上海社会科学院哲学社会科学
创新工程学术前沿丛书·第二辑

主 编：黄仁伟 叶 青

统计学理论前沿

编 著：朱平芳 等

上海社会科学院出版社

上海社会科学院哲学社会科学创新工程学术前沿丛书·第二辑

主　编：黄仁伟　叶　青

统计学理论前沿

编　著：朱平芳　等

上海社会科学院出版社

图书在版编目(CIP)数据

统计学理论前沿/朱平芳等编著.—上海:上海社会科学院出版社,2016

ISBN 978-7-5520-1430-3

Ⅰ.①统… Ⅱ.①朱… Ⅲ.①统计学-研究 Ⅳ.①C8

中国版本图书馆 CIP 数据核字(2016)第 136795 号

统计学理论前沿

编　　著:朱平芳　等
责任编辑:杜颖颖
封面设计:黄婧昉
出版发行:上海社会科学院出版社
上海顺昌路 622 号　邮编 200025
电话总机 021-63315900　销售热线 021-53063735
http://www.sassp.org.cn　E-mail:sassp@sass.org.cn
照　　排:南京理工出版信息技术有限公司
印　　刷:上海颛辉印刷厂
开　　本:710×1010 毫米　1/16 开
印　　张:15
插　　页:2
字　　数:244 千字
版　　次:2016 年 7 月第 1 版　2016 年 7 月第 1 次印刷

ISBN 978-7-5520-1430-3 / C·111　　定价:60.00 元

上海社会科学院哲学社会科学
创新工程学术前沿丛书·第二辑

编审委员会

上海社会科学院哲学社会科学创新工程学术前沿丛书概述

（代序）

当前，社会科学领域正面临大量理论和实践问题，需要理论界的证明和创新。上海社会科学院在“创新工程”的机制下，结合研究生教学和高端智库建设方向，于2015年初正式启动《上海社会科学院哲学社会科学创新工程学术前沿丛书》项目（下称“丛书”）。本丛书力图反映本学科最新研究成果和理论探索前沿，为研究生理论积累和博士阶段学习提供引导，同时也为授课教师提供基础性材料。

此次组织出版的丛书为2015年院“创新工程”和研究生院共同资助的第一批集中成果。丛书以马克思主义、毛泽东思想、邓小平理论、“三个代表”重要思想、科学发展观和“四个全面”为根本指导思想，以我院首批38个创新团队为骨干编辑撰写。整个申报评审过程秉承了院“创新工程”公开竞争，择优选取、差额资助原则，所有立项申请均委托第三方组织评审，根据申报质量进行差额资助，确定通过名单向全院公示。为确保编撰质量，成立了院领导牵头、各研究所所长组织、创新团队首席专家领衔、院部相关处室协同“四位一体”的组织框架和工作机制，为丛书的顺利出版提供了保障。

在此基础上，2015年到2018年期间，我院将组织编写出版60部左右上海社会科学院创新工程学术前沿重点教材，作为上海社科院“创新工程”建设中的重要成果展示平台，也为建院60周年献上一份厚礼。整个项目将分两阶段陆续完成。第一阶段，第二至四年，每年编辑、审定和正式出版学术前沿教材15本左右；第二阶段，后一至二年，结合院“创新工程”各团队标志性成果，新增若干部国内顶级、国际一流的重要系列成果，并对已经出版的前期学术前沿进行必要修订与再版。

本丛书得到王战院长、于信汇书记的全面指导，黄仁伟副院长和叶青副院长策划监督执行，参与本次组织工作的人员包括：朱平芳、佘凌、胡晓鹏、汤蕴懿、王晓丰、杨璇。

上海社会科学院哲学社会科学创新工程
学术前沿丛书编委会
执笔：汤蕴懿
2016 年元月 1 日

引　言

统计学作为一门应用型学科，其发展受其在经济社会等应用领域面临的新问题、新思路、新办法的推动。同时，统计学理论也在这个过程中得以不断更新和拓展。近年来，随着信息技术在数据处理领域的广泛应用，以及自身方法的不断创新和完善，统计学理论与应用得到了跨越式的发展。在统计学研究领域中，数据、统计指标、估计方法和假设检验是四大重要内容。本书围绕这四个方面，撷取了一些在国内外统计学理论和应用领域前沿中广受关注的焦点，逐一介绍。这些前沿内容包括二元离散选择模型和面板数据的非线性模型的非(半)参数估计方法、分位数理论前沿方法和应用、金融高频数据的统计分析、大数据分析和机器学习技术和科技统计与方法应用，以飨读者。

