面板数据计量模型适应性的比较研究①

刘莉亚[1], 刘 平[2], 覃 筱[3], 代 飞[1]
(1. 上海财经大学金融学院, 上海 200433; 2. 上海财经大学现代金融研究中心, 上海 200433;
3. 上海交通大学安泰管理学院, 上海 200052)

摘要: 近年来在经济管理研究中采用面板数据的实证性论文日益增多, 同时关于面板数据处理的理论成果也并不少见. 然而, 问题在于: 一方面已有的关于面板数据处理的理论方法多基于无穷大样本, 而现实中仅能获得有限样本; 另一方面已有的实证研究成果或对面板数据不做任何特殊处理或并不说明处理方法选用的依据. 基于此, 按只存在个体效应, 只存在时间效应和同时存在两种效应3种不同的数据结构, 模拟比较了现有各种理论方法的适用性. 并根据各种处理方法在小样本情况下估计结果存在较大差异这一事实, 反推出用于判断实际数据中存在何种效应的准则.

关键词: 面板数据; 固定个体效应; 固定时间效应; 聚类回归标准差系数; Fama-MacBeth方法
中图分类号: F22 F83 文献标识码: A 文章编号: 1007-9807(2011)02-0086-11

0 引 言

众所周知, 当经典计量经济模型的4个基本假设满足时, 估计参数贝塔的标准差是无偏且一致有效的. 但是, 当不同时期的观测值与残差之间存在相关关系时, 直接应用OLS回归有可能对贝塔的标准差产生有偏的估计. 同时, 检验贝塔参数的置信度需要构造 统计量, 在贝塔标准差估计有偏的情况下, 会相应地减小或增大对应的 值, 从而增加估计结果出错的概率. 此外, 自变量和残差之间存在显著的相关关系, 表明模型中还有重要的信息没有被充分挖掘, 有时贝塔标准差估计甚至会产生较大偏差, 甚至无法确定所建立的方程究竟是否有效. 面板数据(panel data), 由于具有时间和横截面两个维度, 能够很好地研究不同公司(地区)在不同时期的特征, 因此近几年在实证研究中获得了广泛的应用. 例如, 统计了《经济研究》、《管理世界》、《金融研究》、《世界经济》等主要学术期刊中运用面板数据的论文数量, 结果发现, 从2002年到2008年, 主要经济管理类学术刊物中运用面板数据的论文呈现出明显的上升趋势(参见图1).

越来越多的论文采用面板数据, 充分说明了用面板数据研究经济问题的优点. 然而, 正因为此, 面板数据比一般一维时间序列或横截面数据更为复杂, 直接应用OLS回归, 同一时期不同公司间的残差之间可能存在相关关系, 同一公司不同时期的残差之间也可能存在相关关系. 然而, 许多论文在应用面板数据进行实证研究时, 往往不考虑对贝塔标准差的调整, 或者只进行了部分调整. 具体来说, 许多文章采用了面板数据中的固定效应模型进行研究, 但是只考虑了个体效应, 即只

① 收稿日期: 2010-01-12; 修订日期: 2010-10-30
基金项目: 国家社会科学基金重大资助项目(08-ZD036); 国家自然科学基金青年科学基金资助项目(71001070); 上海市哲学社会科学规划课题资助项目(2010BJB008); 上海财经大学211第3期资助项目(2009330046).
作者简介: 刘莉亚(1976-), 女, 山西长治人, 博士, 副教授, 博士生导师. Email: liuliya0112@163.com

针对同一时期不同公司的残差相关关系进行了调整，很少文章考虑时间效应，这样的结果往往是不能令人信服的. 本文将显示，如果没有考虑数据中存在时间效应，即使采用了固定效应模型，其估计结果也是有较大偏误的.

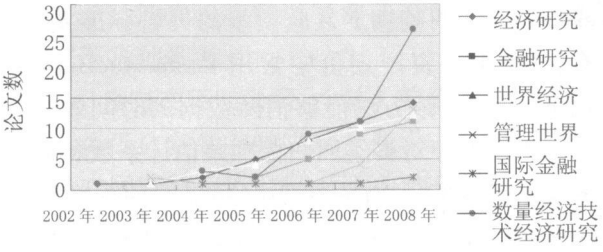

图 1　近年主要学术期刊发表基于面板数据的论文数量

Fig. 1　Paper number based on panel data published in main academic Chinese journals in the period of 2002—2008

下面，做一个简单的回归来说明直接应用 OLS回归可能产生的错误.

针对上市公司成长性与规模的关系研究，本文选择销售收入增长率代表公司的成长性，选择公司总资产的对数值作为公司规模的衡量指标. 并选择了商业百货上市公司 (来自证监会行业分类的零售业)1997至 2006年共 10年的数据，为了消除异常数值的影响，又去除了带有 * ST②，ST以及部分数据不全的上市公司数据，共 45家公司. 所有数据均来源于 wind咨询. 下文以 asset代表资产规模，以 xxsrzz代表销售收入增长率，各变量的统计性描述见表 1.

表 1　各变量的描述性统计

Table 1　Descriptive statistics of various variables

变量	均值	标准差	最小值	最大值
asset	9. 095 5	0. 282 8	8. 254	10. 077 4
xxsrzz	13. 536	34. 3	-64. 994 9	329. 513

设立如下模型来分析上市公司成长性与规模的关系

$$g_{it} = c + \beta \log(size_{it}) + \varepsilon_{it} \qquad (1)$$

其中，下标 i和 t分别表示上市公司和时间；g表示公司成长性，数据为各公司每年的销售收入增长率；size表示公司规模，数据为各公司的总资产. 以上是基准模型，根据研究目的的不同，在具体的研究中会采用不同的形式.

