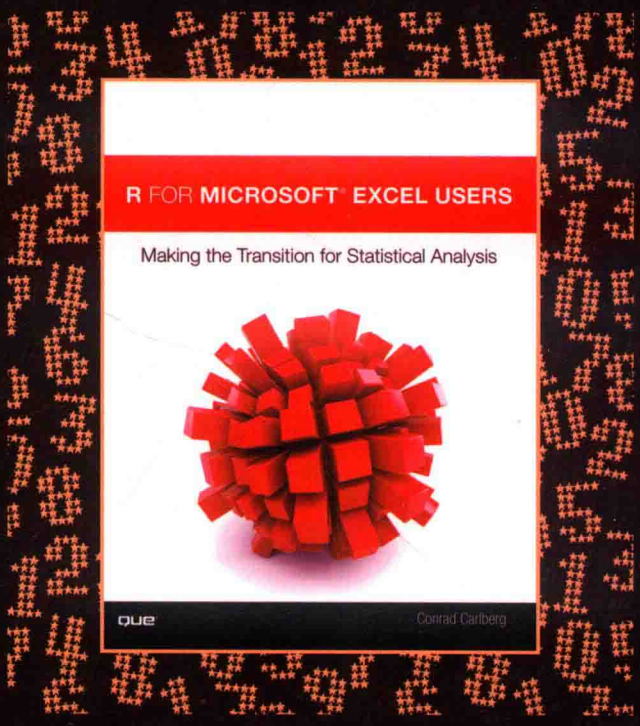
HZ BOOKS 华章IT
Pearson
数据科学与工程技术丛书

统计分析

以R与Excel为分析工具

[美] 康拉德·卡尔伯格（Conrad Carlberg） 著
程豪 译



R FOR MICROSOFT EXCEL USERS
MAKING THE TRANSITION FOR STATISTICAL ANALYSIS

机械工业出版社
China Machine Press

数据科学与工程技术丛书

R FOR MICROSOFT EXCEL USERS
MAKING THE TRANSITION FOR STATISTICAL ANALYSIS

统计分析
以R与Excel为分析工具

[美] 康拉德·卡尔伯格（Conrad Carlberg）著
程豪 译

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

统计分析：以R与Excel为分析工具/（美）康拉德·卡尔伯格（Conrad Carlberg）著；程豪译. —北京：机械工业出版社，2018.10
（数据科学与工程技术丛书）
书名原文：R for Microsoft Excel Users: Making the Transition for Statistical Analysis

ISBN 978-7-111-61001-4

I. 统… II. ① 康… ② 程… III. 统计分析－应用软件 IV. C819

中国版本图书馆CIP数据核字（2018）第221117号

本书版权登记号：图字 01-2017-0482

统计分析能够帮助人们发掘有利于生产生活的规律和价值，为相关决策提供依据和参考。本书为熟悉Excel操作的人士提供通向R语言的实用性指南，借助R与Excel工具系统阐述统计分析方法、技术。全书共6章。第1章介绍如何顺利完成从Excel到R的过渡；第2章介绍描述性统计；第3章介绍回归分析；第4章介绍方差和协方差分析；第5章介绍logistic回归；第6章介绍主成分分析。书中详细列举出所需函数及代码，可有效帮助读者在类比中掌握R语言，实现从Excel到R的过渡，适合于从事统计分析工作的专业人士，以及高等院校相关专业师生。

出版发行：机械工业出版社（北京市西城区百万庄大街22号 邮政编码：100037）
责任编辑：冯秀泳　　责任校对：李秋荣
印　　刷：北京市兆成印刷有限责任公司　　版　　次：2018年10月第1版第1次印刷
开　　本：185mm×260mm　1/16　　印　　张：15
书　　号：ISBN 978-7-111-61001-4　　定　　价：59.00元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88378991　88361066　　投稿热线：（010）88379604
购书热线：（010）68326294　88379649　68995259　　读者信箱：hzjsj@hzbook.com

