格致方法·定量研究系列　吴晓刚　主编

应用logistic回归分析（第二版）

[美] 斯科特·梅纳德（Scott Menard）著
李俊秀 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

13

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

应用 logistic 回归分析（第二版）

[美] 斯科特 · 梅纳德(Scott Menard) 著
李俊秀 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

应用 logistic 回归分析:第二版/(美)梅纳德(Menard, S.)著;李俊秀译. —上海:格致出版社:上海人民出版社,2016
(格致方法·定量研究系列)
ISBN 978-7-5432-2616-6

Ⅰ. ①应… Ⅱ. ①梅… ②李… Ⅲ. ①线性回归-回归分析 Ⅳ. ①O212.1

中国版本图书馆 CIP 数据核字(2016)第 062748 号

责任编辑　王亚丽　高　璇

格致方法·定量研究系列

应用 logistic 回归分析(第二版)

[美]斯科特·梅纳德 著
李俊秀 译

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　2016 年 4 月第 1 版
印　次　2016 年 4 月第 1 次印刷

ISBN 978-7-5432-2616-6/C·146　　定价:25.00 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

线性回归模型是一个非常有效且重要的数据分析方法。研究人员主要着重解释因变量，将因变量看做由多个自变量 X_1 至 X_k 所组成的函数。当所有线性回归假设都符合该模型时，模型辨识、变量测量、普通最小二乘法（Ordinary Least Squares，OLS）估计方程，这一切都很顺利。可是，当因变量有两个或三个分类的话，有几项假设就不符合了。以二分因变量为例，同方差、线性和正态性的假设都不能成立，OLS 的估计也无效。logistic 回归的最大似然估计法（maximum likelihood estimation）就能解决这一问题，即将 $Y(1, 0)$ 转化成 logit（发生比的对数 log）。

梅纳德教授全面地解释了 logistic 回归模型的估计、解释和诊断结果。为了令读者能够从熟悉的事件过渡到新事物，他系统地把 logistic 回归与线性回归模型的 OLS 的 R^2、估计标准误差（standard error of estimate）、t 比率（t ratio）和斜率（slope）做比较。传统回归诊断和学生化残差（studentized residual）、杠杆（leverage）、dbeta 都包括在创新的 logistic 诊断法内。最后仔细说明了多选项和不排序多分类

因变量的问题。

本书讨论了对最新计算机软件的应用,如 SPSS10 NOMREG 用以分析类别变量(nominal variable)比较好,SAS lOGISTIC 分析定序因变量比较好。本书更新了现今应用的计算机软件,同时深入地评论了不同的拟合优度(goodness of fit)。梅纳德博士还提出令人信服的论据去说明 R_L^2 的优势,至少这能直接与 OLS 的 R^2 比较。他同时增加了新内容:分组数据(grouped data)、预测效率(predictive efficiency)和风险比(risk ratios)。

大量著作证明了线性回归的广泛应用,可是由于现实中的因变量很少会是连续的或定距的(interval),因此,logistic 回归开始备受关注。首先出版的是德马里斯(DeMaris)的《对数模型》;接着是梅纳德的《应用 logistic 回归》(第一版);以及潘帕(Pamper)的《logistic 回归简介》)。本书从基本原理到技术应用都做了介绍,除此之外,还提及了当今最复杂的问题和方法。社会科学家要熟悉日新月异的知识,千万别错过梅纳德的这本书。

迈克尔·刘易斯—贝克

目 录

第1章

线性回归和应用 logistic 回归模型

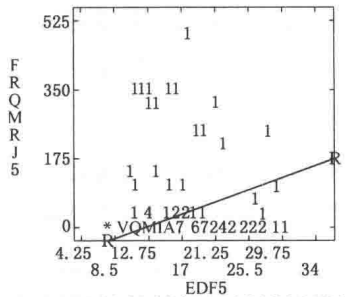
只要两个变量的关系能用方程式 $Y=\alpha+\beta X$ 来表达,那就有可能用线性回归分析去检视两者是否有线性相关,同时计算此相关的强度。Y 是被预测的变量,称为因变量(dependent variable)、基准变量(criterion variable)、结果变量(outcome variable)或内生变量(endogenous variable);X 是用来预测 Y 的变量,称为自变量(independent variable)、外生变量(exogenous variable)或预测变量(predictor variable)[1];α 和 β 是总体参数(population parameters)的估计值,参数 α 称为截距(intercept),代表 $X=0$ 时 Y 的数值;参数 β 代表 X 增加一个单位时 Y 数值的变化,或是 X 预测 Y 的最佳直线斜率。多元回归(multiple regression)含有几个自变量,假设 K 是自变量的数量,方程式便是 $Y=\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k$,而 β_1,β_2,…,β_k 称为偏斜系数(partial slope coefficient),即任何一个自变量 X_1,X_2,…,X_K 只对 Y 值提供部分预测。有时方程式会明确地显示出 X 对 Y 的预测并不精确,$Y=\alpha+\beta X+\varepsilon$,数个自变量的方程式 $Y=\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k+\varepsilon$,$\varepsilon$ 代表 X 预测 Y 时出现的误差,是一个随机变量。就个别个案 j,$Y_j=\alpha_j+\beta_j X_j+\varepsilon_j$ 或 $Y_j=\alpha_j+\beta_1 X_{1j}+\beta_2 X_{2j}+\cdots+\beta_k X_{kj}+\varepsilon_j$,$j$ 表示指定的第 j 个个案($j=1$ 代表第一个个案,$j=2$

