



格致方法·定量研究系列

吴晓刚 主编

多元广义线性模型

[美] 理查德·F.哈斯 (Richard F. Haase) 著
臧晓露 译 王佳 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

格致出版社 上海人民出版社

60

格致方法·定量研究系列 吴晓刚 主编

多元广义线性模型

[美] 理查德·F.哈斯 (Richard F. Haase) 著

臧晓露 译 王 佳 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

多元广义线性模型/(美)理查德·F.哈斯
(Richard F. Haase)著;臧晓露译.—上海:格致出
版社:上海人民出版社,2017.7
(格致方法·定量研究系列)
ISBN 978-7-5432-2756-9

I. ①多… II. ①理… ②臧… III. ①线性模型-研
究 IV. ①0212

中国版本图书馆 CIP 数据核字(2017)第 093558 号

责任编辑 顾悦

格致方法·定量研究系列

多元广义线性模型

[美]理查德·F.哈斯 著
臧晓露 译 王佳 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 上海商务联西印刷有限公司
开本 920×1168 1/32
印张 9.25
字数 186,000
版次 2017 年 7 月第 1 版
印次 2017 年 7 月第 1 次印刷

ISBN 978-7-5432-2756-9/C·177

定价:48.00 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

目 录

第 1 章	一元广义线性模型的简介与回顾	1
第 1 节	一元线性模型分析回顾	4
第 2 节	识别一元回归模型	9
第 3 节	模型的参数估计	12
第 4 节	证实最小二乘估计的有效性所需要的假设	18
第 5 节	分解平方和以及定义拟合优度的测量	20
第 6 节	全模型、限制模型以及半偏相关系数的平方	22
第 7 节	回归系数和判定系数的假设检验	24
第 8 节	广义线性假设检验	26
第 9 节	模型整体假设 $\beta_1 = \beta_2 = \beta_3 = 0$ 和 $\rho_{Y \cdot X_1 X_2 X_3}^2$ 的检验	28
第 10 节	用广义线性检验方法评估 X_1 , X_2 和 X_3 的单独贡献	30
第 11 节	用广义线性检验检验更为复杂的假设	32
第 12 节	从一元到多元广义线性模型的一般化	36

第 2 章	多元广义线性模型的结构识别	37
第 1 节	模型的数学识别	40
第 2 节	定义预测变量和标准变量的实质作用	43
第 3 节	示例数据和模型识别	46
第 3 章	广义多元线性模型的参数估计	55
第 1 节	例 1: 性格特征与成功的工作申请	61
第 2 节	用标准得分的形式估计多元线性模型中的参数	67
第 3 节	例 2: 多氯联苯——心血管疾病的风险因素: 认知功能数据	71
第 4 节	对多元线性模型分析的电脑程序的一个说明	77
第 5 节	本章小结与回顾	78
第 4 章	多元 SSCP 分解、关联强度的测量和 检验统计量	79
第 1 节	在多元广义线性模型中 SSCP 的分解	82
第 2 节	例 1: 性格与工作申请	90
第 3 节	例 2: PCB 数据	93
第 4 节	SSCP 矩阵的进一步分解: 全模型、限制模型 以及定义 Q_H	96
第 5 节	一些关联强度的多元测度的概念定义	101
第 6 节	一个不对称的 R^2 的多元测度 ——Hooper 迹相关系数平方	104

第 7 节	例子:性格数据和 PCB 数据中 Hooper's $\bar{r}^2$	107
第 8 节	一元和多元 R^2 之间的关系和它们的检验 统计量	109
第 9 节	Pillai 迹 V 和相应的关联强度测度 R_V^2	113
第 10 节	Wilks' Λ 及其关联强度测度 R_Λ^2	118
第 11 节	Hotelling 迹 T 及其关联强度测度 R_T^2	122
第 12 节	Roy 最大特征根及其关联强度度量 $r_{C_{\max}}^2$	125
第 13 节	通过一元回归模型建立 Pillai 迹 V 和 Wilks' Λ	127
 第 5 章	 多元广义线性模型中的假设检验	 137
第 1 节	多元广义线性检验	140
第 2 节	多元检验统计量及其近似 F 检验	143
第 3 节	对 Pillai 迹 V 的近似 F 检验	145
第 4 节	Wilks' Λ 的近似 F 检验	148
第 5 节	Hotelling 迹 T 的近似 F 检验	151
第 6 节	Roy 最大特征根 θ 的近似 F 检验	153
第 7 节	对一个或一组预测变量的广义线性检验	155
第 8 节	对一个预测变量的多元假设检验:性格数据	156
第 9 节	一个预测变量的多元假设检验:PCB 数据	163
第 10 节	一组预测变量的多元假设检验和其他 复杂假设	167
第 11 节	检验其他的复杂的多元假设	170
第 12 节	适用于所有多元线性模型分析的假设	173

