CAMBRIDGE

经济科学译库

面板

Analysis of Panel Data
(Second Edition)

面板数据分析（第二版）

萧 政 /著
Cheng Hsiao

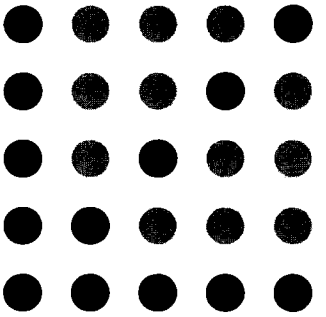
李 杰 /译

中国人民大学出版社

经济科学译库

面板数据分析

（第二版）



萧 政 /著
Cheng Hsiao

李 杰 /译

Analysis of Panel Data
(Second Edition)

中国人民大学出版社
·北京·

图书在版编目（CIP）数据

面板数据分析/萧政著；李杰译．—北京：中国人民大学出版社，2012.12
（经济科学译库）
ISBN 978-7-300-16708-4

Ⅰ.①面… Ⅱ.①萧… ②李… Ⅲ.①计量经济模型 Ⅳ.①F224.0

中国版本图书馆 CIP 数据核字（2012）第 282499 号

/

经济科学译库
面板数据分析（第二版）
萧 政 著
李 杰 译
Mianban Shuju Fenxi

出版发行	中国人民大学出版社		
社　　址	北京中关村大街 31 号	**邮政编码**	100080
电　　话	010－62511242（总编室）	010－62511398（质管部）	
	010－82501766（邮购部）	010－62514148（门市部）	
	010－62515195（发行公司）	010－62515275（盗版举报）	
网　　址	http://www.crup.com.cn		
	http://www.ttrnet.com（人大教研网）		
经　　销	新华书店		
印　　刷	三河市汇鑫印务有限公司		
规　　格	185 mm×260 mm　16 开本	**版　　次**	2012 年 12 月第 1 版
印　　张	20　插页 2	**印　　次**	2012 年 12 月第 1 次印刷
字　　数	436 000	**定　　价**	45.00 元

中文版序

“面板数据”（又称“纵列数据”）是由大量个体的时序观测构成的数据集。也就是说，面板数据集包含样本中各个体的多个观测值。与只在给定时点有一次观测值的横截面数据集或只有单个对象各时期观测值的时序数据集相比较而言，面板数据集具有很多优势，譬如自由度更多，更能有效降低解释变量间的共线性程度，以及得到精度更高的参数估计量。更重要的是，由于面板数据融合了个体间差异和个体内在动态信息，故很多复杂的行为假设——这些行为假设比用横截面数据集或时序数据集研究的行为假设复杂得多——都可以用面板数据进行研究。面板数据一般重点关注个体结果，而影响个体结果的因素何其之多，但经济模型或计量模型不是对现实的完美镜像，而是对现实世界的简化分析，故在模型中引入所有的影响因素既不可行也无必要。面板数据分析面临的一个重要挑战是如何控制各横截面单元和各时期不可观测的异质性因素以得到对可观测因素影响的推断。本书的目的就是简要介绍面板数据分析中的这些基本问题，重点在于如何应用面板数据的模型和方法。我希望本书的中文版有助于推动这一重要领域的研究。

感谢译者李杰将本书翻译为中文。虽然我没有完整地阅读中文译稿，但我们之间有过多次交流，并且他发现了英文原版中我没有注意到的许多错误，所以我有理由相信本书中文版的翻译质量是可靠的。

萧　政

2012 年 3 月 2 日

第二版序

自 1986 年本书第一版出版以来，讨论面板数据的文章显著增多。根据社会科学引文索引（Social Science Citation Index，SSCI）的数据，1989 年有 29 篇文章与面板数据有关，而在 1997 年这个数字是 518，1998 年是 553，1999 年是 650。面板数据日益受到重视，一方面是因为借助面板数据集比利用单纯的横截面数据集或时间序列数据集能更好地回答重点关注的问题，且这种数据越来越容易获取；另一方面是因为研究人员计算能力的快速提升；当然更是受到了该领域内在方法逻辑体系发展的推动［参见 Trognon（2000）］。

