



格致方法·定量研究系列 吴晓刚 主编

删截、选择性样本及截断数据的回归模型

[英] 理查德·布林 (Richard Breen) 著
郑冰岛 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

12

格致方法·定量研究系列 吴晓刚 主编

删截、选择性样本 及截断数据的回归模型

[英] 理查德·布林(Richard Breen) 著
郑冰岛 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

删截、选择性样本及截断数据的回归模型 (英)布林(Breen, R)著,郑冰岛译 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 978-7-5432-2114-7

I ①删… II ①布… ②郑… III ①回归分析-研究
Ⅳ ①O212.1

中国版本图书馆CIP数据核字(2012)第12655号

责任编辑 王亚丽

格致方法·定量研究系列

删截、选择性样本及截断数据的回归模型

[英]理查德·布林 著

郑冰岛 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
(20001 上海福建中路133号21层)



编辑部热线 021 63911988

市场部热线 021 63911081

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 1
字数 75,000
版次 2012年7月第1版
印次 2012年7月第1次印刷
ISBN 978-7-5432-2114-7 C·66
定价 15.00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,硕士研究生贺光桦、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寿;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、贺光桦、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

本书是作者多年从事社会科学研究工作的经验总结，也是作者多年从事社会科学研究工作的经验总结。

在非实验社会科学研究中，回归分析是最常用的方法。在数据收集和录入以后，研究者无一例外地开始尝试回归模型，对其定义的等式使用最小二乘法(OLS)进行估计。但OLS这一强大的工具却并不总是正确的。其一便是某类特殊形式的数据可能导致OLS估计量的偏误。布林(Breen)教授在本书中讨论的数据形式包括删截(censored)数据、选择性样本(sample-selected)数据以及截断(truncated)数据。麻烦的是，该领域的术语运用并不统一，但相信本书的例子会帮助我们澄清这些概念。

假设城市政策学者芭芭拉·布朗(Barbara Brown)希望研究这一问题：为何美国城市比其他城市在空气污染控制上花费更多？她以 Y 表示其因变量污染控制开支，并以 X_1 到 X_{10} 表示各城市从预算到社会经济指标的十项解释变量，然后从标准城市年鉴中搜集数据。设想第一种情况：在其城市样本中，年度污染开支只在超过10万美元时才被记录在案，否则就是缺失值——即 Y 是截断的。然而由于 X 并未被截断，而是包含所有城市的信息，因而构成删截样本。若布朗博士

仍然使用 OLS 方法去估计模型,则结果如何呢?为构成数据集,她只能使用 $Y > 10$ 万美元的个案,或者她可以对所有无记录的城市假设一个小于 10 万美元的取值,如 9 万美元。无论怎样处理,OLS 都会提供有偏的参数估计。

在上面的例子中,数据的删截性(censoring)是由于因变量 Y 的截断(truncation)。而另一类更复杂的截断则是由于因变量 Y 的观测受另一变量 Z 影响。我们稍微改动空气污染的例子,以设想第二种情况:假设其他一切条件不变,但年鉴却只包含通过了空气清洁法令的城市。则变量 Z 在通过空气清洁法令时取值 1,未通过则取值为 0。这即为选择性样本问题。布林教授提示通过两个步骤以回应该问题:首先,某城市通过空气清洁法令的概率有多大;其次,在通过空气清洁法令的前提下,城市的污染开支为多少,那么该模型的参数要怎样估计呢,如果不是用 OLS 模型,那么是应该使用 Tobit 模型,还是赫克曼两步骤方法,还是最大似然估计方法呢。布林教授分别对这些估计方法的弱点和优点进行讨论,如以删截回归为例,他解释了最大似然 Tobit 估计一般来说优于赫克曼两步骤方法的原因。

如布林教授所言,删截数据、选择性样本数据以及截断数据涉及“社会科学中的广泛议题”,而詹姆斯·托宾(James Tobin)1958 年的论文引发了对这类议题的现代研究。因此我们的丛书非常需要这样一本关于删截数据的著作。其次,本书也是对丛书中更早的一本《事件史分析》的有效补充,后者主要处理另一种类型的删截数据。