第 6 章	编码设计矩阵和方差模型的多元分析	175
第 1 节	变量和向量的差异	178
第 2 节	用编码向量来表示一个分类变量	180
第 3 节	通过广义线性检验来检验 MANOVA 假设	188
第 4 节	分解 SSCP 矩阵和 MANOVA 里的假设检验	192
第 5 节	身材估计数据的单项 MANOVA	193
第 6 节	更高阶的 MANOVA 设计:对身材估计数据的一个 2×3 阶 MANOVA	206
第 7 节	关于 MANOVA 分析假设的备注	221
第 7 章	多元线性模型的特征值求解:典型相关系数和多元检验统计量	223
第 1 节	典型相关系数的概念定义	226
第 2 节	2×2 相关系数矩阵的特征值	229
第 3 节	$\mathbf{R}_{(2 \times 2)}$ 的特征向量	235
第 4 节	$\mathbf{R}_{YY}^{-1} \mathbf{R}_{YX} \mathbf{R}_{XX}^{-1} \mathbf{R}_{XY}$ 的特征值	240
第 5 节	特征值、典型相关系数的平方和四个多元检验统计量	243
第 6 节	$\mathbf{R}_{YY}^{-1} \mathbf{R}_{YX} \mathbf{R}_{XX}^{-1} \mathbf{R}_{XY}$ 的典型相关系数的平方的特征向量	250
第 7 节	检验典型相关系数和典型系数上的进一步假设	255
参考文献		270
译名对照表		275

第 **1** 章

一元广义线性模型的
简介与回顾

在社会科学、行为科学以及自然科学中,很少有数据分析技术比多元回归分析更为重要。在各个领域,包括人类学(Cardoso & Garcia, 2009)、经济学(Card, Dobkin & Maestas, 2009)、政治学(Baek, 2009)、社会学(Arthur, Van Buren & Del Campo, 2009),以及心理学的各个分支(Ellis, MacDonald, Lincoln & Cabral, 2008; Pekrun, Elliot & Maier, 2009)中,都可见多元回归分析的示范性应用。

在以上每个领域中,研究者的目的是研究变量之间的关系。用数据拟合回归模型可以使分析者用一个或多个预测变量来解释一个因变量内的变化。广义线性模型是回归模型的一个延伸,用来处理定量和定性的变量分析。众所周知,多元回归分析是一个涵盖所有线性模型的数据分析系统(Cohen, 1968),包括了处理连续变量的分析模型(经典回归分析)、处理分类变量的模型(经典方差分析),以及同时处理连续和分类预测变量的模型。^①这些模型共同定义了广义线性模型。回归模型在处理许多不同类型的预测变量方面是

① 一些作者更喜欢用“定量”或“定性”这类术语去描述连续或分类预测变量。在本书中,我们用“连续”这个术语来指代潜在度量是连续的或离散的变量,用“分类”这个术语来指代名义性的分组结构。这种结构除了指示分组外,没有其他实际意义。

非常灵活的,包括连续变量的交互作用、分类变量的交互作用,以及连续和分类变量的交互作用。这些组合提供了在更广泛的范围内进行分析的可能性,这解释了为什么这项技术在所有科学领域内,包括从人类学到动物学,都有如此广泛的运用。