新版在第一版的基础上进行了重大修订，主要补充了离散选择（第 7 章）和样本选择（第 8 章）等非线性面板数据模型的内容；新增的第 10 章包括模拟技术、大 N 和大 T 理论、单位根及协整检验、多水平结构以及截面相依等若干主题；新增了关于动态模型的估计（4.5 节～4.7 节）、固定系数和随机系数模型的 Bayes 诊断（6.6 节～6.8 节）以及重复横截面数据（或伪面板数据）等若干小节。此外，还更新了原有章节中的若干论述。譬如，引入了严格外生性的概念，在广义矩法框架下介绍估计量，以便建立识别各种模型所必需的假设之间的联系；根据对非观测特异效应假设的约束条件，更新了对固定效应和随机效应的论述；等等。

与第一版一样，新版的目标仍是为分析各种类型的数据提供最新的综合分析

框架，重点在于系统地阐述如何对受重大政策影响的问题进行恰当的统计推断。新版既不是关于面板数据计量经济学的百科全书，也不是这方面的发展史，故有很多重要贡献在本书中都没有提到，对此我深感抱歉。关于面板数据计量经济学发展史的论述可以参见 Nerlove（2000）。一些别的出版物和参考资料可参见 Arellano 和 Honoré（2001）的综述，或者参见近来由 Matyás 和 Sevester（1996）；Hsiao，Lahiri，Lee 和 Pesaran（1999）；Hsiao，Morimune 和 Powell（2001）以及 Krishnakumar 和 Ronchetti（2000）编撰的四卷资料。Blanchard（1996）对相关软件有过评论。

我要感谢编辑 Scott Parris 在准备修订期间对我的鼓励和帮助，还有 Andrew Chesher 和两位匿名读者对初稿的建设性批评。我也非常感谢 E. Kyriazidou 对第 7 章和第 8 章细心而又细致的批评，感谢 S. Chen 和 J. Powell 对第 8 章、H. R. Moon 对大面板这一节有益的批评和建议，感谢 Sena Schlessinger 用她娴熟的技能输入了除第 7 章外的全部手稿，Yan Shen 仔细校对了全部手稿并熟练地输入了第 7 章的手稿，还有 Siyan Wang 绘制了第 8 章的图表。当然，书中所有的遗留错误都由我负责。James Heckman、C. Manski、Daniel McFadden、Ariel Pakes、《计量经济学》（*Econometrica*）、《美国统计学会杂志》（*Journal of the American Statistical Association*）、《计量经济学杂志》（*Journal of Econometrics*）、《区域科学与城市经济学》（*Regional Science and Urban Economics*）、《经济研究评论》（*Review of Economic Studies*）、芝加哥大学出版社、Elsevier Science 等允许我复制其文章中的部分内容，在此一并感谢。

第一版序

近来，经济学领域的实证研究因利用一类样本容量较大的新型数据——不同时期观测到个体的横截面数据——飞速发展。利用这种数据，我们可构建并检验更贴近现实的行为模型，而这些模型仅用单纯的横截面数据集或时间序列数据集却无法识别。但新型数据的利用也产生了新的问题。新方法不断引入，观念也在发生变化。在准备著述一部导论性专著时，作者必须选择所涵括的主题。我选择了控制不可观测的个体和/或时间属性以避免设定偏误并提高估计的效率。虽说作品的选题范围在某种程度上属于作者的个人风格问题，但本书主要还是讨论最基本和最常用的方法。书中也给出了一些应用这些方法的案例，并对（模型的）用途、计算方法以及（对模型的）解释进行了讨论。

非常感谢 C. Manski 和剑桥大学出版社的一位读者，以及 G. Chamberlain 和 J. Ham 建设性的批评与建议。也感谢 Mario Tello Pacheco，他通读全稿并对表述中存在的问题提出了大量的建议，还更正了许多大大小小的错误。还要感谢 V. Bencivenga，A. C. Cameron，T. Crawley，A. Deaton，E. Kuh，B. Ma，D. McFadden，D. Mountain，G. Solon，G. Taylor 和 K. Y. Tsui 等人的有益批评，感谢 Sophia Knapik 和 Jennifer Johnson 耐心地输入初稿以及反反复复的修改稿。当然，此类事务很容易产生错误，请读者就仍然存在的问题对作者提出批评指正。

