整理，至此读者应该可以很好地完成数据管理的工作。然后介绍数据的初步描述和分类汇总和图形绘制，为后续的统计分析打下基础。再后为了让读者能够更好地学习，将进一步介绍R语言的流程控制与函数编写、R软件程序包的设计与开发、数据管理的程序包、R与其他编程语言的相互调用等与编程密切相关的内容。最后将讲解自动化统计分析报告的生成技术，让读者接触可重复研究的内容。作为统计学系列的书籍，不是写假设检验的种种方法，而是专注于数据管理，主要有两个原因：一是笔者作为一名应用统计学家，深知数据管理的重要性，没有很好的数据管理技能作为根基，很难成为一名优秀的统计学家；二是受复旦大学管理学院汪嘉冈教授《SASv8基础教程》的影响，笔者不仅仅在科研上得到了汪教授的帮助(如第一篇SCI收录论文的发表)，更重要的是作为曾经的SAS狂热爱好者，在当年阅读该书时，经历了“读第一遍感觉太简单，读第二遍才认为不错，可以解决实际工作中的绝大部分问题，而读第三遍则深深体会其功力与水平”。综合上面的两个原因，因此笔者将选题定为基础，内容定为数据管理，希望也能给读者带来笔者当年的感觉。若如此，则成矣。

本书适用于与数据打交道行业的师生、科研人员以及业务工作者，如经济学、医学、统计学、会计学、地理学等相关领域。参与本书编写的人员还有胡艺和李锐两位博士。封面图片由李锐设计。由于编者水平有限，书中难免有疏漏与不足之处，还望读者提出宝贵意见。本书能够顺利出版，得到了很多人的帮助，如学院的陈晓敏老师等，在此一并感谢，没有您们的付出，也不会有本书的面世。

丰满的理想从未在心灵深处消失，但骨感的现实让人一次次地放弃，在一切重新开始之时，期盼着我所倡导的开源统计学能一领风骚。诚然推动开源统计学的发展，我们还有很长的

路要走，我们将以此为基础，继续撰写相关书籍、建立技术交流平台、规范培训课程以及举办相关会议等。如果您对此感兴趣或愿意与我们一起为此努力，可关注我们的网站、博客等，并加入我们的QQ群、微信公众平台等，衷心地欢迎您加入我们的团队，一起为实现理想而奋斗。

本书数据：开源统计学网站获取(http://www.osstat.org/)。

开源统计新浪微博：http://weibo.com/u/3299903767。

开源统计腾讯微博：http://t.qq.com/osstat。

QQ群名称：开源统计(294823415)。

微信公众平台：扫描下面的二维码加入。

张志杰

2015年3月

目 录

第一章

R 软件概述与简介

第一节 R 软件的发展历史

R 软件是用于数据处理、科学计算、统计分析与图形绘制的统计计算机程序，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）、丰富的统计分析与建模工具、优秀的统计制图以及简单而强大的编程语言等，可实现条件分支、循环以及用户自定义的功能。R 软件有 Linux，Mac OS 和 Windows 版本。在 R 的默认安装程序中只包含了 30 个基础包，其他程序包可以通过 CRAN(Comprehensive R Archive Network) 网站免费下载安装，在通用公共许可协议(General Public License，GPL)下可以自由使用、复制与传播 R 软件，只要保证接受者拥有同样的权利，并且相应的源代码是公开的即可(http://www.gnu.org/copyleft/gpl.html)。与其说 R 是一个软件，不如说 R 是一种数学/统计计算的“环境”，因为 R 提供的不仅仅是若干统计程序(使用者只需指定数据库和若干参数便可进行一个简单的统计分析)，其核心思想是提供一个集成各种统计工具的平台环境，以期兼容不同编程语言(如 C，C++，Java，Python 等)，实现各种数学计算、统计计算的函数，从而让使用者能直接调用函数，实现更加灵活方便的数据分析，甚至开发出符合需要的新的统计计算方法。它是基于语言的计算，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型和绘图非常有用，因此受到了世界各地统计爱好者的关注。

