格致方法·定量研究系列　吴晓刚　主编

缺失数据

[美] 保罗·D.埃里森（Paul D. Allison）著
林毓玲 译

- 革新研究理念
- 丰富研究工具
- 最权威、最前沿的定量研究方法指南

29

格致出版社　上海人民出版社

格致方法 · 定量研究系列　吴晓刚　主编

缺失数据

[美]保罗 · D.埃里森(Paul D.Allison) 著
林毓玲 译

SAGE Publications ,Inc.

格致出版社　上海人民出版社

图书在版编目(CIP)数据

缺失数据/(美)保罗·D.埃里森著;林毓玲译.
—上海:格致出版社:上海人民出版社,2018.6
(格致方法.定量研究系列)
ISBN 978-7-5432-2867-2

Ⅰ.①缺… Ⅱ.①保… ②林… Ⅲ.①统计数据-数据处理-研究 Ⅳ.①C819

中国版本图书馆CIP数据核字(2018)第089168号

责任编辑 裴乾坤

格致方法·定量研究系列
缺失数据
[美]保罗·D.埃里森 著
林毓玲 译

出　　版　格致出版社
　　上海人民出版社
　　(200001　上海福建中路193号)
发　　行　上海人民出版社发行中心
印　　刷　浙江临安曙光印务有限公司
开　　本　920×1168　1/32
印　　张　5.25
字　　数　103,000
版　　次　2018年6月第1版
印　　次　2018年6月第1次印刷
ISBN 978-7-5432-2867-2/C·198
定　　价　30.00元

出版说明

由香港科技大学社会科学部吴晓刚教授主编的“格致方法·定量研究系列”丛书，精选了世界著名的SAGE出版社定量社会科学研究丛书，翻译成中文，起初集结成八册，于2011年出版。这套丛书自出版以来，受到广大读者特别是年轻一代社会科学工作者的热烈欢迎。为了给广大读者提供更多的方便和选择，该丛书经过修订和校正，于2012年以单行本的形式再次出版发行，共37本。我们衷心感谢广大读者的支持和建议。

随着与SAGE出版社合作的进一步深化，我们又从丛书中精选了三十多个品种，译成中文，以飨读者。丛书新增品种涵盖了更多的定量研究方法。我们希望本丛书单行本的继续出版能为推动国内社会科学定量研究的教学和研究作出一点贡献。

总 序

2003 年，我赴港工作，在香港科技大学社会科学部教授研究生的两门核心定量方法课程。香港科技大学社会科学部自创建以来，非常重视社会科学研究方法论的训练。我开设的第一门课“社会科学里的统计学”（Statistics for Social Science）为所有研究型硕士生和博士生的必修课，而第二门课“社会科学中的定量分析”为博士生的必修课（事实上，大部分硕士生在修完第一门课后都会继续选修第二门课）。我在讲授这两门课的时候，根据社会科学研究生的数理基础比较薄弱的特点，尽量避免复杂的数学公式推导，而用具体的例子，结合语言和图形，帮助学生理解统计的基本概念和模型。课程的重点放在如何应用定量分析模型研究社会实际问题上，即社会研究者主要为定量统计方法的“消费者”而非“生产者”。作为“消费者”，学完这些课程后，我们一方面能够读懂、欣赏和评价别人在同行评议的刊物上发表的定量研究的文章；另一方面，也能在自己的研究中运用这些成熟的方法论技术。

上述两门课的内容，尽管在线性回归模型的内容上有少

量重复，但各有侧重。“社会科学里的统计学”从介绍最基本的社会研究方法论和统计学原理开始，到多元线性回归模型结束，内容涵盖了描述性统计的基本方法、统计推论的原理、假设检验、列联表分析、方差和协方差分析、简单线性回归模型、多元线性回归模型，以及线性回归模型的假设和模型诊断。“社会科学中的定量分析”则介绍在经典线性回归模型的假设不成立的情况下的一些模型和方法，将重点放在因变量为定类数据的分析模型上，包括两分类的 logistic 回归模型、多分类 logistic 回归模型、定序 logistic 回归模型、条件 logistic 回归模型、多维列联表的对数线性和对数乘积模型、有关删节数据的模型、纵贯数据的分析模型，包括追踪研究和事件史的分析方法。这些模型在社会科学研究中有着更加广泛的应用。

修读过这些课程的香港科技大学的研究生，一直鼓励和支持我将两门课的讲稿结集出版，并帮助我将原来的英文课程讲稿译成了中文。但是，由于种种原因，这两本书拖了多年还没有完成。世界著名的出版社 SAGE 的“定量社会科学研究”丛书闻名遐迩，每本书都写得通俗易懂，与我的教学理念是相通的。当格致出版社向我提出从这套丛书中精选一批翻译，以飨中文读者时，我非常支持这个想法，因为这从某种程度上弥补了我的教科书未能出版的遗憾。

