

普通高等教育
“十二五”规划教材



R & ASReml-R Statistical Analysis Tutorial

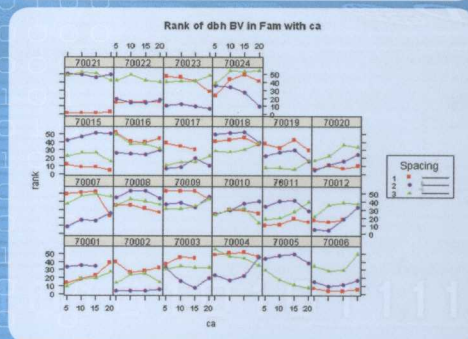
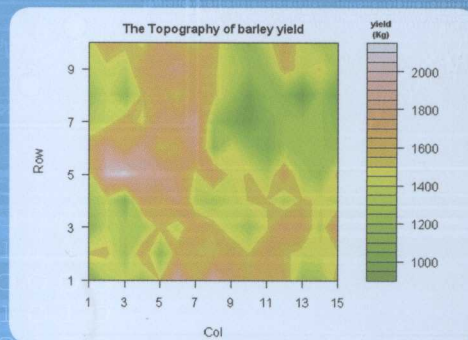
R与ASReml-R 统计分析教程

林元震 陈晓阳 主编

了解学习R,
玩转数据分析与绘图

全面掌握R,
随时进军大数据时代

熟悉ASReml-R,
轻松评估遗传试验



中国林业出版社

014035850

0212.1-43
06

普通高等教育“十二五”规划教材

R 与 ASReml-R 统计分析教程

林元震 陈晓阳 主编



中国林业出版社

0212.1-43



北航

C1723243

06

028280210

内 容 简 介

R 语言以其免费开源且强大统计与绘图的功能,已成为数据挖掘应用中最受欢迎的编程语言之一,也是大数据时代中的重要统计软件之一。本书面向农林业试验,系统介绍了 R 及 ASReml-R 在农林业试验数据分析及绘图中的应用,全书共分 6 章,内容包括 R 语言简介、数据集创建、数据操作、常用统计分析、统计图形绘制以及 ASReml-R 包与遗传参数估算。本书内容新颖,适用面广,可操作性强。此外,章节合理,结构清晰,行文规范,适合于 R 统计分析和生物统计的课程讲授。

本书适用于农林资源类、植物生产类、生物科学类、园艺学类、金融学类等专业本科生的统计分析教材,也可供相关专业的研究生和科研工作人员参考使用。

图书在版编目 (CIP) 数据

R 与 ASReml-R 统计分析教程 / 林元震, 陈晓阳主编. —北京: 中国林业出版社, 2014. 1
ISBN 978-7-5038-7375-1

I. ①R… II. ①林… ②陈… III. ①统计分析 IV. ①O212.1

中国版本图书馆 CIP 数据核字 (2014) 第 019702 号

中国林业出版社·教材出版中心

策划编辑: 肖基洪

责任编辑: 张东晓

电话: 83221489 83220109

传真: 83220109

出版发行 中国林业出版社 (100009 北京市西城区德内大街刘海胡同 7 号)

E-mail: jiaocaipublic@163.com 电话: (010) 83224477

http: //lycb. forestry. gov. cn

经 销 新华书店

印 刷 中国农业出版社印刷厂

版 次 2014 年 3 月第 1 版

印 次 2014 年 3 月第 1 次印刷

开 本 850mm × 1168mm 1/16

印 张 17.75

字 数 421 千字

定 价 45.00 元

未经许可, 不得以任何方式复制或抄袭本书之部分或全部内容。

版权所有 侵权必究

《R 与 ASReml-R 统计分析教程》

编委会

主 编：林元震(华南农业大学)

陈晓阳(华南农业大学)

副主编：郭 海(水利部水土保持中心)

张 谦(广东省林业科学研究院)

郑会全(广东省林业科学研究院)

编 委：(以姓氏笔画为序)

苏 艳(华南农业大学)

杨会肖(华南农业大学)

杨晓慧(北京林业大学)

李义良(广东省林业科学研究院)

张 磊(东北林业大学)

周 玮(华南农业大学)

赵曦阳(东北林业大学)

Preface

I was extremely pleased to receive this new book of *R&ASReml-R Statistical Analysis Tutorial* by Lin Yuanzhen and Chen Xiaoyang, and the book describes data analyses and graphic facilities for agriculture and forestry experiments in Chinese.