本书的若干内容是在我与莫雷山的贝尔实验室、普林斯顿大学、斯坦福大学、南加州大学以及多伦多大学的交流过程中完成的。我非常感谢这些机构为我提供文秘和研究上的便利，非常荣幸与那里富有激情的同事们一同工作。感谢美国国家科学基金和加拿大社会科学与人文科学研究委员会的资助。

目　录

第 1 章　导　论

1.1　面板数据的优点

纵列数据集（longitudinal data set），又称**面板数据集**（panel data set），是在不同时期跟踪由给定个体组成的样本而获取的数据集，它包含样本中每个个体的多个观测值。无论是在发达国家还是在发展中国家，面板数据都已经很常见。譬如，美国有两个最著名的面板数据集：NLS（National Longitudinal Surveys of Labor Market Experience）数据集和密歇根大学的 PSID（Panel Study of Income Dynamics）数据集。

NLS 对数据的收集始于 20 世纪 60 年代中期，它由五个不同的纵列数据库组成，涵括了劳动力的不同组成部分：1966 年年龄在 45～59 岁的男子、1966 年年龄在 14～24 岁的青年男子、1967 年年龄在 30～44 岁的女子、1968 年年龄在 14～24 岁的青年女子、1979 年年龄在 14～21 岁的青年男女。1986 年扩展后的 NLS 还包含对参加 1979 年调查的青年组女子所生孩子的调查。调查的变量上千个，重点在于了解劳动力市场的供给信息。表 1.1 对 NLS 调查的组、初始样本

容量、每个组已经被调查的年数，还有每个组当前的调查状态进行了归纳［详细信息见《NLS手册2000》（*NLS Handbook* 2000），美国劳动部，劳动统计局］。

PSID从1968年开始收集数据并持续至今，它在全国范围内收集具有代表性的6 000多个家庭和15 000多位个人的年度经济信息。该数据集有5 000多个变量，包括就业、收入、人力资本，以及住房、是否乘车上班、流动性等方面的信息。除NLS和PSID之外还有一些经济学家感兴趣的其他面板数据，Borus（1981）和Juster（2000）将这些数据编录在目并进行了论述；读者还可参见Ashenfelter和Solon（1982），以及Becketti等人（1988）的工作来了解这些数据。[1]

表1.1　　NLS：受访组、样本容量、调查年份以及调查状态

受访组	年龄组	出生年份组	初始样本容量	起始年份/结束年份	调查次数	最后受调查人数	状态
老年男子	45～59	4/2/07—4/1/21	5 020	1966/1990	13	2 092[1]	结束
成年女子	30～44	4/2/23—4/1/37	5 083	1967/1999	19	2 466[2]	持续中
青年男子	14～24	4/2/42—4/1/52	5 225	1966/1981	12	3 398	结束
青年女子	14～24	1944—1954	5 159	1968/1999	20	2 900[2]	持续中
NLSY79	14～21	1957—1964	12 686[3]	1979/1998	18	8 399	持续中
NLSY79儿童	出生～14	—	—[4]	1986/1998	7	4 924	持续中
NLSY79青年人	15～22	—	—[4]	1994/1998	3	2 143	持续中
NLSY97	12～16	1980—1984	8 984	1997/1999	3	8 386	持续中

[1]1990年的调查对象包含2 206名寡妇或别的近亲故去的受访者。

[2]初始样本容量。

[3]样本中去掉军人（1985年）和经济贫困的非黑人、非西班牙裔（1991年）之后，包含9 964名有效受访者。

[4]NLSY79中儿童和青年人样本容量与NLSY79中女性受访者所生孩子的数量有关，该数量随着时间递增。

资料来源：NLS Handbook，2000，U. S. Department of Labor，Bureau of Labor Statistics.

