



格致方法·定量研究系列 吴晓刚 主编

定序因变量的logistic 回归模型

[美] 安·A.奥康奈尔(Ann A.O'Connell) 著
赵亮员 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

格致出版社  上海人民出版社

15

格致方法·定量研究系列 吴晓刚 主编

定序因变量的 logistic 回归模型

[美] 安·A.奥康奈尔(Ann A.O'Connell) 著

赵亮员 译

图书在版编目(CIP)数据

定序因变量的 logistic 回归模型 / (美) 奥康奈尔
(O'Connell, A. A.) 著; 赵亮员译. —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 978-7-5432-2133-8

I. ①定… II. ①奥… ②赵… III. ①回归分析-统
计模型-研究 IV. ①0212.1

中国版本图书馆 CIP 数据核字(2012)第 135162 号

责任编辑 高 璇

格致方法·定量研究系列 定序因变量的 logistic 回归模型

[美] 安·A. 奥康奈尔 著
赵亮员 译

出版 世纪出版集团 格致出版社
www.ewen.cc 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 5.75
字数 111,000
版次 2012年7月第1版
印次 2012年7月第1次印刷
ISBN 978-7-5432-2133-8/C·81
定价 18.00 元

Logistic Regression Models for Ordinal Response Variables

Copyright © 2006 by SAGE Publications, Inc.

All rights reserved. No part of this book may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission in writing from the publisher.

This simplified Chinese edition for the People's Republic of China is published by arrangement with SAGE Publications, Inc. © SAGE Publications, Inc. & TRUTH & WISDOM PRESS 2012.

本书版权归 SAGE Publications 所有。由 SAGE Publications 授权翻译出版。
上海市版权局著作权合同登记号:图字 09-2009-548

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,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

在过去 60 年中,logit 模型已经成为社会科学中最流行的统计方法。自从伦西斯·利克特(Rensis Likert)1932 年出版《态度测量的方法》(*A Technique for the Measurement of Attitudes*)以来,人们对人类态度的测量方法一直都在变化着。实际上,今天的任何社会调查都已将李克特类型的量表作为一种主要的方法。一个典型的李克特量表有五个类别(例如,非常不同意、同意、不知道、同意、非常同意),用以测量个体对问题的响应,尽管它在别的地方也许还有三到七个或更多的响应类别。如果我们将五个类别编码为从 1 到 5,那么我们可以用线性回归模型来估计李克特量表,这也是早些时候分析这样的数据时所采用的方法。然而,这种方法也存在诸多明显的问题。首先,经典线性回归假设连续因变量是距离相等的定序因变量类别(ordered response categories)。一个李克特类型的量表,或者任何其他的定序量表,虽然是定序的,但并不一定在类别间等距。其次,也许是更重要的,这样的量表所得出的分布不是经典线性回归所假定的数据呈现的正态分布。

为了分析定序数据的这种属性,也有其他的方法,经常

使用的是列联表和对数线性模型的形式,具体可以参见:《定序数据分析》(*Analysis of Ordinal Data*),作者海德普兰德、莱恩和罗森塔尔(Hiderbrand, Laing, Rosenthal);《对数线性模型》(*Log-linear Models*),作者诺克和伯克(Knoke, Burke);《潜变量的对数线性模型》(*Loglinear Models With Latent Variables*),作者哈根纳斯(Hagenaars);《定序对数线性模型》(*Ordinal Log-Linear Models*),作者石井坤茨(Ishii-Kuntz);《列联表分析中的比数比》(*Odds Ratios in the Analysis of Contingency Tables*),作者鲁达斯(Rudas)。然而,这些方法都不在回归的框架下,而回归方法正是社会科学中广为人知且被广泛应用的量化方法。