First encountered ASReml in a 1999 Melbourne workshop conducted by Dr. Arthur Gilmour for forest geneticists. At that time I was programming using SAS to analyze disconnected half-diallel mating design with many missing mating cells. At the workshop I learned that ASReml only needed one line to overlay the mating design matrix for the incomplete diallel mating structure in contrast to several dozens or even hundreds of lines in SAS to do the same job before Proc glimmix implemented. However, one main issue with the ASReml was the lack of a comprehensive data management system. Therefore SAS and ASReml was a perfect combination for my analyses for forest genetic trials at that time. In 2008, I was in sabbatical leave in France, and noticed that many colleagues including my host in INRA were using R packages for their data management, analyses and plot of sophisticated graphics. I learned that R was free software with a versatile capability of data management, statistical analyses and graphics, and was getting popularity by each day in research community. One particular advantage in R is user-developed or contributed packages that were developed from different projects and for various needs of analyses, including packages for population, quantitative genetics, bioinformatics, and recently, for genome-wide association, and genomic selection. ASReml was also adapted into the R system by developing an ASReml-R package. The fact that a recent ASReml-R workshop by Dr. Brian Cullis using ASReml-R package in Vancouver, Canada in association with IUFRO Forest Genetics (2013) conference is an excellent example to the expanded interest in agriculture and forestry community to use the R and ASReml packages. This demonstrates that the R and ASReml packages could be a great combination for many data analyses in agriculture and forestry research community.

Working in forest genetics and tree breeding, I and my students benefited greatly from using the R and ASReml packages in genetic analyses involving multiple-generations, mixed mating structure, multiple-sites and multiple-traits with heterogeneous genetic and residual variances, and with spatial patterns of variation, etc. I sincerely wish students and practitioners use this book as a starting document to get into the beautiful world of the R and ASReml.

Harry Wu (吴夏明)

Senior Principal Research Scientist, CSIRO, Australia.

Professor and Chair of Forest Genetics and Tree Breeding, Swedish University of Agriculture Sciences, Sweden

注：吴夏明教授是国际知名的林木遗传育种专家，现是澳大利亚 CSIRO Plant Industry 首席科学家、瑞典农业大学教授、国际林联(IUFRO)南方松工作分会主任、澳大利亚林学会林木遗传分会主任和我国多家大学特聘教授，在林木数量遗传学以及分子遗传学和数量遗传学相结合的领域内做出了许多原创和公认的卓越贡献。

序

很高兴收到林元震等人主编的新书《R 与 ASReml-R 统计分析教程》，该书以中文介绍农林试验的数据分析与图形绘制。

我最早接触 ASReml 软件是 1999 年由亚瑟·吉尔摩博士在墨尔本为林业遗传研究人员讲授的 ASReml 课程培训。当时，我正用 SAS 编程来分析不连续半双列杂交试验。在那次培训上，我知道了 ASReml 软件只需一行代码即可囊括不完全双列交配设计的矩阵结构，而 SAS 则需要几十行甚至上百行代码，那时 SAS 的 Proc glimmix 程序还不可用。然而，ASReml 的主要问题在于缺乏一个综合的数据处理系统。因此，我联合使用 SAS 和 ASReml 来分析林业遗传试验。2008 年，我在法国学术休假时，注意到许多同行，包括在法国农科院 (INRA) 上班的房东，正在使用 R 软件来管理数据、分析数据以及绘制复杂图形。我这才知道 R 是免费软件，并且具有数据处理、统计分析以及图形绘制的强大功能。R 软件的一个突出优点是来自不同领域的用户可自由开发、共享程序包，以满足不同数据分析的需求。这些程序包已涵盖了群体遗传学、数量遗传学、生物信息学，以及最近的基因组关联分析与基因组选择等研究领域。ASReml 软件也已开发出适合 R 环境的程序包 ASReml-R。最近的一次 ASReml-R 培训在加拿大的温哥华举办，由布雷·卡尔斯博士讲授，该培训与 2013 年国际林联 (IUFRO) 林木遗传学大会联合进行。此次 ASReml-R 培训非常成功，并激起了与会人员对使用 R 及 ASReml 包来分析农林试验数据的广泛兴趣。这说明了 R 与 ASReml 包可以成为一个强大的组合，用于农林业研究领域的数据分析。