本书的目的是介绍广义线性模型的多元形式,以及展示它的几种应用。多元模型的特点是具有不止一个因变量,通过拟合一个模型来同时分析这些变量。很多多元线性模型分析的概念和统计学基础是对一元回归分析的直接推广,我们将在本章中简单回顾一元回归分析,来为之后的章节做铺垫。在第2章中,我们介绍了会一直用到的示例样本数据,并对广义线性模型(GLM)分析中的第一步——模型识别——进行讨论。第3、4、5章的内容涉及了模型参数的估计、模型拟合优度的评价及相应的多元检验统计量,以及对模型的假设检验。第6章介绍了多元方差分析的线性模型解决方法。第7章用对典型相关分析的介绍来结束本书。典型相关分析涵盖了之前章节介绍过的所有线性模型。本书最重要的目的是从一个整合的视角把所有不同的技术用一个模型框架展现出来。

第 1 节 | 一元线性模型分析回顾

线性模型的主要目标是评价变量间的关系,用某个被识别的模型和误差来解释一个因变量的变化:

$$\text{响应变量} = \text{模型} + \text{误差}$$

对于一元的例子来讲,回归模型就是只包括一个标准变量、响应变量、因变量或结果变量的模型。^①对于只有一个预测变量的简单模型,数学上,一元回归模型可以用下面的回归函数来表示:

$$Y = \beta_0 + \beta_1 X_1 + \epsilon \quad [1.1]$$

对于有多个预测变量的更为复杂的模型,表达式可以写作:^②

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_q X_q + \epsilon \quad [1.2]$$

在公式 1.1 和公式 1.2 中, Y 表示一系列响应变量向量,需要用回归系数 $\beta_0, \beta_1, \cdots, \beta_q$ 和解释变量 $X_1, X_2, \cdots, X_q$,

① 本书中,我们把因变量、标准变量、响应变量和结果变量这几个术语作为同义词使用,来描述模型中的 Y 变量。模型中的 X 变量将会用预测变量、解释变量或自变量这几个同义词表述。这些术语在回归分析的文献中反复出现。在有控制条件的实验设计中,一些作者更喜欢使用“因变量”这个术语。

② 我们不用下角标来指定从第 1 个到第 n 个观察值的序列顺序,以此来识别响应变量和解释变量 Y 或 X 。本书中,所有的模型都是基于全部 n 个观察值而做出的,加法和乘法的运算指示都假定是基于全部 n 个参与者而进行的。

以及代表所有其他引起 Y 值波动的变化来源的干扰项或误差项(系统的或随机的)的加权线性组合来解释。解释变量 $X_j, j=1, 2, \dots, q$, 既可以是连续的, 也可以是分类的。^①许多当前的教科书都强调用这种一元的线性模型来进行回归分析和方差分析(例子见 Cohen, Cohen, West & Aiken, 2003; Myers & Well, 2003)。

虽然在本章节中我们简单回顾了一元回归(线性)模型分析的基本思路, 但我们的目的是为处理连续和分类变量的多变量多元回归分析(广义线性模型分析)做铺垫。多元模型可以被定义为一元对应模型的推广。一元回归模型的 Y 值由单列向量来定义, 多元模型的识别则在很大程度上同时包括不止一个因变量。解释变量集 $X_1, X_2, \dots, X_q$, 在一元和多元模型中可以是相同的; 只有 Y 变量的数目、回归系数的向量列数和误差项 ϵ 的数目会有差别。

随着模型变得更为复杂, 用矩阵代数形式来表达模型和它们的应用会更加方便。虽然在本书的讨论中我们用基本矩阵去表示线性模型, 但我们并不会对这一话题进行全面的讨论。德雷珀和史密斯(Draper & Smith, 1998: 第4章)花了一个章节来介绍相关的许多细节; 南布迪里(Namboodiri, 1984)或肖特(Schott, 1997)对此著有专门的教科书做介绍。

公式 1.1 和公式 1.2 中的单变量(Y)多元回归模型可以方便地用矩阵表示为:^②

① 我们将会在后面的部分更详尽地介绍分类变量的编码体系。

② 我们用斜体字来代表标量(例如, X, Y, Z, β, ϵ), 黑体小写字母来指代行或列向量(例如, $\mathbf{a}, \mathbf{b}, \mathbf{y}, \mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\epsilon}$), 黑体大写字母来表示矩阵(例如, $\mathbf{X}, \mathbf{Y}, \mathbf{B}, \mathbf{E}, \boldsymbol{\Gamma}$)。如果一个列向量或行向量是特意用矩阵来表示的, 它的向量状态将会用矩阵的阶数来表示, 例如, $(n \times 1)$ 或 $(1 \times p)$ 。