R 软件最早来自新西兰 Auckland 大学 Ross Ihaka 和 Robert Gentleman 两人的开创性工作，其语法(syntax)类似于 S 语言，语义(semantics)则源自 Scheme 语言。它是函数编程语言的变种并且和表处理语言(LISt Processor，LISP)以及数组处理语言(Array Processing Language，APL)有很强的兼容性。它的源代码主要是由 C，Fortran 和 R 写的。R 软件的使用与商业软件 S-PLUS 有很多类似之处，两个软件有很高的兼容性，在使用上几乎可以不加修改地从 S-PLUS 移植到 R，而 S-PLUS 的使用手册，仅需要经过不多的修改就能成为 R 的使用手册。因此，有研究者称 R 是 S-PLUS 的一个“克隆”版本，这是有一定道理的。S-PLUS 是基于 S 语言开发的，该语言在 1975—1976 年间由 AT&T 贝尔实验室 Rick Becker，John Chambers 和 Allan Wilks 共同开发出来，是一种用来进行数据探索、统计分析、绘图的解释型语言。第一个可运作的 S 语言版本在 1976 年正式发布，在 GCOS 操作系统上运作。当时它还没有正式名称，曾经被称为交互式统计计算子程序(interactive

statistical computing subroutines, ISCS)、统计计算系统(statistical computing system)、统计分析系统(statistical analysis system, SAS)等。直到1979年,它才被正式命名为S语言。于1998年美国计算机协会(ACM)将软件系统奖授予了对S语言作出主要贡献的John Chambers,ACM评价其永久地改变了人们分析、可视化以及处理数据的方式(For the S system, which has forever altered how people analyze, visualize, and manipulate data),这是迄今为止被ACM授予的唯一一个统计软件系统(http://awards.acm.org/software_system/year.cfm)。1993年MathSoft公司(2001年MathSoft总部迁到西雅图,并改名为Insightful公司)买下S语言的使用许可,并于2004年买断。基于S语言的S-PLUS商业软件诞生,并由该公司的统计科学部发展与完善,2008年被TIBCO公司收购,整合进TIBCO Spotfire软件。R语言正是由于继承了S语言的优秀血统,因此也就拥有了很多优势。2009年《纽约时报》发表了"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应用以及使用者的评价。2010年,美国统计协会(American Statistical Association)又将第一届"统计计算及图形奖"授予了Robert Gentleman和Ross Ihaka,以表彰其在发起统计计算的R项目上的卓越贡献(http://stat-computing.org/awards/comp-graphics/winners.html)。命名为"R"有两个原因:一是两人名字的首字母都为"R",为了体现他们的贡献,该软件的名字取为"R";二是该软件被看作S语言的一个"早期"简化实现版本,结合字母顺序该软件可看作"S"的前一个版本,因此"R"也符合了其刚好位于"S"前一位置的特点。

R软件的发展历史较短,主要的标志性事件如下。

(1) 1993年8月,Ross Ihaka和Robert Gentleman将R软件的部分二进制代码(binary copies)放到了Statlib网站上,并在s-news邮件发送清单(mailing list)中发布公告。这一"开源"进展在交流平台的互动下使他们受到很大鼓舞。很多人开始尝试他们的二进制代码,并对使用情况进行反馈,其中最持之以恒的一位就是苏黎世联邦理工大学(ETH Zurich)的Martin Maechler,并且Martin开始支持并说服他们以自由软件(free software)的形式发布R源代码(source code)。

(2) 1995年6月,R软件终于踏出了"开源"的历史性一步,在自由软件基金会的GNU通用公共许可协议(Free Software Foundation's GNU general public license)下以FTP的形式公开了其源代码。之后,Martin加入他们的工作,通过email接收使用者的错误报告,并偶尔发布更新的R软件版本。他们很快认识到R软件使用者没有真正的一个交流平台,一个人工维护的小邮件发送清单应运而生,最后由Martin建立了一个自动关于讨论R软件使用和开发的邮件发送清单;同年5月他们向 *Journal of Computational and Graphical Statistics* 杂志提交了第一篇R软件的论文,并于1996年4月正式发表。

(3) 1996年3月,r-testers邮件发送清单开始运行,约1年后它被r-announce, r-help和r-devel 3个新闻组(news group)所取代。R软件的发展开始加速,一个比较好的软件/文档的存档分配机制(archive distribution mechanism)被提上日程,这一艰巨任务由维也纳工业大学(TU Wien)的Kurt Hornik完成。基于此,他们开始收到越来越多的错误报告、建议、修复补丁以及使用者贡献的代码等,这都大大提高了R软件的功能与性能。但是,仅他们3个人来修正错误以提高R软件的工作根本无法满足日益突出的需求,需要更多人加入他们的工作中。

