格致方法·定量研究系列　吴晓刚　主编

关联模型

[美] 黄善国 著
肖东亮 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

18

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

关联模型

[美] 黄善国 著
肖东亮 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

关联模型/(美)黄善国著;肖东亮译.—上海:
格致出版社:上海人民出版社,2012
(格致方法·定量研究系列)
ISBN 978-7-5432-2127-7

Ⅰ.①关… Ⅱ.①黄… ②肖… Ⅲ.①定量社会学-
研究方法 Ⅳ.①C91-03

中国版本图书馆 CIP 数据核字(2012)第 132266 号

责任编辑 高 璇

格致方法·定量研究系列
关联模型
[美]黄善国 著
肖东亮 译

出 版 世纪出版集团 www.ewen.cc
格致出版社 www.hibooks.cn
上海人民出版社
(200001 上海福建中路193号24层)
编辑部热线 021-63914988
市场部热线 021-63914081
格致出版
发 行 世纪出版集团发行中心
印 刷 浙江临安曙光印务有限公司
开 本 920×1168 毫米 1/32
印 张 6
字 数 116,000
版 次 2012 年 7 月第 1 版
印 次 2012 年 7 月第 1 次印刷
ISBN 978-7-5432-2127-7/C·77
定 价 18.00 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中的35种，翻译成中文，集结成八册，于2011年出版。这八册书分别是：《线性回归分析基础》、《高级回归分析》、《广义线性模型》、《纵贯数据分析》、《因果关系模型》、《社会科学中的数理基础及应用》、《数据分析方法五种》和《列表数据分析》。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欢迎，他们针对丛书的内容和翻译都提出了很多中肯的建议。我们对此表示衷心的感谢。

基于读者的热烈反馈，同时也为了向广大读者提供更多的方便和选择，我们将该丛书以单行本的形式再次出版发行。在此过程中，主编和译者对已出版的书做了必要的修订和校正，还新增加了两个品种。此外，曾东林、许多多、范新光、李忠路协助主编参加了校订。今后我们将继续与SAGE出版社合作，陆续推出新的品种。我们希望本丛书单行本的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

往事如烟，光阴如梭。转眼间，出国已然十年有余。1996年赴美留学，最初选择的主攻方向是比较历史社会学，研究的兴趣是中国的制度变迁问题。以我以前在国内所受的学术训练，基本是看不上定量研究的。一方面，我们倾向于研究大问题，不喜欢纠缠于细枝末节。国内一位老师的话给我的印象很深，大致是说：如果你看到一堵墙就要倒了，还用得着纠缠于那堵墙的倾斜角度究竟是几度吗？所以，很多研究都是大而化之，只要说得通即可。另一方面，国内（十年前）的统计教学，总的来说与社会研究中的实际问题是相脱节的。结果是，很多原先对定量研究感兴趣的学生在学完统计之后，依旧无从下手，逐渐失去了对定量研究的兴趣。

我所就读的美国加州大学洛杉矶分校社会学系，在定量研究方面有着系统的博士训练课程。不论研究兴趣是定量还是定性的，所有的研究生第一年的头两个学期必须修两门中级统计课，最后一个学期的系列课程则是简单介绍线性回归以外的其他统计方法，是选修课。希望进一步学习定量研

究方法的可以在第二年修读另外一个三学期的系列课程，其中头两门课叫“调查数据分析”，第三门叫“研究设计”。除此以外，还有如“定类数据分析”、“人口学方法与技术”、“事件史分析”、“多层线性模型”等专门课程供学生选修。该学校的统计系、心理系、教育系、经济系也有一批蜚声国际的学者，提供不同的、更加专业化的课程供学生选修。2001 年完成博士学业之后，我又受安德鲁·梅隆基金会资助，在世界定量社会科学研究的重镇密歇根大学从事两年的博士后研究，其间旁听谢宇教授为博士生讲授的统计课程，并参与该校社会研究院(Institute for Social Research)定量社会研究方法项目的一些讨论会，受益良多。

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”(Statistics for Social Science)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，这两本书拖了四年多还没有完成。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。中山大学马骏教授向格致出版社何元龙社长推荐了这套书，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港台地区的二十几位研究生参与了这项工程，他们目前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。