译 者 序

统计分析能够帮助人们发掘有利于生产生活的规律和价值，为相关决策提供依据和参考。统计分析工具的正确选择和使用，既能体现出数据处理硬件技术的进步，又能满足信息数字化和网络迅速发展的实际需求。作为基础分析软件，Excel 简单易懂、方便快捷，在基础研究、行政管理等领域应用广泛。但随着数据急速膨胀，统计分析的内容不断扩大，不仅需要完成数据整理、模型构建、可视化等环节，还需要借助功能强大的分析工具，丰富方法模型的内容，加强研究学习的深度，呈现分析结果的特色。作为一种功能强大的开源编程语言，R 语言包含丰富的软件包和绘图技术，能帮助我们完成数据分析，构建统计模型，展现研究结果。

本书为熟悉 Excel 操作的人士提供通向 R 语言的实用指南。通过两种软件的比较，围绕描述性统计、回归分析、方差分析、logistic 回归、主成分分析几大模块，详细举出所需函数及代码，有效帮助读者在类比中学习掌握 R 语言，实现从 Excel 到 R 的过渡。

从大学开始，本人花了 9 年时间学习统计学。《 The Elements of Statistical Learning 》《复杂数据统计方法——基于 R 的应用》以及《 An Introduction to Statistical Learning with Application in R 》对我的影响很大。让我在深入学习数据挖掘与机器学习、社会网络分析、结构方程模型、分位回归和缺失数据理论方法的同时，关注 R、Python、SAS 等软件的编程与实现。这类编程软件不但可以帮助我们训练大脑的逻辑，验证改进方法的效果，而且有助于我们获得研究结论。也正因如此，我才致力于向广大读者推荐编程类软件，灵活多样地实现个性化需求，完成深度研究任务。

虽然我独立翻译过《Python 数据可视化》《预测分析建模：Python 与 R 语言实现》和《 R 统计应用开发实战》，参与翻译过《商务与经济统计》和《 R 语言编程艺术》，也参与编写过《大数据挖掘与统计机器学习》，但本次翻译与以往不太一样。它让我重新认识了 Excel 和 R 之间的区别与联系，用不同软件示范处理同一问题，为广大读者提供多种选择。

记得在中国人民大学“毕业十星”之“学术之星”的获奖感言中，我曾写道：对我而言，翻译是一种特殊的休息方式。与音乐一样，知识的传播没有国界。因此，翻译不仅是知识表达语言的转换，更是一次学习和交流的机会。与原作者对话，高山仰止，受益匪浅；与读者对话，高山流水，闻过则喜。我喜欢这种自由创作的休息方式，更乐意在翻译中发现自己的不足。

最后，非常感谢机械工业出版社的编辑。感谢刘钰洁参与第 4 章初稿的校对工作，程悦参与第 5 章、第 6 章初稿的校对工作。

感谢中国科协创新战略研究院的各位领导和同事。感谢我的博士导师——中国人民大学的易丹辉教授。感谢我的爷爷奶奶、爸爸妈妈以及各位亲朋好友，是他们给了我前行的动力和勇气。

鉴于个人时间与水平有限，如有纰漏，向您致歉，还望海涵。同时也请各位读者予以反馈，不吝赐教！

程豪

作者简介

康拉德·卡尔伯格（Conrad Carlberg）是美国定量分析、数据分析和管理应用程序（Microsoft Excel、SAS 和 Oracle）领域的知名专家。曾获科罗拉多大学统计学博士学位，是微软 Excel 的 MVP。个人网站是 www.conradcarlberg.com。

Carlberg 是南加利福尼亚本地人。大学毕业后，他搬到科罗拉多州，在那里就职于一些创业公司，并进入研究生院继续深造。他还在中东待过两年，从事计算机科学教学工作。研究生毕业后，Carlberg 在美国西部公司（从 AT&T（美国电话电报公司）解体时分拆出来的“小贝尔公司”（Baby Bell））的产品管理部门和摩托罗拉公司工作。

1995 年，他创办了一家小型咨询公司，为那些想通过定量分析指导商业决策的公司提供设计与分析服务。如今，这些定量分析方法统称为“分析学”。他喜欢有关这些分析技术，尤其是有关如何通过最广泛使用的数值分析应用程序 Microsoft Excel 来实现这些技术的写作。

前　言

父亲曾经告诉我，在学术界，研究问题像木桩，各种应用程序像刀，因为木桩很小，所以伐木的刀需要格外锋利，否则难以砍下木桩，解决棘手的问题。这曾是高校教职人员不断争论的话题。我还听过很多不同的其他版本。当看见人们在讨论应用程序 R 和 Microsoft Excel 的区别时，我又想起了这句话。那种感觉异常强烈。