在林木遗传育种的工作中，我和我的学生们非常受益于 R 与 ASReml 包在遗传试验分析中的使用，这包括了多世代材料试验、混合交配设计试验、多地点试验和多性状分析，涉及了异质的遗传方差和残差结构，以及变异的空间模型等。我衷心地希望学生们和研究人员能将此书作为进入 R 与 ASReml 优美世界的入门书籍。

吴夏明

澳大利亚联邦科学与工业研究组织 (CSIRO) 首席科学家
瑞典农业大学林木遗传育种系主任、教授

注：此序为吴夏明教授所作序的中文版，由林元震博士翻译，如有不当之处，敬请读者批评指正。

前言

2012年3月22日,美国奥巴马政府宣布投资2亿美元启动“大数据研究和发展计划”,旨在增强收集海量数据、分析萃取信息的能力,加快科学研究的步伐,加强美国国家安全,并引发教育与学习方式的变革。这暗示着大数据时代的春天来了!

虽然R语言开发于20世纪90年代初,还是一门很年轻的语言,但是越来越多人喜欢使用R语言,尤其是大数据产品销售商,如数据仓库与Hadoop数据过滤器。部分原因在于,R作为开源语言吸引了大量的统计学家与定量分析师,由这些精英们构成的R社区能够引领R语言开发。这也使得R语言具备解决大数据时代统计需求的能力和潜力。2012年,英国KDNuggets网站做了一个“数据挖掘与分析工具”的调查,结果显示,R语言超越了SQL和Java,在最受欢迎的数据挖掘应用编程语言排行榜中排名第一。R语言在数据挖掘和可视化应用领域的快速崛起意味着R语言已经为大数据时代做好准备。

与其他R语言的书籍一样,阅读本书的读者,不需要具备统计编程或R语言知识背景,读者只要有数据处理和可视化的需求,即可在本书中获取相关的方法、代码和结果。R语言的书籍虽已出版不少,但未见有针对农林领域的R工具书,因此本书特面向农林业试验,介绍R及ASReml-R在农林业试验数据分析及绘图中的应用,并且所用方法、代码同样适用于其他领域。

本书的第1~3章简单介绍R语言、数据创建和数据操作,主要是让读者了解R语言的特点,并熟悉数据录入与数据操作,为后续的数据分析及图形绘制打好基础。

第4章较全面、系统地演示了数据处理中常用的各种统计分析,包括描述性统计分析、方差分析、协方差分析、 t 检验、卡方检验、线性回归、相关分析、主成分和因子分析以及聚类分析,其中,还专门介绍了常用试验设计和数据分析。

第5章介绍了R语言的基础绘图和高级绘图,基础绘图包括条形图、直方图、散点图、热图、等高图等常见图形,以及R语言高级绘图包lattice和ggplot2的数据可视化方法。本章主要是让读者了解R语言在绘图方面的优势和强大功能。

第6章特别展示了遗传评估软件ASReml-R包在农林业遗传测定中的应用。ASReml是当前国际上遗传测定评估方面的知名软件,国内尚未见有书籍介绍,因此本书专门设置此章,主要以林业试验数据为对象,演示ASReml-R包的基本语法及其常用模型在农林业遗传测定中的应用,内容涉及子代测定、无性系测定,并介绍了遗传力、育种值、遗传相关等遗传参数的估算方法。尤其介绍了空间分析和多地点分析($G \times E$),其他类似软件目前还尚未能分析。我们希望此章对从事林木遗传改良的研究人员能有一定的参

考价值。

附录部分介绍了一些优秀的 R 语言和 ASReml 软件相关网站,我们相信,这些网站将有助于读者进一步深入学习 R 语言和 ASReml 软件。此外,本书所有的代码和数据以及本书彩图,也将无偿地存放在网盘 <http://yzhlin-asrem1.ys168.com/>,供读者免费下载、自由使用。

最后,笔者要衷心感谢澳大利亚联邦科学与工业研究组织植物所(CSIRO Plant Industry)的合作导师吴夏明教授(Professor Wu Harry)和 Ivković Miloš 博士,是他们把我带入了 R 语言和 ASReml 软件的殿堂,并教会我使用它们处理与分析科研试验数据,为自己今后的科研教学工作增添一种强大的新工具。同时,也要特别感谢 ASReml 软件开发者 Arthur Gilmour 对笔者在遗传参数估算问题上给予的耐心和详细解答。

吴夏明教授是国际著名林木遗传育种学家,也是 R 与 ASReml 软件的资深用户,特别感谢他百忙之中欣然为本书作序。同时,还感谢中国科学院植物研究所赖江山博士审阅本书,并给予不少宝贵意见。