(4) 1997年4月,Linux系统的R-0.49的源代码包正式发布。

(5) 1997年8月,他们建立了一个由11人组成的有权修改当前版本系统R源代码的核心团队,成员包括:Doug Bates, Peter Dalgaard, Robert Gentleman, Kurt Hornik, Ross Ihaka, Friedrich Leisch, Thomas Lumley, Martin Mächler, Paul Murrell, Heiner Schwarte和Luke Tierney。现在该核心团队的成员数已经扩充为20人(http://www.r-project.org/contributors.html)。

(6) 1997年12月,R软件正式成为GNU项目(https://directory.fsf.org/wiki/R#tab=Overview)。

(7) 2000年2月,Linux系统相对稳定的R-1.0.0版本发布,同时,Windows系统的R-1.0.0版本发布,之后约6个月发布新版本。

(8) 2004年11月,MAC系统的R.app_1.01.src.tgz版本发布;2005年4月R-2.1.0.dmg版本发布;2006年1月,Mac-GUI-1.14.tar.gz版本发布;之后不定期地进行更新。

(9) 2010年4月,R软件开始支持64位的Windows操作系统。

截至2014年10月,R软件的最新版本为R-3.1.1,其发展已得到越来越多人的关注。

第二节 R软件的优缺点

作为一款"开源"的免费软件,R软件具有许多吸引统计学者的地方,当然对于非专业统计学者则有很多缺陷。一般来说,R软件具有以下7个特点:

(1) 程序代码对于大小写字母敏感,如A不同于a, mean不同于Mean;

(2) 命令间使用分号(";")分割或者在不同行显示,我们推荐使用后者;

(3) 相关的一组命令使用大括号组合在一起("{}"),这在使用者编写自定义的函数时经常用到;

(4) 注释默认以"#"开头,从"#"开始到一行的结束均被看作注释,可出现在程序的任何位置;

(5) 如果命令内容较多、一行显示不全,那么将在下一行继续显示命令,但在新行的开始默认将有符号"+",如:

```
>mean(c(1,3,5,6))          >mean(c(1,3,
                            +5,6))
 [1]3.75                    [1]3.75
```

(6) 默认一行的最长长度限制为4 095个字节(不是字符,一个英文字符2个字节,一个中文字符4个字节);

(7) 键盘上的向上和向下箭头可以用来向前和向后翻看命令历史,当找到需要的命令后,可以使用向左和向右箭头进行位置定位,进而修改相关内容。

根据我们自己的使用经验,认为R软件的主要优点包括:

(1) 相对于商业软件而言,免费是其最大优势之一,虽然免费,但能完成甚至超越商业

软件所能实现的功能；

(2) 在更新、整合最新统计学方法的速度上非常出色，越来越多的研究者愿意把他们的研究方法写成R软件程序包，这使得诸如SAS等商业统计软件望尘莫及；

(3) 新的统计方法程序包的涌现，使得研究者在开发算法上的时间大大缩短，并可以让应用者将更多时间放在具体的分析过程本身，而不是方法的研究上；

(4) 图形绘制展示功能非常优秀，可以做出极其精美的图形；

(5) 拥有非常活跃的用户群体，如每年举行一次useR! 大会和两年举行一次DSC (Directions in Statistical Computing)大会；

(6) 借助于Sweave(=R+LaTeX)，R软件实现了自动生成统计分析报告的功能，对于经常要进行重复性分析工作的人非常方便；

(7) 数据连接的扩展性能非常优秀，例如提供接口连接众多关系型数据库(如SAP HANA，DB2，MySQL，Access等)、提供API接口连接网络数据(如Google，Twitter及微博等)；

(8) 方便同其他编程语言(如C，C++，Python，Java等)互相调用，从而借助于其他语言的优势进一步增加其性能；

(9) 与其他优秀传统商业统计软件可以方便地进行程序间内嵌使用，如SAS，SPSS，Stata等。

R软件的主要缺点如下：

(1) 学习掌握R软件的使用需要花费较多时间，周期较长，主要是因为要掌握方法背后的统计理论与统计知识需要较长的时间，这就急需一套从最基础的软件使用介绍到高级的统计分析方法的系列丛书，以帮助使用者有效提高学习技能；

