格致方法·定量研究系列　吴晓刚　主编

广义线性模型：一种统一的方法

[美] 杰夫·吉尔（Jeff Gill）　著
王彦蓉　译　　许多多　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

44

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

广义线性模型：一种统一的方法

[美] 杰夫 · 吉尔 (Jeff Gill) 著
王彦蓉 译　　许多多 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

广义线性模型:一种统一的方法/(美)吉尔(Gill, J.)著;王彦蓉译.—上海:格致出版社:上海人民出版社,2014

(格致方法·定量研究系列)

ISBN 978-7-5432-2450-6

Ⅰ.①广… Ⅱ.①吉… ②王… Ⅲ.①线性模型-研究 Ⅳ.①O212

中国版本图书馆 CIP 数据核字(2014)第 243554 号

责任编辑 顾 悦

美术编辑 路 静

格致方法·定量研究系列

广义线性模型:一种统一的方法

[美]杰夫·吉尔 著

王彦蓉 译 许多多 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)

格致出版

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 4.75
字 数 92,000
版 次 2015 年 1 月第 1 版
印 次 2015 年 1 月第 1 次印刷

ISBN 978-7-5432-2450-6/C·115 定价:22.00 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

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

用普通最小二乘法(OLS)估计基本的线性模型是对数据进行分析的一个好的起点,但可能不会是一个好的终点。普通最小二乘法建立在高斯-马尔科夫假设基础上,从概念上讲,它是强有效的,但是在实际中却难以满足。一方面,如果因变量是偏态的或者是分类变量,分析者就需要把 OLS 转为 logit 模型来估计。另一方面,分析者可能会诉诸概率模型(probit),并查阅专著来进行比较。又或者,变量可能是删截的或者包含事件计数,因此布林(Breen)所著的《回归模型:删截的、样本抽选的或是截断的数据》和艾利森(Allison)的《事件史分析(第二版)》值得更多的关注。

这些不同于 OLS 的新的尝试很好,但问题是它们是支离破碎的。通过具体的操作,可以对每一条被违背的假设都单独做出处理,但这么做却忽视了关注问题和方法间的有机联系。在这里,吉尔教授在《广义线性模型》中提供了统一的分类方法,涵盖了基本的线性模型、logit 模型和其他概率模型。首先,吉尔教授把常见的概率密度和概率质量函数归在一起,正如指数族一样。然后,发展了指数族分布的最大似

然函数。接下来的部分就是这部专著的核心(第 4 章)——通过兼容离散和有界因变量的连接函数将线性模型一般化。

至于软件,目前大部分权威的软件包都支持广义线性模型以及它迭代加权最小二乘法(IWLS)的算法。为了说明IWLS的应用和对系数的解释,作者提供了大量基于现实世界中社会科学的数据而进行的原始模型的练习。他审视了以下问题:美国各州的死刑、新苏格兰议会的税务投票、加利福尼亚州的标准教育考试、国会委员会条例草案的工作分配、世界铜价。例子的广泛与多样可以使读者对这种方法产生的结果感到满意。

在广义线性模型中,残差和模型拟合等问题与其在基本线性模型中一样重要,尽管在前者中达到可接受的表现基准要更难。尽管这些残差通常不是正态分布的,正态性依然是一项有用的诊断标准。五种不同的残差有待检验:响应残差、皮尔森残差、工作残差、安斯库姆残差和偏差残差。在吉尔教授看来,偏差残差是最有用的。对于模型拟合度,这里也有五个选择:卡方近似值、赤池信息准则(AIC)、施瓦茨准则,图以及概括的偏差统计(这是作者所偏好的)。

在社会科学研究中,奥卡姆剃刀原理有着相当高的价值。OLS的一般性假设很少,可以使研究者在研究的路上走得很远。但是,区间测量和正态性的要求却是这条路上的障碍。吉尔教授曾经谨慎地表明,广义线性模型可以帮助我们移除那些障碍,与此同时,保持简约的原则。

迈克尔·S.刘易斯-贝克

目 录

第1章

介 绍

社会学家采用广泛的数据分析方法来探究和解释各种经验性的现象。这其中相当大一部分的工具是从应用统计学批量引进的。这种方法富有成效,因为社会科学研究者遇到的大多数问题都可以由发展成熟和随时可得的统计方法解决。不幸的是,在这些知识的传播过程中,技术和方法有时会不必要地被当成是独一无二的。这一点尤其适用于那些回归方法:logit 模型和概率模型、截断分布模型、事件计数模型、概率结果模型和基本线性模型,所有这些(以及更多的模型)都是广义线性模型的特殊情况:一种产生模型参数估计的单一方法。

典型的社会科学研究生的方法教育始于学习线性模型(事实上,有时也止于此),接着介绍离散选择模型、生存模型、计数模型及其他。这导致了一种分隔且有限的世界观。同时,它也意味着很多特殊的步骤、模型设定和诊断必须分开学习。相反,在本书中,所有这些方法都可以看成一种广义方法的特殊情况。因此,这本书通过整合看似不同的方法来达到对其他课本的补充。