当直接采用 OLS对上式进行回归时，β的估计值为 10. 325，标准差 5. 710 1，t值为 1. 808 3，在 10% 的置信水平下显著. 然而，如果考虑了固定个体效应，情况就大不相同. 采用固定个体效应模型估计，得到 β的估计值为 7. 847 5，标准差为 9. 263 3，其值约为 OLS估计的 1. 6倍，t值只有 0. 847 16. 采用固定时间效应模型估计，得到 β的标准差为 6. 085 7，与 OLS估计结果差别并不大. 因此，如果没有考虑固定个体效应而直接估计，那么很有可能得出的是完全相反的结果. 面板数据模型扰动项可能存在异方差性、序列相关与截面相关，许多学者提出了不同的方法应用于面板数据的估计. 针对固定效应模型，Kiefer[1] 提出了 Kiefer标准差，针对面板数据估计中残差存在序列相关的情况，该估计方法在面板数据不存在异方差的时候是稳健的. White[2] 提出了针对计量模型中存在异方差的一般估计方法. 在两者的基础上，Arellano[3] 提出了 Arellano标准差，该估计系数在存在异方差和序列相关时都是稳健的. 尽管如此，Arellano系数在提出后并没有被广泛采用，可能的原因在于该估计系数在小样本情况下的有效性问题. 然而，Kezdi[4] 通过模拟证明了在只存在固定个体效应的情况下，只要个体数目大于 50，Arellano标准差依然能够得到比较稳健的结果.

Fama和 MacBeth[5] 提出了 Fama-MacBeth回归方法，该方法能够很好地处理资产价格的时间效应，因此该回归方法在资产定价的学术论文中得到了广泛应用. 由于资产价格的时间序列数据也可以看作特殊的面板数据，在下文的模拟和实证分析中，也包括了文献中常用的 Fama-MacBeth方法. 此外，Newey和 West[6] 提出了 Newey-West方法，并在后来被改进用于面板数据的回归分析，Driscoll和 Kraay[7] 提出了 Driscoll-Kraay标准差以调整面板数据中存在个体效应和时间效应时的 β标准差. 但是这些方法并没有得到广泛应用.

② ST是英文 special treatment的缩写，意即 "特别处理 ". 该政策自 1998年 4月 22日起实行，针对的对象是出现财务状况或其他状况异常的上市公司.

尽管面板数据的处理在理论上有了较大的发展,但在应用上仍受到一定限制.主要原因在于理论推导是基于大样本进行的,而实际研究中人们可以得到的面板数据样本量却很小,因此,许多理论上完美的方法在实际应用中就会大打折扣,在小样本情况下,不同方法对同一面板数据的应用结果可能差别很大.但是,差别究竟有多大?各种方法分别适用何种结构的数据?这就需要采用模拟方法来回答了.本文试图通过对实证中常用的面板数据研究方法进行比较研究,通过模拟指出采用不同方法进行估计的偏误,尤其是针对运用面板数据研究中样本过小的问题,较为深入地比较在只存在个体效应,只存在时间效应和同时存在个体效应和时间效应时,各种方法在小样本情况下的估计结果.结果表明:在数据中存在不同的效应时,运用不同的方法会产生较大差异的结果.然而,这也提醒人们,在产生较大差异时,需要注意数据中可能存在着某一类型的效应.

1 3种不同情形下面板数据处理方法在有限样本下的比较分析

在实际的面板数据中,常常有3种类型的因素影响残差的相关性.Hsiao[8]把它们称为个体时期恒量(individual time-invariant),时期个体恒量(period individual-invariant)和个体时期变量(individual time-varying).在存在个体时期恒量(individual time-invariant)的情况下,某个公司在不同时期的残差可能是相关的,伍德里奇把它称为不可观测个体效应(unobserved firm effect);在存在时期个体恒量(period individual-invariant)的情况下,某个时期不同公司之间的残差可能是相关的,伍德里奇把它称为不可观测时间效应(unobserved time effect);另外,不同公司在不同时期的残差可能存在相关关系,这叫做不可观测个体时间效应(unobserved firm-time effect).

正因为面板数据受到多种因素的影响,针对不同的影响因素,计量经济学家提出了不同的方法对贝塔的标准差进行调整.最常用的有Arellano[3]在Kiefer[1]、White[2]等人的基础上引入聚类回归分析(cluster analysis),称为聚类回归标准差估计系数(cluster standard error estimator),调整固定效应模型的贝塔标准差.再有,学术上在资产定价中常用Fama-MacBeth方法进行回归,该方法能够消除数据的时间效应.

为了便于对聚类回归标准差估计系数和OLS方法进行比较③,首先简要回顾基于面板数据的OLS回归模型.一般地,面板数据的OLS回归模型可以写为

$$y_{it}=x_{it}\beta+\varepsilon_{it} \tag{2}$$

其中x_{it}表示第i个公司在第t时期的观测值,ε_{it}表示回归的残差.x_{it}和ε_{it}是相互独立的,ε_{it}的期望为零,方差为σ_ε^2,不失一般性,可假设x_{it}的期望为零,方差为$\sigma_x^2$④.根据以上假设,可以推出OLS中贝塔的估计值为

$$\hat{\beta}_{OLS}=\frac{\sum_{i=1}^{N}\sum_{t=1}^{T}x_{it}y_{it}}{\sum_{i=1}^{N}\sum_{t=1}^{T}x_{it}^2}=\frac{\sum_{i=1}^{N}\sum_{t=1}^{T}x_{it}(x_{it}\beta+\varepsilon_{it})}{\sum_{i=1}^{N}\sum_{t=1}^{T}x_{it}^2}$$
$$=\beta+\frac{\sum_{i=1}^{N}\sum_{t=1}^{T}x_{it}\varepsilon_{it}}{\sum_{i=1}^{N}\sum_{t=1}^{T}x_{it}^2} \tag{3}$$

如果满足OLS回归的经典假设,那么可以推出β_{OLS}的渐进方差为

$$\mathrm{Var}(\hat{\beta}_{OLS}-\beta)$$
$$=\operatorname*{plim}_{\substack{N\to\infty\\ T\text{固定}}}\left(\frac{1}{N^2}\left(\sum_{i=1}^{N}\sum_{t=1}^{T}x_{it}\varepsilon_{it}\right)^2\left(\frac{\sum_{i=1}^{N}\sum_{t=1}^{T}x_{it}^2}{N}\right)^{-2}\right)$$
$$=\operatorname*{plim}_{\substack{N\to\infty\\ T\text{固定}}}\left(\frac{1}{N^2}\left(\sum_{i=1}^{N}\sum_{t=1}^{T}x_{it}^2\varepsilon_{it}^2\right)\left(\frac{\sum_{i=1}^{N}\sum_{t=1}^{T}x_{it}^2}{N}\right)^{-2}\right)$$
$$=\frac{1}{N}\left(T\sigma_x^2\sigma_\varepsilon^2\right)\left(T\sigma_x^2\right)^{-2}=\frac{\sigma_\varepsilon^2}{\sigma_x^2NT} \tag{4}$$

③ 实证研究中POOL OLS亦较为常用,本文对OLS方法也进行了理论和模拟分析,以便对OLS方法和其他方法进行比较.