具体而言，本书共有五章：

第一章介绍二元选择模型的各种估计模型和相应的估计方法。针对现实研究中离散二元选择数据建模中存在的内生性问题，对非参数面板二元选择模型的新估计方法进行介绍。首先介绍二元选择变量的内生性非参数模型，并提出估计方法；其次，将该方法推广到面板二元选择模型；然后，对估计量的一致性进行论证，推导其极限分布；最后，用随机模拟的方法给出小样本性质。

第二章主要介绍分位数回归的前沿方法，包括贝叶斯分位数回归估计方法、分位数处理效应方法和无条件处理效应方法。在已有的分位数估计方法中，频率学派估计方法的小样本性质表现不佳，且收敛速度较慢，而贝叶斯估计方法正好能够弥补这些缺陷，所以贝叶斯分位数估计方法在实际应用中受到了广泛的重视。随着数据可得性的提高，分位数上的异质性影响备受关注，而且在均值回归中的一些方法和思想正在逐步地向分位数上推广，分位数方法的应用也在不断地深入和拓展。在均值回归中，常用平均处理效应来分析一项政策或者干预对结果变量的因果效应，如果要考察不同分位数上政策干预的影响，则要估计分位数处理效应。在分位数处理效应的基础上，本章还将介绍无条件分位数处理效应估计方法，该方法得到的处理效应更符合政策评价的经济学解释。

第三章介绍金融统计领域中有关波动率、资产价格过程、资本市场高频数

据的处理和建模方法方面一些新发展。在瞬息万变的资本市场，采用金融高频数据，基于一阶矩、二阶矩建模的收益率、波动率的变化规律是什么，如何通过高频数据识别金融市场微观结构，一直是资本市场关注的主要议题。就此，本章将对金融数据中的跳跃行为、波动率，以及微观结构噪声估计的前沿方法和应用进行介绍。

第四章是关于大数据分析技术的研究综述。互联网、物联网、云计算技术的快速发展，各类应用的层出不穷引发了数据规模的爆炸式增长，使数据渗透到了当今每一个行业，以及相应的业务领域，成为重要的生产因素。大数据因此也成为社会各界关注的新焦点，大数据时代已然来临。如何运用现有的统计技术进行数据可视化分析和属性的选择，如何用统计测试方法对大多数机器学习模型进行验证与评估，是应用大数据进行统计分析的前沿领域。本章将从大数据分析的应用与发展趋势出发，对大数据前沿统计分析技术和应用技术，以及大数据分析在中国的研究现状和前景进行介绍，期望对现有统计和计量方法有所启发。

第五章介绍统计学的分支学科科技统计技术的前沿理论和应用研究。科技统计是用统计的方法对科学技术活动的规模和结构进行定量测定的研究。它是用一套可以有效测度国家科技系统复杂机制的指标，对一国范围内科技活动的规模、结构及功能进行连续的年度数量测定，为国家科技政策的制定与评价提供准确、系统的年度科技统计数据及统计分析报告。科技统计研究的对象是科学技术活动的总体的数量特征和数量关系，其主要任务是通过对科技活动的有关数据的收集、处理、分析，反映科技活动的规模、结构和布局的总体数量特征和关系，从而为评价和制定科技政策和发展规划提供依据。本章将对国内外科技统计相关的前沿技术方法、指标体系构建方法、R&D 活动科技统计、科技进步与经济增长关系、科技投入与产出效率评价、中外对比等方面进行介绍。