欧洲的许多国家有年度或更高频率的全国调查数据，如荷兰的SEP（Socio-Economic Panel），德国的GSOEP（German Social Economics Panel），卢森堡的PSELL（Luxembourg Social Economic Panel），英国的BHPS（British Household Panel Survey）等。“为满足欧盟内部对各成员国关于收入、工作与就业、贫穷与社会排斥、住房、健康，以及若干其他有关私有财产和人员居住条件的社会指标等可比较信息不断增长的需求”［Eurostat（1996）］，欧盟统计局在1994年建成的国家数据采集器已经开始用统一设计的标准化多功能年度纵列数据调查表来协调和链接现有的国家面板数据。譬如，分别始于1993年和1995年的MIP（Manheim Innovation Panel）和MIP-S（Manheim Innovation Panel-Service Sector）包含德国制造业和服务业中有5个及以上雇员企业的创新活动（产品创新、创新支出、研发支出、阻碍创新的因素、资本收益、工资和雇员的技能结构等

等）的年度数据。调查严格按照经合组织和欧盟统计局的《奥斯陆手册》（*OSLO Manual*）中推荐的创新调查方法进行，因此获取的德国企业创新活动的数据可在国家之间进行比较。1993 年和 1997 年的数据也成为了欧盟创新调查数据 CIS Ⅰ和 CIS Ⅱ的一部分［详情参见 Janz 等（2001）］。类似地，ECHP（European Community Household Panel）计划推出家庭和个人级别的欧盟人口统计数据。ECHP 包含人口特征、劳动力行为、收入、健康、教育与培训、住房、移民等信息。ECHP 现在覆盖欧共体 15 国中除瑞典之外的 14 国［Peracchi（2000）］。ECHP 的详细统计数据公布在欧盟统计局的参考数据库 New Cronos 中，发布的数据涉及三个领域，即健康、住房以及收入与居住条件（ILC）。[2]

虽然多数发展中国家尚未形成收集统计数据的传统，但面板数据在这些国家也越来越常见。获取原始调查数据对回答许多有意义的重大问题显得尤为重要。世界银行已经发出倡议并协助设计了许多面板数据调查。譬如，中国国务院农村发展研究中心发展研究所与世界银行合作，从 1984 年至 1990 年对中国的 200 家大型乡镇企业进行年度调查［Hsiao 等（1998）］。

在经济研究中，与传统的横截面数据集或时间序列数据集相比，面板数据集具有多方面的优势［参见 Hsiao（1985a，1995，2000）］。面板数据通常能为研究人员提供大量的数据点，因此增加了数据的自由度并降低了解释变量间的共线性程度，故而提高了计量模型估计的有效性。更重要的是，纵列数据允许研究人员分析大量用横截面数据或时间序列数据无法处理的重大经济问题。譬如，我们考虑一个来自 Ben-Porath（1973）的案例：假设从某已婚女士的横截面样本中发现样本中的女士们平均每年有 50%的劳动参与率。对该发现的一种极端解释是：样本中的女士们来自同一个总体，在任意给定的年份，样本中的每位女士正处于就业状态的几率为 50%。而另一种极端解释是：样本中的女士们来自不同的总体，其中的 50%一直在工作，而另外 50%从不工作。在第一种情形，人们预期每位女士婚后一半的时间在劳动力市场之外度过，而另一半时间花在劳动力市场内，并以平均两年的工作期限频繁地变动工作。在第二种情形，女士们没有工作状态的变化，她们的当前工作状态就是将来工作状态的完美预测。要对这两个模型进行区分，我们需要利用个体的劳动力历史信息来估计在生命周期的不同时段参与劳动力市场的概率。这只有在我们对大量的个体有连续的观测时才可能实现。

利用横截面数据对变化的动态情况进行推断时遇到的困难，在劳动力市场研究的其他方面也会碰到。我们以工会组织对经济行为的影响为例进行说明［参见 Freeman 和 Medoff（1981）］。对于有工会的企业（或雇员）和没有工会的企业（或雇员）进行比较而发现的差别，有些经济学家倾向于相信其基本真实性，并认为工会及集体议价方式从根本上改变了雇佣关系中最重要的方面，包括补偿金、劳动力内部与外部的流动性、工作规则和（工作）环境等。而另一些经济学家则认为工会影响不过是些假象而已，他们认为现实世界几乎满足完全竞争的所有条件；上述差别主要根源于这些企业（或工人）在工会成立之前的差异，或工会成立之后在发挥协调作用之前的差异。长远来看，工会无法帮助工人提高工资，因为企业会雇用更高技能的工人来应对所支付的（被工会强迫的）高工资。如果我们认同前一种观点，那么工资或收入方程中表示工会有无的虚拟变量前的