关于每一位译者的学术背景，书中相关部分都有简单的介绍。尽管每本书因本身内容和译者的行文风格有所差异，校对也未免挂一漏万，术语的标准译法方面还有很大的改进空间，但所有的参与者都做了最大的努力，在繁忙的学习和研究之余，在不到一年的时间内，完成了三十五本书、超过百万字的翻译任务。李骏、叶华、张卓妮、贺光烨、宋曦、於嘉、郑冰岛和林宗弘除了承担自己的翻译任务之外，还在初稿校对方面付出了大量的劳动。香港科技大学霍英东南沙研究院的工作人员曾东林，协助我通读了全稿，在此

我也致以诚挚的谢意。有些作者，如香港科技大学黄善国教授、美国约翰·霍普金斯大学郝令昕教授，也参与了审校工作。

我们希望本丛书的出版，能为建设国内社会科学定量研究的扎实学风作出一点贡献。

吴晓刚

于香港九龙清水湾

序

社会和公共舆论调查通常询问的问题的答案多为类别。这些答案的类别可以完全是离散的或者定序的。让我们考虑一个由古德曼·克洛格和其他学者分析过的经典数据表——曼哈顿中城(Midtown Manhattan)精神健康和父母社会经济地位(SES)数据。

父母的社会经济地位	精神健康状况			
	良好	一般症状	中度症状	重度症状
A(高)	64	94	58	46
	(48.5)	(95.0)	(57.1)	(61.4)
B	57	94	54	40
	(45.3)	(88.8)	(53.4)	(57.4)
C	57	105	65	60
	(53.1)	(104.1)	(62.6)	(67.3)
D	72	141	77	94
	(71.0)	(139.3)	(83.7)	(90.0)
E	36	97	54	78
	(49.0)	(96.1)	(57.8)	(62.1)
F(低)	21	71	54	71
	[illegible]	[illegible]	[illegible]	[illegible]

尽管社会科学家的专业出身不同，但他们分析以上数据的一个通常做法是检验两个变量或者多向表(multiway tables)中多于两个变量的情况，在这里即精神健康状况和父母的社会经济地位是否相关。而在统计学用语中，我们会关注一个关于独立性的虚无假设能否被拒绝。一些基本的统计学课程通常会讲解皮尔逊卡方检验和似然比检验的应用。简单地观察频率表(上表括号里的数字为期望频率)将无法得出结论。如果我们使用 F_{ij} 指代表中第 i 行和第 j 列的观察频率 f_{ij} 的期望值，那么，F_{ij} 在独立模型(父母的 SES 和精神状态)中则可表示为：

$$F_{ij} = \frac{f_{i+} f_{+j}}{f_{++}}$$

这里，f_{i+} 表示第 i 行的列总和，f_{+j} 表示第 j 列的行总和，而 f_{++} 则表示整个表格的总和。为了检验独立性假设，我们计算皮尔逊 χ^2 系数和似然比系数 L^2：

$$\chi^2 = \sum_i \sum_j \frac{(f_{ij} - F_{ij})^2}{F_{ij}} \text{ 和 } L^2 = 2\sum_i \sum_j \ln\left(\frac{f_{ij}}{F_{ij}}\right)$$

应用这些公式，我们获得了一个皮尔逊 χ^2 值 45.985 和似然比 L^2 值 47.418。在自由度为 15(行数减去 1 的差乘以列数减去 1 的差)的情况下，我们在所有常用的显著性水平上拒绝关于独立性的虚无假设，并且结论是精神健康状态和父母的 SES 不是彼此独立的，或者换一种说法，它们以某种方式相互关联。

然而，尽管我们知道它们以某种方式相互关联，但是我们并没有充分利用已有的信息来进一步探索它们相互关联的形式。作为对数线性模型中的一种，关联模型正是为这一

目的而建立的。之前我们进行的检验相当于对数线性模型中的主效应估计。使用单一性关联模型(也被称为"线性相关关联模型"):

$$\ln F_{ij} = \lambda + \lambda_i^A + \lambda_j^B + \beta U_i V_j$$

这里的前三项代表主效应对数线性模型,其附加项则表示两个变量各自的观测数值组之间的关联程度,这样,我们就获得了自由度为 14 的皮尔逊 χ^2 值 9.732 和似然比 L^2 值 9.895。因此,引入 β 参数,我们仅仅损失一个自由度便可以保留"存在线性相关卡方值"这个虚无假设。至此,读者一定对关联模型的检验能力有了深刻的印象。

