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

非外借

- 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

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

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)乔治·H.邓特曼,(加)何满镐著;林毓玲译.—上海:格致出版社:上海人民出版社,2019.3
(格致方法·定量研究系列)
ISBN 978-7-5432-2985-3

Ⅰ.①广… Ⅱ.①乔… 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(2019)第 031242 号

责任编辑 张苗凤

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
[加]何满镐 著
林毓玲 译

出　　版 格致出版社
上海人民出版社
(200001 上海福建中路 193 号)
发　　行 上海人民出版社发行中心
印　　刷 浙江临安曙光印务有限公司
开　　本 920×1168 1/32
印　　张 4.25
字　　数 83,000
版　　次 2019 年 3 月第 1 版
印　　次 2019 年 3 月第 1 次印刷
ISBN 978-7-5432-2985-3/C·213
定　　价 32.00 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”，[illegible]经成功举办三届。研修课程设计体现“化整为零、[illegible]经典正态回[illegible]教学、学以致用”的方针，吸引了一大[illegible]

第 3 节　logistic 回归
第 4 节　比例风险生存模型

[illegible]研究的博士生和青年学[illegible]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。

目 录

序 1

第 1 章 广义线性模型 1

第 2 章 一些基础的模型化概念 9
第 1 节 作为类别变量的自变量 12
第 2 节 回归模型的必要成分 16

第 3 章 经典多元回归模型 19
第 1 节 假设与模型方法 22
第 2 节 回归分析结果 25
第 3 节 多元相关 26
第 4 节 假设检验 27

第 4 章 广义线性模型的基本原则 33
第 1 节 指数家族分布 37
第 2 节 经典正态回归 40
第 3 节 logistic 回归 41
第 4 节 比例风险生存模型 43

第 5 章 最大似然估计 45

第 6 章 离差和拟合优度 55
第 1 节 使用离差进行假设检验 59
第 2 节 拟合优度 60
第 3 节 通过残差分析衡量拟合优度 61

第 7 章 logistic 回归 65
第 1 节 logistic 回归概述 66
第 2 节 logistic 回归实例 72

第 8 章 泊松回归 77
第 1 节 泊松回归概述 78
第 2 节 泊松回归模型实例 81

第 9 章 生存分析 91
第 1 节 生存时间分布 94
第 2 节 指数生存模型 98
第 3 节 指数生存模型实例 101

第 10 章 结论 103

附录 107
参考文献 116
译名对照表 117

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，许多形态的 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

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。

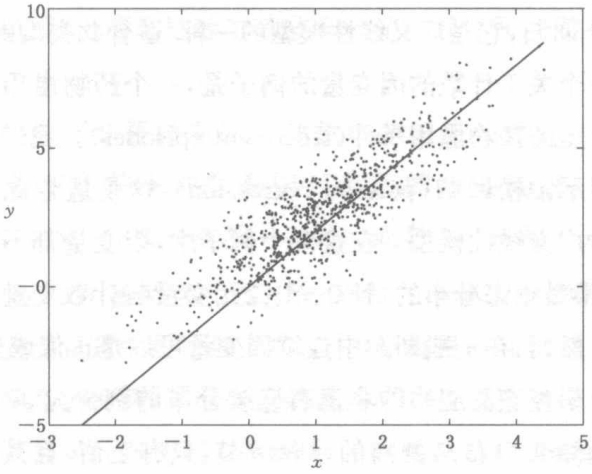


图 1.1　线性回归模型

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

正值的因变量，如收入、生存时间及雨量。因变量为 gamma 分布的模型可以被置于广义线性模型的架构之下。

要注意，对于一个给定分布的因变量而言，自变量可以有许多种分布形式，且它们并不需要和因变量具有相同的分布。例如，与一正态分布因变量相关联的自变量可以包含许多不同的非正态分布，如均匀分布和多峰分布。如前所述，一般回归假设 y 的平均值会随着自变量变化，但关于条件平均数的 ε 的变异则维持不变。对于二元变量和计数变量来说，条件平均数的方差为其平均数的函数。如二元变量，因变量条件平均数为概率 p（如事件 1 发生的概率），而此平均数的方差为 $p(1-p)$，是平均数 p 的一个函数。因为平均数 p 会作为自变量的一个方程而变化，此二元变量的方差也会如此。对计数变量而言，泊松分布常被使用，而此项分布的方差等于其平均数。因此，当泊松分布的条件平均数作为一自变量方程而变化，其方差亦是如此。广义线性模型在 logistic 和泊松回归模型这种情形下，在模型的公式化及估计回归参数里，已明确地通过其概率分布体现了平均数和方差的关系。

经典回归亦假设在回归参数中模型是线性的，即其假设期望值或条件平均数是回归参数的线性函数。例如，$E(y \mid X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$，或者 $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1 X_2$。需注意，在第二个模型中，参数是线性的但自变量是非线性的。事实上，经典线性回归是一个广义线性模型的特例，其因变量的条件平均数直接被模型化而没有对条件平均数进行某种转换。对其他的广义线性模型而言，条件平均数无法被写成回归参数的一个线性函数，但