安·A. 奥康奈尔的这本书填补了空白。即使有的著作也将定序因变量变量在 logit 模型中加以处理,但本书则完全通过呈现三种形式的因变量来集中关注该类 logit 模型,这些因变量抓住了响应的定序本质。本书以展示来自儿童早期追踪研究的实例开始,其中的主要因变量虽然不是李克特量表,但其对早期读写和算术熟练度的测量还是定序的。作者进而回顾了 logistic 回归,并接着呈现三个核心主题章节:累积或比例比数模型(cumulative or proportional odds model),连续比例模型(continuation ratio model)以及相邻类别模型(adjacent categories model)。与此同时,书中给出了 SAS® 和 SPSS® 的案例。尽管比例比数模型也许是三种方法中最常用的,但在结论部分,读者可以领会到关于替代方法尤其是什么时候使用替代方法的提示。

目 录

序	1
第 1 章 概述	1
第 1 节 本书的目的	5
第 2 节 软件和句法	7
第 3 节 本书结构	8
第 2 章 背景介绍：儿童早期追踪研究	11
第 1 节 儿童早期追踪研究概述	12
第 2 节 定序结果的实际重要性	15
第 3 节 模型中的变量	16
第 3 章 背景：logistic 回归	19
第 1 节 logistic 回归概述	20
第 2 节 模型拟合的评估	25
第 3 节 模型的解释	27
第 4 节 关联的测量	30
第 5 节 logistic 回归	31
第 6 节 比较统计软件间的结果	40

第 4 章	定序结果的累积（比例）比数模型	45
第 1 节	累积比数模型概述	46
第 2 节	单个解释变量的累积比数模型	50
第 3 节	累积比数的全模型分析	64
第 4 节	logit 中的比例比数和线性假设	68
第 5 节	累积比数模型的替代方法	73
第 6 节	偏比例比数	76
第 5 章	连续比例模型	83
第 1 节	连续比例模型概述	84
第 2 节	关联函数	88
第 3 节	相关的概率	90
第 4 节	因变量的向性和连续比例的形成	92
第 5 节	使用 logit 关联和重构数据的连续比例模型	94
第 6 节	使用互补双对数关联的连续比例模型	103
第 7 节	关联的选择和两个互补双对数模型的等价	109
第 8 节	连续比例模型的方法选择	112
第 9 节	ECLS-K 数据的全模型连续比例分析	114
第 6 章	相邻类别模型	119
第 1 节	相邻类别模型概述	120
第 2 节	仅有性别的模型	123
第 3 节	两个解释变量的相邻类别模型	129

第 4 节 全相邻类别模型分析	131
第 7 章 结论	133
第 1 节 本书内容总结	134
第 2 节 进一步研究的考虑	138
附录	141
注释	152
参考文献	154
译名对照表	161

第 **1** 章

概 述

对于很多在教育和社会科学中的因变量,定序尺度提供了一种将可能的结果区分开来的简单且便捷的方法,而这些可能结果最好被看成是等级定序的(rank-orderd)。定序数据的主要特点,是指定给所测量变量的相继类别的数字代表的量级上的差异,或者“大于”或者“小于”的性质(Stevens, 1946, 1951)。一些定序数据的例子包括:对开放式答案或文字进行编码的一类题目,以及解决以能力改善为基础的答案的计算方案(例如,0 = 差,1 = 可以接受,2 = 优秀)。对比之下,名义水平的数据发生在数值用于通过认定类别之间简单的质的差别来测量一个变量的情况(例如,1 = 男性,或者 2 = 女性;所在学校的地理位置描述:1 = 乡村,2 = 城市,3 = 郊区,等等);名义变量没有定序数据的方向特征。另一方面,在定距或者定比水平尺度上测量的变量的确使用了表明定序数据“大于”或者“小于”属性的尺度值,但还保持了在尺度间相邻值距离均等或者间隔均等长度的属性。以摄氏尺度测量的温度是我们熟知的定距水平变量的例子。然而,定距水平的变量的零点相对主观,而非绝对。拥有所有定距尺度属性但也有真实零点的变量,被称为定比水平。例如,对一项任务的反应时间、重量以及距离等是熟悉的定比水平的变量。^[1]