WILEY

大数据应用与技术丛书

Statistical Data Cleaning with Applications in R

R统计数据清洗及应用

[荷兰] 马克·范德鲁(Mark van der Loo) 著
埃德温·德荣格(Edwin de Jonge)
杨小冬 译

清华大学出版社

大数据应用与技术丛书

R统计数据清洗及应用

[荷兰] 马克·范德鲁(Mark van der Loo)
埃德温·德荣格(Edwin de Jonge) 著
杨小冬 译

清华大学出版社
北 京

Mark van der Loo, Edwin de Jonge
Statistical Data Cleaning with Applications in R
EISBN: 978-1-118-89715-7

北京市版权局著作权合同登记号　图字：01-2018-5172

图书在版编目(CIP)数据

R 统计数据清洗及应用 / (荷)马克·范德鲁(Mark van der Loo)，(荷)埃德温·德荣格(Edwin de Jonge) 著；杨小冬 译. —北京：清华大学出版社，2019
(大数据应用与技术丛书)
书名原文: Statistical Data Cleaning with Applications in R
ISBN 978-7-302-52662-9

Ⅰ. ①R… Ⅱ. ①马… ②埃… ③杨… Ⅲ. ①统计分析—统计程序 Ⅳ. ①C819

中国版本图书馆 CIP 数据核字(2019)第 053590 号

责任编辑：王　军
封面设计：孔祥峰
版式设计：思创景点
责任校对：牛艳敏
责任印制：沈　露

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：三河市少明印务有限公司
经　　销：全国新华书店
开　　本：170mm×240mm　　印　　张：18.5　　字　　数：374 千字
版　　次：2019 年 6 月第 1 版　　印　　次：2019 年 6 月第 1 次印刷
定　　价：79.80 元

产品编号：080801-01

译 者 序

我们生活在一个信息大爆炸的时代，通过各种媒介获取的数据不仅数量巨大，而且纷繁复杂、形式各异。其中既包含我们需要的有用信息，也充斥着很多冗余的、无用的甚至错误的信息。这种情况为数据分析带来了非常大的困难。面对这种困境，数据清洗适时出现。

数据清洗是对数据进行重新审查和校验的过程，目的在于删除重复信息、校正存在的错误、处理无效值和缺失值，并提供数据一致性。顾名思义，数据清洗就是把“脏”的东西“洗掉”，它是发现并校正数据文件中可识别的错误的最后一道程序。其原理是利用数理统计、数据挖掘或预定义的清洗规则等技术将脏数据转换为满足相应质量要求的数据。

本书既包含丰富的理论知识，同时又列举了大量示例，让我们可以轻松地了解如何在实践中应用相关知识。当然，书中涉及很多专业性很强的概念和术语，要求读者对数学和统计学有基本的了解，同时还应该具备一定的编程经验。其中期望值、方差、基础微积分和线性代数方面的知识是必不可少的。此外，如果具备一定的 R 语言知识，对理解本书中的内容会有很大的帮助。除了理论知识介绍和示例之外，书中还提供了一些练习，可以进一步加强读者对软件和相应方法的理解和运用。如果正好面临数据分析的难题，相信本书一定能够为你提供必要的帮助，助你顺利完成工作。

本书的两位作者均来自荷兰统计局统计方法部，在数据清洗、统计和分析方面具备丰富的理论和实践经验，同时还融入了很多相关领域专家的真知灼见，也得到了 R 核心团队、软件包开发人员的大力帮助以及 R 社区的大力支持，这使得本书在所述领域具备很高的权威性。学完本书，一定会让你受益匪浅。

在这里我要对清华大学出版社的李阳编辑及其他相关人员表示感谢，感谢你们的辛勤付出，没有你们就没有本书的顺利出版。

在本书的翻译过程中，译者本着“信、达、雅”的原则，力求能够准确表达出作者的原意，让读者能够轻松理解所讲述的内容。当然，由于译者本身的水平有限，书中难免会出现一些错误，欢迎读者不吝指正，在此向你表示感谢。本书内容主要由杨小冬翻译，参与翻译工作的还有彭明珍、王亚、王浩、程建福、彭林、王欢、彭征肖、武彦龙、杜娟、汪春燕、白德、王亚卫。