翻译是一件吃力不讨好的事。不但要有对中英文两种语言的精准把握能力，还要有对实质内容有较深的理解能力，而这套丛书涵盖的又恰恰是社会科学中技术性非常强的内容，只有语言能力是远远不能胜任的。在短短的一年时间里，我们组织了来自中国内地及香港、台湾地区的二十几位

研究生参与了这项工程，他们当时大部分是香港科技大学的硕士和博士研究生，受过严格的社会科学统计方法的训练，也有来自美国等地对定量研究感兴趣的博士研究生。他们是香港科技大学社会科学部博士研究生蒋勤、李骏、盛智明、叶华、张卓妮、郑冰岛，硕士研究生贺光烨、李兰、林毓玲、肖东亮、辛济云、於嘉、余珊珊，应用社会经济研究中心研究员李俊秀；香港大学教育学院博士研究生洪岩璧；北京大学社会学系博士研究生李丁、赵亮员；中国人民大学人口学系讲师巫锡炜；中国台湾“中央”研究院社会学所助理研究员林宗弘；南京师范大学心理学系副教授陈陈；美国北卡罗来纳大学教堂山分校社会学系博士候选人姜念涛；美国加州大学洛杉矶分校社会学系博士研究生宋曦；哈佛大学社会学系博士研究生郭茂灿和周韵。

参与这项工作的许多译者目前都已经毕业，大多成为中国内地以及香港、台湾等地区高校和研究机构定量社会科学方法教学和研究的骨干。不少译者反映，翻译工作本身也是他们学习相关定量方法的有效途径。鉴于此，当格致出版社和 SAGE 出版社决定在“格致方法·定量研究系列”丛书中推出另外一批新品种时，香港科技大学社会科学部的研究生仍然是主要力量。特别值得一提的是，香港科技大学应用社会经济研究中心与上海大学社会学院自 2012 年夏季开始，在上海（夏季）和广州南沙（冬季）联合举办《应用社会科学研究方法研修班》，至今已经成功举办三届。研修课程设计体现“化整为零、循序渐进、中文教学、学以致用”的方针，吸引了一大批有志于从事定量社会科学研究的博士生和青年学者。他们中的不少人也参与了翻译和校对的工作。他们在

繁忙的学习和研究之余，历经近两年的时间，完成了三十多本新书的翻译任务，使得“格致方法·定量研究系列”丛书更加丰富和完善。他们是：东南大学社会学系副教授洪岩璧，香港科技大学社会科学部博士研究生贺光烨、李忠路、王佳、王彦蓉、许多多，硕士研究生范新光、缪佳、武玲蔚、臧晓露、曾东林，原硕士研究生李兰，密歇根大学社会学系博士研究生王骁，纽约大学社会学系博士研究生温芳琪，牛津大学社会学系研究生周穆之，上海大学社会学院博士研究生陈伟等。

陈伟、范新光、贺光烨、洪岩璧、李忠路、缪佳、王佳、武玲蔚、许多多、曾东林、周穆之，以及香港科技大学社会科学部硕士研究生陈佳莹，上海大学社会学院硕士研究生梁海祥还协助主编做了大量的审校工作。格致出版社编辑高璇不遗余力地推动本丛书的继续出版，并且在这个过程中表现出极大的耐心和高度的专业精神。对他们付出的劳动，我在此致以诚挚的谢意。当然，每本书因本身内容和译者的行文风格有所差异，校对未免挂一漏万，术语的标准译法方面还有很大的改进空间。我们欢迎广大读者提出建设性的批评和建议，以便再版时修订。

我们希望本丛书的持续出版，能为进一步提升国内社会科学定量教学和研究水平作出一点贡献。

吴晓刚

于香港九龙清水湾

序

在经验社会科学研究中缺失数据的问题很普遍，大多非实验性研究所报告的统计结果都立足于较小的样本数，有时比初始选择的个案数目还要小。在一些变量上相对多缺失观察值会降低有效的 N。假设有个意见调查，多变量分析中有效样本只有原来的一半，这种情况在现实中并不少见。假设商学院的玛丽·罗斯(Mary Rose)教授在一个消费者态度及行为调查中检验一个 $N=1000$ 受访者的概率样本。她使用一般计算机选项成列删除(也就是任一受访者有缺失任一模型变量即被排除)，对支出估计一个合理设定的多元回归模型。结果实际可得的个案降至 $N=499$。这就产生了严重的问题。这 499 个受访者是否仍"代表"了总体？要拒绝零假设，样本是否太小？为了保持样本数，是否应该尝试成对删除？抑或，有其他新的方法值得考虑？这些问题及其他问题都在保罗·埃里森这本杰出的专题著作中讨论到。