编著本书的目的是，为统计学专业的研究生、研究人员提供现代统计学的前沿方法和研究思路，以便迅速掌握该领域的前沿动态，并开展相应的理论研究和应用研究。因此，我们在本书的编著过程中，会突出三个方面的内容。一是该领域内的研究综述。通过综述以往的研究成果，厘清研究脉络，便于读者对该领域有一个全局的把握；尤其是在综述的基础上指出有待进一步研究的内容，为读者提供可选的研究方向。二是介绍该领域内前沿的核心方法。从建模、统计量的构造、渐进分布、蒙特卡罗模拟等基本的手段和工具入手，对该领域的前沿理论模型和估计方法有一个清晰的认识和理解。三是穿插一些经典文献研究。在每一个前沿理论领域都聚集了众多有趣的应用研究，我们会着重介绍一些具有代表性的论文，从问题的提出、研究的背景、建模的思路，以及技术细节上的处理技巧等方面进行分析，使读者能够对前沿方法的应用研究有深入的理解，举一反三，触类旁通。

目　录

第一章　非参数面板二元选择模型的内生性研究

本章我们将介绍二元选择模型的各种估计模型和相应的估计方法。针对现实研究中存在的内生性问题，提出非参数面板二元选择模型的估计方法，推导极限分布，最后用随机模拟的方法给出小样本性质。

第一节　非参数面板二元选择模型及内生性

一、参数二元选择模型

二元选择问题在实证经济学研究中经常出现，例如，人们关于婚姻状态的选择，是否读研究生，是否购置房屋，是否出游，是否换工作等；公司选择是否上市，是否开发新产品，是否投资某项目等；政府选择是否加息，是否提高税率，是否增加信贷等。类似这些问题的研究都可以用二元选择模型来刻画，令 y 代表被解释变量，y 可能的取值是 0 或 1。例如，如果选择结婚 $y=1$，不结婚 $y=0$。在二元选择模型中，我们通常感兴趣的是做出某一选择的概率

$$P(y=1 \mid x)=P(y=1 \mid x_1, x_2, \cdots, x_k),$$

其中 x 是解释变量。例如，如果 y 表示个人是否选择工作，x 可能包含各种影响个人做出该选择的特征，如教育、年龄、婚姻状态以及其他影响就业的二元随机变量，如是否最近参加就业培训项目、过去是否有犯罪记录等。

二元选择模型可以利用线性概率模型(Linear Probability Model)来估计，也就是令

$$P(y=1 \mid x)=x'\theta,$$

然而除非严格限制 x 的范围，线性概率模型并不能很好地描述选择概率 $P(y=1 \mid x)$，因为对于给定的参数 θ，总存在 x 使得选择概率在区间[0，1]之外。另外，线性概率模型意味着其他变量保持不变的条件下，自变量每增加一单位，无论自变量的初值大小，$P(y=1 \mid x)$ 总对应变化相同的量，且自变量的持续增加将会使选择概率低于 0 或大于 1。线性概率模型的这些特征限制了其在阐述实际问题中的应用，因而研究者们又陆续提出了新的解决办法。

Domencich 和 Mcfadden(1974)提出的基于效用最大化的二元选择模型，将经济学理论引入选择问题，为微观计量模型的发展做出了先驱性的工作，推动了二元选择模型的广泛应用。具体来讲，令

$$y^* = x'\theta + u,\ y = 1\{y^* > 0\}, \tag{1.1}$$

其中 y^* 通常代表个人的效用函数，是潜在变量，只有 y 能够被研究者观测到。假定随机误差项 u 与解释变量 x 相独立，且概率密度函数关于原点对称，如果记随机误差项 u 的概率分布函数为 G，则

$$P(y=1 \mid x) = P(u > -x'\theta) = 1 - G(-x'\theta) = G(x'\theta),$$

令 (y_i, x_i)，$i=1, 2, \cdots, n$ 代表总体(1.1)独立同分布的样本，则样本的极大似然函数为

$$\prod_{i=1}^{n} [G(x_i'\theta)]^{y_i} [1 - G(x_i'\theta)]^{1-y_i},$$

通过极大化样本的似然函数可得未知参数 θ 的估计。如果 e 服从标准正态分布函数，则模型(4.1)就是 Probit 模型，如果 e 服从 logistic 分布，即

$$G(z) = \frac{\exp(z)}{1 + \exp(z)},$$

则模型(1.1)就是 Logit 模型。

二、半参数二元选择模型

含有参数的极大似然估计和其他参数估计方法要求设定误差项 u 的概率分布。然而，一旦误差项的分布假定不成立，将会导致估计有偏。为了解决这些问题，不需要假定误差项 u 的分布类型的半参数估计方法应运而生。Cosslett(1983)，Han(1987)，Stoker(1986)提出了估计未知参数 θ 的相合半