④ 面板数据估计模型分为变截距模型,变系数模型和联合回归模型3种.这里采用的是无截距的POOL OLS模型设定,一方面是因为本文的重点在于斜率标准差的估计问题,在固定效应模型中,截距并不是重要的估计量;另一方面,采用无截距的POOL OLS模型设定,可以使下文的推导更为简洁,并且不失一般性.

上述 OLS回归方程是建立在假设残差独立同分布的基础上的，残差独立的假设用于从第 1行推出第 2行，残差同分布的假设（即同方差）用于从第 2行推出第 3行⑤. OLS回归可以得到无偏的贝塔标准差. 然而，正如前文所述，在面板数据中关于残差独立的假设常常被违背，下面分别考虑以下几种情形.

1.1　只存在固定公司（个体）效应

为了放宽残差独立这一假设，本文首先假设只存在固定的个体效应，不存在时间效应，因此残差就可以写为

$$\varepsilon_{it} = d_i + \mu_{it} \tag{5}$$

同时假定 x_{it}也包含某一特定个体效应，从而 x_{it}可以表示为

$$x_{it} = c_i + v_{it} \tag{6}$$

上述 c_i、v_{it}、d_i、μ_{it}都是具有零期望、有限方差的变量，并且相互独立. 这样，就能保证下文模拟中对模型的系数估计结果是有效的. 因为自变量和残差在同一个体中都是相关的，但是在不同个体之间是相互独立的，即

$$\mathrm{corr}(x_{it}, x_{js}) = \begin{cases} 1, & i = j,\ t = s \\ \rho_x = \sigma_c^2/\sigma_x^2, & i = j,\ t \neq s \\ 0, & i \neq j \end{cases}$$

$$\mathrm{corr}(\varepsilon_{it}, \varepsilon_{js}) = \begin{cases} 1, & i = j,\ t = s \\ \rho_\varepsilon = \sigma_d^2/\sigma_\varepsilon^2, & i = j,\ t \neq s \\ 0, & i \neq j \end{cases}$$

通过上式可以看出，由于 c_i和 d_i的存在，OLS回归对贝塔的标准差估计是有偏的. 一般来说，OLS低估了真实的贝塔标准差，并且随 T的增加而增加[10].

1.1.1　固定效应模型的估计系数

在只存在公司固定效应时，可以用下面的固定效应模型进行回归，即

$$y_{it} = \alpha_i + x_{it}\beta + \varepsilon_{it} \tag{7}$$

其中，α_i表示固定个体效应，代表个体的特征，并且不随时间而变化. 把式（7）写成向量的形式，为

$$y_i = \alpha_i e + x_i\beta + \varepsilon_i \tag{8}$$

其中，e为 T阶单位列向量. 如果把式（8）左右同乘矩阵 $Q = I - e(e'/T)$（e'为 e的转置向量），就可以消去 α_i，从而式（8）转化为

$$\ddot{y}_i = \ddot{x}_i \cdot \beta + \ddot{\varepsilon}_i \tag{9}$$

其中 $\ddot{y}_i$、$\ddot{x}_i$和 $\ddot{\varepsilon}_i$分别等于 $Q \cdot y_i$、$Q \cdot x_i$和 $Q \cdot \varepsilon_i$. 这样一来，通过乘以 Q矩阵消除个体效应的方法，实际上是对 $\ddot{y}_i$和 $\ddot{x}_i$进行回归. 可得

$$\hat{\beta}_{FE} = (\ddot{x}'\ddot{x})\ddot{x}'\ddot{y} \tag{10}$$

其中 $\ddot{x}' = (\ddot{x}_1' \cdots \ddot{x}_N')$，$\ddot{y}' = (\ddot{y}_1' \cdots \ddot{y}_N')$.

虽然固定效应模型中对 β_{ov}的估计是一致的，但却有不同的估计其标准差方法. 一般地，如果不存在异方差，并且 $\ddot{x}_i$和 $\ddot{\varepsilon}_i$不存在序列相关时，则有

$$\mathrm{Var}(\hat{\beta}_{FE}) = \sigma_\varepsilon^2 (\ddot{x}'\ddot{x})^{-1} \tag{11}$$

而 Kiefer[11] 提出了下面的 β标准差的估计量，该估计量在满足不存在异方差情况时是有效的.

$$\mathrm{Var}(\hat{\beta}_{FE}) = \sigma_\varepsilon^2 (\ddot{x}'\ddot{x})^{-1} \left(\sum_{i=1}^{N} \ddot{x}_i' \hat{\Omega} \ddot{x}_i \right) (\ddot{x}'\ddot{x})^{-1} \tag{12}$$

其中 $\hat{\Omega} = N^{-1} \sum_{i=1}^{N} \hat{\varepsilon}_i \hat{\varepsilon}_i'$.

继而，Arellano[3] 提出了更广义的 β标准差的估计量，即

$$\mathrm{Var}(\hat{\beta}_{FE}) = (\ddot{x}'\ddot{x})^{-1} \times \left(\sum_{i=1}^{N} \ddot{x}_i' \hat{\varepsilon}_i \hat{\varepsilon}_i' \ddot{x}_i \right) (\ddot{x}'\ddot{x})^{-1} \tag{13}$$

该估计量在存在异方差或者残差存在序列相关时都是有效的.

虽然理论上在大样本的情况下（$N \to \infty$）完全可以采用 Arellano估计量来代替 Kiefer估计量，但是并不能确保在小样本的情况下也可以这样做，因为小样本下两种估计量的精确度是不同的.

