



格致方法·定量研究系列 吴晓刚 主编

解释概率模型： Logit、Probit 以及其他广义线性模型

[美] 廖福挺 (Tim Futing Liao) 著
周穆之 译 陈伟 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

39

格致方法·定量研究系列 吴晓刚 主编

解释概率模型： Logit、Probit 以及 其他广义线性模型

[美] 廖福挺 (Tim Futing Liao) 著
周穆之 译 陈伟 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

解释概率模型:Logit、Probit 以及其他广义线性模型/(美)廖福挺著;周穆之译. —上海:格致出版社,2014

(格致方法·定量研究系列)

ISBN 978-7-5432-2459-9

I. ①解… II. ①廖… ②周… III. ①概率-数学模型-解释 IV. ①0211

中国版本图书馆 CIP 数据核字(2014)第 269881 号

责任编辑 顾悦
美术编辑 路静

格致方法·定量研究系列

解释概率模型:Logit、Probit 以及其他广义线性模型

[美]廖福挺 著
周穆之 译 陈伟 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
字数 96,000
版次 2015 年 1 月第 1 版
印次 2015 年 1 月第 1 次印刷

ISBN 978-7-5432-2459-9/C·118

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

某事件发生的概率是多少？某个自变量与此事件概率之间有什么关系呢？还有些问题，研究人员其实也经常涉及。举一个常见的政治学的例子，我们有一个涉及一些投票者样本的调查数据。关注的因变量(Y)是供投票者选择的两个政党——自由党派(记为“0”)和保守党派(记为“1”)。问题就是，选择保守党派的可能性会因为收入的增加而增加吗(收入自变量 X, 测量单位是以 1 000 美元作为一个单位)? 使用线性模型是一个选择, 用普通最小二乘法(OLS)回归得出的结果如下:

$$Y = -0.02 + 0.01X + e$$

根据对线性概率模型的解释, 斜率的估计(0.01)说明收入上每增加一个单位(1 000 美元), 投票给保守党派的概率平均增加了 1 个百分点。同时, 我们看到对于收入 X 等于 3 万美元的人来说, 预测的投票给保守党派的概率是 0.28。尽管 OLS 的结果很简洁, 但它缺乏某些估计量应该具有的特点。由于因变量的二元特性, 误差项一定是异方差的(het-

eroscedastic),也就意味着估计是无效率的。此外,这种模型也会得出一些概率在(0, 1)之外的无意义的预测。比如,对于没有收入的投票者,预测的概率等于-0.02;对于收入是103000的投票者,预测出的概率是1.01。最后,投票与收入之间的关系也许完全是非线性的,在极端收入的情况下,每个单位变化的影响会变小,也说明 OLS 的斜率估测是有偏误的。

因此,尽管最初从 OLS 得出的线性概率模型看上去很有吸引力,但在解释概率的时候还是应该尽量避免。取而代之的是一个更加适合的技巧,廖教授会对此做出详细的说明。对目前的例子来说,logit 模型是一个更好的选择,因变量在此进行了转化,最大似然的评估因此是:

$$\log \left[\frac{\text{Prob}(y=1)}{1-\text{Prob}(y=1)} \right] = c + dX + v$$

廖教授对概率模型给出了四个基本的解释。他指出在 logit 模型中最有用的解释就是比数和比数比。为了说明这一点,他使用了来自全国儿童调查的数据,创立了一个未成年人性行为的 logit 模型。通过取 logit 系数的指数(例如,以 e 为底,系数作为 e 的指数)他能够得出,比如未成年男性有过性行为的比数高于女性,为 1.9 倍。第二种常用的解释 logit 结果的方法就是基于一系列预定的 X 值来预测某事件的概率。继续他的阐述,我们可以得出,未成年人是白人女性时,有过性行为的概率是 0.146。

logit 模型只是廖教授所涵盖话题的其中一个。他也评估了另外一个可替代的分析方法——probit 模型。probit 的累积概率函数和 logit 模型很相似,所以它们得出的概率通常

都是一样的。然而 logit 的优点是这些预测的概率用普通计算器就能得出。此外,在有许多极端的观测分布时,logit 还是优于 probit 的。

尽管大部分 logit 和 probit 模型集中在二元因变量上,但这并非概率模型的前提。在阐述了序列 logit 和 probit 模型之后,廖教授介绍了如何来解释有序 logit 和 probit 模型。接着他介绍了多类别 logit 模型,也就是因变量是无法排序的(有趣的是,他指出计算上的难度让多类别 probit 十分罕见)。对模型的讨论以用来估测罕见事件概率的泊松回归结束。廖教授提醒我们,这也正是另外一种错误使用 OLS 回归的情况。因为本书的全面性以及它对解释概率模型的主题进行了统一,本书在读者掌握了单个等式的回归方法后值得一读。

迈克尔·S.刘易斯-贝克

目 录

序	1
第 1 章 介绍	1
第 1 节 为什么要用概率模型?	3
第 2 节 为什么需要解释?	4
第 2 章 广义线性模型和对其系数的解释	7
第 1 节 广义线性模型	9
第 2 节 解释参数估计	13
第 3 章 二分的 Logit 和 Probit 模型	19
第 1 节 Logit 模型	22
第 2 节 解释 Logit 模型	24
第 3 节 Probit 模型	35
第 4 节 解释 Probit 模型	37
第 5 节 Logit 还是 Probit 模型呢?	41

第 4 章	序列 Logit 和 Probit 模型	43
第 1 节	模型	45
第 2 节	解释序列 Logit 和 Probit 模型	50
第 5 章	有序 Logit 和 Probit 模型	61
第 1 节	模型	63
第 2 节	解释有序 Logit 和 Probit 模型	68
第 6 章	多类别 Logit 模型	77
第 1 节	模型	79
第 2 节	解释多类别 Logit 模型	82
第 7 章	条件 Logit 模型	93
第 1 节	模型	96
第 2 节	解释条件 Logit 模型	99
第 8 章	泊松回归模型	109
第 1 节	模型	111
第 2 节	解释泊松回归模型	114
第 9 章	总结	123
第 1 节	概括	125

第 2 节 概率模型的重要文献	127
第 3 节 解释概率模型的进一步评论	130
注释	132
参考文献	133
译名对照表	138

第 **1** 章

介 绍

社会科学家用概率模型解释很多关键的问题。对这些模型的解释却常常被忽略而且对许多人来说很困惑。本书的目的就是系统介绍对社会科学家常用的多种概率模型的解释。为了进一步说明概率模型的重要性以及它们的解释,我在介绍部分回答两个问题:(1)为什么要用概率模型?(2)为什么需要解释?

第1节 | 为什么要用概率模型？

在社会学家的工具箱里，最基本的统计方法就是对一个连续的线性的因变量（或者可以转化成线性的）进行回归分析。然而很多社会科学家研究的对象是无法用经典的回归模型来分析的，因为很多的态度、行为、特点、决定以及事件（无论本质上是连续的或者不连续的）是用离散的、虚拟的、序列的或者简单来说，非连续的方法来测量的。

划分处理此类数据的一些统计模型常常根据数据的种类来代表和讨论，比如“二分数据分析”、“序列数据分析”、“类别数据分析”或者“离散选择分析”，或者作为一个特别的模型，比方说 logit 或者 probit 模型。这些相关联的统计方法的共同特点就是它们都是对某事件的概率，即这个事件究竟有多大可能发生来建模。因此，在本书里，我将所有分析事件概率的统计模型统一称为“概率模型”。我们讨论的概率模型包括二分的、序列的、有序的 logit 和 probit，多类别 logit，条件 logit，以及泊松回归模型。