由于编者的知识水平有限,书中难免存在疏漏和不足之处,欢迎读者批评指正。对本书如有任何意见或建议,请发邮件至: yzhlin@163.com。

林元震

2013 年 10 月

目 录

第1章 R语言简介	(1)	3.2 变量的重命名	(19)
1.1 R语言	(1)	3.3 缺失值的处理	(20)
1.2 R的特点	(1)	3.4 数据类型的转换	(21)
1.3 R的资源	(2)	3.5 数据的排序	(21)
1.4 R的安装与运行	(2)	3.6 数据集的合并	(22)
1.4.1 R软件的安装、启动与关闭	(2)	3.6.1 列合并	(23)
1.4.2 R程序包的安装与使用	(3)	3.6.2 行合并	(24)
1.5 Rstudio的安装与运行	(4)	3.7 数据子集的提取	(25)
第2章 数据集创建	(9)	3.7.1 根据位置选取子集	(25)
2.1 数据的创建	(9)	3.7.2 根据列名选取子集	(26)
2.1.1 向量的建立	(9)	3.7.3 使用 subset() 函数	(26)
2.1.2 数组的建立	(10)	3.8 数据的整合与重构	(27)
2.1.3 矩阵的建立	(11)	3.8.1 数据转置	(27)
2.1.4 数据框的建立	(12)	3.8.2 数据整合	(28)
2.1.5 列表的建立	(13)	3.8.3 使用 reshape2 包	(28)
2.1.6 因子的建立	(13)	3.9 处理数据常用的函数	(30)
2.1.7 对象的模式和属性	(14)	3.10 数据探索	(31)
2.2 数据的输入	(15)	3.10.1 数据结构查看	(31)
2.2.1 键盘输入	(15)	3.10.2 查看单个变量	(31)
2.2.2 使用 scan() 函数	(15)	3.10.3 查看多个变量	(31)
2.2.3 导入 csv 文件	(16)	3.10.4 其他查看方法	(32)
2.2.4 导入 Excel 数据	(16)	3.10.5 保存统计图形	(33)
2.2.5 导入 SAS 数据	(17)	3.11 循环控制和条件操作	(33)
2.2.6 导入 SPSS 数据	(17)	3.11.1 条件语句	(33)
2.2.7 其他方式导入	(17)	3.11.2 循环语句	(34)
2.3 数据的存储	(17)	3.12 自编函数	(35)
第3章 数据操作	(19)	第4章 常用统计分析	(36)
3.1 变量的创建与删除	(19)	4.1 描述性统计分析	(36)
		4.1.1 使用 summary() 函数	(36)
		4.1.2 使用 stat.desc() 函数	(37)