⑤　panel data极限行为仅仅依赖于单位数 N和时间长度 T趋于无穷的方式. 例如一种是固定 N，让 T趋于∞，接着 N趋于∞，它们用（N, T趋于∞）表示；另一种是 T= T(N)，表示 T的大小受 N控制，N趋于∞，T(N)趋于∞，记为（T(N), N趋于∞）；第 3种是 T, N分别趋于∞，没有相互约束，记为（N, T趋于∞）. 这 3种方式极限分别称为序贯、对角和联合极限. Phillips和 Moon[9] 主要对序贯极限理论和联合极限理论进行了研究，认为序贯极限在寻求极限行为快速渐近性上是有益的. 即使在更强一些的条件下，联合极限理论也是很难得到并加以应用，但幸运的是，在所面临的 T很大、N适中的情况下，联合极限理论研究和应用并没有多大困难.

为了对其进行比较，本文通过固定 T，针对不同数目的 N进行了模拟研究.

在不考虑序列相关的情况下，本文模拟了一组面板数据，并且估计了 β和 β的方差. 通过模拟多次，就可以得到一系列其估计值，这样一来就可以计算出真实的 β方差和 β方差估计的平均值. 在这里，首先加入了固定个体效应. 在不同的模拟中，固定 T= 10，改变个体的数目，从 10增加到 250. 模拟 5 000次的结果如表 2. (以下模拟实验均是来自 5 000次蒙特卡洛模拟的结果，均采用 MATLAB完成. 模拟中设定数据真实的斜率 β= 1，由于无论在存在个体效应或者时间效应时 OLS方法的 β估计量都是无偏的，通过 5 000次模拟可以得到 5 000个 β值，并可以计算出 β的方差，这是该估计值的真实的方差.)

表 2　不同个体数情况下各种方法的估计结果 (不存在序列相关)

Table 2　Estimation results of various methods under different individual number (no serial correlation)

β估计量计算方法	N= 10	N= 25	N= 50	N= 250
truevarbeta	0.046 2	0.017 9	0.009 1	0.001 8
Avar	0.040 3	0.017 0	0.008 7	0.001 8
Kvar	0.044 6	0.017 8	0.008 9	0.001 8
Wvar	0.043 9	0.017 6	0.008 9	0.001 8
Ovar	0.021 8	0.008 3	0.004 1	0.000 797

注：T= 10.

表 2中，truevarbeta表示真实的 β方差，Avar表示基于 Arellano估计量计算的 β方差的平均值，Kvar表示基于 Kiefer估计量计算的 β方差平均值，Wvar表示基于 White[⑥] 估计量计算的 β方差平均值，Ovar表示基于 OLS估计量计算的 β方差平均值. 可以看出，直接用 OLS估计明显低估了 β的真实方差，在 N= 10的情况下，OLS估计的 β方差仅仅是真实值的 47.19% (0.021 8/0.046 2)，随着 N的增加，OLS估计量与真实值的比值分别为 46.37%、45.05%和 44.27%，偏差有所增加，但增加幅度并不大. 同时，在样本较小的情况下 (N=10)，Arellano估计量与真实方差还是有较大差异的，偏差为 12.77% ((0.046 2－0.040 3)/0.046 2)，而 Kiefer估计量的偏差仅为 3.46% ((0.046 2－0.044 6)/0.046 2)，White估计量的偏差为 4.98% ((0.046 2－0.043 9)/0.046 2)，这表明在不存在序列相关和小样本的情况下，Kiefer估计量和 White估计量的计算结果都比 Arellano估计量的精确度高. 然而，随着样本规模的逐渐增大，Arellano估计量与真实值之间的偏差逐渐缩小，在本文的模拟实验中，只要 N足够大 (N≥250)，这几种估计量之间的差异就可以忽略不计.

由于许多论文在实证过程中往往更关注异方差的情况，而忽略了序列相关这一问题，因此，按照上面的数据结构，本文加上序列相关，重复进行了模拟实验，以期能够得到上述几种估计方法在不同情况下的比较结果，模拟结果见表 3.

P表示 x_{it}和 μ_{it}的自相关系数. 当 P= 0.3时，表示 x_{it}的一阶自相关系数为 0.3，μ_{it}的一阶自相关系数亦为 0.3. 表格中第 1行数据表示真实的 β方差，第 2行数据表示基于 Arellano方法估计的 β方差，第 3行数据表示基于 Kiefer方法估计的 β方差，第 4行数据表示基于 White方法估计的 β方差，第 5行数据表示 OLS估计的 β方差.

⑥　White[12] 提出了处理回归数据中存在异方差的方法，在面板数据中常采用如下方法估计 $\mathrm{Var}(\hat{\beta}_{FE}) = (x_+'x_+)^{-1} \times \left(\sum_{i=1}^{N} \hat{\varepsilon}_i^{+\prime} \hat{\varepsilon}_i^{+} x_i^{+\prime} x_i^{+}\right) (x_+'x_+)^{-1}$.

表 3　不同个体数情况下各种方法的估计结果(存在序列相关)

Table3 Estimation results of various methods under different individual number(serial correlation)

样本数	N= 10	N= 25	N= 50	N= 250
β估计量计算方法	P= 0.3			
truevarbeta	0.070 0	0.028 6	0.014 3	0.002 8
Avar	0.063 1	0.026 6	0.013 6	0.002 8
Kvar	0.069 3	0.027 9	0.013 9	0.002 8
Wvar	0.060 1	0.024 3	0.012 1	0.002 4
Ovar	0.023 7	0.008 9	0.004 4	0.000 857
β估计量计算方法	P= 0.5			
truevarbeta	0.800 3	0.286 3	0.133 5	0.026 3
Avar	0.594 6	0.250 5	0.126 0	0.025 6
Kvar	0.720 4	0.271 3	0.131 2	0.025 8
Wvar	0.411 3	0.159 8	0.078 7	0.015 7
Ovar	0.029 6	0.010 8	0.005 2	0.001 0
β估计量计算方法	P= 0.8			
truevarbeta	0.026 8	0.010 4	0.005 3	0.001 1
Avar	0.022 7	0.009 9	0.005 1	0.001 0
Kvar	0.025 4	0.010 4	0.005 3	0.001 1
Wvar	0.012 0	0.004 7	0.002 4	0.000 5
Ovar	0.012 7	0.005 0	0.002 5	0.000 499