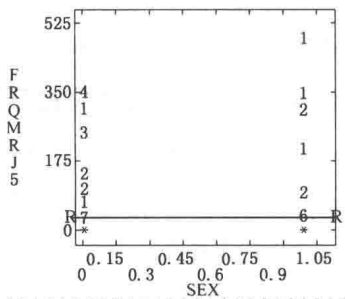
是第二个个案，如此类推）。Y_j、X_{1j}、X_{kj} 等代表因变量和自变量的特定值。上述的方程式主要用来计算在个案 j 中，Y 的数值，多用于描述变量间的关系。

截距 α 和回归系数 β（regression coefficient）的估计值是通过普通最小二乘法（OLS）计算出来的，这点在许多基础统计学的书中都讨论过（Agresti & Finlay，1997；Bohrnstedt & Knoke，1994）。这些估计值所形成的方程式 $\hat{Y} = a + bX$（一个自变量）或 $\hat{Y} = a + b_1X_1 + b_2X_2 + \cdots + b_kX_k$（多个自变量）中，$\hat{Y}$ 是 Y 的线性回归方程式的预测值，a 是截距 α 的普通最小二乘法的估计值，b（或 b_1，b_2，…，b_k）是斜率 β（或 β_1，β_2，…，β_k）的普通最小二乘法的估计值。每个个案的残差（residuals）$e_j = (Y_j - \hat{Y}_j)$，其中 $\hat{Y}_j$ 是 Y 在个案 j 的估计值。二元回归（bivariate regression）的残差可以用二元散点图表示，即每一点和回归直线的垂直距离。多元回归的残差就比较难以用图形展示，因为它需要多维空间。

图 1.1 是二元回归模型的一个例子，其数据来自 1980 年第五次美国家庭调查，访问对象是 16 岁的青少年。图 1.1A，因变量 FRQMRJ5，即被访者自我报告的每年使用大麻的频率（在过去一年，你曾经吸过多少次大麻?）；自变量 EDF5，即接触违法朋友的程度[2]。量度接触违法朋友的方法是将八个问题加起来，问受调查的青少年到底有多少个朋友曾被牵涉在不同的罪行当中（如偷窃、侵犯他人、药物滥用）。每题有五个选项，由 1（没有朋友）至 5（所有朋友），因而 EDF5 的总和是由 8 至 40。图 1.1A 显示接触违法朋友与大麻使用成正相关，方程式如下：

$$(\text{FRQMRJ5}) = -49.2 + 6.2(\text{EDF5})$$





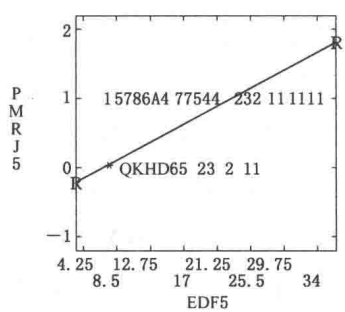


图 1.1 二元回归图

换言之,接触违法朋友每增加一个单位,每年大麻使用频率就会随之增加约六倍,或每两个月增加一次。

决定系数(coefficient of determinant)R^2 是指自变量能多准确地预测因变量。通过知道青少年接触违法朋友的数量,我们可根据 EDF5 的值和回归方程去表现它与 FRQMRJ5 之

间的关系。这一方法可减少大约 12%的预测平方误差总和(sum of the squared errors of prediction)。

$$\sum e_j^2 = \sum (\hat{Y}_j - Y_j)^2, \ R^2 = 0.116$$

当解释结果时,必须考虑因变量和自变量的真实数值。截距指的是没有接触过违法朋友的人,其大麻使用频率为负数。出现这种不合理的数值是因为接触违法朋友的数值由 8(完全没有朋友涉及任何一项违法活动)至 40(所有朋友都涉及违法活动)。因此,当个别人的接触违法朋友的数值最小时,大麻使用的期望值应为 $-49.2+6.2(8)=0.4$,接近于零,表示即使没有接触违法朋友的人,偶尔也会吸食大麻。这一研究样本的 EDF5 最大值是 29,相应的大麻使用的期望值为 $-49.2+6.2(29)=130.6$ 或每三天吸食一次大麻。这样,无论是从统计学或现实世界的角度看,这一数值都算合理。