最后，要对我的家人表示深深的感谢，谢谢你们的理解和支持，没有你们做坚强的后盾，本书就无法按时翻译完成。

前　言

在数据分析中，数据清洗往往是最为耗时的部分。在“官方统计”(Official Statistics)社区，很久以前就已经开始将数据清洗作为一个单独的学科进行研究(在该社区，数据清洗被称为“数据编辑”)。此外，在研究中还引入了数据库的相关知识，尽管如此，针对大型统计社区的文献著作还是非常有限。正是因为这个原因，当出版方邀请我们对之前为 *useR*!2013 大会编写的《R 数据清洗简介》教程进行扩展，进而编纂成一本书时，我们毫不犹豫就答应了，这也是我们的心声。一方面，我们认为，过去 50 年中在“官方统计”社区发布的一些方法应该为更多的用户所了解和使用，而本书或许可以为此助一臂之力。另一方面，我们正在从基于调查的数据源过渡到管理型“大”数据源，希望本书能够帮助为“官方统计”社区增加一些(通常是预先存在的)相关技术。

对于我们来说，通过编写本书也可以帮助我们系统地梳理相关知识，进一步完善之前针对这一主题所编写的软件。回过头来看，我们最终不仅成功完成了本书的编写，还重新开发并普及了很多之前编写的数据清洗 R 软件包。为什么要这样做呢？其中一个原因是，我们发现了一些很好的方法，能够普及和扩展我们的软件和方法；另一个原因就是，我们希望将最近出现的“tidyverse”接口风格与 R 功能联系起来。

本书包含的内容

本书包含一系列精选的主题，我们认为这些主题对于开发数据清洗(也称为数据编辑)系统非常有用。主题范围非常广泛，与计算机科学、数字方法、技术标准、统计以及数据建模和编程等相关的主题，全部涵盖其中。

本书涵盖“技术数据清洗”方面的主题，包括数字、文本和日期类型的转换和解释。同时对与这些数据类型相关的技术标准也做了较为详细的介绍。在对象的数据内容方面，相关主题包括数据验证(数据检查)、错误定位、各种错误校正方法以及缺失值插补方法。

对于本书中讨论的理论知识，为了便于用户理解，我们会尽可能地提供可执行的 R 代码进行举例说明。此外，我们还提供了相应的练习，希望可以指导读者进一步加

强对软件和对应方法的理解。

广泛的主题既反映出这一课题涵盖范围之广，同时也体现了作者广博的专业知识。当然，还有很多主题并未在书中进行介绍，其中，最重要的主题可能要算清洗时间序列对象和离群值检测。

本书面向的读者

本书的读者应该对数学和统计学有基本的了解，同时还应该具备一定的编程经验。我们假定读者已经了解期望值、方差、基础微积分和线性代数方面的知识。如果具备一定的 R 语言知识，那么对理解本书中的内容会有很大的帮助，因为本书就是使用 R 语言进行说明介绍的。不过，为了便于读者理解和参考，我们还是利用一章内容简要介绍了相关的基础知识。

致谢

本书最终能够顺利出版，离不开很多人的辛勤工作。在这里，我们要感谢荷兰统计局的同事们，他们在百忙之中抽出时间与我们就数据验证、插补和错误定位进行了卓有成效的讨论，为我们提供了很多真知灼见。本书中的部分章节参考了合著者所发表的论文和报告。我们要感谢 Jeroen Pannekoek、Sander Scholtus 和 Jacco Daalmans 的帮助，没有你们的密切合作，就没有本书的成功出版。此外，R 核心团队、软件包开发人员也为我们提供了非常大的帮助，当然，还有 R 社区的大力支持，在此，一并表示感谢。

最后，还要感谢我们的家人，感谢他们的关爱与支持。

2017 年 6 月

Mark 和 Edwin

关于本书配套网站

可以通过以下网址访问本书的配套网站：

www.data-cleaning.org

在本书的配套网站中，可以找到很多极具价值的参考资料，从而更好地学习本书的相关内容，其中包括一些补充资料。

目　录

第1章

数据清洗

1.1 统计价值链

之所以要进行数据清洗，目的在于提升数据质量，使其能够可靠地用于生成统计模型或统计报表。创建部分统计输出必须达到的质量要求由下面这个简单的成本效益问题决定：统计输出什么时候适合使用，为使数据达到这一质量水平需要付出多大的努力？

要想准确回答这个问题，一种非常有用的方法就是按照价值链进行数据分析。粗略地说，价值链由能够逐步增加产品价值的一系列活动组成。在过去 20 年中，统计价值链已经成为“官方统计”社区的常用术语，不过，似乎并未对此形成一个通用的定义。大致来讲，统计价值链通过定义一些有意义的中间数据产品构造而成，针对这些中间数据产品，使用一组精选的质量属性对其进行描述[1]。构造统计价值链的方法有很多，但对于这些作者来说，图 1.1 中所示的内容已经被证明是非常通用的，并且有助于围绕统计生产过程组织整理各种想法。