(2) 同一种方法有多个不同的函数可以使用，由于不同函数设计的算法可能不同，结果有所差异，准确性也不同，对于非专业人员而言，选择合适的函数相对困难；

(3) 受到算法架构的通用性和速度性能方面的影响，其初始设计完全基于单线程和纯粹的内存计算，采用的是“Call by Value”的评价模式，在参与计算时数据通常被多次拷贝，如拟合线性模型过程中，其设计矩阵将反复被拷贝6次，因此在大数据等的处理上存在一定困难，必须与其他软件系统结合(如R+Hadoop)；

(4) 主要采用编程的方法完成相关工作，而缺乏像IBM SPSS Statistics那样简单、方便的拖拉式的操作界面；

(5) 作为免费软件，其技术支持相对较差，有时候靠个人努力要弄清楚某个函数的正确使用存在较大困难，因此经常会犯一些使用者未意识到的错误。

第三节 R软件的下载与安装

一、R软件的下载

登录网站http://cran.r-project.org/，在“Download and Install R”下面选择对应的操作系统，这里我们下载Windows版本的R软件，左键单击“Download R for Windows”(图1-1)。

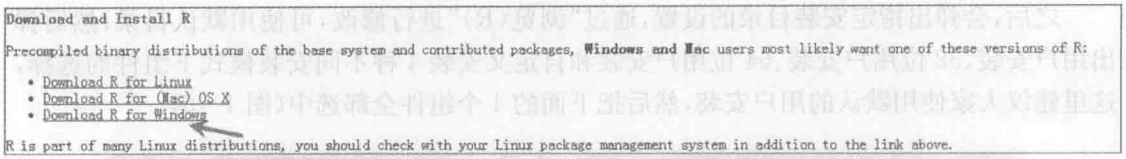

图 1-1　CRAN 上不同操作系统 R 软件的下载地址

然后，左键单击“Base”(contrib 里包含了不同版本的程序包；Rtools 是 Windows 系统下开发 R 程序包的一些必需辅助工具集合，见图 1-2)。

