格致方法·定量研究系列　吴晓刚　主编

广义线性模型导论

[美] 乔治·H. 邓特曼（George H. Dunteman）
[加] 何满镐（Moon-Ho R. Ho）　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

31

格致出版社　上海人民出版社

广义线性模型导论

[美] 乔治·H.邓特曼(George H.Dunteman)
[加] 何满镐(Moon-Ho R.Ho) 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

广义线性模型导论/(美)邓特曼(Dunteman, G. H.),(加)何满镐著;林毓玲译. —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 978-7-5432-2161-1

Ⅰ. ①广… Ⅱ. ①邓… ②何… ③林… Ⅲ. ①线性模型-研究 Ⅳ. ①O212

中国版本图书馆 CIP 数据核字(2012)第 214038 号

责任编辑 高 璇

格致方法·定量研究系列
广义线性模型导论
[美]乔治·H. 邓特曼 著
[加]何满镐
林毓玲 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
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 4.25
字 数 83,000
版 次 2012 年 10 月第 1 版
印 次 2012 年 10 月第 1 次印刷
ISBN 978-7-5432-2161-1/C·87
定 价 15.00 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

本书编辑过程并不寻常：作者及编者都有所改变。我的前一任编辑，迈克尔·刘易斯·贝克，非常睿智地看到《广义线性模型导论》的价值。在 2004 年初从主编岗位退下之前，他看遍了计划书及先前的手稿版本。令人难过的是，乔治·H. 邓特曼在完成他所认为的终稿后就过世了。进一步的修改由何满镐接手，他非常勇敢地接受挑战，并对原稿作出许多重要的修正。

社会科学家所分析的结果变量可以是连续的或是离散的。在已出版的丛书中，有许多书目涉及需要处理一个连续的因变量（及一些重要假设）的模型，经典线性回归为这类模型的代表。除此之外也涉及因变量是非连续的，通常统计模型的对象为事件发生几率，但也可能是频率或是对数频率。在过去 20 年中，许多型态的 logit、probit（及对数线性）模型已经成为社会科学家众多分析方法中的标准，并且该丛书中也有多本涉及这些主题。

连续结果变量及离散的因变量这两种模型间的关系，在广义线性模型的架构下变得清晰。在社会科学中，研究者对

在方程右方以 x 和 β 线性组合所表示的可线性化的自变量比较熟悉。然而，位于这两种模型左方的因变量 y 可以是多种形式的，包括 metric、二元的、序列的、multinomial 和计数的。再者，两种模型中的随机结果 y 可能服从正态、二项、泊松、gamma 分布，且所有这些分布都属于指数家族分布。一旦我们对于 y 的随机分布作出服从指数分布的适当假设后，剩余的任务便是指明随机变量的期望以及 x 和 β 线性组合间的关系。将期望的随机结果变量对应到 x 和 β 的线性组合，是广义线性模型的一部分。

本书的根本目标是：对于熟悉经典线性回归的普通社会科学研究者，要如何从线性回归模型推广到非连续自变量的其他模型，而不失两种模型间的共同根基及相似性？本书两位作者陪着读者走访这一过程，并在沿途中启蒙不识此道者，这也对丛书提供了有益的增补。

廖福挺

目 录

第1章

广义线性模型

广义线性模型，顾名思义，为经典线性回归模型的普遍化。经典线性回归模型假设因变量为一组自变量的线性方程，且因变量为连续且正态分布的，有固定的方差。自变量则可以是连续的、类别的或两者的组合。多元回归分析、方差分析及协方差分析皆为线性模型的经典例子。它们皆可被写成：

$$y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \varepsilon$$

其中 y 是连续性因变量，X_j 是自变量，ε 为假设正态分布的误差。因变量由两部分组成：系统性(systematic)成分 $\beta_0 + \sum_{j=1}^{p} \beta_j X_j$；误差成分 ε。系统性成分即在任意组给定的 X_j 的值之下，y 的期望值 $E(y)$，即：

$$E(y \mid X_1, \cdots, X_p) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$$

