第一部分

R 语言基础知识

第 1 章
欢迎使用 R

1.1 R 是什么

市场分析师无疑都听说过 R。你或许曾尝试过使用 R 但是觉得摸不着头脑，之后你转而使用其他“好用”的工具。你或许知道 R 使用的是命令行，你并不喜欢这点。抑或你确信 R 对于分析领域专家来说是很有效的工具但你没有时间学习如何使用。

我们可以帮你！我们的目标是在最短的时间里展示最关键的部分，你可以按照指导亲自动手实践，尽快学会有效使用 R。此外，我们还会涉及一些高阶话题，在展示 R 的能力同时教高阶用户一些 R 的新技能。

记住关键的一点，R 是一门编程语言。它不是像 SPSS、SAS、JMP 或 Minitab 这样的“统计程序”，也并不想成为它们中的一员。官方 R 项目对其描述是“用于统计计算和图形的语言和环境”。注意这里着眼点是 R 是一门“语言”。“统计”和“图形”处于同等地位作为修饰语。R 是用于统计的很好的编程语言。其底层语言的发明者 John Chambers 获得了 1998 年美国计算机协会的软件系统奖（Association for Computing Machinery (ACM) Software System Award），该奖项针对的是能够“永久改变人们分析，可视化和处理数据的方式……”的系统 [6]。

R 基于 Chambers 于 20 世纪七八十年代在贝尔实验室发明的 S 语言（S 是英文单词“统计（Statistics）”的首字母），UNIX 操作系统和 C 编程语言也是在该实验室产生的。S 语言的实现商业版本 S-PLUS 在 20 世纪 90 年代吸引了分析师和学术界的注意。Robert Gentleman 和 Ross Ihaka 希望扩大 S 的使用范围并且在 1997 年推出了开源项目 R。

自那以后 R 的普及度呈几何级增长。R 真正神奇的地方在于用户可以对 R 的发展做贡献，提升 R 的方式从添加核心函数到高度特异化的方法等不一而足。许多使用者都有对 R 社区做贡献！现在有 6000 多个 R 包，为 R 添加了许多功能（最新的统计数字见 http://cran.r-project.org/web/packages）。

如果你有编程的经验，那么会立刻喜欢上 R 的一些关键特性。如果你是编程新手，本章将介绍 R 的特殊之处并且在第 2 章介绍 R 编程基础知识。

1.2 为什么用 R

使用 R 的理由有很多。R 是统计学家使用最广泛的平台，统计专业人士提出的最新方法首先得以用 R 实现。R 迅速发展成为大学教育的默认平台，且扩展到其他学科，如经济学和心理学。

R 为分析师提供了最多最广的分析工具和统计方法集。它使你能够重复使用分析代码，且扩展 R 系统本身。R 能在大多数操作系统上运行并且能和其他数据系统交互，如在线数据、SQL 数据库。R 提供优美且强大的绘图函数，相较于电子表格，R 能大量绘制更加贴合当前情况更具信息量的图形。把这些特点加起来，R 能够极大地提高分析师效率。Elea 认识一个企业分析师使用 R 将下载数据和创建固定形式月度报表的过程自动化。该自动化过程每个月为他节省了几乎 40 小时……他直到几个月之后才告诉他的经理！

R 有一个社区。许多 R 用户都是 R 的狂热者，他们乐于助人，并且享受解决问题带来的喜悦感以及总是学到新东西的过程。R 是用户建立的动态系统，关于 R 总有新的东西要学习。R 相关知识是越来越多顶级公司分析工作需要的宝贵技能。

R 代码是开放的，你可以选择相信函数背后的代码，也可以随时核实。所有的核心代码和人们贡献的大部分 R 包都是开源的。你能检查其中的代码，看看究竟是如何分析的，以及函数运行背后到底发生了什么。

最后，R 是免费的。这是 R 核心开发组成员对职业的热爱和自豪感的结果，其中包括杰出的统计学家和计算机科学家。与所有杰作一样，建造者的专注明显反映在最终的作品上。

1.3　为什么不用 R

R 有哪些不讨人喜欢的地方？你无疑已经看到并非人人都使用 R。对我们而言不用 R 是无法想象的，但是其他一些分析师或许有不用 R 的理由。

其中一个不用 R 的理由是：在你掌握了该语言的基本知识之前，用 R 做一些简单分析也困难重重。如果你是 R 的新手，想要用 R 得到均值表格，交叉表格或者 t 检验会令你抓狂。R 是用来进行强大、灵活、可控、可迭代的分析的，是用最前沿的方法，而非仅仅是用鼠标点击得到结果。

另外一个不用 R 的理由是你不喜欢编程。如果你是编程新手，R 是很好的入口。但是如果你已经尝试过编程并且一点也不喜欢，那么 R 对你而言也会是一个挑战。我们的任务是最大限度地帮助你，我们会努力尝试教你使用 R。然而不是每个人都喜欢编程的。从另一个角度，如果你是一个有经验的程序员，R 对你而言会很简单（或许只是表面上简单实质并非如此）。我们会帮助你避免一些错误。