系数就是对工会影响的度量。如果我们接受后一种观点，那么表示工会有无的虚拟变量可能仅仅只是工人技能的代理变量。通常我们不能依靠单纯的横截面数据集在这两种假设之间直接作出选择，因为估计量很可能反映的是不同的人或企业间在任何比较中都存在的个体差异。但利用面板数据，我们就可以研究工人从无工会企业转到有工会企业（或从有工会企业转到无工会企业）后其工资的差异来区分这两种假设。如果我们接受工会（对工资）没有影响的观点，那么工人从无工会企业转到有工会企业后，只要他的工作技能在各时期保持不变，则他的工资将不受影响。另一方面，如果工会确实有助于提高工人的工资，则在工人技能保持不变的情况下，工人从无工会企业转到有工会企业后其工资会得到提升。当给定的工人或企业改变状态时（譬如说从无工会企业到有工会企业，或从有工会企业转到无工会企业），我们跟踪调查他们一段时间，就可以构建一个合适的递归结构来研究前后的效应。

然而微观动态效应和宏观动态效应通常不能用横截面数据集进行估计，单纯的时间序列数据集通常也无法给出动态系数的准确估计。譬如，考虑分布滞后模型

$$y_t = \sum_{\tau=0}^{h} \beta_\tau x_{t-\tau} + u_t, \quad t = 1, \cdots, T \tag{1.1.1}$$

的估计问题，其中 x_t 是外生变量，u_t 是随机扰动项。通常情况下，x_t 接近 x_{t-1}，且更接近 $2x_{t-1} - x_{t-2} = x_{t-1} + (x_{t-1} - x_{t-2})$；这 $(h+1)$ 个解释变量 x_t，x_{t-1}，…，x_{t-h} 中存在非常严重的多重共线性。因此，如果不假定每个滞后项系数都只是少量参数函数的先验信息，就不能获取滞后项系数精确估计的充分信息［参见 Almon 滞后，有理分布滞后（Malinvaud (1970)］。如果借助面板数据，我们就可以利用个体之间 x 值的差分缓解共线性问题，因此我们可放弃约束滞后项系数 $\{\beta_\tau\}$ 的传统方法，而添加一个不同的先验约束来估计无约束分布滞后模型。

另一个例子是在通常情况下测量误差可能导致模型不可识别的问题。但利用对给定个体或在给定时点的多重观测值，研究人员便可识别用其他数据无法识别的模型［参见 Biørn (1992)；Griliches 和 Hausman (1986)；Hsiao (1991b)；Hsiao 和 Taylor (1991)；Wansbeek 和 Koning (1989)］。

与单纯的横截面数据或时间序列数据相比，我们可利用面板数据构建并检验更复杂的行为模型。除此优点外，对实证研究中常出现的一个重要计量经济学问题，即我们发现（或没有发现的）某些效应产生的真实原因是存在与解释变量相关的遗漏（被误测或不可观测）变量，利用面板数据，我们有降低或消除该问题影响的方法。如果同时利用跨期动态信息和调查对象的个体信息，则我们可用更自然的方式更好地控制遗漏变量或不可观测变量的影响。譬如，考虑简单的回归模型：

$$y_{it} = \alpha^* + \boldsymbol{\beta}' \mathbf{x}_{it} + \boldsymbol{\rho}' \mathbf{z}_{it} + u_{it}, \quad i = 1, \cdots, N, t = 1, \cdots, T \tag{1.1.2}$$

其中 $\mathbf{x}_{it}$ 和 $\mathbf{z}_{it}$ 分别是 $k_1 \times 1$ 和 $k_2 \times 1$ 的外生变量向量；α^*，$\boldsymbol{\beta}$ 和 $\boldsymbol{\rho}$ 分别是 1×1，$k_1 \times 1$ 和 $k_2 \times 1$ 的常数向量；误差项 u_{it} 对所有的 i 和 t 独立同分布，其均值为零，方差为 σ_u^2。众所周知，y_{it} 对 $\mathbf{x}_{it}$ 和 $\mathbf{z}_{it}$ 的最小二乘回归可得到 α^*，$\boldsymbol{\beta}$ 和 $\boldsymbol{\rho}$ 的无偏一致估