参数估计方法。但是前两种方法的极限分布未知，第三种方法对解释变量 x 的分布做了参数化假定。Ichimura(1988)，Klein 和 Spady(1989)，Powel 等(1986)，Ai(1997)也提出未知参数 θ 的半参数 $\sqrt{N}$ 相合渐近正态估计。Manski(1975，1985)发现 $x_i'\theta$ 与 $E(y^* \mid x)$ 符号相同。基于以上发现，在假定随机误差项 u 关于解释变量 x 的中位数为 0 条件下，可通过极大化如下样本得分函数得到未知参数的估计

$$\hat{\theta} = \arg\max_{\theta'\theta=1} \frac{1}{n}\sum_{i=1}^{n} y_i \times \operatorname{sgn}(x_i\theta),$$

其中 sgn(·)表示符号函数。相对于 Cosslett(1983)，Han(1987)，Stoker(1986)，Ichimura(1988)，Klein 和 Spady(1989)，Powel 等(1986)，Ai(1997)的方法，Manski 的方法假定更宽松允许异方差，而且计算更加简便。然而，由于得分函数中的符号函数是阶梯函数，不可导，因此，最大得分函数的极限分布也无法推导，Horowitz(1992)通过光滑 Manski 的得分函数修正了最大得分函数估计，即

$$\hat{\theta} = \operatorname{argmax} \frac{1}{n}\sum_{i=1}^{n} [2 \cdot 1(y_i = 1) - 1] K\left(\frac{x_i'\theta}{h_n}\right)$$

其中 $K(\cdot)$类似于分布函数，有界，而且 $\lim\limits_{\tau \to -\infty} K(\tau) = 0$，$\lim\limits_{\tau \to \infty} K(\tau) = 1$，$h_n$ 代表窗宽。

类似于 Manski 的估计，光滑最大得分函数假定误差项 u 分布未知，允许误差项 u 的分布以未知的方式依赖于解释变量 x(允许异方差)，光滑最大得分函数可导，因此可以利用泰勒展开的方法得到其极限正态分布。

三、半参数面板二元选择模型

我们从面板二元选择模型开始：

$$y_{it} = 1\{x_{it}'\theta_0 + c_{1i} + u_{it} \geqslant 0\},\ t = 1,\ 2,\ i = 1,\ 2,\ \cdots,\ n。 \qquad (1.2)$$

其中 x_{it} 代表解释变量，u_{it} 代表未观测的误差项，c_{1i} 代表未观测到的个体效应项。注意到，对任意两个时间点 t 和 s，简单的差分有：

$$y_{it} - y_{is} = 1\{x_{it}'\theta_0 + c_{1i} + u_{it} \geqslant 0\} - 1\{x_{is}'\theta_0 + c_{1i} + u_{is} \geqslant 0\},$$

取期望，我们有：

$$E(y_{it}-y_{is}\mid x_{it},x_{is},c_i)=\Pr(u_{it}\geqslant -x'_{it}\theta_0-c_i\mid x_{it},x_{is},c_i)$$
$$-\Pr\{u_{is}\geqslant -x'_{is}\theta_0-c_i\mid x_{it},x_{is},c_i\}。$$

显然,简单的关于时间的差分并不能消掉个体效应项,除非 $x'_{it}\theta_0=x'_{is}\theta_0$。Manski(1987)观察到上式右端的概率差与 $x'_{it}\theta_0-x'_{is}\theta_0$ 具有相同的符号,$y_{it}-y_{is}$ 与 $\mathrm{sgn}(x'_{it}\theta_0-x'_{is}\theta_0)$ 正相关。基于上述发现,Manski(1987)提出了面板二元选择模型的最大得分函数估计:

$$\hat{\theta}=\arg\max_{\theta'\theta=1}\frac{1}{n}\sum_{i=1}^{n}\sum_{s<t}(y_{it}-y_{is})*\mathrm{sgn}(x'_{it}\theta_0-x'_{is}\theta_0),$$

