格致方法 · 定量研究系列　吴晓刚　主编

对数线性模型的关联图和多重图

[美] 哈里 · J.哈米斯（Harry J. Khamis）著
王彦蓉 译　曾东林 校

- 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

52

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

对数线性模型的关联图和多重图

[美] 哈里 · J.哈米斯（Harry J. Khamis）　著
王彦蓉　译　曾东林　校

SAGE Publications, Inc.

格致出版社　上海人民出版社

图书在版编目(CIP)数据

对数线性模型的关联图和多重图/(美)哈里·J.哈米斯著;王彦蓉译.—上海:格致出版社:上海人民出版社,2016.12
(格致方法·定量研究系列)
ISBN 978-7-5432-2674-6

Ⅰ.①对… Ⅱ.①哈… ②王… Ⅲ.①对数-线性模型-研究 Ⅳ.①O122.6②O212

中国版本图书馆 CIP 数据核字(2016)第 248972 号

责任编辑 张苗凤

格致方法·定量研究系列

对数线性模型的关联图和多重图

[美]哈里·J.哈米斯 著
王彦蓉 译 曾东林 校

出 版 世纪出版股份有限公司 格致出版社
世纪出版集团 上海人民出版社
(200001 上海福建中路 193 号 www.ewen.co)
编辑部热线 021-63914988
市场部热线 021-63914081
www.hibooks.cn
发 行 上海世纪出版股份有限公司发行中心

印 刷 浙江临安曙光印务有限公司
开 本 920×1168 1/32
印 张 5.75
字 数 114,000
版 次 2016 年 12 月第 1 版
印 次 2016 年 12 月第 1 次印刷

ISBN 978-7-5432-2674-6/C·159 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

此为试读，需要完整PDF请访问：www.ertongbook.com

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

对数线性模型(LLM)在利奥·古德曼(Leo Goodman)等人的努力下,于20世纪60年代末期至70年代初得以发展并流行,而毕晓普(Bishop)、费恩伯格(Fienberg)和霍兰(Holland)于1975年写的《离散多元分析》一书堪称该模型发展和推广中的一个里程碑,它对多向列联表(multi-way contingency tables)的分析产生了革命性的作用。包括LLM在内的分类数据分析的课程目前是社会科学专业研究生教育的一个重要内容。哈里·哈米斯(Harry Khamis)的专著《对数线性模型的关联图和多重图》展示了如何动用数学资源中的图论来理解复杂LLM表明的关联结构。

哈米斯回顾了二向与多向列联表的关系模式,以及这些表的LLM。在介绍了图论中的一些关键概念之后,他紧接着将这些思想应用到LLM的两个图形典型中:关联图和生成多重图。在更为直观并且广为人知的关联表中,顶点(点)代表列联表中的变量,边(线)对应于模型中的关联项。相反,在多重图中,顶点代表一个分级LLM中的高阶关联项(生成因子),边对应由成对生成因子共享的变量。

利用经过精挑细选的丰富示例以及清晰的解释，哈米斯展示了 LLM 的这两个图形典型如何说明模型的结构，揭示了它们的一些特征，如条件性独立（比如说，变量 A 和变量 B 在变量 C 的分类中是独立的），可分解性（比如说，模型中的 A、B 和 C 联合单元格的概率是 AB 边际概率和 C 边际概率的乘积），以及可压缩性（比如说，A 和 B 的边际关联与 A 和 B 在 C 的分类中的关联是一致的）。总的来讲，这一专著中描述的图论有助于构想和解释多向列表中的复杂 LLM。对于想要深入理解 LLM 结构的学生和在研究中应用这些模型的研究者，这一专著都是有价值的。

编辑注：这一专著是在前一任丛书编辑廖福挺的指导下开始着手的。

约翰·福克斯

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：研究“人生”
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

此为试读，需要完整PDF请访问：www.ertongbook.com

目 录

序 1

第 1 章 介绍 1

第 2 章 关联结构 9

第 1 节 离散变量的统计独立 11

第 2 节 比值比:二向表 12

第 3 节 比值比:三维列表 16

第 4 节 模型拟合:三维表 20

第 5 节 多向表 22

第 3 章 对数线性模型回顾 23

第 1 节 二向列联表 25

第 2 节 三维列联表 29

第 3 节 三维表 LLM 之间的关系 33

第 4 节 LLM 和列联表属性 36
第 5 节 多向表 41

第 4 章 对数线性模型的关联图 43
第 1 节 基本图论法则 45
第 2 节 三向表的关联图 52
第 3 节 多维表的关联图 56
第 4 节 可分解的 LLM 61
第 5 节 总结 64

