格致方法·定量研究系列　吴晓刚　主编

分位数回归模型

[美] 郝令昕（Lingxin Hao）
丹尼尔·Q.奈曼（Daniel Q. Naiman）著
肖东亮 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

10

格致出版社　上海人民出版社

图书在版编目（CIP）数据

肛肠疾病研究进展/刘仍海，姜春英，韩平主编．—北京：中医古籍出
（当代肛肠疾病临床与研究）
ISBN 978－7－5152－0275－4

Ⅰ．①肛…　Ⅱ．①刘…②姜…③韩…　Ⅲ．①肛门疾病－研究进展②

中国版本图书馆 CIP 数据核字（2012）第 213052 号

肛肠疾病研究进展

刘仍海　姜春英　韩　平　主编

责任编辑　李艳艳　贾蓉荣

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”(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

40年来，经典的线性回归模型一直是社会科学定量研究方法论中重要的组成部分。目前已有的关于定量方法应用的书籍，涉及许多关于线性回归的各种延伸话题，例如logit、probit、事件史、广义线性、广义非参数模型和处理删截、样本选择、截断和缺失数据的线性回归模型；此外，还包括许多其他相关的模型，例如方差分析、协方差分析、因果模型、对数线性模型、多重比较和时间序列分析等。

经典回归的主旨在于基于解释变量来估计因变量的均值。当回归假设成立时，这一方法是有效的；但当出现非标准情况时，它就会失效（关于线性回归假设的详细讨论，见《理解回归假设》，William Berry 著）。其中两个是正态性假设和方差齐性假设。通常的社会科学数据无法满足这两个关键的假设。例如，（条件）收入分布一般不是正态的，并且首席执行官的年度分红分布随着公司规模的增大而上升，这意味着存在异方差性问题。这正是分位数回归可以处理的问题，因为它放松了这些假设。另外，分位数回归为研究者提供了一个（无法从经典回归中获得的）新视角，研究解释变

量对因变量分布中位置、尺度和形状的效应。

分位数回归的思想并不新颖，事实上它起源于 1760 年，当时一个游历学者、克罗地亚基督徒 Rudjer Josip Boscovich——他拥有许多头衔：物理学家、天文学家、外交官、哲学家、诗人和数学家——来到伦敦讲授他尚未成熟的中位数回归方法。然而，这一回归方法计算的复杂性直到最近依然是一大挑战。由于今日快速的计算功能和统计软件的广泛应用（如可执行分位数回归程序的 R、SAS 和 Stata），使得拟合分位数回归模型变得更加容易。但是，至今我们仍未提供任何关于分位数回归是什么的介绍。在本书中，Hao 和 Naiman 提出了分位数和分位数函数的概念，并阐述了分位数回归模型，讨论了它们的估计和推断方法，并通过具体例子演示了对分位数回归估计值（是否转换）的解释。同时，他们也提供了应用分位数回归分析美国 1991 年和 2001 年收入不平等的完整例子，以此确定这一方法的思想和步骤。本书填补了丛书的空白并且有助于社会科学研究者更加熟悉分位数回归。

廖福挺

目 录

第1章

引　言

回归分析的目的在于揭示因变量和自变量的关系。在实际应用中，自变量并不能精确地估计因变量。相反，与每个自变量的特定值相对应的因变量是一个随机变量。因此，我们常常使用集中趋势的测量方法，来概括自变量特定值域下的因变量变化情况，主要包括均值、中位数和众数。

传统的回归分析主要关注均值，即采用因变量条件均值的函数来描述自变量每一特定数值下的因变量均值，从而揭示自变量与因变量的关系。模型化和拟合条件均值函数(conditional-mean function)是回归模型法大族谱中的核心思想，具体包括常见的简易线性回归模型、多元回归、加权最小平方数下的异方差误差模型和非线性回归模型。