4.1.3 使用 describe() 函数	4.9.2 随机区组设计
..... (37)	 (106)
4.2 频数表分析	4.9.3 拉丁方设计
..... (38)	 (115)
4.3 方差分析	4.9.4 平衡不完全区组设计
..... (39)	 (118)
4.3.1 单因素方差分析	4.9.5 正交试验设计
..... (39)	 (120)
4.3.2 无重复试验的双因素方差	4.9.6 裂区设计
分析	 (131)
..... (44)	4.9.7 巢式设计
4.3.3 等重复试验的双因素方差	 (135)
分析	4.9.8 全因子设计
..... (45)	 (138)
4.3.4 多元方差分析	4.10 功效分析
..... (48)	 (142)
4.4 协方差分析	4.10.1 功效分析
..... (51)	 (142)
4.5 显著性检验—— t 检验	4.10.2 绘制功效分析图
..... (57)	 (145)
4.5.1 单样本检验	4.11 主成分和因子分析
..... (57)	 (147)
4.5.2 独立的双样本 t 检验	4.11.1 主成分分析
..... (58)	 (147)
4.5.3 非独立的双样本 t 检验	4.11.2 因子分析
..... (60)	 (152)
4.6 χ^2 检验	4.12 聚类分析
..... (61)	 (155)
4.6.1 适合性检验	4.12.1 系统聚类法
..... (61)	 (156)
4.6.2 独立性检验	4.12.2 动态聚类法
..... (62)	 (159)
4.7 线性回归	第5章 统计图形绘制
..... (64)	 (161)
4.7.1 简单线性回归	5.1 基础绘图
..... (65)	 (161)
4.7.2 多项式回归	5.1.1 条形图
..... (66)	 (161)
4.7.3 多元线性回归	5.1.2 饼图
..... (68)	 (163)
4.7.4 有交互项的多元线性回归	5.1.3 直方图
..... (72)	 (164)
4.7.5 回归诊断	5.1.4 核密度图
..... (74)	 (166)
4.7.6 异常值判断	5.1.5 散点图
..... (79)	 (166)
4.7.7 回归模型的改进措施	5.1.6 热图
..... (81)	 (169)
4.7.8 最佳回归模型的选择	5.1.7 等高图
..... (82)	 (172)
4.7.9 广义线性模型	5.1.8 三维透视图
..... (85)	 (173)
4.8 相关分析	5.1.9 小提琴图
..... (93)	 (174)
4.8.1 相关的类型	5.1.10 颜色等高图
..... (93)	 (174)
4.8.2 相关显著性的检验	5.1.11 散点图矩阵
..... (100)	 (175)
4.8.3 相关关系的可视化	5.1.12 条件图
..... (102)	 (178)
4.9 常用试验设计与数据分析	5.1.13 相关图
..... (103)	 (180)
4.9.1 完全随机设计	5.1.14 地图
..... (103)	 (182)
4.9.2 随机区组设计	5.2 高级绘图
..... (106)	 (184)
4.9.3 拉丁方设计	5.2.1 lattice 包
..... (115)	 (184)
4.9.4 平衡不完全区组设计	5.2.2 ggplot2 包
..... (118)	 (187)
4.9.5 正交试验设计	5.3 基础绘图常用参数
..... (120)	 (192)
4.9.6 裂区设计	5.3.1 颜色
..... (131)	 (192)
4.9.7 巢式设计	
..... (135)	
4.9.8 全因子设计	
..... (138)	
4.10 功效分析	
..... (142)	
4.10.1 功效分析	
..... (142)	
4.10.2 绘制功效分析图	
..... (145)	
4.11 主成分和因子分析	
..... (147)	
4.11.1 主成分分析	
..... (147)	
4.11.2 因子分析	
..... (152)	
4.12 聚类分析	
..... (155)	
4.12.1 系统聚类法	
..... (156)	
4.12.2 动态聚类法	
..... (159)	

5.3.2	符号和线条	(193)	6.4.1	半同胞子代测定.....	(219)
5.3.3	标题	(195)	6.4.2	全同胞子代测定.....	(224)
5.3.4	图例	(195)	6.4.3	无性系测定	(233)
5.3.5	坐标轴	(196)	6.4.4	空间分析	(235)
5.3.6	多图组合	(196)	6.4.5	多地点试验—— $G \times E$ 分析	(244)
第 6 章 ASReml-R 包与遗传参数估算			6.4.6	遗传相关分析	(251)
6.1	ASReml 程序包的介绍	(198)	6.5	综合案例分析	(255)
6.2	ASReml-R 的基本语法	(200)	常用网址		(262)
6.2.1	单性状分析	(201)	参考文献		(264)
6.2.2	双性状分析	(212)	索引		(266)
6.3	林业试验常用模型	(217)	程序包名单		(268)
6.4	遗传参数估算	(219)			

第 1 章 R 语言简介

1.1 R 语言

R 既是软件,也是语言,在 GNU 协议 General Public Licence 下免费发行,是 1995 年由新西兰奥克兰大学统计系的 Ross Ihaka 和 Robert Gentleman 基于 S 语言基础上共同开发的一种统计软件。现在由 R 开发核心小组(R Development Core Team)负责维护与更新,并将全球优秀的统计应用程序包免费提供给大家使用、共享。

R 是一套由数据操作、计算和图形展示功能整合而成的软件系统,包括:

- ①有效的数据存储和处理功能;
- ②一套完整的数组(特别是矩阵)计算操作模块;
- ③拥有完整体系的数据分析工具;
- ④为数据分析和显示提供强大的图形功能;
- ⑤一套完善、简单、有效的编程语言(包括条件、循环、自定义函数和输入输出功能等)。

1.2 R 的特点

为何现在越来越多的人开始学习和使用 R,笔者认为 R 有以下几大显著的优点:

(1) 免费且开源

目前市面上有许多流行的统计和制图软件,如 Microsoft Excel、SAS、SPSS、Mini-Tab 以及 Original 等,但这些软件多属商业软件,并且费用昂贵。然而 R 是一个免费且开源的统计分析软件,对于教学工作来说,这无疑是一个极大的优点。