迈克尔·刘易斯—贝克

目 录

序	1
第 1 章 概论	1
第 1 节 删截、选择性样本和截断数据	4
第 2 节 两步模型	7
第 3 节 社会科学中的删截、选择性样本以及截断问题	9
第 4 节 理论基础	12
第 5 节 本书内容	15
第 2 章 删截数据的 Tobit 模型	17
第 1 节 删截的潜在变量	19
第 2 节 两步骤模型	21
第 3 节 最大似然估计	25
第 4 节 Tobit 模型的最大似然估计	32
第 5 节 Tobit 模型的参数解释	37
第 6 节 一个实际例子	41
第 3 章 选择性样本模型和截断回归模型	43
第 1 节 选择性样本模型	45
第 2 节 参数解释	54

第3节	一些实际问题	56
第4节	实证例子	58
第5节	截断回归模型	60
第4章	基本模型的扩展	63
第1节	多重阈值的選擇过程	66
第2节	内生性选择和结果	71
第5章	应注意的问题	75
第1节	对分布假设的敏感性	77
第2节	模型辨识和稳健性	84
第3节	评估研究中的样本选择模型	88
第4节	删截模型和选择性样本模型的使用指南	91
第5节	结论	94
附录		95
注释		100
参考文献		102
译名对照表		108

第 **1** 章

概 论

请考虑如下问题。某次校级考试的及格成绩为 40%，所有参加考试的学生皆被授予证书，但只有及格的学生才会同时获知考试成绩。某位研究考试成绩之影响因素的社会学家抽出一部分学生样本，试图考察一系列解释变量诸如阶级、性别、父母教育程度对学生考试成绩的影响。但其关于学生考试成绩的信息来自学生自己的考试证书。因此若以 y_i 表示第 i 位学生的考试成绩，则仅当 $y_i > 39$ 时，研究者才会得知学生的具体分数。否则（对于那些考试未及格的学生），研究者仅仅知道 $y_i \leq 39$ 。因而研究者面临这样的问题：如何使用这种样本数据去估计考试成绩和解释变量之间的关系？有两种简单的办法。一是使用最小二乘法（OLS）对 y 进行所有解释变量的回归，该方法使用所有样本，并且对所有不及格的学生指定其 $y = 39$ ^[1]。这种方法有许多不妥之处，而其中最重要的是 OLS 的回归系数（它本应告诉我们 y 和解释变量之间的关系）显然是总体真值的偏误估计。

第二种解决办法是仅仅使用 $y > 39$ 的样本信息对 y 进行 OLS 回归。但这种方法不仅舍弃了 $y \leq 39$ 的所有样本信息，而且由于其估计来源于一个并不是随机选择的子样本，

因而不能很好地代表总体。此处的 OLS 估计同样是总体参数的偏误估计。虽然并不那么显而易见,但更重要的是, OLS 回归系数甚至也不是 $y > 39$ 的部分总体的无偏误估计(第 2 章将做解释)。

第 1 节 | 删截、选择性样本和截断数据

为了解决这一问题(这也是本书要讨论的方法),我们需要采取两个步骤。首先是测量个体通过考试的概率。换言之,我们使用一系列相关的解释变量来拟合 y 大于 39 的概率,即 $\text{pr}(y > 39)$ 。然后我们再使用一系列相关变量,拟合通过者的期望成绩,即 $E(y | y > 39)$,其中 E 代表期望值。在模型拟合中,这两个步骤可以分开进行,也可更有效率地共同进行。

我们描述的此例在统计学文献中被称做删截样本问题。我们可以引入一些名称来更准确地说明其含义。若对于随机变量 y 有某数值 c ,对于 $y > c$ 的所有样本,我们知道 y 的确切数值,但对于其他样本,我们则仅仅知道 $y \leq c$,则称为由下截断(左截断)。这正是我们开始时使用的例子所描述的情况。同时我们还有由上截断(右截断),表示我们知道所有 y 小于某一阈值 c 时 y 的确切值,但对于所有其他样本,我们仅知道 $y \geq c$ 。收入是一个典型的例子,对于样本中的高收入群体,我们可能仅仅知道其年收入是 10 万美元或以上。若存在两个或更多阈值,则还有可能出现多截断的情况。如两个阈值 $d > c$,若 $c < y < d$,则已知 y 的具体数值;而当 $y \leq c$ 时, $y \leq c$ 即为全部已知信息;而对 $y \geq d$,我们仅知