第 5 章 可压缩性条件和关联图 67
第 1 节 三维列联表的可压缩性 69
第 2 节 压缩性定理和关联图 73
第 3 节 结论 78

第 6 章 生成多重图 83
第 1 节 构建多重图 85
第 2 节 三维表格的多重图 86
第 3 节 多维表的多重图 88
第 4 节 最大生成树 92
第 5 节 可分解性 98
第 6 节 分解可分解 LLM 的联合概率 103

第 7 节 可分解 LLM 的基本条件独立 107

第 7 章 不可分解对数线性模型的基本条件独立 115

第 1 节 边割集 117

第 2 节 不可分解 LLM 的 FCI 122

第 3 节 使用多重图的压缩条件 127

第 4 节 FCI：总结 128

第 8 章 结论及附加实例 129

第 1 节 关联图和多重图的比较 131

第 2 节 附加实例 136

第 3 节 最后要点 155

数据集 157

参考文献 159

译名对照表 163

第1章

介　绍

对连续结果变量的复杂统计方法研究始于19世纪80年代弗朗西斯·高尔顿(Francis Galton),以及其后费舍尔(R. A. Fisher)和卡尔·皮尔逊(Karl Pearson)等人的著作,然而直到20世纪60年代,分类响应模型才有重大突破,即便卡尔·皮尔逊和尤尔(G. Udny Yule)早在1900年就已经发表了极有影响力的文章。在过去的50年间,处理分类响应模型的方法激增。关于分析多维列联表的方法,我们可以列一个简要的历史性大纲,包括如下内容:

- 尤尔在20世纪早期将比值比作为属性间关联的一个关键测度;60多年后,比值比已经成为对数线性模型(loglinear models, LLMs)的基石(参考第3章对LLM的回顾)。
- M. S. Bartlett(1935)首次论述了对三向列联表的分析,包括一阶和二阶的交互效应分析。
- 自20世纪60年代以来,关于多向列联表的LLM研究开始兴盛。M. W. Birch(1963)解决了三维列联表LLM最大似然估计的问题。Grizzle、Starmer & Koch(1969)展示了分类数据模型的加权最小二乘法。Bishop, Fienberg & Holland(1975)写了第一本多向列联表分析的权威教材。从此以后,很多教材开始分析分类响应模型。例如C. C.

Clogg, D. R. Cox, J. N. Darroch, I. J. Good, L. A. Goodman, S. J. Haberman, P. E. Lazarsfeld, R. L. Plackett, N. Wermuth和其他很多人对分类数据的研究也做出了广泛贡献。

● Darroch, Lauritzen & Speed(1980)的文章具有里程碑的意义,它介绍了用数学的图形来展示 LLMs 并用其属性分析和解释列联表的数据。自从 1980 年以来,研究人员,如 D. R. Cox, D. Edwards, S. L. Lauritzen 和 N. Wermuth,对所谓的一阶交互图,也被 Agresti(2002)叫做关联图,做了大量的工作。数学的图论和 LLM 理论的联姻为分析和解释分类变量间复杂的结构性关系提供了革命性的方法。Edwards (1995)和 Lauritzen(1996)提供了这一方法的两本好书。

在心理学、社会学、政治科学、教育学和其他研究领域的两大发展要求更加完善的工具来分析和解释庞大复杂的 LLM。

1. 首先,数据搜集和数据录入技术在近几年使研究者得以获得非常大的数据,尤其得益于功能强大的计算机和电子扫描仪器。在包含分类变量的研究中,这导致了大样本量的庞大多向列联表。事实上,研究者只要通过互联网就可以获得大量的数据。很多数据由国家机关提供并公开使用。一个很好的例子是瑞典统计局(Statistics Sweden),他们可以提供以研究为导向的数据(参看 www.scb.se)。又如美国劳工统计局(http://bls.gov/nls/nlsy79.htm)也是如此。

2. 第二,在列联中"稀疏"的概念已经发生了变化。在早年的列联表分析中,统计学家建议只有在每个单元格的期望频数不小于 5 的情况下才能使用卡方分析(Fisher, 1925)。

此为试读,需要完整PDF请访问: www.ertongbook.com