作为不仅在自己的研究领域中应用关联模型,并且为关联模型的发展作出贡献的关键学者之一,黄善国撰写了这本"社会科学定量研究方法"丛书非常需要的著作。他将带领我们走进一段旅程,领略更多不同形式的关联模型,例如,行效应模型、列效应模型、行列效应模型、行列乘法效应模型和其他多种不同的形式,包括为涉及多种因素的多向表设计的模型。

在上述例子中,我们引入了一个统计值表示任意分配的两组数值的关联情况,但这个统计值并非固定不变,它可以由模型估计得来。为了我们的研究工作而学习这种模型和其他不同的、令人兴奋而且有用的关联模型,只要进入书中描述的关联模型的奇境即可。

廖福挺

目 录

第1章

简　介

许多社会科学的数据都会很自然地以交互表格的形式被组织起来。比如，在社会学方面，教育水平和职业的关系存在性别差异和/或者种族差异、社会网络中的友谊模式、择偶中的跨国因素和/或者时间变化；在地理学方面，存在城市邻居关系随时间变化的特点和省际或区域内移民流动情况的时间变化；在经济学方面，存在全球经济系统中进出口贸易的动态变化趋势；在政治科学方面，存在阶级地位、政党认同以及选举之间随时间变化的关系；最后，在心理学方面，存在关于刺激识别和刺激类化的实验数据。尽管我们所关心的是探索它们的系统性联系，但是有时应用适当的统计工具来诠释和理解其中关系的意义及其复杂性却存在困难，对于研究新手而言更是如此。

过去，学者尝试过各种各样的方法来计算表格形式下行与列的关联情况。例如，如果假设研究中涉及的变量在本质上是排序的，那么便可以使用一些测量关联程序的定序测量方法。然而，这些关于关联程度的测量方法不仅无法提供优比(odds ratio)(或者它们的对数)的自然变换功能，而且它们同样无法避免受到边缘分布的影响(Clogg & Shihadeh, 1994:19)。因此，拥有相同优比的表格由于不同的边缘分布

将产生不同的关联测量值(Agresti, 2002; Bishop, Fienberg & Holland, 1975; Fienberg, 1980; Rudas, 1997:第2章)。更重要的是,这些测量关联程序的单一方法,尤其当行和列的类别数量比较多的时候,经常无法提供交互表中关于关联程度的全面描述。

除了使用以上这些描述型方法之外,另一个替代性的策略是发展源于实证研究并可进行正式检验的关联程度测量方法。对数线性模型的发展(Bishop et al., 1975; Fienberg, 1980; Haberman, 1978)为我们提供了理解多个定类或定序变量之间关系的重要途径。但是在多维交互表中,当每个变量的类别数量增多而出现许多需要诠释的参数时(Goodman, 2007,在定类数据分析中使用对数线性模型的非技术但富有洞见的介绍),我们便需要解释那些非结构性交互项系数。经过古德曼的努力及其后的克洛格和邓肯及其他合作者的发展,我们现在拥有了一系列非常丰富及适合此类分析的统计模型,尤其是关联模型。虽然关联模型被社会分层研究学者广泛使用,尤其在社会流动和择偶的研究领域(Breen, 2004; Grusky & Hauser, 1984; Hout, 1988; Smits, Ultee & Lammers, 1998、2000; Wong, 1990、1992、2003b; Xie, 1992; Yamaguchi, 1987),但是这种统计技术仍未能普及至其他社会科学领域。我想部分原因是,那些最重要的统计和应用文献分散在不同的杂志里,而且它们的讲解主要集中在简单的二维交互表中。除了克洛格等人的努力外,至今还没有人系统地将各种关联模型整合到一个一以贯之的框架中。

呈现在您面前的这本《关联模型》尝试填补这个重要的