如果说我对 R 和 Excel 存在个人偏好，那么你可能认为我更倾向于选择 Excel。自 20 世纪 80 年代末以来，我一直使用 Excel 作定量分析。无论是金融分析还是统计推断，Excel 都能帮助我很好地解决问题。作为一名顾问，如果客户的系统中安装了 Excel 并且他们能熟练操作，那么这对我来说意义非凡。

Excel 可以展示出很多统计分析内部（“黑匣子”）的细节。客户尽管没必要掌握从原始数据到最终概率表达的所有细节，但也需要知道这些细节可查，以便应对不时之需。

此外，Excel 还是一种功能强大的学习工具。Excel 的工作表函数和求解器 Solver 可以构建二元 logistic 回归模型。完全理解统计分析的最佳方法就是从头开始完成整个操作。

从更技术的角度来说，Excel 并不是理想的统计应用程序。（Excel 从不在考虑范围内。）这是因为自从 30 年前 Excel 首次发布以来，你还是会发现它在统计性能上的一些缺陷和错误，但 SAS、SPSS、Stata、Minitab 等软件不存在这些不足。在此期间，Microsoft 已经解决和修正了很多统计功能方面的问题。但是，解决 LINEST() 函数中常数为 0 的问题比较麻烦，需要对传统代数矩阵进行 QR 分解。从 Excel 2016 的分析功能来看，这些问题仍然存在。

但是，Excel 确实有助于统计分析，尤其是用 VBA 新增功能修复本地工作表函数时，Excel 的帮助更大。另一方面，Excel 能够处理的统计问题有限。比如，习惯于分析损益表和资产负债表的 Excel 用户，很容易达到初级、中级的统计分析水平（如多元回归）。Excel 在处理统计问题方面也仅限于此。

R 则有所不同。你很难举出 R 无法处理的统计问题。作为另一种免费开源软件，学会使用 R 完全是另一回事。我们主要通过命令行界面和菜单结构实现 R 的操作。（也可以通过一些图形用户界面使用 R，在我看来这些界面都不令人满意。）下面列出

R 的一些特征：

- **R 语言是区分大小写的**，使用时要确保正确使用大小写字母。例如，Anova 和 anova 在 R 中是两个不同的函数。尽管这两个函数都可返回方差分析表（方差分析的首字母缩略词即为函数名），但只有一个函数可以正确处理单元格观测数不同的因子分析。

  再比如，函数 XLGetRange 可以直接导入 Excel 工作表数据，为后续分析做准备。但是，最好不要输入 xlgetrange，因为 R 会显示无法找到目标函数 xlgetrange。

- **R 不存在明确的数据格式管理规则**。存在这样一类函数，需要通过设置一些函数参数来决定函数结果的小数位数。还有一类函数，需要通过 options 语句或 print 语句来提供这些信息。在某些情况下，可以将字符作为整数中的千位分隔符，对于分数等数值，需要再次使用字符作为分隔符。

- **R 中反斜杠的作用与文件地址中的反斜杠不同**。以前，可能常用反斜杠指定一个路径，比如，csv 文件的地址如下：C:\Users\Fred\Desktop\jr.csv。

  但如果在 R 的 read.csv 函数中用反斜杠读入文件，则会出现错误。

  R 不用单个反斜杠分隔子文件夹和文件夹。R 中的单个反斜杠解释为一个转义符。如果想要指定文件路径，则必须输入两个反斜杠：

  C:\\Users\\Fred\\Desktop\\jr.csv

  或者使用斜杠：

  C:/Users/Fred/Desktop/jr.csv

现在，这些规则可以称为一些“小麻烦”，而不是“错误”或者“缺陷”。R 与 Excel 在 LINEST() 中返回回归系数的顺序问题类似，R 与 Excel 中的函数 CORREL() 和 PEARSON() 等价。然而，这些代表着成功学会用 R 进行统计分析的阻碍。

上面提到的问题仅仅是一些例子。那么，如何充分利用这一免费且功能广泛的应用程序，而不受这些“小麻烦”的影响呢？在我看来，唯一的方法是多加练习，熟能生巧。

但是，如果你习惯用 Excel 做统计分析，我知道你会做哪些分析。你会得到均值、标准差、中位数等描述性统计量和置信区间等推断统计量，以便更好地理解数据的分布特征。这些统计分析工作会用到诸如 AVERAGE() 的工作表函数和数据分析插件等应用工具。