模拟结果表明,加入序列相关后的模拟结果与没有序列相关的情况有着很大的不同. 对比 N=10的情况可以看出,当序列自相关系数分别为 0.3、0.5和 0.8时, Arellano估计量、Kiefer估计量和 White估计量的偏差分别为 9.86%、1%、14.14%; 25.70%、9.98%、48.60%; 15.30%、5.22%、55.22%.显然,存在序列相关的小样本情况下, Kiefer估计量要显著优于 Arellano估计量和 White估计量. White估计量严重偏离了真实值,无论在小样本还是大样本下均如此. 在 N= 250时, Arellano估计量和 Kiefer估计量与真实值的偏差分别为 0.0(P=0.3); 2.66%、1.90%(P=0.5); 9.09%、0(P= 0.8).这表明在大样本下, Arellano估计量与 Kiefer估计量的差异已经很小,这与上面不存在序列相关的模拟结果是一致的.

为了更形象表示上述结果,进一步给出了上述结果的图表形式,以便能够一目了然小样本情况下各种估计方法的精确情况.

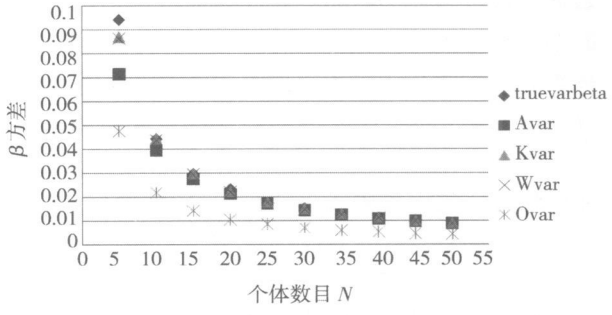

图 2　不同个体数情况下各种指数的估计结果(不存在序列相关)

Fig. 2 Various index simulation results under different individual number(no serial correlation)

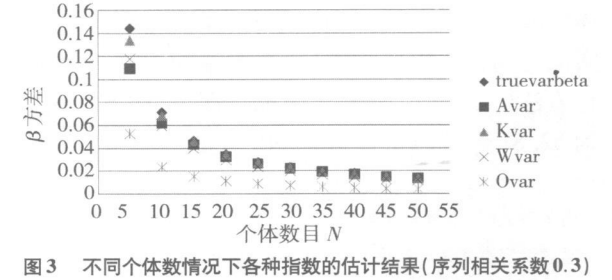

图 3　不同个体数情况下各种指数的估计结果(序列相关系数 0.3)

Fig. 3 Various index simulation results under different individual number (serial correlation coefficient = 0.3)

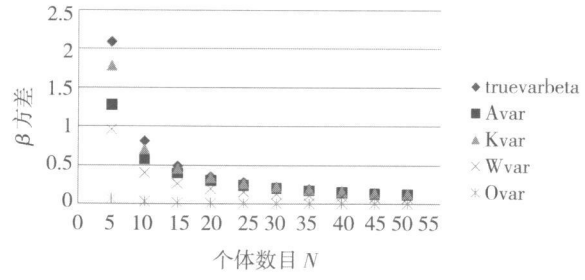

图 4　不同个体数情况下各种指数的估计结果(序列相关系数 0.5)

Fig. 4 Various index simulation results under different individual number(serial correlation coefficient = 0.5)

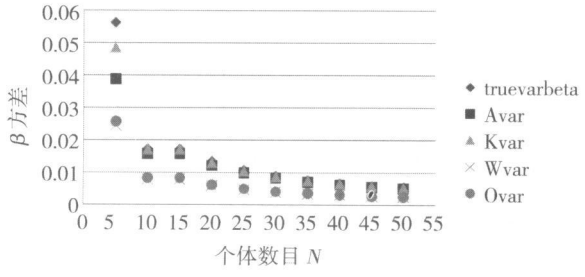

图 5　不同个体数情况下各种指数的估计结果(序列相关系数 0.8)

Fig. 5 Various index simulation results under different individual number(serial correlation coefficient = 0.8)

1.1.2　Fama-MacBeth误差估计

现有文献中常见的另外一种估计回归系数和标准误差的方法是 Fama-MacBeth方法. 该方法的具体运用可表述为研究者首先对面板数据的每一时期的数据进行回归, 一共回归 T次得到 T个 β(T代表时期), 最后得到 Fama-MacBeth的 β估计值

$$\beta_{FM}=\sum_{t=1}^{T}\frac{\beta_t}{T}$$

而该估计值的方差则用下式进行计算

$$S^2(\beta_{FM})=\frac{1}{T}\sum_{t=1}^{T}\frac{(\beta_t-\beta_{FM})^2}{T-1}$$

用上文的数据结构, 模拟检验了 Fama-Macbeth方法的结果, 如表 4.

从上面模拟结果可以看出, 当只存在固定个体效应时, Fama-Macbeth方法估计的 β方差与真实值之间存在着较大差异, 差异甚至大于 OLS估计方法(见表 1). 在 N= 10的情况下, OLS的偏差为 52.81%, 而 Fama-Macbeth方法的偏差达到了 83.48%, 这表明 Fama-MacBeth方法并不适合只存在固定个体效应的估计.

表 4　只存在固定个体效应的 Fama-MacBeth方差估计

Table 4 Fama-MacBeth variance estimator only existing individual effect

样本数	N= 10	N= 25	N= 50	N= 250
True	0.100 5	0.037 5	0.017 5	0.003 6
FM estimator	0.016 6	0.005 6	0.002 6	0.000 505
偏差(%)	83.48	85.07	85.14	85.97

1.2　只存在固定时间效应

与只存在个体效应类似, 可假设

$$\varepsilon_{it}=d_t+\mu_{it}$$

$$x_{it}=c_t+v_{it}$$

其中 d_t、μ_{it}、c_t、v_{it}都是具有零期望、有限方差的变量, 并且相互独立.

