DOI:10.13546/j.cnki.tjyjc.2013.08.010

非平衡面板数据模型的估计方法及应用

吴　勇[1,2],林　悦[2]

(1.安徽工程大学 管理学院,安徽 芜湖 241000;2.合肥工业大学 管理学院,合肥230009)

摘　要:面板数据中,如果每个时期在样本中的个体不完全一样,则被称为非平衡面板数据。文章整理了非平衡面板数据估计方法的原理和思路,并采用2004~2011年中西部省际非平衡面板数据建立模型对影响中西部引进内资的主要因素进行了实证研究,结果显示,集聚效应因素、地区创新能力与中西部省份引进内资规模显著影响正相关。

关键词:非平衡面板数据;中西部;内资

中图分类号:F224　　**文献标识码:**A　　**文章编号:**1002-6487(2013)08-0076-03

1　非平衡面板数据的概念

面板数据因具有更多的信息,更大的变异等优点,在近年经济管理的实证研究中得到广泛的应用。迄今为止绝大多数的研究都是基于"平衡面板"进行的,即每个时期在样本中的个体完全一样,然而,有些时候某些个体的数据可能有缺失,如企业倒闭、个体不再参与调查,有的时候又有一些新的个体后来才加入调查中来,再或是一些地区的历史数据要比其他地区更久远,在这些情况下每个时期观测到的个体数并不相同,这就是所谓的"非平衡面板数据"(unbalanced panel)或"不完全面板"incomplete panel。考虑到非平衡面板更符合经济管理问题的实际情况,更有可能是实证研究中被经验设定的标准形式,而在非平衡面板中提取平衡面板,无论是最大化该平衡面板数据中被观测的个体数量还是该平衡面板中总的观测值数量,都会损失样本容量,降低估计效率。更进一步,人为剔除的观察值并非随机,也会破坏样本的随机性。因此,考察与非平衡面板数据相关的计量问题,并比较它们与平衡面板数据的差异受到越来越多研究者的重视。

2　非平衡面板数据的估计方法

面板数据模型的一般形式为:

$$Y_{it}=\alpha+X_{it}^{'}\beta+\mu_i+\nu_{it}\quad i=1,\ldots,N;t=1,\ldots,T_i \quad (1)$$

(1)式中,下标i和t分别代表个体和时间。横截的N是个体数,T是时间序列的维数。α为一个标量,β是K×1的待估系数矩阵,X_{it}是第k个解释变量的第i个个体在第t时期的观测值,μ_i表示不可观测到的个体的特殊效应,ν_{it}表示随机扰动。一般而言,平衡面板数据模型有两种处理方法:如果μ_i与解释变量相关,就将所有变量进行去均值处理然后再进行估计,从而得到固定效应模型;如果μ_i与解释变量不相关,可以采用随机效应模型。对于固定效应模型,将方程(1)两边对时间取平均可得组间回归式:

$$\bar{Y}_i=\alpha+\overline{X_i}^{'}\beta+\mu_i+\bar{\nu}_i \quad (2)$$

(1)式减去(2)可得离差形式的组内回归式:

$$Y_{it}-\bar{Y}_i=(X_{it}-\overline{X_i})^{'}\beta+(\nu_{it}-\bar{\nu}_i) \quad (3)$$

由于(3)式中已将μ_i消去,因此,只要$(X_{it}-\overline{X_i})^{'}$和$(\nu_{it}-\bar{\nu}_i)$不相关,就可以用OLS一致地估计β。显然,非平衡面板数据并不影响计算离差形式的组内估计量(within estimator),因此,固定效应模型仍然可以使用。

随机效应模型假设μ_i与解释变量不相关,由于μ_i的存在,同一个体不同时期的扰动项之间存在自相关,即:

$$\rho=Corr(\mu_i+\nu_{it},\ \mu_i+\nu_{is})=\sigma_\mu^2/(\sigma_\mu^2+\sigma_\nu^2)\quad t\neq s \quad (4)$$

平衡面板数据随机效应方法是先以OLS的残差来估计($\sigma_\mu^2+\sigma_\nu^2$),以FE的残差来估计$\sigma_\nu^2$,再用广义最小二乘法(FGLS)来估计原模型,即用OLS来估计下面的广义离差模型,

$$Y_{it}-\hat{\theta}\bar{Y}_i=(X_{it}-\hat{\theta}\overline{X_i})^{'}\beta+[(1-\hat{\theta})\mu_i+(\nu_{it}-\hat{\theta}\bar{\nu}_i)] \quad (5)$$