对于简单的相关关系和不同因子水平下数值变量的双变量分析，通常会用到 Excel 工作表函数，如 CORREL()、带趋势线的散点图和数据透视表。

可以用多元回归分析多变量的样本数据。对于这类统计推断问题，Excel 中的

TREND() 和 LINEST() 函数，以及数据分析插件中的回归工具，都是有用的方法。

你可能不想止步于对不同因子水平下数值变量的简单统计分析，也不想仅仅完成对数据总体的统计推断。这时可以用方差分析法（ANOVA），即用标准的工作表函数完成 ANOVA，同时得到上述的统计分析和推断结果。数据分析插件中的工具同样能够达到相同的效果。

或许，你还想进一步研究二分类结局变量（如购买 / 不购买）的概率，作为以*生产线*等为因子、*页面停留时间*等为协变量的函数。那么你需要使用前面提到的 logistic 回归，使用 LN() 和 EXP() 以及求解器 Solver 来确定方程表达式，预测二分类结局变量。

甚至还有可能，Excel 的统计分析功能已无法满足你的需求，需要用 VBA 代码从相关矩阵提取主成分。主成分分析法是处理数据集中可测变量过多的一种标准方法，它可以在 Excel 工作表中将这些变量降维为少数几个潜变量。

此外，你可能还常常在 Excel 中进行一些其他的统计分析工作，但上面列出的应该是你会在 Excel 中进行的绝大多数分析工作。这些都可为学习 R 打下理想的基础。

假定你一开始关注的是 R 中与 Excel 处理任务相同的函数。那么随后，你可以关注与 Excel 操作最为类似的 5 或 10 个 R 程序。通过比较这两个应用程序的运行结果，你可以像熟悉 Excel 中类似功能一样熟悉这 5 或 10 个 R 函数。

通过上述学习方法可以减少学习 R 的难度。这样，你就突破了现有的分析限制。你的数据集可能至少包括两个因子并且每个单元格的观测数不同，或者包括一个因子和一个协变量，你需要使用方差分析法。尽管这些分析过程需要付出很大努力，但是你仍有充足的理由用 Excel 解决这些问题。

但如果你已经尝试用 R 的 ANOVA 函数处理平衡因子设计，那么在处理不平衡因子设计时只需要验证需要设置的选项。接下来的一小步是知道如何通过方差分析检验因子和协变量交互效应。尽管 Excel 能够展示分析的内部过程，但在细节设置方面有些不足。相比之下，R 看起来更具吸引力。

这些就是我在本书中采取的方法。能够用 Excel 进行统计分析的你，应该至少熟悉前面提到的一些常用分析方法：单变量描述性统计、双变量分析、一元回归和多元回归、方差和协方差分析、logistic 回归和主成分分析。

作为引言或者综述，我会给出这些分析在 Excel 中的实现过程。我也会展示如何用 R 得到相同的正确结果，包括安装哪些软件包以及如何获取这些软件包。这样你就可以在特定情况下做出选择：也许，Excel 适合于需要逐步解释分析过程的你，而 R 适合于对 Excel 持怀疑态度，直接完成运行，得到翔实分析结果的你。

从未用过 R 的读者可能会在第 1 章讨论的内容中发现一些不同于 Excel 的新知识。

致谢

感谢 Charlotte Kughen 和 Michael Turner。Charlotte 过去一直指导我写书，Michael 提供了简化清楚的技术建议。我很高兴感受到他们为本书的付出——因为本书的目的是覆盖两种应用程序，而不仅仅是一种，所以本书看起来有些难度。也要感谢 Trina MacDonald 将这些内容整理到一起。

目　录

第 1 章

从 Excel 到 R 的过渡

有时很多有经验的 Excel 用户出于各种原因，决定尝试使用 R。可能你已经习惯于用 Excel 比较生产线，预测网站点击率或医院病人数，按照性别和政治关系划分选民调查。对于这些统计分析，Excel 几乎总是表现良好。