1.2.1　固定时间效应模型的估计系数

在只存在时间效应情况下, β系数及其方差的估计与只存在固定个体效应情况下的估计方法类似, 在公式的推导过程中, 只需将相应的下标 i替换为 t. 为了考察小样本情况下的适用性, 本文同样做了模拟. 结果显示, 当只有 10年的数据时, FE模型估计系数产生了一定的偏差. 这与本文之前在只存在固定个体效应中的小样本情况下的估计结果是一致的. 但是, 如果固定 N增加 T, 随着样本数的增加, FE模型估计将会逐渐产生无偏的结果(与只存在个体效应类似, 限于篇幅结果未列出). 从而有理由相信, 随着样本数的增加, 估计 β标准差能够依概率收敛于真实值. 本文的模拟结果显示, 250个样本就足够产生无偏的 β标准差估计, 而 10个样本则显得太少.

1.2.2　Fama-MacBeth误差估计

与上面只存在固定个体效应的模拟结果不同, 当只存在时间效应时, Fama-MacBeth方法显示出非常好的效果. 固定 T= 250, 不断改变 N的模拟结果如表 5⑦:

⑦　当 T较大时, 不可忽视的问题是平稳性问题. 本文模拟数据自变量 x_{it}和残差 ε_{it}均是服从期望为零, 有限方差的正态分布, 并且相互独立, 通过设定真实 β = 1, 由 $y_{it}=x_{it}+\varepsilon_{it}$可知 y_{it}亦服从期望为零, 有限方差的正态分布. 因此本文实际上假设了模拟数据是平稳的, 这样做是为了防止数据非平稳对 β以及 β标准差的估计造成影响. 在实际研究中, 该假设并不一定成立, 因此在实际研究中当 T比较大时需要考虑和检验数据的平稳性.

表 5　只存在时间效应的 Fama-MacBeth方差估计 (固定 T= 250)

Table 5　Fama-MacBeth variance estimator only existing time effect(fixing T= 250)

样本数	N= 10	N= 25	N= 50	N= 250
True	0. 100 8	0. 077 1	0. 068 5	0. 063 6
FM Estimator	0. 101 5	0. 076 5	0. 069 7	0. 062 8
偏差 (%)	-0. 69	0. 78	-1. 75	1. 26

既然小样本情况下 Fama-MacBeth方法都能产生较好的结果，那么在大样本下情况如何呢？通过固定 N= 10，不断增加样本数 T进行模拟实验，也得到了较好的结果，见表 6.

模拟结果表明，无论样本数大小 (T)，在只存在时间效应的情况下，Fama-MacBeth方法都能产生较好的估计结果.

1. 3　同时存在个体效应和时间效应

1. 3. 1　固定效应模型的估计系数

表 6　只存在时间效应的 Fama-MacBeth方差估计 (固定 N= 10)

Table 6　Fama-MacBeth variance estimator only existing time effect(fixing N= 10)

时间效应	T= 25	T= 50	T= 250
True	0. 040 9	0. 020 7	0. 004 2
FM Estimator	0. 041 4	0. 020 5	0. 004 2
偏差 (%)	-1. 22	0. 97	0. 00

注：N= 10

以上模拟主要考虑了小样本的情况，因为实际中有些面板数据样本很小，比如公司数目只有几十个，或者更少. 根据上面的模拟，可以发现一些有趣的现象，尽管聚类回归方法调整贝塔标准差是有偏的，但是偏离的程度很小，事实上，在聚类数目为 50时，偏离的程度只有 4. 40% (见表 2 Arellano估计量. (0. 087 - 0. 091) /0. 091 = -0. 043 96). 这表明在小样本情况下，依然有足够的理由使用聚类回归方法进行调整. 另一方面，由于 Fama-MacBeth方法在只存在固定时间效应时是无偏的，在确定面板数据中只存在时间效应时，可以用该方法进行调整.

此外，很多作者在进行回归模拟时，引入了虚拟变量 α_i，这说明作者考虑了数据中可能存在的个体效应. 但是却很少有考虑时间效应的. 如果数据中不存在个体效应，只存在时间效应，这样做往往不能得到正确的结果. 如果同时存在个体效应和时间效应，使用 α_i 仅仅能够吸收个体效应，并不能吸收时间效应.

在上面的模拟中，由于事先设定了数据结构，这便能够采用正确的方法处理数据. 然而，实际中并不了解真实的数据结构. 如果数据中只存在时间效应，却认为存在个体效应而采用了吸收个体效应的方法，或者同时存在两种效应，但是却只考虑了个体效应，又会出现什么情况呢？为了进一步分析方法使用错误而造成的偏误，基于上述模型，同时考虑了个体效应和时间效应

$$\varepsilon_{it} = d_i + d_t + \mu_{it} \tag{14}$$

$$x_{it} = c_i + c_t + v_{it} \tag{15}$$

上式中的每个变量 (c、d、μ、v) 都是零均值、有限方差，并且相互独立的变量. 进一步，以 $\sigma^2_{c_i} + \sigma^2_{c_t}$ 和 $\sigma^2_{d_i} + \sigma^2_{d_t}$ 来度量总效应，控制其不变，把 $\sigma^2_{c_t}/(\sigma^2_{c_i} + \sigma^2_{c_t})$ 和 $\sigma^2_{d_t}/(\sigma^2_{d_i} + \sigma^2_{d_t})$ 作为时间效应大小的度量，显然，当 $\sigma^2_{c_t}/(\sigma^2_{c_i} + \sigma^2_{c_t})$ 和 $\sigma^2_{d_t}/(\sigma^2_{d_i} + \sigma^2_{d_t})$ 等于零时，表示不存在时间效应，这种情况上面已经考虑过；当 $\sigma^2_{c_t}/(\sigma^2_{c_i} + \sigma^2_{c_t})$ 和 $\sigma^2_{d_t}/(\sigma^2_{d_i} + \sigma^2_{d_t})$ 等于 100% 时，表示不存在个体效应. 另外，尽管同时存在个体效应和时间效应，对数据的处理方法是按照只存在个体效应来进行的，这是为了模拟实际中不了解数据的真实结构而采用了不当方法造成的错误. 为了研究时间效应逐渐增大对模型估计系数误差的影响，分别考虑了 $\sigma^2_{c_t}/(\sigma^2_{c_i} + \sigma^2_{c_t})$ 和 $\sigma^2_{d_t}/(\sigma^2_{d_i} + \sigma^2_{d_t})$ 等于 25%、50%、75%、100% 的情况. 举例来说，在 $\sigma^2_{c_t}/(\sigma^2_{c_i} + \sigma^2_{c_t})$ 和 $\sigma^2_{d_t}/(\sigma^2_{d_i} + \sigma^2_{d_t})$ 等于 100% 时表示只存在时间效应，但是却使用了消除个体效应的方法进行回归.

