格致方法
定量研究系列

- 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

吴晓刚 / 主编
【美】保尔·D. 埃里森(Paul D. Allison) / 等著

高级回归分析

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
【美】保尔·D. 埃里森(Paul D. Allison) / 等著

高级回归分析

SAGE Publications ,Inc.

格致出版社 上海人民出版社

图书在版编目(CIP)数据

高级回归分析/(美)埃里森(Allison, P. D.)等著;吴晓刚主编. —上海:格致出版社:上海人民出版社,2011

(格致方法·定量研究系列)

ISBN 978-7-5432-1899-4

Ⅰ. ①高… Ⅱ. ①埃… ②吴… Ⅲ. ①回归分析-研究 Ⅳ. ①O212.1

中国版本图书馆 CIP 数据核字(2010)第 254745 号

责任编辑 罗 康

封面装帧 人马艺术工作室·储平

格致方法·定量研究系列

高级回归分析

[美]保尔·D. 埃里森 等著

吴晓刚 主编

出 版 世纪出版集团 www. ewen. cc

格 致 出 版 社 www. hibooks. cn

上海人民出版社

(200001 上海福建中路193号 24 层)

格致出版

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 上海图宇印刷有限公司

开 本 787×1092 毫米 1/16

印 张 34.75

插 页 1

字 数 504,000

版 次 2011 年 8 月第 1 版

印 次 2011 年 8 月第 1 次印刷

ISBN 978-7-5432-1899-4/C·44

定 价 78.00 元

出版说明

《高级回归分析》是“格致方法·定量研究系列”之一种，由5本讨论高级回归分析的小册子组成，分别是《固定效应回归模型》、《现代稳健回归方法》、《删截、选择性样本及截断数据回归模型》、《分位数回归模型》及《空间回归模型》。

《固定效应回归模型》介绍了多种形式的固定效应回归模型，讨论了如何在固定效应模型及随机效应模型之间作出选择；《现代稳健回归方法》通过一套统一的符号系统，介绍了不同来源的多种稳健回归方法，以及它们彼此之间的联系；《删截、选择性样本及截断数据回归模型》是一本有关删截数据、选择性样本数据及截断数据的最新研究；《分位数回归模型》提出了分位数和分位数函数的概念，阐述了分位数回归模型，讨论了它们的估计和推断方法，并通过具体的例子演示了对分位数回归估计值的解释；《空间回归模型》介绍了两种应用最广泛的空间回归模型：空间定距因变量和空间性误差模型。

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。

由于所选每本书都有一篇序言，对相关方法的背景和应用作了很好的介绍，我们均予以保留，内容在此不再赘述。为了方便起见，我们将内容相似的书目集册出版，每册三至五本不等，共八册，它们分别是：《线性回归分析基础》、

《高级回归分析》、《广义线性模型》、《列表数据分析》、《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》和《数据分析方法五种》。所冠书名未必能精准涵盖其中的内容，读者可自行参阅每本书的序言或目录。

我们希望本丛书的出版，能为推动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

目 录

固定效应回归模型

现代稳健回归方法

删截、选择性样本及截断数据的回归模型

分位数回归模型

空间回归模型

固定效应回归模型

作 者 简 介

保尔·D. 埃里森

保尔·D. 埃里森(Paul D. Allison)是宾夕法尼亚大学社会学教授,讲授的研究生高级课程包括事件史分析、分类数据分析以及能够处理潜变量的结构方程模型。他有 7 本专著,并发表了 50 多篇论文。每年暑假,他都会召开一个为期 5 天的工作坊,向来自全美的约 100 名研究者讲授生存分析和 Logistic 回归分析。曾经荣获古根海姆学者奖(Guggenheim Fellow),并于 2001 年获得拉扎斯菲尔德奖(Lazarsfeld Award),以奖励他在社会学研究方法领域的突出贡献。

译 者 简 介

李丁

李丁,北京大学社会学系本科和硕士毕业,现为北京大学社会学系博士研究生。

序

在最近的一次会议上，我聆听了某研究者分析国家一年度(country-year)数据的报告，理应使用固定效应模型，他用的却是随机效应模型。而那篇文章却受到了来自不同社会科学背景的学者的热烈欢迎。显然，在诸多社会科学专业里，就如何选用固定效应模型和随机效应模型还存在很多疑惑，很多人甚至还不清楚这些模型有何用处。无疑，埃里森讨论的是这两种模型更重要和一般的方面。本书将很好地满足“社会科学定量研究方法丛书”在这一主题上的需要，尤其是考虑到现在获得跟踪调查数据(panel data)越来越容易的现实①。

上述国家一年度数据代表着这样一种数据类型，在这种数据中，个体案例得到了历时的(多次)观察。动态跟踪调查(panel survey)之所以近年来非常流行，一个重要的原因是跟踪数据允许研究者把握社会的发展变化，而把握这种变化是真正理解社会机制的必要条件。尽管有些跟踪调查每年都会观测一次，例如英国家户跟踪调查(British Household Panel Study Survey，简称 BHPS)开始于 1991 年，目前仍在持续进行；其他一些则只有少数几轮调查，例如全美青少年健康跟踪调查(National Longitudinal Study of