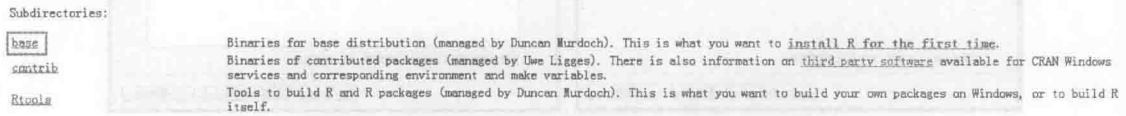

图 1-2　R 软件下载中的 3 个不同内容

最后，右键单击“目标另存为”下载最新版本的 R 软件；另外两个链接，一个是软件安装的指导，一个是介绍该版本与之前的一款旧版本相比新增加的特征(图 1-3)。

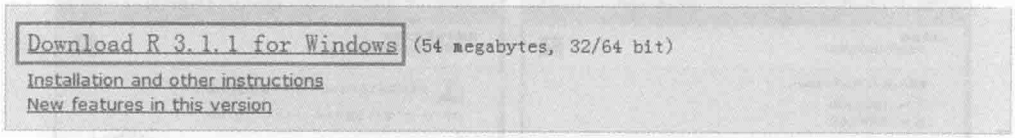

图 1-3　Windows 版本的真正下载链接

二、R 软件的安装

双击 R 软件安装程序，首先弹出的是软件安装过程中使用的语言，使用较多的是中文或英文，这里我们使用默认的中文设置(与操作系统地语言系统有关，见图 1-4)。

随后，正式进入软件安装的向导，在开始的两个界面不需要任何设置，直接点击“下一步(N)”即可(图 1-5)。

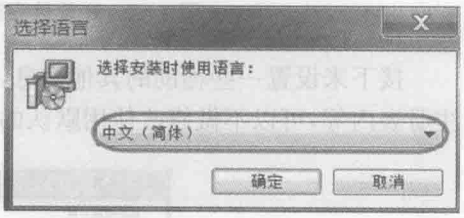

图 1-4　R 软件安装时的语言选择界面

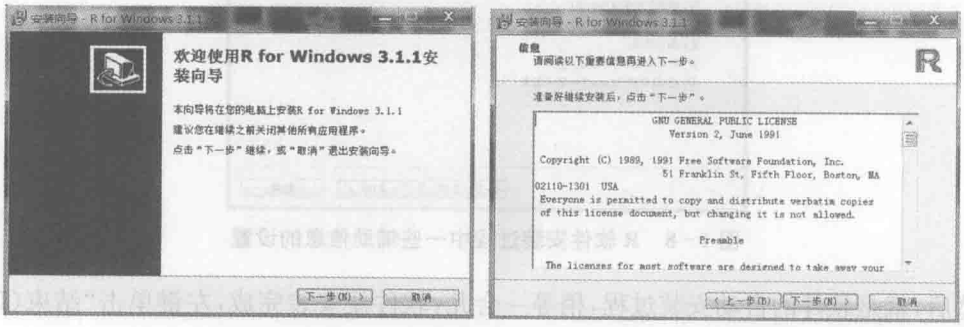

图 1-5　R 软件安装过程中的界面

之后，会弹出指定安装目录的设置，通过“浏览(R)”进行修改，可使用默认目录；然后弹出用户安装、32位用户安装、64位用户安装和自定义安装4种不同安装模式下组件的选择，这里建议大家使用默认的用户安装，然后把下面的4个组件全部选中(图1-6)。

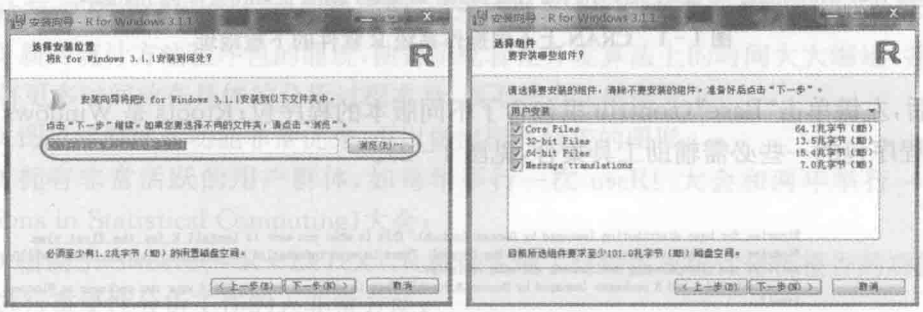

图1-6 R软件安装过程中安装目录和组件的设置

然后，弹出是否要自定义启动选项的对话框，使用默认选项即可；随后，在开始的所有程序中添加软件快捷方式的对话框，默认的文件夹名字是“R”，可通过右侧“浏览”按钮更改(图1-7)。

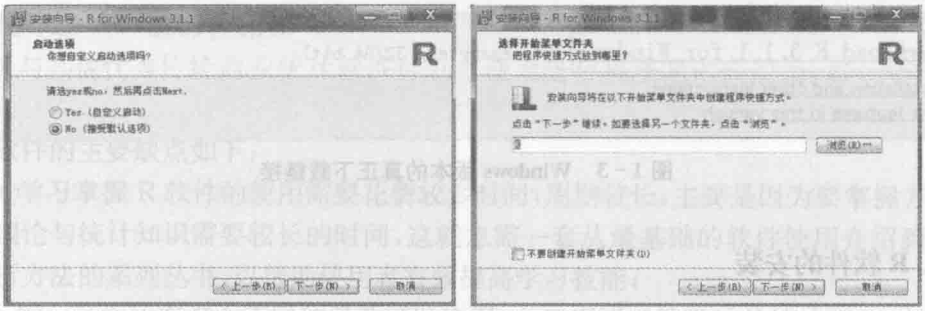

图1-7 R软件安装过程中启动选项和快捷方式的设置

接下来设置一些辅助的其他信息，如是否创建桌面快捷方式、是否要把版本信息保存在注册表内等，可以不做修改使用默认的设置，也可以全部选中(图1-8)。

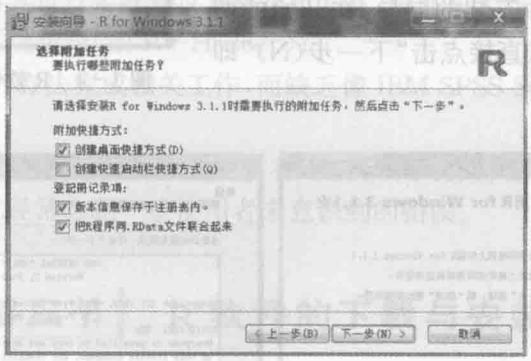

图1-8 R软件安装过程中一些辅助信息的设置

然后，就是软件的自动安装过程，稍等一会儿，软件便安装完成，左键单击“结束(F)”按钮退出软件的安装过程(图1-9)。