而且，Excel 通常很容易完成数值分析并得到结果。在大多数商业、教育和政府环境中，你很难找到没有安装 Excel 的电脑，也很难找到不会使用 Excel 的人。

尽管如此，你还是可以遇到 Excel 无法完成的统计工作。Excel 只是一个普通的分析包，不可能包括所有的统计应用工具。

比如，多元分析通常需要你从相关矩阵提取主成分。Excel 在诸如 VBA 子程序等的外部帮助下可以实现主成分分析。抑或在方差分析 (ANOVA) 中，你可能想运行基于 q 分布而不是更普遍的 t 分布或 F 分布的多元比较程序。在大多数情况下，Excel 无法实现这些分析。同样，Excel 也无法帮助你得到诸如区间、中位数和分位数的所谓“稳健”统计量：你可以很轻松地在 Excel 数据透视表中计算并得到均值，但第一分位数的计算和获得就没那么容易。

更复杂地，很多办公单位存在一些对应用程序的偏见：这暗示着任何不会用 R 完成统计分析的人可能都无法胜任这里的工作。

1.1　调整预期

在此，我由衷地说明：将 R 加入到 Excel 工具箱并不简单。R 与 Excel 存在很多不同之处，比如：

- 我们习惯于菜单式的应用程序（比如 Office 办公应用程序），R 的用户界面看起来完全不同。R 需要你输入命令行。命令行所在位置是 R 控制台。图 1.1 是打开 R 后立即可以看到的控制台。
- Excel 在单元格中保存计算公式，在工作表中展示结果，当引用单元格数据发生变化时会自动更新计算结果。R 展示的是静态结果，这意味着当引用单元格数据发生变化时需要人工重新计算从属单元格数据。
- Excel 中的函数名称和参数不需要区分大小写。无论输入 "=average(A1:A10)"、"=AVERAGE(A1:A10)" 还是 "=AvErAgE(A1:A10)"，都可以得到相同的结果。但 R 可以识别 "DescTools"，却无法识别 "desctools"。