计量。如果假定$\mathbf{z}_{it}$的值不可观测，且$\mathbf{x}_{it}$与$\mathbf{z}_{it}$的协方差不为零，则y_{it}对$\mathbf{x}_{it}$的最小二乘回归系数是有偏的。但如果能够获取一组个体的重复观测值，则我们可清除$\mathbf{z}$变量的影响。譬如，如果对所有的t都有$\mathbf{z}_{it}=\mathbf{z}_i$（即每个个体有自己的$\mathbf{z}$值，且该值在各时期保持不变），则我们可以求个体观测值在时间上的一阶差分，并得到

$$y_{it}-y_{i,t-1}=\boldsymbol{\beta}'(\mathbf{x}_{it}-\mathbf{x}_{i,t-1})+(u_{it}-u_{i,t-1}),\quad i=1,\cdots,N,t=2,\cdots,T \tag{1.1.3}$$

类似地，如果对所有的i都有$\mathbf{z}_{it}=\mathbf{z}_t$（即每个时期有各自的$\mathbf{z}$值，且该值对所有个体的影响都一样），则我们可以在给定的时点求个体对当期均值的偏离值，并得到

$$y_{it}-\bar{y}_t=\boldsymbol{\beta}'(\mathbf{x}_{it}-\bar{\mathbf{x}}_t)+(u_{it}-\bar{u}_t),\quad i=1,\cdots,N,t=1,\cdots,T \tag{1.1.4}$$

其中$\bar{y}_t=(1/N)\sum_{i=1}^{N}y_{it}$，$\bar{\mathbf{x}}_t=(1/N)\sum_{i=1}^{N}\mathbf{x}_{it}$，$\bar{u}_t=(1/N)\sum_{i=1}^{N}u_{it}$。现在可由式（1.1.3）或式（1.1.4）的最小二乘回归导出$\boldsymbol{\beta}$的无偏一致估计。但如果前一种情形（$\mathbf{z}_{it}=\mathbf{z}_i$）仅有横截面数据（$T=1$），或后一种情形（$\mathbf{z}_{it}=\mathbf{z}_t$）仅有时间序列数据（$N=1$），则我们无法进行这样的数据转换。除非存在与$\mathbf{x}$相关，但与$\mathbf{z}$和$u$无关的工具变量，否则我们不能得到$\boldsymbol{\beta}$的无偏一致估计。

麦柯迪（MaCurdy，1981）关于确定条件下盛年期男子的生命周期劳动力供给的研究就使用这种方法。在某种简化的假设下，麦柯迪证明工人的劳动供给方程可以表示成式（1.1.2），其中y是工作小时数的对数，x是实际工资率的对数，而z是工人初始财富的边际效用（不可观测）的对数，初始财富是工人一生中工资和财产性收入的综合衡量指标，假定它在工人的一生中都不变，但每个工人有各自的初始财富（即$\mathbf{z}_{it}=\mathbf{z}_i$）。在该问题中，不仅$\mathbf{x}_{it}$与$\mathbf{z}_i$相关，而且每个可以作为$\mathbf{x}_{it}$工具变量的经济变量（譬如教育）也与$\mathbf{z}_i$相关。故一般情况下，我们不能利用横截面数据集一致地估计$\boldsymbol{\beta}$[3]，但如果使用面板数据集，我们对式（1.1.2）进行一阶差分后就能一致地估计$\boldsymbol{\beta}$。

还有一个例子是增长率的“条件收敛”问题［参见 Durlauf（2001）；Temple（1999）］。通常认为，动态转移路径已知后，增长回归应该控制稳态收入水平［参见 Barro 和 Sala-i-Martin（1995）；Mankiew，Romer 和 Weil（1992）］。故增长率回归模型的回归元一般包括投资率、初始收入、政策效果的度量（比如注册学生数和黑市汇率溢价）等。但国家技术效率的初始水平$\mathbf{z}_{i0}$这一重要成分因不可观测而被遗漏。由于越是低效的国家越可能有更低的投资率或入学率，人们很容易认为$\mathbf{z}_{i0}$与回归元相关，所以导出的横截面参数估计量将产生遗漏变量偏误。但利用面板数据，通过对各国不同时期的观测值求如式（1.1.3）的一阶差分就可以消除初始效率水平的影响。