(2) 运算快且功能强大

R 可以作为一台高级科学计算器,不需要编译即可执行代码。

(3) 操作系统依赖性低

R 可以运行于 UNIX、Linux、Windows 和 Mac 各大操作系统上。

(4) 帮助功能完善

R 嵌入了一个非常实用的帮助系统,软件自带的 pdf 或 html 帮助文件可以随时通过主菜单打开浏览或打印。通过 help 命令可随时了解 R 所提供的各类函数的使用方法和例子。

(5) 作图功能强大

其内嵌的作图函数能将产生的图片展示在一个独立的窗口中,并能将之保存为各种

格式的图形文件(如 jpg, png, bmp, ps, pdf, tiff, metafile, svg)。

(6) 统计分析能力突出

R 内嵌了许多实用的统计分析函数, 统计分析的结果可直接显示出来, 也可保存到文件中, 用于进一步的分析。此外, R 还囊括了其他软件尚不可用的、先进的统计计算例程。有关统计新方法的更新速度是以周来计算的。

(7) 兼容性强

许多常用的统计分析软件(如 SPSS, SAS, Stata 及 Excel)的数据文件都可以读入 R, 并作进一步的分析。

(8) 较强大的拓展与开发能力

R 是开发新的交互式数据分析方法一个非常好的工具。R 开发周期短, 有大量的程序包(packages)可以使用。当然读者也可以编制自己的函数来扩展现有的 R 语言, 或制作相对独立的统计分析包。

1.3 R 的资源

R 的核心开发与维护小组通过 R 的主页, 即 R 工程网站(<http://www.r-project.org>), 及时发布有关信息, 包括 R 的简介、R 的更新及程序包信息、R 常用手册、已经出版的关于 R 的图书、R 通信和会议信息等。

R 的 CRAN 社区(<http://cran.r-project.org/web/packages/>)是获得软件(及源代码)和资源的主要场所, 通过它或其镜像站点就可以下载 R 的最新版本以及大量的统计程序包(packages)。R 的最新版本为 3.0.2。截至 2013 年 10 月 16 日, 在 CRAN 社区上可使用的程序包就有 4920 个, 而且还在不停地增加和更新。

1.4 R 的安装与运行

1.4.1 R 软件的安装、启动与关闭

1.4.1.1 R 的安装

从 CRAN 社区下载最新的 R 安装程序, 运行安装文件, 缺省的安装目录为 C:\Program Files\R\R- $x.x.x$, 其中 $x.x.x$ 为版本号。虽然从 2.2.0 版本以后可以选择中文作为基本语言, 但建议还是以英文为安装语言, 主要是为了避免其他扩展的程序包在中文环境下使用出现问题。

1.4.1.2 R 的启动

安装完成后点击桌面上 R 快捷图标就可启动 R 的交互式用户窗口(R-GUI)。R 是按照问答的方式运行的, 即读者在命令提示符“>”后键入命令并回车, R 就完成一些操作。例如输入命令:

```
> plot(rnorm(1000))
```