其中,$\hat{\theta}$是$\theta=1-\sigma_\nu/(T\sigma_\mu^2+\sigma_\nu^2)^{1/2}$的一致估计量。

对于非平衡面板数据,只要让$\theta_i=1-\sigma_\nu/(T_i\sigma_\mu^2+\sigma_\nu^2)^{1/2}$($T_i$为第i个个体的时间维度),可照样进行可行广义最小二乘法(FGLS)估计。但进行非平衡面板随机效应的可行广义最小二乘法(FGLS)估计必须找到合适的方法对其方差组合进行一致的估计(Baltagi and Chang,1994)。非平衡面板的单因素误差回归模型可表示为:

$$Y_{it}=\alpha+X_{it}^{'}\beta+u_{it}\quad i=1,\ldots,N;t=1,\ldots,T_i \quad (6)$$
$$u_{it}=\mu_i+\nu_{it}$$

用向量形式表示,该模型为:

基金项目:安徽省教育厅人文社科一般项目(2010sk314);安徽省哲学社会科学规划项目(AHSK11-12D58)

作者简介:吴　勇(1977-),男,安徽合肥人,博士研究生,讲师,研究方向:区域经济。

$$Y=\alpha\iota_n+X\beta+u=Z\delta+u$$
$$u=Z_\mu\mu+\nu \tag{7}$$

其中，Y 和 Z 分别 n×1 和 n×K 维矩阵，$Z=(\iota_n,X)$，$\delta^{'}=(\alpha,\beta^{'})$，$n=\sum T_i$，$Z_\mu=\mathrm{diag}(\iota_{T_i})$，其中 ι_{T_i} 是元素为1的 T_i 维向量。

Searle（1971）指出，平衡面板数据模型的方差分析（ANOVA）方法非平衡面板仍然适用且具有无偏性，ANOVA是通过令平方和二次型等于期望值并求解线性方程组得到，可定义组内和组间平方和的两种二次型形式：

$$q_1=u^{'}Qu,\ q_2=u^{'}Pu \tag{8}$$

其中，$Q=\mathrm{diag}[E_{T_i}]$，$P=\mathrm{diag}[\bar{J}_{T_i}]$，$\bar{J}_{T_i}=J_{T_i}/T_i$，$E_{T_i}=I_{T_i}-\bar{J}_{T_i}$，$I_{T_i}$ 代表 T_i 阶的单位矩阵，J_{T_i} 代表 $T_i\times n_j$ 维元素都为1的矩阵。根据Swamy和Arora(1972)的建议，我们使用用组间和组内使用两段回归来估计方差分量，即将组内方差和组间方差带入式(8)中的 q_1 和 q_2 得到 $\tilde{q}_1=\tilde{u}^{'}Q\tilde{u}$，$\tilde{q}_2=\tilde{u}^{'}P\tilde{u}$，$\tilde{q}_1$ 和 $\tilde{q}_2$ 的期望值为：

$$E(\tilde{q}_1)=(n-N-K+1)\sigma_\nu^2$$
$$E(\tilde{q}_2)=[n-\mathrm{tr}((Z^{'}PZ)^{-1}Z^{'}Z_\mu Z_\mu^{'}Z)]\sigma_\mu^2+(N-K)\sigma_\nu^2 \tag{9}$$

令式(9)中的 $\tilde{q}_i$ 与其期望值 $E(\tilde{q}_i)$ 相等，可得到方差分量的Swamy-Arora估计量：

$$\hat{\sigma}_\nu^2=\tilde{u}^{'}Q\tilde{u}/(n-N-K+1)$$
$$\hat{\sigma}_\mu^2=(\tilde{u}^{'}P\tilde{u}-(N-K)\hat{\sigma}_\nu^2)/[n-\mathrm{tr}((Z^{'}PZ)^{-1}Z^{'}Z_\mu Z_\mu^{'}Z)] \tag{10}$$

Jennrich和Sampson（1976）认为，极大似然估计方法（MLE）也能够非平衡面板数据随机效应模型的方差组合进行估计，其对数似然函数为：

$$\ln L=-(n/2)\ln(2\pi)-(n/2)\ln\sigma_\nu^2-0.5\ln|\sum|-(Y-Z\delta)^{'}\sum{}^{-1}(Y-Z\delta)/2\sigma_\nu^2 \tag{11}$$

其中，$\sum=I_n+\rho Z_\mu Z_\mu^{'}=\mathrm{diag}(E_{T_i})+\mathrm{diag}[(1+\rho T_i)\bar{J}_{T_i}]$，$\rho=\sigma_\mu^2/\sigma_\nu^2$。由于 ρ 的一阶条件是非线性的，参数的估计值必须通过迭代法进行数值求解。然而，由于同时给出回归系数估计量而损失了相应的自由度。Patterson和Thompson（1971）提出了受约束的极大似然估计方法（REML）弥补了这个缺点。