面板数据包含两个维度：横截面维度N，时间维度T。我们预期面板数据估计量的计算比仅用横截面数据（$T=1$）或者时间序列数据（$N=1$）时更复杂。但在某些场合使用面板数据事实上简化了计算和推断。譬如，考虑如下动态 Tobit 模型

$$y_{it}^{*}=\gamma y_{i,t-1}^{*}+\beta x_{it}+\epsilon_{it} \tag{1.1.5}$$

其中y^{*}不可观测，我们观测到的是y，如果$y_{it}^{*}>0$，则$y_{it}=y_{it}^{*}$，否则$y_{it}=0$。计

算 y_{it} 在 $y_{i,t-1}=0$ 下的条件密度比计算 $y^*_{i,t-1}$ 已知时的条件密度更复杂，因为（y_{it}，$y_{i,t-1}$）的联合密度函数包含对 $y^*_{i,t-1}$ 从 $-\infty$ 到 0 的积分。此外，不同时期有大量截取的观测值时，完全利用最大似然原理几乎不可能。但利用面板数据，只需集中关注 $y_{i,t-1}>0$ 的数据子集，就可在很大程度上简化对 γ 和 β 的估计，因为联合密度函数 $f(y_{it}, y_{i,t-1})$ 可以表示成条件密度函数 $f(y_{it} \mid y_{i,t-1})$ 和 $y_{i,t-1}$ 的边缘密度函数的乘积。但如果 $y^*_{i,t-1}$ 可观测，则 y_{it} 在 $y_{i,t-1}=y_{i,t-1}{}^*$ 下的条件密度就是 ϵ_{it} 的密度［Arellano，Bover 和 Labeager（1999）］。

最后再看一个非平稳数据时序分析的例子。如果数据是非平稳的，则 $T\to\infty$ 时，最小二乘估计量或最大似然估计量的极限分布不再是正态分布［参见 Dickey 和 Fuller（1979，1981）；Phillips 和 Durlauf（1986）］。因此，常用检验统计量的行为通常要靠计算机模拟进行推断。但如果使用面板数据，且各横截面单元的观测相互独立，则我们可用各横截面单元的中心极限定理证明许多估计量的分布仍是渐近正态的，Wald 型检验统计量是渐近 χ^2 分布的［参见 Binder，Hsiao 和 Pesaran（2000）；Levin 和 Lin（1993）；Pesaran，Shin 和 Smith（1999）；Phillips 和 Moon（1999，2000）；Quah（1994）］。

与纯粹的时序数据相比，面板数据还能对个体结果作出更精确的预测。如果控制某些变量后个体的行为具有相似性，则面板数据不仅提供个体行为信息，还使得通过观察其他个体的行为来研究某个体的行为成为可能。因此，通过混合数据可以得到对个体行为更为精确的描述［参见 Hsiao 和 Mountain（1994）；Hsiao 和 Tahmiscioglu（1997）；Hsiao 等（1989）；Hsiao，Applebe 和 Dineen（1993）］。

1.2 使用面板数据时的问题

1.2.1 异质性偏误

面板数据因理论上可分离特殊行为、实验处理，或更一般的政策影响而备受推崇。这种理论能力以“经济数据由受控制的实验生成”的假设为基础，在这些实验中，实验结果可用服从某概率分布的随机变量表示，而该分布是关于多个用来描述实验环境的变量的光滑函数。如果获取的数据确实是由可控的简单实验生成，则可直接使用标准的统计方法。但是，大多数面板数据来自非常复杂的日常经济生活。通常情况下，不同的个体可能受不同因素的影响。解释个体行为时，我们可以列出无穷多个影响因素。但建模的目的不是模仿现实，而是刻画影响结果的基本力量，所以，在模型设定中包含所有影响个体结果的因素既不可行也不必要。典型的做法是剔除那些没有显著影响的或个体所特有的因素。

但剔除个体特有的重要因素后，经济变量 y 由包含参数的概率分布函数 $P(y \mid \boldsymbol{\theta})$（其中 $\boldsymbol{\theta}$ 是 m 维实数向量，该分布函数在任何时候对所有的个体都是一样的）生成的标准假设可能不合实际。如果忽略了存在于横截面单元中的个体特