



格致方法·定量研究系列 吴晓刚 主编

分析复杂调查数据 (第二版)

[美] 李殷嵩 (Eun Sul Lee) 著
罗纳德·N.福索佛 (Ronald N. Forthofer)

张卓妮 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

格致出版社  上海人民出版社

23

格致方法·定量研究系列 吴晓刚 主编

分析复杂调查数据 (第二版)

[美] 李殷嵩(Eun Sul Lee) 著
罗纳德·N. 福索佛(Ronald N. Forthofer) 著
张卓妮 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

分析复杂调查数据:第二版/(美)李殷嵩,(美)
罗纳德·N.福索佛著;张卓妮译.—上海:格致出版
社:上海人民出版社,2018.4
(格致方法·定量研究系列)
ISBN 978-7-5432-2847-4

I. ①分… II. ①李… ②罗… ③张… III. ①抽样调
查统计-研究 IV. ①C811

中国版本图书馆 CIP 数据核字(2018)第 044092 号

责任编辑 张苗凤

格致方法·定量研究系列

分析复杂调查数据(第二版)

[美]李殷嵩 罗纳德·N.福索佛 著
张卓妮 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 4 月第 1 版

印 次 2018 年 4 月第 1 次印刷

ISBN 978-7-5432-2847-4/C·192

定 价 32.00 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

当乔治·盖洛普(George Gallup)正确预测到富兰克林·罗斯福(Franklin D. Roosevelt)成为 1936 年美国总统选举赢家时,公众民意调查进入了科学抽样的时代,那时使用的方法是配额抽样,一种代表目标总体的非概率抽样方法。但是,同样的方法却错误地预测托马斯·杜威(Thomas Dewey)将成为 1948 年的总统,但实际上哈利·杜鲁门(Harry S. Truman)才是最终的赢家。这种方法之所以失败,是因为配额抽样是非概率的,而且盖洛普的配额抽样框基于 1940 年的人口普查,忽略了第二次世界大战期间的城市移民。

今天的调查抽样方法和早些时候相比已有了很大进步,现在我们依赖于复杂的概率抽样设计。一个关键的特征就是分层:目标总体被分成一系列不同阶层的子总体,阶层内的样本量由抽样者控制且通常与该阶层的人口规模成比例。另一个特征是整群和多阶段抽样:先抽取组群——即在不同阶层选到的一系列不同等级的集群,到最后才在这些“最后阶层”的集群内部选取个体成员。比如,美国综合社会调查使用的就是分层多阶段集群抽样设计。卡尔顿(Kalton,

1983)为调查抽样提供了很好的入门介绍。

当调查设计具有这个复杂的性质时,数据的统计分析就不再是一个简单的运行回归(或任何其他模型)分析的事情了。现在的调查都伴随着抽样权重以支持正确的统计推断,大部分关于统计分析的教材,通过假定简单随机抽样而没有处理抽样权重的问题,而这个被忽略的抽样权重可能对统计推断产生重要的影响。在最近二三十年间,数据分析的统计方法也已取得了巨大进展。这些很可能就是迈克尔·S.刘易斯-贝克(Michael S. Lewis-Beck)选择为《分析复杂调查数据》(*Analyzing Complex Survey Data*)出第二版的原因。

李(Lee)和福索佛(Forthofer)这本书的第二版为我们在调查抽样设计和调查数据分析的联合方面提供了最新的情况。作者在本书开头先回顾了调查抽样设计常见的类型,并通过解释什么是抽样权重及其如何产生和调整,揭开了抽样权重的神秘面纱。然后他们详细讨论了方差估计以及考虑抽样权重后的复杂截面调查数据的初级和多变量分析等主要问题。他们重点关注了基于设计的方法,这种方法直接在分析中涉及抽样设计(虽然他们也讨论了基于模型的视角,该视角在某些分析中能扩充基于设计的方法),他们还以流行的软件为例阐释了这种方法的使用。读者将会在以样本为基础作出统计推断的实践中发现本书的巨大益处。

廖福挺

Analyzing Complex Survey Data, 2nd edition

Copyright © 2006 by SAGE Publications, Inc.

All rights reserved. No part of this book may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission in writing from the publisher.

This simplified Chinese edition for the People's Republic of China is published by arrangement with SAGE Publications, Inc. © SAGE Publications, Inc. & TRUTH & WISDOM PRESS 2018.