一些公司和他们的信息技术或者法律部门因为 R 是开源的而对其抱有怀疑。经理们常常问：“免费的东西质量能好吗?”该问题有很多种回答，可以列举出成百上千的关于 R 的教科书，学术杂志中对 R 的引用，以及一些杰出的 R 贡献者（在 R 中运行 contributors() 命令可以得到贡献者的名字，可以在网上搜索他们的背景）。你或许可以尝试工程师的格言：“质量、速度和价格三选其二。”R 有好的质量和价格，而非速度，在目前情况下掌握 R 需要努力和时间。

至于 R 是免费的这一点，你要明白，R 的贡献者的确从中得到了回报，只是非金钱回报。对他们工作结果的报酬是尊敬和名誉，并且每个贡献者也从其他贡献者的工作中获益。这是一个理性的经济学模型，尽管货币价格是 0。

关于 R 的最后一点担忧在于该生态系统的不可预测性。成千上万的作者贡献无数 R 包，有些含金量极高，有些表现平平，有些甚至有错误。能够使用最前沿的方法的负面效应源于有些方法并不能经受时间的检验。使用者需要自行决定某方法是否能够满足其需求，而不能永远靠时间或者权威告诉你该方法是不是适合你（但是你很快会知道哪些作者和专家的建议

是可信的）。如果你相信自己的判断，使用 R 和使用所有其他软件在这点上没有区别：一经出售概不负责。

希望我们成功地说服你在很多情况下，R 能带来的好处远超过使用的困难。

1.4 什么时候用 R

常见的使用 R 的情形有：

- 你想要使用比现有方法更有效的新方法。许多 R 用户正是因为这个原因开始使用 R 的。他们在学术期刊，会议论文或者报告中看到某种方法，然后发现该方法只能用 R 实现。
- 你需要重复某分析很多次。这是 Chris 开始使用 R 的原因。在毕业论文中，他需要在现有方法应用 bootstrap 以便将它们的结果和一种新的机器学习的方法进行比较。R 对于模型迭代很有效。
- 你需要对多个数据集进行分析。因为 R 是脚本语言，所以很适合需要在不同数据集上重复分析的情况。R 还有自动报告的功能。
- 你需要开发新的分析技术或者希望更好地理解和控制当前方法的使用。对于许多统计过程，R 强过许多其他编程语言。
- 你的经理、教授或者同事鼓励你使用 R。我们已经影响了一批学生和同事开始使用 R，我们很高兴看到他们中大部分现今是 R 的狂热使用者。

通过展示 R 的强大功能，我们希望说服你们意识到现有的工具并不令人满意（与 R 比较起来）。更进一步的，我们希望重写你们对“令人满意的工具”的定义。

1.5 如何使用本书

本书的写作目标是启发性和实践性。我们想要用简洁的语言介绍 R 和我们使用的模型，我们期待读者的互动，实践书中的 R 代码。书中的代码是为读者能够在阅读过程中自行键入命令而设计的（我们还提供了代码文档，读者可以从本书的网站下载这些文档，见 1.5.3 节）。

1.5.1 关于本书编排

读者可以自己键入运行的 R 命令通过如下代码框展示：

```
> citation()

To cite R in publications use:
  R Core Team (2014). R: A language and environment for statistical computing.
     R Foundation for Statistical
  Computing, Vienna, Austria. URL http://www.R-project.org/.
...
```

我们在第 2 章中描述代码框和 R 的使用。这些代码总体按照 Google 的 R 代码格式（具体格式指南见 http://google-styleguide.googlecode.com/svn/trunk/Rguide.xml），为了更清晰地展示代码和文本，有个别地方我们没有按照 Google 的格式。（进一步学习 R，你会希望让你的代码更具有可读性；Google 指南对代码格式非常有帮助。）

当涉及 R 命令、附加 R 包或者代码框外的文本时，我们会用这样的等宽字体：citation()。函数名称后面保留括号是为了表明其是个 R 函数，如 summary() 函数（2.4.1 节），而对象的表示方法则不同，如 Groceries 数据集（12.2.1 节）。

当引入或者定义全新的概念时，我们会用斜体，如向量（vector）。斜体有时也用来表示强调。

我们将在本书中循序渐进地教 R，且大部分的 R 教学内容将穿插在各个介绍市场营销和统计模型的章节中。在这些章节中的语言简述部分讨论程序语言话题（如 3.4.5 节）。为了尽可能多地学习 R 语言，你需要阅读语言简述部分的内容，即便只是稍稍浏览相关统计模型的部分。

一些小节覆盖了更多细节或者更高阶的话题，读者可以跳过这些部分。这些小节用星号标记，如*知识拓展 **。

1.5.2　关于数据

本书中用于分析的大部分是模拟数据集。这些数据都是用 R 代码生成的并且有特定的结构。这么做有如下好处：

- 其使我们在无法找到合适的真实市场营销数据的情况下也可以展示分析过程。这是非常有用的手段，因为很少有公司会愿意分享数据，如客户分组数据。
- 其使得本书更独立而不需要依赖很多下载数据。
- 其使我们能够改变数据并且重新运行分析过程并且观察结果的变化。
- 我们能够通过这种方式教授重要的 R 技能：处理数据，生成随机数以及写循环代码。
- 其展示了如何在拿到真实数据前准备好分析代码。这样一来，当你拿到实际数据时只要在该数据上运行代码即可。