---

① Panel Data 在经济学文献中通常被翻译为面板数据(如《面板数据计量经济学》[*Panel Data Econometrics*]，曼纽尔·阿雷拉诺著，朱平芳、徐伟民译，上海财经大学出版社，2008 年 10 月)或综列数据(如《计量经济学：现代观点》，[美]J. M. 伍德里奇著，费剑平译，中国人民大学出版社，2003 年 3 月)。面板数据分析(Panel Data Analysis)与时间序列分析(Time Series Analysis)及横截面数据回归分析(Regression Analysis with Cross-Sectional Data)构成计量经济学的三大内容。其中横截面数据是一个时点上收集的不同观察对象的数据，比方说，就一次人口普查来说就是一个截面研究；时间序列数据通常是一个观察单位在不同时点的观察结果构成的数据，如我国 1978 年以来，每一年的 GDP 的增长速度数据就构成一个时间序列数据。而 Panel Data 将这两种数据的特性综合在一起，首先在同一时点上对不同的案例(通常为总体中的一个规模相对较小的子样本)的多个特征进行了观测；其次，对每一个案例在不同时点进行了多次观测；由此所得到的数据就是 Panel Data。面板数据这一翻译尽管因计量经济学而流传甚广，但“面板”的中文字义与英文 panel 含义相差甚远。Panel Survey 翻译成为面板调查显得笨拙，可翻译为小样本重复调查、固定样本长期追踪调查、追踪调查、纵贯调查或者历时调查等等。其中追踪调查或跟踪调查即带有在一段时间内对固定样本进行多次调查的含义。此外，根据艾尔·巴比的观点，Longitudinal Data 为历时研究，它包括趋势研究、队列研究和追踪研究三类。——译者注

Adolescent Health in the United States)只在 1994 年到 2002 年间进行了 3 轮调查。

不管分析单位是个人、单位还是国家，回归模型中每个案例在不同时点上的残差都将存在一定的相关或互相依赖，这通常是因为不同案例在某些未被观察到的特征上存在差异造成的。此时，回归模型有关误差项相互独立的假定被违背(尽管这个一般规律同样适应于限值因变量[limited dependent variable]回归，但这里我们将讨论限定在线性模型上)。

固定效应模型和随机效应模型都能解决残差相关问题。但固定效应模型做得彻底得多。用埃里森的话说，这些模型"将每个个体作为其自身的控制。"经此处理，它们实际上就控制了所有稳定的、未被观测到的变量，就像这些变量实际得到了观测并被纳入模型一样。就此而言，这些模型所起的作用和实验设计中的随机分配如出一辙。

本书作者在过去 30 年间为社会科学研究方法作出了持续的贡献，涉及诸多重要的主题。他撰写的《事件史分析》(Event History Analysis, 1984)，至今仍是社会科学领域介绍事件史数据分析著作的榜样和标准。确切地说，在本书中，埃里森介绍了多种形式的固定效应模型——可以用于连续因变量的、分类因变量的、计数因变量的甚至结构方程情境等——并且讨论了如何在固定效应模型及随机效应模型之间作出选择，这一讨论对于本序言开始时提及的那位报告人将大有裨益。

廖福挺(Tim Futing Liao)

第 1 章 | 绪言

多年以来，统计学领域最具挑战性的议题，是如何创造一些方法以从非实验数据中进行有效的因果推论。而在这一议题内最难的问题，是如何从统计上控制无法观测的变量。对于实验主义者而言，问题的解决方案非常简单：随机分配（random assignment）。通过将研究对象随机分配到实验组（treatment group），可使这些小组在研究对象各属性上几乎相似，不管这些属性是可观测的还是不可观测的。但是在非实验研究中，控制这些潜在干扰变量的传统办法就是测量它们，并把它们放到回归模型里。没有测量就没有控制。

在本书中，我描述了一些被称为固定效应模型的回归模型，这些模型使得我们有可能对那些没有或无法被测量的变量进行控制。基本的思想非常简单：用每个个体作为其自身的控制（因素）。例如，如果你想弄清婚姻是否能减少惯犯们（chronic offenders）的再犯行为（recidivism），可以通过对个体结婚前后遭拘捕的比率进行比较。假定其他情况都不变（这是一个很大的假定），前后两个时期拘捕率的差异可以作为婚姻对该个体产生的效果的估计。如果我们将人群中不同个体的这一差异进行平均，就能得到“平均处置效应”（average treatment effects）的估计值。这一估计控制了惯犯们所有的稳定属性。它同时控制了容易被测量的变量，诸如性别、民族、种族、出生地，以及更难被控制的变量，如智商、儿童期父母的照料情况、遗传结构等。虽然它不控制诸如就业状况、收入之类的时变变量，但这些变量通过常规的办法——对其进行测量并放入回归模型——就可以得到控制。

再举一个例子，假如你想研究打电脑游戏的时间是否会影响小孩的学习成