Manski(1987)不需要假定随机误差项的具体分布,只要求关于 x_{it},x_{is},c_i,u_{it} 与 u_{is} 分布相同,而且允许异方差。在一定的条件下对参数正则化,Manski(1987)说明最大得分函数估计相合,然而,该估计不是 $\sqrt{N}$ 相合,而且极限分布非正态。如果在这里运用 Horowitz(1992)的光滑技术,得到的估计极限分布是正态的,但收敛速度仍然低于 $\sqrt{N}$。面板二元选择模型的其他估计方法见 Honore 和 Kyriazidou(2000),Anderson(1970),Chamberlain(1993),以上这些估计方法都要求解释变量与随机误差项独立。对于具有内生解释变量的面板二元选择模型,目前文献还没有提供满意的估计方法,本文将填补这一研究空白。

四、内 生 性

具有内生的解释变量的问题是我们讨论的重要内容。下面,我们从最基本的线性模型开始讨论内生性及处理内生性的方法。考虑如下线性模型

$$y=x'\beta+u,\tag{1.3}$$

如果 $E(xu)=0$,$E(xx')$ 满秩,则最小二乘估计(OLS)是模型(1.3)的相合估计。因为解释变量 x 包括常数项,假设 $E(xu)=0$ 相当于 $\mathrm{cov}(x,u)=0$ 即解释变量与随机误差项不相关。一旦该假设不成立,即 $\mathrm{cov}(x,u)\neq 0$,则称解释变量 x 内生(endogenous variables),内生性(endogeneity)通常来源于被忽略的解释变量,测量误差,联立性等。例如研究教育对收入的影响时,令

$$y=\beta_0+x_1\beta_1+x_2\beta_2+\cdots+x_k\beta_k+\lambda q+e,\tag{1.4}$$

其中 y 代表收入，$(x_1, x_2, \cdots, x_{k-1})$ 代表影响收入的解释变量(例如经验，孩子个数，年龄等)，x_k 代表教育水平，q 代表个人能力，但个人能力是不能观测的。假设 x_k 与 q 之间有如下关系，

$$x_k = \delta_0 + \delta_1 q + v$$

将上式带入(模型 4.4)，我们实际使用的估计模型是

$$y = \left(\beta_0 - \frac{\lambda\delta_0}{\delta_1}\right) + x_1\beta_1 + x_2\beta_2 + \cdots + x_k\left(\beta_k + \frac{\lambda}{\delta_1}\right) + u,$$

其中 $u = -\frac{\lambda}{\delta_1}v + e$，$x_k$ 与 v 相关进而与误差项 u 相关。如果不考虑(模型 1.4)中 x_k 的内生性，实际得到是参数 $\beta_k + \frac{\lambda}{\delta_1}$ 的估计，这将过高估计教育对收入的影响。

处理内生性的常用方法包括工具变量方法(instrument variables approach)和控制函数方法(control function approach)。令模型(1.3)中的 $x' = (x_1', x_2')'$，其中 $E(x_1 u) = 0$，$E(x_2 u) \neq 0$，即 x_2 是内生的解释变量。所谓工具变量的方法是假设存在一组工具变量 $z = (x_1', z_2')'$，即

$$E(zu) = 0,$$

则

$$0 = E(P[x \mid z]u) = E(\Pi' zu) = E(\Pi' z(y - x'\beta))$$

其中 $P[x|z]$ 是 x 关于 z 的总体最小二乘投影，也就是

$$\Pi = \{E(zz')\}^{-1}E(z'x)$$

用样本均值代替模型(1.3)中的总体均值，则有两阶段最小二乘估计(2SLS)，

$$\hat{\beta}_{2SLS} = (\hat{X}'\hat{X})\hat{X}'Y$$

其中

$$\hat{X} = Z\hat{\Pi},\ \hat{\Pi} = (Z'Z)^{-1}Z'X,$$

X，Z，Y 分别为对应于 x，z，y 的数据矩阵。当用非线性参数形式的残差项 $u = m(y, x, \beta)$ 代替线性形式的残差项 $u = y - x'\beta$，并假设

$$E(u \mid z)=0,$$

运用工具变量的方法可以得到 Sargan(1959)的广义工具变量估计(generalized instrumental variable estimator)，Amemiya(1974)的非线性两阶段最小二乘估计(nonlinear two-stage least squares)及 Hansen(1982)的广义矩估计(generalized method of moments)。