第 12 章的交易数据是一个例外，这样的数据很复杂且难以制造，并且我们可以找到合适的公开数据 [20]。

我们建议读者自己阅读数据模拟的部分。这些部分可以教读者使用 R，并体现了典型市场营销数据的特征。但若你为了继续某章的学习而要马上得到数据，可以下载每章的数据（下一节会提到）。

只要可能，读者应该分析自己的数据。我们在每个章节中都会用数据作为例子，但最好的学习方式是将书中介绍的分析应用到其他数据上，然后自己解决在应用过程中遇到的问题。由于本书是教科书性质的，而不是学习手册，且由于 R 学习起来可能较慢，我们建议读者在没有截止日期压力的情况下平行实践不同的分析任务。

一开始你可能会觉得在你自己的数据上重复书中介绍的分析实在太简单了，但当你尝试对另外一个数据集使用高阶模型时，最好还是事先有一些用相同模型分析不同数据的经验。你越早能用 R 分析自己手上的数据，就能越快地有效使用 R。

1.5.3　在线资源

本书有网站：http://r-marketing.r-forge.r-project.org 。建立该网站的主要目的是存放供下载的 R 代码和数据集。虽然我们鼓励你们偶尔使用这些资源，但你如果自己写代码模拟数据则会学得更快。

你会在网站上找到：

- 一个欢迎页面，上面有一些更新的消息：http://r-marketing.r-forge.r-project.org。
- 以“.R”为后缀的 R 代码文本文档：http://r-marketing.r-forge.r-project.org/code。
- 书中模拟的一些数据集副本：http://r-marketing.r-forge.r-project.org/data 这些数据集可以通过 R 函数 read.csv() 直接下载（2.6.2 节有相关代码，3.1 节中有下载数据的代

码样本）。

- 一个打包了所有数据集和 R 代码文件的 ZIP 压缩包：http://r-marketing.r-forge.r-project.org/data/chapman-feit-rintro.zip。

数据链接是简化版的 goo.gl 链接，以节省空间。附录 D 中有更多关于在线资料和不同数据获取方式的细节。

1.5.4 遇到问题时的解决方案

当你学习如 R 一样复杂的技能或者新的统计模型时，会遇到大大小小的问题（警告和错误提示）。不仅如此，R 生态系统是动态的，本书出版后的情况和写书时可能又不一样。我们并不想用一大堆可能的问题来吓唬你，只是想让你对一些小问题有所准备并且知道如何应对一些较大的故障。当你的结果和书中的不符合时你可以尝试如下方法。

- **检查代码本身**。使用 R 时基本的差错过程是仔细检查代码的每个细节，特别是圆括号、方括号和大小写字母。如果面对一条很长的命令，将其分解成为不同的部分然后重组起来（我们之后会举例介绍）。
- **检查 R 包**（附加包）。R 包时不时会有更新。有时一些包会改变工作的方式，或者会在一段时间里不工作。一些包很稳定，另一些则不然。如果你的安装包遇到问题，上网搜索相关的错误提示信息。如果输出或者有些细节和书中展示的略有不同，不要担心。错误提示“没有名为……的 R 包”（"There is no package called..."）表明你需要安装该包（2.2 节）。其他问题见下面的建议或者查阅包的帮助文档（2.4.2 节）。
- **检查 R 的警告和错误信息**。R 的“警告”通常富含信息且不需要纠正。之后我们展示代码的时候会指出。有时候这些警告信息在包升级后又会消失。如果 R 给你“错误”提示，那意味着有些地方出了问题且需要纠正。在这种情况下，再尝试运行该代码，或者在网上搜索相应错误提示。
- **检查数据**。我们的数据是随机抽取的，受随机种子的影响。如果你生成的数据略有不同，再次从头开始运行一遍，或者从本书的网站上下载数据（1.5.3 节）。
- **检查模型**。导致统计估计的原因可能有三：数据略有不同（参看前一条），包有所更改导致估计结果略有不同，统计模型中使用了随机抽样。如果你运行模型得到的结果仅和书中有细微的差别，可以假定上面三件事中发生了某件。尽管继续就好。
- **检查输出**。包有时会改变输出的内容。本书中展示的是在写书时相应包的输出。但你使用这些包时可能有所改变。
- **检查无法找到的函数名**。有时 R 包中函数的名字或者结果的结构会变化。如果你用一条错误的代码试图从统计模型中提取某信息，那么检查模型的帮助文档（2.4.2 节）。可能一些函数名字改变了。

我们的总体建议是，如果差异很小，如均值分别为 2.08 和 2.076，或者 p 值分别为 0.726 和 0.758，那么不要太担心，通常情况下可以将其忽视。如果你发现差异很大，如统计估计 0.56 而不是 31.92，那么再次运行书中的该段代码（见 1.5.3 节）。

1.6 关键点

我们在每个章节的末尾都会总结关键点。本章只有一条：如果你准备好了学习 R，我们从第 2 章开始吧！