这本专著解释并说明了一种在社会学里应用回归模型的统一方法。一旦理解了一般框架,就可以很容易地通过结

果变量的结构与其离散性质来选择合适的模型结构。这一过程不仅引导我们对模型的理论基础有一个更好的理解，而且可以增强研究者对于新数据类型的灵活性使用。广义线性模型背后的基本原则是线性模型的系统成分可以被转化，从而形成近似于标准线性模型的研究框架，但它也适应于非正态和非区间的结果变量。高斯-马尔科夫假设作为线性模型理论的基础，要求误差项符合均值为零并且方差恒定的独立分布。如果结果变量不服从正态分布，即使线性模型在轻微偏差的情况下稳健，这些假设也常常不能被满足，而且会使估计的有效性出现严重的错误。广义线性模型采用“连接函数”（定义为数据的系统成分和结果变量之间的关系）使得渐进正态性和方差恒常不再成为必然要求（但是，假设观测值之间的不相关仍然很重要）。这就使得许多模型的产生可以不再受标准线性理论的限制。

为了可以统一貌似多样的概率模型，一般性的方法是首先将普通的概率密度函数和概率质量函数改为一个统一的指数族形式。这有助于发展用于潜在转换线性模型的原理，使之成为更加严密和全面的理论处理方式。由内尔德和韦德伯恩（Nelder & Wedderburn, 1972）阐述的统一处理方法表明，对应用统计工作所得结论的理解可以通过广义理论的进一步发展被大大加强。

这里所强调的是广义线性模型的理论基础，而不是一堆应用。因此，我大部分的精力会花在支撑这个结构的数学统计理论上。会用一些熟悉的分布作为例子，但是理论的强调就意味着读者将需要能够针对自己数据分析的具体应用发展出相应的广义线性模型设定。

第 1 节 | 模型设定

乔治·博克斯(George Box)曾经宣称所有的模型都是错的,只有一些是有用的。这是一种基于对提供的信息进行必要的简化并删截后而发展统计模型的观察。模型设定是一个决定数据的哪些特征是重要的而哪些又是不需要的过程。这一过程的着眼点就在于决定哪些解释变量应该被包括,哪些应该被忽略,在解释变量和因变量之间设定数学和概率性的关系,并且设定关系成立的标准。由模型设定和执行产生的汇总统计数据很有希望为那些未知总体的主要参数提供充足的统计上和通俗意义上的了解。

事实上,模型设定比科学更具有艺术性,因为从些许因素中便可以衍生出数量巨大的潜在具体模型。[1]通常,研究者对模型设定的其中一部分有理论依据,而且在很多领域,模型的囊括是基于传统的。这一过程的基础是在简约和拟和之间进行权衡。设定简约的模型之所以有效是因为它忽视了那些不是很重要的效应。这些模型可以高度概括是因为应用的条件更容易达到(宽口径)。但是,模型变得越简单,在保持其他条件恒定的情况下,由均值 μ 所描述的因变量的表现也会变得越极端,误差项也更有可能包含重要的系统信息。最糟的情况是,这会导致有偏误的估计值。并且,

一旦模型中有任何随机的因素，不论它会不会导致偏差，这个模型从上面提到的博克斯的角度来讲都可以说是错误的。

我们也可以建立一个完全正确的模型，即使它会局限于不能描述数据的潜在结构。这种模型叫做饱和模型或者全模型，它的一组参数等于数据点的个数，每一个都由一个指标函数作出索引。因此每一个参数都是精准无误的，因为它完美地描述了一个观测数据点的位置。但是，这种模型没有进行数据归约，推论的价值也是有限的。[2]饱和模型是非常有用的探索工具，它允许我们以此为基准来检测假设的模型设定(参看 Lindsey, 1997:214—215; Neter, Kutner, Nachtsheim & Wasserman, 1996:586—587)。稍后，我们会看到，在检验模型设定的拟合质量时，饱和模型是产生类似残差的偏差所必不可少的。典型的统计模型不同于饱和模型，因为它们试图缩减观测数据的数据量和复杂程度以总结出少量的汇总统计。这些模型用简洁性来换取确定性，并对潜在的总体值做出推论性的结论。由此估计产生的参数值从表面上看是错误的，但是通过提供不确定性的关联程度，可靠度是可以被检测的。

总体上来讲，模型设定的目的是要产生一组由模型而来的拟合值$\hat{\mathbf{Y}}$，它近似于观测的因变量值 $\mathbf{Y}$。$\hat{\mathbf{Y}}$和 $\mathbf{Y}$ 越接近，我们会感觉到模型越准确地描述了现实。但是，这个目标不是唯一的，我们也不能简单地满足于饱和模型。因此，一个好的模型会在简洁性和拟合度这两个相冲突的目标之间进行权衡。

广义线性模型与常规的线性模型在模型设定的过程方面没有什么大的不同，只是广义线性模型包含了可以适应非