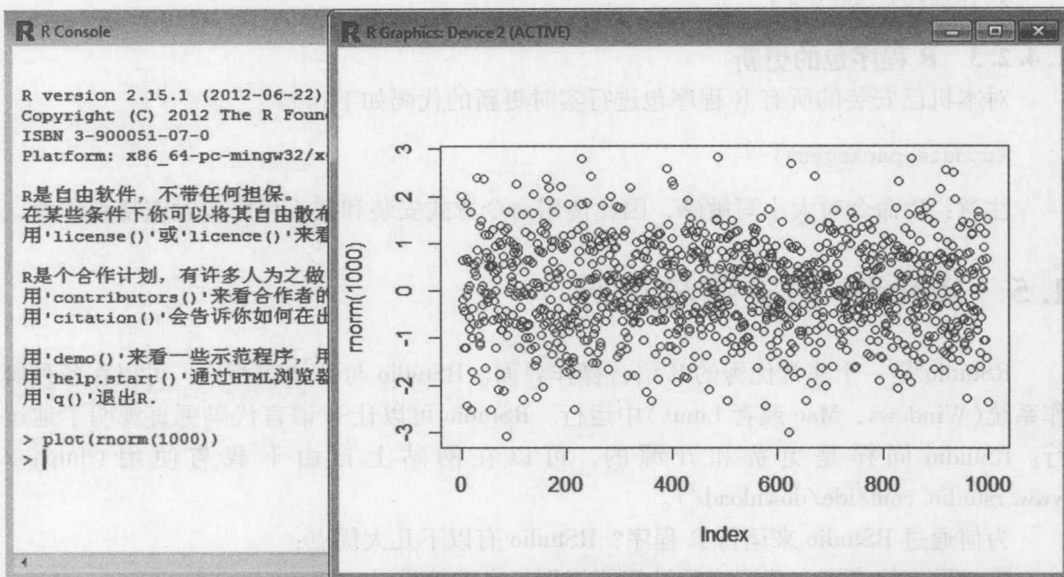


图 1-1 R 的启动

就可得到图 1-1。

1.4.1.3 R 的关闭

在命令行键入 `q()` 或直接关闭 R-GUI 窗口。退出时可选择保存工作空间，缺省文件名为 R 安装目录的 `bin` 子目录下的 `R.RData`。以后可通过命令 `load()` 或通过菜单“文件”下的“载入工作空间”加载，从而继续前一次的工作。

1.4.2 R 程序包的安装与使用

1.4.2.1 R 程序包的安装

(1) 菜单方式

在联网的条件下，按菜单栏【程序包】下拉选择【安装程序包】，选择所需的程序包进行实时安装。

(2) 命令方式

在联网的条件下，在命令提示符后键入：

```
> install.packages("PKname")
```

完成程序包 `PKname` 的安装。

(3) 本地安装

从 CRAN 社区下载需要的程序包，按菜单栏【程序包】下拉选择【从本机 zip 文件安装程序包】选定本机上的程序包(zip 文件)进行安装。

1.4.2.2 R 程序包的使用

除 R 的标准程序包(如 `base` 包)外，新安装的程序包(如 `lattice`)在使用前必须先载入，比如要使用 `lattice`，则需输入如下代码：


```
>library(lattice)
```

1.4.2.3 R 程序包的更新

对本机已安装的所有 R 程序包进行实时更新的代码如下：

```
>update.packages()
```

注意：R 命令对大小写敏感，因此使用命令方式安装和载入程序包时应特别注意。

1.5 Rstudio 的安装与运行

RStudio 是一个非常优秀的 R 语言操作界面。RStudio 与 R 语言相似，可以在各种操作系统(Windows, Mac 或者 Linux)中运行。RStudio 可以让 R 语言代码更直观明了地运行。RStudio 同样是免费和开源的，可以在网站上自由下载与使用(<http://www.rstudio.com/ide/download/>)。

为何通过 RStudio 来运行 R 程序？RStudio 有以下几大优点：

- ①代码字体高亮，代码完整性智能识别、自动缩进；
- ②可直接执行 R 程序代码；
- ③可运行多个 R 程序；
- ④可直接浏览工作表和和数据；
- ⑤可随意缩放绘制的图形，并且有多种输出格式；
- ⑥整合 R 帮助和 R 使用文档；
- ⑦可查看 R 命令的运行记录。