图 1.1 中的架构有一个显著的特点，那就是它很自然地将通常分类为“数据清洗”的活动引入统计生产过程。从左侧开始，首先是原始数据。必须对其进行处理，使之满足(足够的)技术标准，从而可以作为一致性检查、数据校正和插补过程的输入。实现这一目标后，数据便可以被认为是有效的，能够(足以)用于生成统计参数。之后，仍要对这些数据进行格式化处理，以准备生成相应的输出。

应该认识到，尽管此架构能够很好地组织数据分析活动，但实际上，此过程基本上不可能是线性的。更常见的过程是清洗数据、创建一些聚合、发现问题，然后返回。统计价值链的目的更多的是简要了解各个活动发生的位置(例如，通过将它们放在单独的脚本中)，而不在于规定实际工作流程的线性顺序。实际上，在统计价值链的后续阶段，工作流程会循环多次，直到生成质量足够理想的输出。在接下来的内容中，我们将对每个阶段进行较为详细的讨论。

---

1 最早引入这一定义的似乎是Willeboordse(2000)。

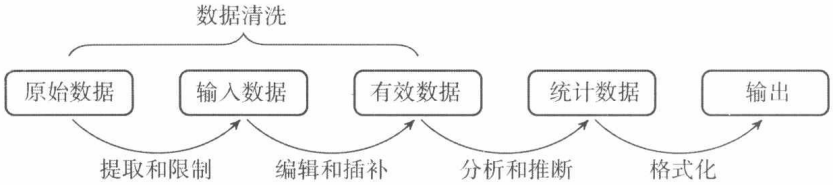

图 1.1 统计价值链的一部分，从原始数据到统计产品，显示 5 个不同的统计价值级别

1.1.1 原始数据

所谓原始数据，指的是最初提供给分析人员的数据。当然，根据数据源的不同，此类数据的状态可能会有非常大的差异。不管是哪种情况，我们都假定一个前提，那就是负责分析的人员对数据的收集方式没有任何影响，或者影响微乎其微，可以忽略。第一步包括使数据易于访问和理解。为了在完成最初的处理步骤后使数据变得更精确，我们要求数据中的每个值都由它们所代表的实际对象(人员、公司等)来标识。对于每个值，已知它所代表的变量(年龄、收入等)，并且值以对应的技术格式(数字和字符串)进行存储。

根据技术上的原始数据格式，实现所需的技术格式所必需的活动通常包括文件转换、字符串规范化(例如编码转换)以及数值的标准化和转换。针对已知统计对象的主干布局注册表连接数据(可能使用能够处理键不精确匹配的过程)也被认为是此类过程的一部分。这些过程将在后面的第 3～5 章中进行更为详细的介绍。

1.1.2 输入数据

输入数据指的是每个值都以正确的类型存储，并以它所代表的变量和所属的统计实体标识的数据。在很多情况下，此类数据集可以通过表格格式来表示，行表示实体，列表示变量。在 R 社区中，这被称为整洁数据(Wickham，2014b)。在这里，我们将格式置于开放状态。很多数据都可以按照树或图表的形式表示(例如，网页和 XML 结构)。只要所有元素都能轻松地标识并且具有正确的格式，那么此数据就可以作为输入数据。

当数据集达到输入数据的级别以后，必须处理缺失值、非置信值以及非置信值组合。这一过程通常称为数据编辑和插补。它与前面各个步骤的不同之处在于，它重点关注领域知识的数据一致性。此类领域知识通常可以表示为一组规则，例如 age >= 0、mean(profit) > 0 或 if (age < 15) has_job = FALSE。我们将在本书后面的部分(第 6～8 章)详细介绍如何定义、应用和维护此类规则，以便可以实现数据清洗的自动化，从而能够以可重复的方式执行。此外，在第 7 章中，我们将介绍一些方法，使得可以在一条记录中拾取最小数量的字段，通过对这些字段进行更改或插补，能够使所有规则都得到满足。

在后面的第 9 章中，我们将正式介绍如何利用知识规则安全地自动执行数据修改，而第 10 章将介绍如何进行缺失值插补。

1.1.3　有效数据

如果数据能够如实地表示它们所代表的变量和对象，它们就是有效的。如果能够确保数据满足通过一组验证规则的形式表达的领域知识，便可以可重复的方式执行此操作。通常情况下，还会在此基础上补充某些形式的专家审阅。例如，根据各种可视化表示形式进行审阅，或者由领域专家对聚合值进行审阅。

数据被认为有效以后，便可以通过已知的建模和推理技术生成统计数据。这些技术可能需要考虑一些过程，例如，在特定插补过程之后估计方差时，具体取决于前面所采用的数据清洗过程。