另外，在μ和n服从正态分布的前提下，Rao（1971）提出了方差组合的两种估计方法，即最小正态二次无偏估计值（MINQUE）和最小方差二次无偏估计值（MIVQUE），使用这两种估计方法，MIVQUE需要一个方差分量的先验值，要得到MINQUE估计量，常用的先验初始值分别为单位矩阵（MQ0）和Swamy和Arora的ANOVA估计量（MQA）。Baltagiet al.（2001）经过蒙特卡罗模拟对这些方法进行比较后发现：简便的ANOVA估计量对回归系数的估计最优，而在进行方差组合估计时，ρ 值不同时各种估计方法的表现有所不同。总体来说，ANOVA方法和极大似然估计方法在方差组合以及标准误差的估计中要优于其他方法。

3 非平衡面板模型的应用

考虑到中西部地区是跨区内资的主要流入地，本文将根据近年来的省级数据构建模型深入探讨中西部地区引进内资的影响因素问题。参照对外商投资区位选择影响因素的研究，我们将考察市场规模、创新能力、基础设施水平、劳动力成本、市场化程度、集聚效应、区位等因素等对中西部地区引进内资的影响，因此本文建立模型如下：

$$CI_{it}=\alpha+\beta_1CONS_{it}+\beta_2INNOV_{it}+\beta_3TRAF_{it}+\beta_4WAGE_{it}+\beta_5GOV_{it}+\beta_6NONPU_{it}+\beta_7FAI_{it}+\beta_8NUM_i+\mu_i+\nu_{it} \tag{12}$$

CI表示每年各地区引进内资实际到位规模；CONS为市场规模，以地区全社会消费品零售总额表示；INNOV为创新能力，用各地每年的专利授权量来表示；TRAF为基础设施水平，本文选择的是交通基础设施变量交通线路密度，交通线路密度=（公路里程+铁路里程+内河里程）/地区面积；劳动力价格WAGE采用相对劳动工资指标，用各省份二三产业的工资总额和增加值之比来衡量；本文选择两个指标来间接地反映各地区市场化程度的情况：一是政府规模GOV，用政府消费支出占总消费支出的比重表示，二是国有经济比重NONPU，用非公有制企业工业产值在地区工业总产值中所占比重表示；集聚效应本文考察的是投资的集聚效应，衡量指标为上一年的固定资产投资规模FAI；SUM为区位控制变量，中部省份取1，西部省份取0。

中西部地区共有20个省级行政区，因为西藏数据不全，江西内资的统计口径和其他省份不统一，本文样本中共包括18个省份，由于每个省份引进内资统计工作的起始时间不同，研究的时间序列也不全一致。其中，湖北和山西分别为2008~2011年和2007~2011年的数据，黑龙江、吉林、内蒙古和甘肃为2005~2011年的数据，其余12个省份是2004~2011年的数据，可以看出总样本是一个非平衡面板数据。根据前文所述，人为构建平衡面板会降低估计的效率并破坏随机性，因此，我们将以非平衡面板数据的估计方法进行参数估计。引进内资数据分别来自相关省份历年政府工作报告及商务厅网站，解释变量数据来源于相关年份各省统计年鉴和《中国统计年鉴》。

我们运用Stata11软件对模型(12)进行回归分析，回归之前，先用方差膨胀系数(vif)判断解释变量的多重共线性问题，当vif值大于10时，回归存在有害的多重共线性。从表1的VIF(1)可以看出lnGDP的vif值最高且高于临界值，因此在回归中先将其剔除。剔除lnGDP后，如表的VIF(2)所示，变量的vif值都在临界值之下，不再存在有害的多重共线性。

表1 多重共线性检验结果

	CONS	FAI	INNOV	TRAF	NONPU	SUM	WAGE	GOV
VIF(1)	12.55	10.13	4.52	3.48	2.64	2.40	1.87	1.69
VIF(2)		5.91	3.48	3.47	2.63	2.03	1.80	1.67

为了确保模型估计结果的准确性和可靠性，并不同估计方法的统计差异，我们分别利用固定效应方法（FE）、ANOVA方法（Swamy-Arora估计量）、极大似然估计方法