表 7是 T为 10年，N为 250的模拟结果. 可以看出，在存在时间效应时，即使很小 (占比 25%)，使用消去个体效应的固定效应模型估计也会产生较大的偏差，而随着时间效应的增加，产生的偏误越来越大. 尽管在模拟实验中显示，时间效应从 25% 增加到 100%，Arellano估计量的偏误仅仅从 95. 52% 增加到 98. 96%，增量似乎并不大，但是，25% 的时间效应就足以产生 95. 52% 的偏误，这就不得不重视实际数据中可能存在的时间效应.

表 7　同时存在个体效应和时间效应情况下各种指数模拟结果(大样本)

Table 7　Various index simulation results existing two effects simultaneously(Large size)

β估计量计算方法	时间效应			
	25%	50%	75%	100%
Truevarbeta	0.026 8	0.052 9	0.076 5	0.073 7
Avar	0.001 2	0.000 918	0.000 770	0.000 766
Kvar	0.001 2	0.000 947	0.000 815	0.000 815
Wvar	0.001 3	0.001 1	0.000 989	0.000 977
Ovar	0.000 802	0.000 810	0.000 836	0.000 926

注: 时间 T= 10,个体 N= 250.

表 8是 T为 10年, N为 50的模拟结果.与大样本情况类似,在存在时间效应时,使用消去个体效应的固定效应模型估计也会产生较大的偏差,并且偏误也随着时间效应的增加而增大.

表 8　同时存在个体效应和时间效应情况下各种指数模拟结果(小样本)

Table 8　Various index simulation results existing two effects simultaneously(Small size)

β估计量计算方法	时间效应			
	25%	50%	75%	100%
Truevarbeta	0.031 4	0.056 8	0.072 5	0.088 0
Avar	0.005 9	0.004 6	0.003 8	0.003 2
Kvar	0.006 2	0.004 9	0.004 0	0.003 4
Wvar	0.006 4	0.005 4	0.004 9	0.004 8
Ovar	0.004 1	0.004 1	0.004 2	0.004 3

注: 时间 T= 10,个体 N= 50.

1.3.2　Fama-MacBeth误差估计

同样对 Fama-MacBeth方法的估计结果进行了比较,从表 9可以看出,随着时间效应的增加,Fama-MacBeth方法估计的偏差迅速减小,当时间效应占总效应的一半时, Fama-MacBeth方法估计的误差只有 4.17%.然而,如果时间效应较小而个体效应较大,估计仍有较大的偏误,本文认为这是个体效应造成的原因.上文已经说过,Fama-MacBeth方法不适合估计只存在个体效应的数据结构.

表 9 同时存在个体效应和时间效应情况下 Fama-MacBeth方法模拟结果

Table 9 Fama-MacBeth simulation results existing two effects simultaneously

时间效应	25%	50%	75%
True	0.005 7	0.014 4	0.031 1
FM Estimator	0.003 7	0.013 8	0.029 8
偏差(%)	35.09	4.17	4.18

注: 时间 T= 10,个体 N= 250.

1.3.3　同时存在个体效应和时间效应的处理

当同时存在个体效应和时间效应时,对回归系数的估计会复杂得多.在实际数据中,由于横向的公司数目 N通常比纵向的时间数 T大,实证研究常常在每一时期使用虚拟变量吸收时间效应,然后按照公司聚类进行回归.如果时间效应是固定的,那么使用时间虚拟变量可以完全吸收时间效应,在公司数目较大的情况下,可以得到较好的结果.

Hsiao介绍了对同时存在个体效应和时间效应的面板数据估计方法.令

$$y_{it} = \alpha_i + \lambda_t + x_{it}\beta + \varepsilon_{it} \tag{16}$$

两边同时乘以

$$Q = I_{NT} - I_N \otimes \frac{1}{T} e_T e_T' - \frac{1}{N} e_N e_N' \otimes I_T + \frac{1}{NT} e_{NT} e_{NT}' \tag{17}$$

其中, I表示 阶单位矩阵; e表示 阶单位行向量; ⊗表示克罗内克积(Kronecker product),这样就能够消去 α_i和 λ_t的影响,从而估计 β和 β标准差.

最近,针对同时存在固定个体效应和固定时间效应的情况, Thompson[11] 和 Cameron等[12] 提出了下面的估计方法

$$V_{firm\&time} = V_{firm} + V_{time} - V_{white} \tag{18}$$

其中, V_{firm}表示按照公司聚类回归的 β方差系数, V_{time}表示按照时间聚类回归的 β方差系数, V_{white}表示考虑了异方差的 β方差系数.为避免重复计算,在等式右边减去 V_{white}.

2　结束语

众所周知,当面板数据中存在固定个体效应

或者固定时间效应时，直接使用 OLS估计的 β标准差是有偏的.然而，部分实证论文并没有考虑数据中可能存在的固定个体效应或者固定时间效应，部分论文虽然考虑了固定效应，但在固定个体效应和固定时间效应的辨别上，往往是比较主观的，换言之，论文中并没有提供可信的证据来表明存在何种效应.本文通过模拟面板数据进行回归分析，证明了在数据中存在不同效应时，采用不合适的估计方法依然会产生较大的偏误.具体来说，本文比较了 POOL OLS、固定效应模型和 Fama-Macbeth方法在不同的数据结构中的处理结果，并且通过模拟可以看到，对某种数据结构，如果采用了不正确的方法，回归结果的偏差将会非常大.