1.1.4　统计数据

统计数据只是对应输出变量的估计值。通常情况下，这些数据是简单的聚合值(总计和平均值)。但是，从原则上来说，它们可以由更为复杂的参数(例如，回归模型系数或训练好的机器学习模型)组成。

1.1.5　输出

输出是分析过程的终点。接收统计数据并使其做好传播准备，即可创建输出。这可能涉及技术格式化，例如，使数字可以通过(Web) API 或布局格式化获取(例如，通过准备报告或可视化表示形式)。对于技术格式化，技术验证可能再次成为必不可少的步骤，例如，通过针对某些(JSON 或 XML)架构检查输出格式。通常情况下，一个分析人员的输出往往是另一个分析人员的原始数据。

1.2　本书使用的表示法和约定

本书讨论的主题与数学中的各种子域、逻辑、统计学、计算机科学和编程有关。由于涵盖的领域如此之广，因此很难对各种不同的变量类型和概念使用一种统一的表示法。不过，我们尽量在整本书中使用一种统一的表示法。

基础数学和逻辑

我们遵循传统的表示法，并使用$\mathbb{N}$、$\mathbb{Z}$和$\mathbb{R}$分别表示自然数、整数和实数。符号$\vee$、$\wedge$、$\neg$分别表示逻辑或、逻辑与和逻辑非。在讲到逻辑的情况下，我们使用$\oplus$来表示“异或”。有时候，区分定义和等式是非常有用的。在这种情况下，定义将使用$\equiv$来表示。

线性代数

向量使用粗体的小写字母表示，通常为 $\boldsymbol{x}$、$\boldsymbol{y}$ 等。除非另有说明，否则本书所说

的向量一般都是指列向量。符号 **1** 和 **0** 分别表示系数全部为 1 或 0 的向量。矩阵使用粗体的大写字母表示，通常为 $\boldsymbol{A}$、$\boldsymbol{B}$ 等。单位矩阵使用 表示。转置通过上标 T 表示，而矩阵相乘通过并置表示，例如，$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}$。向量的标准欧几里得范数通过 $\|\boldsymbol{x}\|$ 表示，如果使用了另一个 L_k 范数，则使用下标表示。例如，$\|\boldsymbol{x}\|_1$ 表示 $\boldsymbol{x}$ 的 L_1 范数。在线性代数中，⊗和 ⊕ 分别表示直积(张量积)和直和。

概率和统计

随机变量使用大写字母表示，通常为 X、Y 等。概率和概率密度都使用 $P(X)$表示。期望值、方差和协方差分别表示为 E(X)、V(X)和 cov(X, Y)。估计值通过 ∧ 符号表示，例如，$\hat{E}(X)$表示 X 的估计期望值。

代码和变量

R 代码使用固定字体宽度的代码段表示。输出内容以两个注释符号为前缀，如下所示：

```
age <- sample(100,25,replace=TRUE)
mean(age)
## [1] 52.44
```

有时区分代码中的变量与逻辑概念是非常有用的。在这种情况下，代码中的变量将表示为 age，而逻辑概念将表示为 *age*。

第2章

R语言简介

接下来，我们将简要介绍 R 语言的一些核心功能。除了安装 R 语言程序以外，我们建议同时安装适用于 R 语言的一种集成开发环境(Integrated Development Environment，IDE)。选择恰当的 IDE 不仅可以为 R 语言编程提供良好的界面，其帮助系统还可以帮助组织项目、代码和数据。

为了充分理解并合理应用本书中的内容，建议亲自运行提供的代码示例，也可以进行一些自定义调整并对生成的结果进行解释。

2.1 命令行中的R语言

在启动 R 程序或适用于 R 语言的 IDE 之后，即可访问一个交互式控制台或命令行界面。第一个应用是替换袖珍计算器。可以键入一个算式，R 程序会返回计算结果(以[1]为前缀)，如下所示：

```
1 + 1
## [1] 2
```

要开始使用 R 语言，可以先试着键入下面的语句。当然，可以使用不同的数字或运算。R 语言支持所有常用的数学函数。

```
1 + 1
3^2
sin(pi/2)
(1 + 4) * 3
exp(1)
sqrt(16)
```

如果想要重复使用特定的结果或值，可以使用<-运算符将它们存储起来。

```
x <- 10
y <- 20
```

R 现在已经记住 10 和 20 这两个值，并将它们命名为 x 和 y。实际上，x 和 y 现在已经正式成为 R 对象。R 语言非常灵活，除了上面介绍的方法以外，还可以通过多种