"观察值是随机缺失的"，这是根据留下的个案以面对处理数据缺失时的通常论点。但这个假设是隐含的。假若观察值"完全随机缺失"，这表示没有任何变量，不论是因变

量(Y)或自变量(X),其缺失分数都不与该变量自身的值相关。例如,上述的支出变量,对于支出多者其未回答率应不比支出少者的未回答率高。对于其他模型变量,假设在相同的条件下,则499个次样本将可代表一次科学的抽取,允许有效的推论。而且,它允许回归估计值是不偏及一致的。无问题派的研究者可能喜欢这种完全随机缺失的(Missing Completely at Random, MCAR)随机性,但这需要有很强的假设来支持。

较为实际一点的假设为观察值是"随机缺失的"(Missing at Random, MAR)。假设在控制了其他变量后,如果Y的值不能预设缺失分数的位置,则Y变量缺失数据为随机的。所以在上述的举例中,职业地位(X)可能与支出的缺失数据相关,高地位的受访者更可能低报支出。一旦X在右手边,那么Y的观察值将会是随机的。在MAR情况下,如埃里森所言,缺失数据产生机制是可忽略的。虽然他也论及不可忽略的缺失数据机制的困难细节,但他这本专题著作着重于在MAR条件下,以改良估计处理的方法。

如果数据是MAR,则估计的质量很大程度地取决于系统性误差的位置。令人鼓舞的是,当相关缺失数据仅限于自变量时,则成列删除仍能产生不偏的估计值。例如,在例子中,职业地位X缺失数据可能与另一个自变量年龄(Z)相关;例如:没有报告年龄的可能年纪较大且地位较高。在年龄较大与报告支出没有相关的条件下,则没有误差。事实上,正如埃里森巧妙论证的,在一些MAR情况下,标准成列删除选项比传统缺失数据修正方法(成对删除,虚拟变量调整或平均值替换)表现更好。

处理缺失数据问题的新策略占用了本专题论著的大部分篇幅。在缺失数据的条件下回顾最大似然估计，即 ML 估计，他以一个仔细筛选的美国大专院校毕业率的数据为例，解释了插补法的 EM 算法。后几章超越了 ML 方法，解释多重插补方法，并讨论了不可忽略的缺失数据。这本书是最新的处理缺失数据的精心杰作，几乎所有的统计书籍都很少涉及这个主题。保罗·埃里森也睿智地提醒我们，缺失数据最佳的解决方法是“没有任何最佳解决方法”。但如果你也有这个问题且在寻求补救方法，那么就请阅读本书的内容。

迈克尔·刘易斯-贝克

格致方法·定量研究系列

1. 社会统计的数学基础
2. 理解回归假设
3. 虚拟变量回归
4. 多元回归中的交互作用
5. 回归诊断简介
6. 现代稳健回归方法
7. 固定效应回归模型
8. 用面板数据做因果分析
9. 多层次模型
10. 分位数回归模型
11. 空间回归模型
12. 删截、选择性样本及截断数据的回归模型
13. 应用 logistic 回归分析（第二版）
14. logit 与 probit：次序模型和多类别模型
15. 定序因变量的 logistic 回归模型
16. 对数线性模型
17. 流动表分析
18. 关联模型
19. 中介作用分析
20. 因子分析：统计方法与应用问题
21. 非递归因果模型
22. 评估不平等
23. 分析复杂调查数据（第二版）
24. 分析重复调查数据
25. 世代分析（第二版）
26. 纵贯研究（第二版）
27. 多元时间序列模型
28. 潜变量增长曲线模型
29. 缺失数据
30. 社会网络分析（第二版）
31. 广义线性模型导论
32. 基于行动者的模型
33. 基于布尔代数的比较法导论
34. 微分方程：一种建模方法
35. 模糊集合理论在社会科学中的应用
36. 图解代数：用系统方法进行数学建模
37. 项目功能差异（第二版）
38. Logistic 回归入门
39. 解释概率模型：Logit、Probit 以及其他广义线性模型
40. 抽样调查方法简介
41. 计算机辅助访问
42. 协方差结构模型：LISREL 导论
43. 非参数回归：平滑散点图
44. 广义线性模型：一种统一的方法
45. Logistic 回归中的交互效应
46. 应用回归导论
47. 档案数据处理：生活经历研究
48. 创新扩散模型
49. 数据分析概论
50. 最大似然估计法：逻辑与实践
51. 指数随机图模型导论
52. 对数线性模型的关联图和多重图
53. 非递归模型：内生性、互反关系与反馈环路
54. 潜类别尺度分析
55. 合并时间序列分析
56. 自助法：一种统计推断的非参数估计法
57. 评分加总量表构建导论
58. 分析制图与地理数据库
59. 应用人口学概论：数据来源与估计技术
60. 多元广义线性模型
61. 时间序列分析：回归技术（第二版）
62. 事件史和生存分析（第二版）
63. 样条回归模型
64. 定序题项回答理论：莫坎量表分析
65. LISREL 方法：多元回归中的交互作用
66. 蒙特卡罗模拟
67. 潜类别分析

目 录

第1章

导　论