本书版权归 SAGE Publications 所有。由 SAGE Publications 授权翻译出版。
上海市版权局著作权合同登记号：图字 09-2009-551

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

序	1
第 1 章 概论	1
第 2 章 抽样设计和调查数据	7
第 1 节 抽样方法的种类	9
第 2 节 调查数据的属性	14
第 3 节 调查数据的不同看法*	18
第 3 章 分析调查数据的复杂性	21
第 1 节 调整不同的代表性:权重	23
第 2 节 用事后分层的方法加权	27
第 3 节 在追踪调查中调整权重	31
第 4 节 评估精确度的得失:设计效应	34
第 5 节 调查数据分析中抽样权重的使用	37
第 4 章 方差估计的策略	41
第 1 节 复合抽样:一种通用的方法	43
第 2 节 对称重复抽样	48
第 3 节 “折叠式”重复抽样	55

第 4 节	自主抽样法	61
第 5 节	泰勒级数法(线性化)	63
第 5 章	调查数据分析的准备	67
第 1 节	调查分析的数据要求	69
第 2 节	预备性分析的重要性	71
第 3 节	方差估计方法的选择	75
第 4 节	可用的计算资源	77
第 5 节	创建复合权重	81
第 6 节	寻找合适的调查数据分析的模型	84
第 6 章	调查数据分析的操作	87
第 1 节	预备性分析的策略	89
第 2 节	描述性分析	92
第 3 节	线性回归分析	100
第 4 节	列联表分析	106
第 5 节	logistic 回归分析	112
第 6 节	其他 logistic 回归模型	118
第 7 节	基于设计和基于模型的分析	125
第 7 章	总结	131
注释		135
参考文献		137
译名对照表		146

第 **1** 章

概 论

调查分析似乎通常是在所有的样本观察值都以同等的概率被独立选中的情况下实行的。如果在数据收集中使用的是简单随机抽样(SRS),这种分析就是对的,但实际上样本选择比 SRS 复杂得多。某些样本观察值可能比其他观察值以更高的概率被选中,某些之所以被包括在样本中,是因为他们属于某个特定组别的成员(比如家庭),而不是被独立选取的。我们可以在调查数据的分析中简单地忽略这些与 SRS 相悖的事实吗?使用调查数据分析统计书中的标准技术是否合适?或者是否有特别的方法和计算机程序更适合复杂调查数据的分析?这些问题将在随后的章节中进行论述。

今天典型的社会调查反映了统计理论和关于社会现象的知识两者间的结合,过去 70 年间从许多不同的调查中获得的经验塑造了它的发展。社会调查的进行是为了满足用以讨论社会、政治和公共卫生议题的信息需要。为了满足这种信息需要,在政府内部和外部都设立了调查机构。但是,在早期提供这些信息的尝试中,调查小组们最关心的是实地调查中的操作问题,如抽样框的建立、员工培训/监督,以及成本的降低等,理论上的抽样议题获得的只是次级重视(Ste-

phan, 1948)。随着这些实际问题得到解决,现代抽样实践方法的发展远远超越了SRS。复杂抽样设计已经走在了前面,随之而来的是一系列分析问题。

因为早期调查通常需要的只是描述性统计,所以很少有人对分析问题感兴趣。更近一些时期,社会和政策科学家对分析研究的需要已经增加,那些并不参与数据收集过程的研究人员,正利用可用的社会调查数据对各种不同的现实议题进行分析。这个传统上被称为二手资料分析(Kendall & Lazarsfeld, 1950)。研究者通常并没有对复杂抽样设计的发展给予应有的关注,并假设这些设计对将要使用的分析程序没有什么影响。

二手资料分析中统计技术使用的增加以及近年来对数线性模型、logistic 回归和其他多变量技术的使用(Aldrich & Nelson, 1984; Goodman, 1972; Swafford, 1980),并没有把设计和分析更紧密地结合起来。这些技术断定数据收集使用的是简单有放回随机抽样(SRSWR),但这个假设在应用了观测单位的分层和整群方法以及不等概率选择法的社会调查中几乎无法得到满足。因此,利用SRSWR假设的社会调查分析可能导致有偏差和误导性的结果。比如,基柯尔特和内森(Kiecolt & Nathan, 1985)在他们关于二手资料分析的书中承认了这个问题,但是他们几乎没有就如何把抽样权重和其他设计特征融入分析之中提供什么建议。最近一篇关于公共健康和流行病学的文献研究表明:对基于设计的调查分析方法的使用正在逐渐增加,但依然处于较低水平(Levy & Stolte, 2000)。

任何对抽样进行了限制且那些限制超越了SRSWR所做