条件均值模型具有以下优点：在理想的条件下，它们可以为我们提供关于自变量和因变量分布关系的完整的和参数的描述。另外，采用条件均值模型可获得具有优越统计特性的估计量(最小二乘法和最大似然法)，它更容易计算，并且更容易解释。这种模型通过不同的方式被推广，从而适用于误差具有异方差性的情况，因此，对于特定的自变量，因变量条件均值和条件单位的模型化可以同时进行。

条件均值模型被广泛应用于社会科学中，尤其在过去的

半个多世纪里，使用最小二乘法及其衍化方法对连续型因变量和自变量的关系进行回归建模被认为是现代重要的统计工具。最近，分析二分因变量的 logistic 和 probit 模型、分析计数因变量的泊松回归模型在社会科学研究中的重要性不断提高。这些方法并没有超出条件均值模型的框架。当社会科学定量研究者已经应用更高级的分析方法来放宽条件均值框架下的一些建模假设时，这个框架本身却很少被质疑。

条件均值框架存在先天的局限性。首先，当归纳自变量特定数值下的因变量情况时，这个条件均值模型并不能轻易地扩展到非中心位置，而非中心位置往往正是社会科学研究的兴趣所在。例如，关于经济不平等和流动的研究对穷人（低尾）和富人（上尾）的情况有浓厚的兴趣。教育研究者会设法在既定的成绩水平下去理解和减少群体差异（如三层次参照标准：基础、熟练和高级）。这样，对中心位置的强调，长期阻碍了学者采用恰当的技术来研究有关因变量非中心位置的课题。而采用条件均值模型来分析以上问题是没有效率的，甚至会偏离研究重点。

其次，这些模型的假设在现实生活中并不总会得到满足。特别是方差齐性假设经常被违反。另外仅仅关注集中趋势会忽视关于因变量分布的有用信息。并且，社会现象中通常会出现重尾分布，从而导致离群值优势。正因为条件均值深受离群值的干扰，所以它对中心位置的测量是不恰当和具有误导性的。

最后，一直以来对中心位置的关注转移了学者对因变量整体分布性质的注意力。我们需要跳出预测变量的位置和数值范围对因变量的效应这一框架，进而探讨预测变量的变

化会如何影响因变量分布的基本形状。例如，许多社会科学研究关注社会分层和不平等，这一领域要求深入分析因变量的分布特征。对分布特征的描绘包括中心位置、数值范围、偏态和其他高阶特性，而不仅仅是中心位置。因此，采用条件均值模型来表述因变量分布与自变量的关系是具有先天性缺陷的。关于不平等主题的例子包括工资、收入和财富等经济不平等；在学业成绩上的教育不平等；在身高、体重、疾病发生概率、毒品上瘾、医疗和预期寿命上的健康不平等和由于社会政策而导致的生活质量的不平等。这些课题通常采用条件均值框架进行分析，从而忽略了其他更重要的分布特征。

条件均值模型的替代方法可以追溯到 18 世纪中期。这一方法被称为条件中位数模型，或简称中位数回归。它解决了一些上面提出的关于集中趋势测量方法的选择问题。这种方法用最小绝对距离估计代替最小二乘估计。最小二乘估计不需要大功率的计算机便可轻松实现，然而最小绝对距离估计必须借助强大的计算机力量。所以，直到 20 世纪 70 年代后期，当计算机技术融合了如线性优化等算法系统时，采用最小绝对距离估计的中位数回归模型才变得实用。

中位数回归模型可以实现与条件均值回归模型同样的目标：表述因变量的中心位置与一组协变量之间的关系。然而，当因变量的分布是高度偏态时，均值在解释的时候就会受到质疑，而中位数依然保有大量信息。因此，条件中位数模型具有更大的应用潜力。

中位数是一个特殊的分位数，它表示一种分布的中心位置。中位数回归是分位数回归的一种特殊情况，在这里，第 0.5 分位数被模型化为一个关于协变量的函数。一般地说，