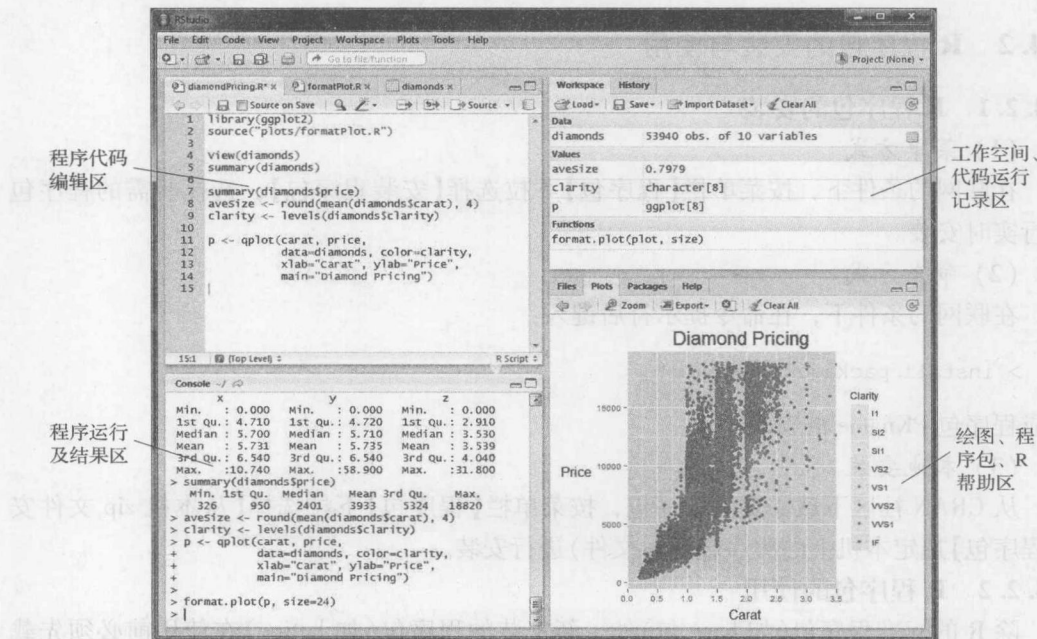


图 1-2 Rstudio 主界面

Rstudio 主界面总共有 4 个工作区域(图 1-2): 左上区是编写代码区; 左下区是程序运行和结果输出区; 右上区是工作空间和代码运行记录区; 右下区是绘图输出、程序包目录和 R 帮助区。其中, 右下区有 4 个主要的功能: Files 是查看当前 workspace 下的文件; Plots 是展示 R 绘制的图形; Packages 是系统已安装的所有程序包, 并且可勾选载入内存; Help 则是帮助文档。

(1) 左上代码编辑区

Rstudio 具有代码高亮的功能, 只是高亮的颜色有点少。点击工具栏上的 File, 选择 New, 总共可以看到 4 种格式的文件, 选择 R Script, 就能建立一个 R 语言的新代码文件。如图 1-3 所示, 写好代码之后, 直接点击右上角的 Run 就可运行当前行, 如果先用鼠标在代码上选好要运行的部分, 比如前面的五行, 然后再点 Run, 就能一下运行这五行代码。点击 Run 右边的按钮 Re-Run, 则是重复运行上次代码。

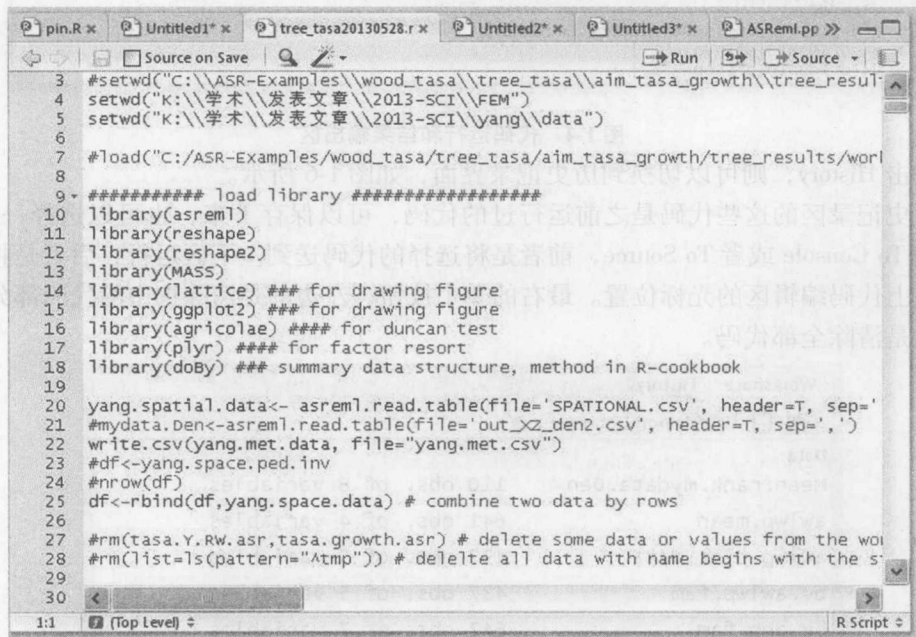


图 1-3 代码编辑区

(2) 左下程序运行和结果输出区

如图 1-4 所示, 该区可以看到程序包的载入、代码运行以及运行结果等。运行的代码均以蓝色字体并以“>”开头标记, 而运行的结果则是黑色字体。

(3) 右上工作空间和代码历史记录区

界面工具栏 Workspace 下的最左边按钮 Load, 可以切换工作区, 如图 1-5 所示, 每个工作区都会有一个隐藏文件 .RData; 旁边的按钮 Save, 可以保存当前工作区; 按钮 import dataset 则可导入数据作为数据集; 最右边的按钮 Clear all, 可将当前工作区的 Value 和 Function 全部清除。RStudio 不会自动更新这个工作区的值, 因此, 建议读者每次运行新代码前, 应先彻底清除工作区残留的旧数据。