它是给定 X_j 值的条件平均数(conditional mean)。回归分析的目的就是寻找一组以拟合优度来衡量具有高度解释力的自变项，即我们能凭借自变量的线性组合来解释 y 大部分的变异。如果回归参数 β_j 很大，当 X_j 的值从一观察值变化至另一观察值时，y 的期望值或 y 的条件平均数也将有很

大的变异。如果在条件平均数或预测值中的变异比在 ε 中的变异更大，我们则能利用一个有用的模型，在给定自变量取值的条件下预测 y 的取值，以及了解不同自变量在解释因变量 y 的变异时的相对重要性。图 1.1 给出了一个简单的线性回归模型($\beta_0 = 1$，$\beta_1 = 1.5$)。我们通过观察对象的一个随机样本，收集 y 的测量值以及 X_1，X_2，…，X_p 来估计回归参数 β_j。就观察目的而言，我们的观察对象通常是人，但在其他应用中，观察对象可以是任何事物，如树、牛，甚至河流。如果我们以 i 标示人，以 j 标示变量，则可以通过最小化误差的平方和来估算 β_j。

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j X_{ij})^2$$

在此，小标 i 被用以强调自变量的值随着个体的不同而变化的事实。此回归参数的估计方法通常被称做普通最小二乘法。

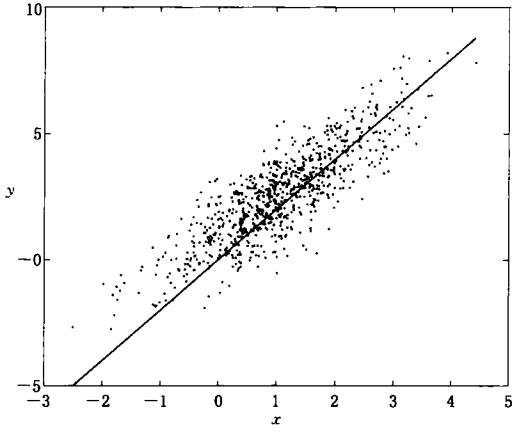

图 1.1　线性回归模型

这个线性回归模型自19世纪初步发展以来，对社会科学及其他科学特别有用。它很易于公式化，易于理解，并且回归系数易于利用普通最小二乘法估计。因此，它至今仍被广泛应用于各学科中。虽然它假设误差是正态分布的，但当误差接近于正态分布时，它仍是稳健的。

然而，在过去几十年中，人们已经广泛地意识到线性回归模型的局限。它假设因变量为连续或至少是准连续的，如考试成绩、个人特质测量等。而且，它假设该连续变量至少是接近于正态分布，并且其方差并不是其平均数的函数。内尔得和韦德伯恩(Nelder & Wedderburn, 1972)提出了广义线性模型，后来发展为应用于非正态因变量的回归模型。

在许多应用中，因变量是类别的或包含计数的，抑或为连续的但并非正态。一个类别的因变量的例子是二元变量，只有两个离散的值0或1，其中，1代表事件的发生(如从大学中退学)，而0代表事件未发生(如未从大学中退学)。目标是要模型化感兴趣的事件的发生概率。在稍后会提及logistic回归，它是广义线性模型的一种，适合此类型的数据。

一个关于计数的因变量的例子是，一个药物滥用者群体在五年里的药物滥用事件(treatment episodes)。我们将再一次地展示泊松回归(poisson regression)，这是适合此情形的另一种广义线性模型。在这两个例子中，因变量都不是连续的，更不是正态分布的，且0—1二元变量与计数变量都为非负数。然而，在一般回归中连续因变量可以是正值或负值。

一个被广泛应用的非正态连续分布的例子为gamma分布。gamma分布是偏斜的(skewed)，只有正值，且其方差为其平均数的函数。它可以用来模型化一般性的、类别的、只有