尽管实证研究中选择正确的模型非常重要，但是，针对实际中面板数据的复杂性，研究中并没有针对模型选择的统一标准.特别是对于数据是否存在时间效应这一问题，实证中往往过于主观.通过本文的模拟，实际上为判断实际数据中存在何种效应提供了一种准则：即以 OLS估计为基准，通过使用不同的模型进行估计，当固定个体效应模型估计与 OLS估计结果有较大的偏差时，（见上文模拟结果和实证数据结果）有理由相信面板数据中存在个体效应；如果固定时间效应模型估计与 OLS估计结果差别较大时，可以认为面板数据中存在时间效应；如果两种估计模型与 OLS估计结果均有较大差别时，可以认为面板数据中同时存在个体效应和时间效应.这样，通过进一步了解实际中面板数据的结构，采用正确的估计方法，减少犯错的概率.

在数据中只存在时间效应时，Fama-Macbeth方法能够产生无偏的 β标准差.通过使用固定时间效应模型也能够产生无偏的结果，但是必须在时间数比较多的情况下才成立.如果时间数较少，用固定时间效应模型仍然会产生较大的偏误.但是，由于实际中往往存在样本过小的问题，特别是许多面板数据的时间维度往往过小.因此，如何在小样本下提供较为精确的判断准则，是本文今后将要关注和研究的问题.

参 考 文 献：

[1] Kiefer N M. Estimation of fixed effect models for time series of cross-sections with arbitrary intertemporal covariance[J]. Journal of Econometrics, 1980, 14(2): 195-202.

[2] White H A. Heteroscedasticity-consistent covariance matrix estimator and a direct test of heteroskedasticity[J]. Econometrica, 1984, 48(1): 817-838.

[3] Arellano M. Computing robust standard errors for within-groups estimators[J]. Oxford Bulletin of Economics and Statistic, 1987, 49(4): 431-434.

[4] Kezdi G. Robust standard error estimation in fixed-effects panel models[C]//Econometrics. EconWPA, 2004, 0508018.

[5] Fama E, MacBeth. Risk, return and equilibrium: Empirical tests[J]. Journal of Political Economy, 1973, 81(3): 607-636.

[6] Newey W, Kenneth W. A simple, positive semi-definite, heteroscedastic and autocorrelation consistent covariance matrix[J]. Econometrica, 1987, 55(3): 703-708.

[7] Driscoll J C, Kraay A C. Consistent covariance matrix estimation with spatially dependent panel data[J]. Review of Economics and Statistics, 1998, 80(1): 549-560.

[8] Hsiao. Analysis of Panel Data[M]. Cambridge: Cambridge University Press, 2003.

[9] Phillips P C B, Moon H R. Linear regression limit theory for nonstationary panel data[J]. Econometrica, 1999, 67(1): 1057-1111.

[10] Petersen M A. Estimation standard errors in finance panel data sets: Comparing approaches[J]. Review of Financial Studies, 2007, 22(1): 435-480.

[11] Thompson S. Simple formulas for standard errors that cluster by both firm and time[C]//Harvard working paper. Available at SSRN, 2006. http://ssrn.com/abstract=914002.

[12] Cameron A C, Gelbach J B, Miller D L. Robust inference with multi-way clustering[C]//University of California-Davis Working Paper No. 327, 2006.

Simulation comparison on adaptability of econometrical models based on panel data

LIU Li-ya[1], DING Jian-ping[2], QIN Xiao[3], DAI Fei[1]

1. School of Finance, Shanghai University of Finance and Economics, Shanghai 200433, China;
2. Modern Finance Research Center, Shanghai University of Finance and Economics, Shanghai 200433, China;
3. Aetna School Management, Shanghai Jiaotong University, Shanghai 200052, China

Abstract: In recent years, there are more and more empirical papers using panel data in economic and management field. At the same time, theoretical results on how to process panel data are also increasing. However, one problem is that most theoretical methodologies on panel data are based on infinite sample, but we can only get limited samples in reality; another problem is that existing papers on using panel data don't do special treatment for panel data sample or don't explain why adopting their processing methodology. Therefore, the two contributions of this paper are as follows: 1) this paper compares the applicability of existing theoretical methodologies by simulating based on the three different data structures — that with individual effect, with time effect and with two effects simultaneously; 2) considering that estimation results have significant differences for various processing methods in small sample, this paper provides a criterion for judging what effects exist in the actual data.

Key words: panel data; fixed individual effect; fixed time effect; clustering regression standard-error coefficient; Fama-MacBeth methodology

附录

简单证明如下.

对于只存在时间效应的面板数据, 有如下式子

$$y_{it} = \lambda_t + x_{it}\beta + \varepsilon_{it}$$

与只存在固定个体效应不同, 在只存在固定时间效应时, 使用时间维度进行分组以估计组内估计量, 需要用到以下假设.

假设 A1: $E(\varepsilon_t | x_t, \lambda_t) = 0$ $(N = 1, 2, \cdots, T)$

假设 A2: $Var(\varepsilon_t | x_t, \lambda_t) = \sigma^2$.

如果假设 A1成立而 A2不成立. 估计组内方差的普通回归公式将会产生不一致的标准误, 该公式如下

$$\hat{Var}(\hat{\beta}) = \hat{\sigma}^2 (\ddot{x}'\ddot{x})^{-1}$$

然而, 既然

$$\left(\frac{1}{T}\ddot{x}'\ddot{x}\right)\sqrt{T}(\beta_{FE} - \beta) = \frac{1}{\sqrt{T}}\sum_{t=1}^{T} \ddot{x}_t'\ddot{v}_t$$

且 $E(\ddot{x}_t'\ddot{v}_t) = 0$, 前面等式右边是一个零均值随机变量的样本均值, 则当 N固定, T趋向于无穷时, 应用多元独立同分布观测的中心极限定理, 有

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} \ddot{x}_t'\ddot{v}_t \rightarrow N(0, E(\ddot{x}_t'\ddot{v}_t\ddot{v}_t'\ddot{x}_t))$$

因此, 当 N固定, T非常大时, 对于任意异方差和序列相关的情形, 可以得到组内估计的近似方差的稳健估计如下

$$\hat{Var}(\hat{\beta}_{FE}) = (\ddot{x}'\ddot{x})^{-1}\left(\sum_{t=1}^{T} \ddot{x}_t'\ddot{v}_t\ddot{v}_t'\ddot{x}_t\right)(\ddot{x}'\ddot{x})$$