多数线性模型的估计采取上述标准工具变量方法，或者两阶段最小二乘估计，又或者采用广义矩估计方法。控制函数方法是另外一种处理内生性的方法。具体来讲，具有内生解释变量的线性模型(1.3)，类似于两阶段的最小二乘估计，令

$$x_2=\Pi_2 z+v,$$

$$E(zv)=0,$$

x_2 内生源于 v 和 u 相关，记 u 关于 v 的投影为

$$u=\rho' v+e, \tag{1.5}$$

这里内生的解释变量可能不止一个，因此设 v 是一个向量。根据定义 $E(ve)=0$，因为 v 和 u 均与 z 不相关，$E(ze)=0$。

将(1.5)带入模型(1.3)，有

$$y=x'\beta+\rho' v+e \tag{1.6}$$

将 v 看作附加的解释变量。注意到 e 与 v，x 均不相关，(1.6)说明以 x 和 v 为解释变量关于 y 回归可以得到未知参数的相合估计。唯一的问题是我们这里并未观测到 v，我们可以利用 v 的估计量代替，令

$$\hat{v}=x_2-\hat{\Pi}_2 z。 \tag{1.7}$$

在线性模型中，利用控制函数的方法也可以得到两阶段的最小二乘估计，但对非线性模型，控制变量方法和工具变量方法通常不同。对于非线性模型，控制变量的方法有很多优点。例如，内生性的检验更加方便。控制函数的主要思想是利用工具变量对解释变量回归的残差项控制内生性(Newey, Powell, & Vella, 1999; Blundell & Powell, 2003; Chesher, 2003; Das, Newey, & Vella, 2003; Blundell & Powell, 2004; Florens, et al. 2008; Imbens & Newey, 2009; Blundell & Powell, 2007; Lee, 2007)。控制函数

方法的缺点是内生解释变量必须连续，但是其他变量包括工具变量可以离散。本文采用控制函数方法处理面板二元选择模型中的内生性。

五、具有连续内生解释变量的二元选择模型

现有文献为具有连续内生解释变量的二元选择模型提供了很多种估计方法（Heckman，1978；Nelson & Olson，1978；Amemiya，1978；Newey，1985；Blundell & Smith，1989；Newey，1987；Rivers & Vuong）。这些估计方法的共同特点是对随机误差项及估计方程做参数化假定。参数化方法中，Rivers 和 Vuong(1988)的两阶段极大似然估计因为计算和内生性检验相对简便，最为常用。

考虑模型(1.1)，令 $x'=(x_1', x_2')'$，如果 $E(u \mid x_1)=0$，$E(u \mid x_2) \neq 0$，则称 x_2 是内生的解释变量，令

$$x_2=\Pi_2 z+v, \tag{1.8}$$

其中 $z=(x_1', z_2')'$ 为 x_2 的工具变量，即 $E(v \mid z)=0$。Rivers 和 Vuong(1988)假定工具变量 z 具有有限的正定协方差矩阵 Σ_{zz}，u 和 v 关于工具变量 z 均值为 0 的协方差矩阵

$$\Omega \equiv \begin{pmatrix} \sigma_{uu} & \Sigma_{uv} \\ \Sigma_{vu} & \Sigma_{vv} \end{pmatrix}$$

联合正态分布，则

$$u=v'\rho+e, \tag{1.9}$$

其中 e 关于 z，x_2 服从 $N(0, \sigma_{uu}-\rho'\Sigma_{vv}\rho)$，并做如下正则化约束 $\sigma_{uu}-\rho'\Sigma_{vv}\rho=1$。将(1.9)带入模型(1.1)有

$$y^*=x'\theta+v'\rho+e,\ y=1\{y^*>0\}。$$

则 y 关于 z，x_2 的联合概率密度函数为

$$f(y \mid x_2, z)=\Phi(x'\theta+\rho'v)^{y_i}\left[1-\Phi(x'\theta+\rho'v)\right]^{1-y_i} \tag{1.10}$$

Rivers 和 Vuong(1988)的两阶段估计方法可表述为，第一步由 z 关于 x_2 回归得到误差项 v 的估计量 $\hat{v}$。第二步将 $\hat{v}$ 带入(1.10)，通过极大化(1.10)