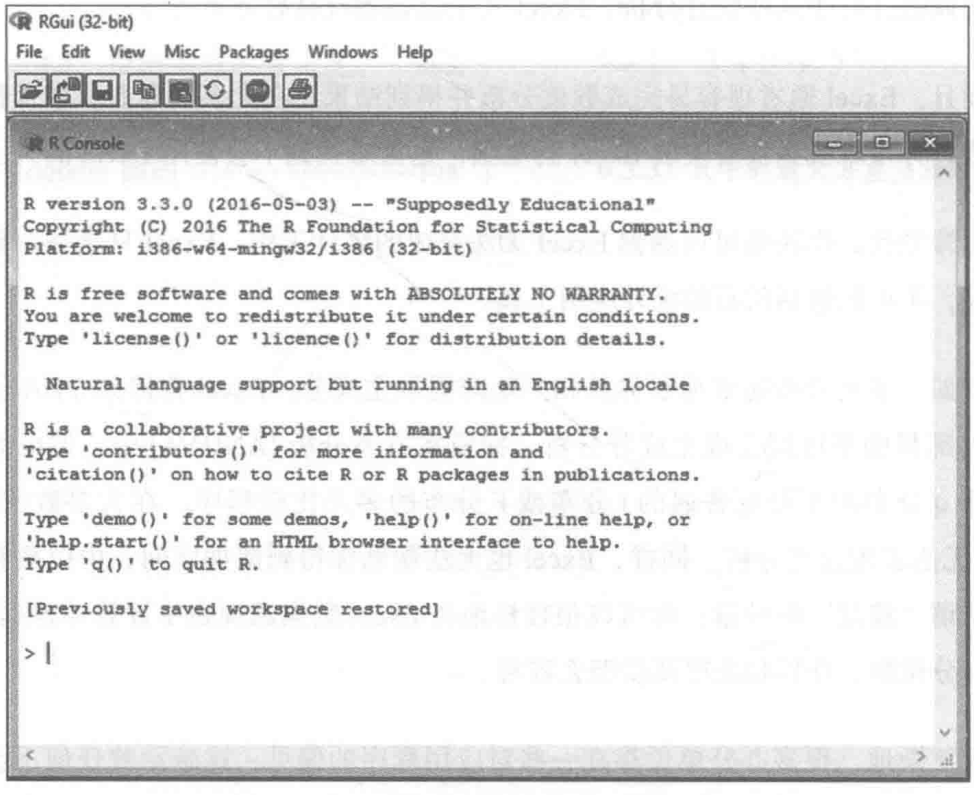


图 1.1　稍大一些的是命令提示符。可以在其右侧立即输入命令行

因此，如果你正从 Excel 或其他菜单式的应用程序向 R 过渡，那么就需要彻底

改掉一些长期存在的习惯。这里将举出一些例子，让你有所认识，并在必要时重复强调。

1.1.1 分析数据：软件包

R 自带一些已安装好的软件包，比如 base（基础）和 stats。这些软件包提供很多函数（比如均值函数），你需要花费一些时间才能熟悉这些统计功能。此外，R 也提供了大量扩展包，用于拓宽 base 应用程序的功能。R 列出的扩展包成千上万。你在使用时可能会读到更多软件包。

例如，一种比较基本的软件包是 DescTools，具体介绍见第 2 章。使用 DescTools 软件包中的 Desc 函数，可以快速获得单个名义变量或者数值变量的分布信息。可以通过一些方法进行双变量分析，比如：

- 数值变量的相关关系（以及散点图）
- 根据名义变量（R 中称为因子）划分数值变量
- 两个因子的列联表，通常叫做交叉表。

你可能希望从 R 的 base 应用程序中得到这类信息，但从 DescTools 软件包得到的结果更为深入：不仅包括 Pearson 相关系数，还有 Spearman 相关系数 r 和 Kendall 相关系数 tau；按秩划分的 Kruskal-Wallis 单因素方法；以及 Pearson 卡方检验中列联表的似然比。

1.1.2 存储和排列数据：数据框

R 强烈依赖于数据框。你如果熟悉 Oracle、DB2、SQL Server 等主流的几个数据库管理系统中的一个，那么就会发现数据框和数据库中的表格类似。数据框是矩形的，即每行有相同的列数，每列有相同的行数。这个特点限制了数据框的结构类型——因此，你不能在数据框中存储一个统计报告。通常，在数据框中存储的是原始数据。

数据框比数据值蕴含更多信息。无论是你提供，还是你不提供时默认 R 提供，数据框都包括变量名信息。它们会自动识别每个变量中的数据值类型：常规文本型、数值型和逻辑型（即布尔值 TRUE 和 FALSE）。

无论出自R的基础系统还是扩展包，数据框通常是R统计程序使用的数据源。生成数据框的方法多种多样。可以通过R控制台生成数据框，尽管这些方法通常是最耗时的且单一的。也可以将Access或者Excel等其他应用程序的CSV（逗号分隔值）格式文件导入到R中。可以用如XLGetRange的R函数，直接从一个开源的Excel工作表抓取数据。


注释　很多函数（包括本书提到的在内）都属于特定的扩展包。因此，只有安装并加载扩展包，R才能识别出相应的函数。比如，在使用XLGetRange前，先运行library(DescTools)语句，这是因为XLGetRange是定义在软件包DescTools中（而不是R中）的函数。

R的基础系统及其扩展包包括你可用的数据框，用来比较已有分析结果和预期结果间的差异。

1.2　用户界面

R的大多数操作在控制台中实现（如图1.1所示）。对控制台越熟练，操作效率越高。R控制台包括一个菜单结构（文件、编辑、视图等），但该菜单结构的排列非常稀疏。发送给R的绝大多数命令都是通过在控制台中输入命令语句来实现的。

R命令语句很容易输错——比如，R语句需要区分大小写，你可能会发现自己输入的是Mean而不是mean。因此，通常需要返回一处命令（有时是多处命令）进行修改。这时，你可能想要复制错误命令，返回命令提示符，粘贴并改正。

上述操作有时很有帮助，尤其当出现问题的命令还在控制台时。还有一种更好的方法是按向上的箭头键，直到命令提示符旁再次出现错误命令行，使用向左的箭头键或者通过单击定位到错误位置进行修改，然后不管光标是否在命令行中间，都按Enter键回车。Enter键不会在命令中间位置回车：它只是让R运行光标所在的命令行。

在R会话过程中，通常可以创建用于存储的对象：数据框、列表和向量，甚至还有将要存储到一个变量中的统计分析结果。这些对象都保存到R的工作空间（workspace）中。