(MLE)和受约束的极大似然估计方法(REML)进行参数估计,得到的结果如表2所示。

表2　　中西部地区引进内资影响因素回归结果

变量	FE	Swamy-Arora	MLE	REML
FAI	0.2587*** (3.39)	0.2952*** (4.36)	0.2934*** (4.40)	0.3590*** (5.61)
INNOV	0.1475*** (4.09)	0.1211*** (3.61)	0.1224*** (3.66)	0.0524 (1.57)
TRAF	0.0162 (0.22)	0.0049 (0.08)	0.0057 (0.09)	-0.0438 (-0.86)
WAGE	-7987.7 (-1.40)	-3787.949 (-0.79)	-3979.2 (-0.83)	3089.1 (0.79)
GOV	-15.289 (-0.71)	-11.84 (-0.63)	-11.92 (-0.65)	-18.926 (-1.26)
NONPU	5.5575 (0.35)	3.8976 (0.32)	3.9377 (0.33)	3.7649 (0.45)
NUM	——	-322.57 (-0.47)	-330.33 (-0.76)	39.568 (0.16)
常数项	786.66 (0.64)	452.10 (0.43)	470.91 (0.46)	156.86 (0.19)
R2(组内)	0.7777	0.7756	——	——

说明:*、**、***分别表示在10%、5%、1%水平上显著,括号内为z值。

根据表2可以发现,固定效应模型估计、Swamy-Arora估计的R^2(组内)均大于0.7,说明模型拟合效果较好。Swamy-Arora估计和极大似然估计的回归结果非常相似,固定效应模型的估计结果与这两种方法也比较接近,有约束极大似然估计的结果和以上三种方法有明显差别。具有统计显著性的变量是,FAI的全部四种方法的估计和INNOV的前三种方法的估计,其余变量在统计上都不具有显著性。根据Baltagiet al.(2001)的观点,ANOVA估计量对回归系数的估计最优,因此,我们主要依据Swamy-Arora估计量来分析中西部引进内资的影响因素。前期固定资产投资对引进内资规模有显著正向影响表明了投资集聚效应在跨区内资区位选择中的重要作用,投资目的地的前期固定资产投入水平越高,越能带来示范效应和配套效应,对外来企业的吸引力就越大。专利授权数的系数显著为正,说明当地创新能力越强对外地资本的吸引力越大,创新能力已成为中西部地区间内资招商竞争的重要力量。

4　结束语

鉴于经济管理实证研究中的面板数据往往并不"完整",非平衡面板数据的估计方法越来越受到重视。本文在介绍了非平衡面板主要处理方法的基础上,使用中西部地区18个省份的省级非平衡面板数据建立模型,对中西部地区引进内资的影响因素进行了实证分析,并比较了不同方法的估计结果,经比较发现ANOVA和MLE估计量的结果很相似,而REML估计的结果则和其他方法差异明显。实证结果显示,前期固定资产投资和专利授权数系数估计显著为正,这说明当地资本集聚效应和创新能力对中西部地区引进内资有积极作用,中西部地区必须要重视集聚因素,对已具备集聚效应的地区,应进一步发挥集聚经济的自我强化作用,形成更大的产业集中区,对于经济基础比较薄弱的地区也要集中有限资源建立有效的集聚增长点。与此同时,中西部地区应加大在研发和教育等领域的投入,积极培养和引进科技人才,不断提高地区自主创新能力,这样才能为招商引资和提高经济发展水平提供不竭动力。

参考文献:

[1][美]巴尔塔基著.白仲林等译.面板数据计量经济分析[M].北京:机械工业出版社, 2010.

[2]陈强.高级计量经济学及Stata应用[M].北京:高等教育出版社, 2010.

[3]李汉君.我国FDI流入的地区差异与影响因素分析_基于1992_2007年省级面板数据[J].国际贸易问题,2011,(3).

[4]Baltagi,B.H. , Chang,Y.J. Incomplete Panels: a Comparative Study of Alternative Estimators for the Unbalanced One-way Error Component Regression Model[J].Journal of Econometrics, 1994,(62).

[5]Searle,S.R.Linear Models[M]. New York:John Wiley,1971.

[6]Swamy, P.A.V.B., S.S. Arora. The Exact Finite Sample Properties of the Estimators of Coefficients in the Error Components Regression Models[J]. Econometrics,1972,(40).

[7]Patterson,H.D., Thompson,R. Recovery of Inter-block Information when Block Sizes Are Unequal[J]. Biometrica,1971(58).

[8]Rao,C.R. Estimation of Variance and Covariance Components-minque Theory[J]. Journal of Multivariate Analysis,1971,(1).

[9]Baltagi,B.H.,Song,S.H., Koh,W. The Unbalanced Nested Error Component Regression Model[J].Journal of Econometrics, 2001,(101).

(责任编辑/易永生)