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

12

格致出版社　上海人民出版社

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)理查德·布林著;郑冰岛译. —上海:格致出版社:上海人民出版社,2018.6
(格致方法·定量研究系列)
ISBN 978-7-5432-2865-8

Ⅰ. ①删… Ⅱ. ①理… 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(2018)第 089170 号

责任编辑 张苗凤

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　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
字　　数　77,000
版　　次　2018 年 6 月第 1 版
印　　次　2018 年 6 月第 1 次印刷
ISBN 978-7-5432-2865-8/C·200
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

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，本书也是对丛书中另一本《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·S.刘易斯-贝克

目 录

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 $y_i \leqslant 39$。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 $y \leqslant 39$ 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章将作解释）。

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 $\text{pr}(y>39)$。然后我们再使用一列相关变量,拟合通过者的期望成绩,即 $E(y \mid y>39)$,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 $y \leqslant c$,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 $y \geqslant c$。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 $y \leqslant c$ 时,$y \leqslant c$ 即为全部已知信息;而对 $y \geqslant d$,我们仅知

$y \geqslant d$。例如高收入和低收入都被截断的例子。

假设我们有一个截断 y 的样本，其中包含一系列变量 x_k，$k=1, \cdots, K$，而 y 是 x_k 的函数。则 x_k（简写为 x）是以 y 为因变量的回归分析中的解释变量。若对所有样本我们都有 x 的观察值，则样本称做删截的。所以在左删截的样本里，我们既能获得所有 $y>c$ 的 x 值（其中 y 有确切值），也可知道 y 小于或等于 c 时的 x 值。相反，如果仅仅对那些 y 有确切值的样本，其 x 才被观察到，则该样本称作截断的。在这种情况下，对于 y 缺乏具体取值的样本，我们没有任何信息。

现在我们对截断的随机变量，以及含有这类变量的整体样本数据进行区分。后者可以是一个删截样本，即使 y 落入其截断区域，我们也有样本的部分信息；它亦可是一个截断样本，当 y 落入截断区域时则我们不具备任何样本信息。此处我们使用了与赫克曼（Heckman，1992：205）相同的术语名称，但在文献中，这类术语的使用却并不一致：类似删截随机变量的说法相当常见，其中 c 被称为“删截”（而不是截断）阈值。但我认为名称反而是第二位的，读者理解删截数据和截断数据的不同才是重点。

接下来我们将区分两大类删截样本，它们之间的区别在于决定因变量 y 是否具有确切观察值的机制有所不同。在类似本书列举的第一个例子的一般删截问题里，y 的观察值的特性取决于其本身，例如大于阈值 c。但在选择性样本问题中（Heckman，1979），y_i 是否能被确切地观察，取决于另一变量 z_i 的值。我们可以举一个简单的例子，比如成年人给予其孩子零花钱的数额（y）。因为不是所有的成年人都有孩子，所以在一个子样本中，我们不具备 y 的观察值。若以