非等间隔动态面板数据模型：基于半差分的估计方法和应用

乔坤元

内容提要：传统的估计方法并不适用于非等间隔动态面板数据模型，本文在总结已有文献的面板、非线性最小二乘和最短距离估计量的基础上，进一步提出了基于半差分方法的估计量以改进估计精度，与此同时着重强调了缺失观测期中遗漏变量的问题。蒙特卡洛模拟试验了这些估计量在有限样本中的表现，发现半差分估计量的精度最高，尤其是在考虑遗漏变量的情况下。本文将新得到的半差分估计用于中国劳动收入过程的研究中，实证结果表明，中国居民的劳动收入差距在拉大，并且劳动收入对收入冲击更加敏感。

关键词：非等间隔动态面板数据模型；半差分方法；缺失观测期的遗漏变量；劳动收入过程
中图分类号：F222.3　　文献标识码：A　　文章编号：1002－4565(2014)01－0098－08

The Unequally Spaced Dynamic Panel Data Model: Estimation and Applications

Qiao Kunyuan

Abstract: This paper provided nonlinear least squares, minimum distance and their one-step estimators for the unequally spaced dynamic panel data models with establishing their consistencies as well as asymptotic normality. The simulation study corroborates the results in the case of finite sample. We also apply these estimators to the real world problems, and the estimation results are basically consistent with the previous literature finally.

Key words: Unequally Spaced Dynamic Panel Data Model; Nonlinear Least Squares Estimation; Minimum Distance Estimation; One-Step Estimator

一、文献综述

估计动态面板数据模型最主要的方法是广义矩方法(GMM) Arellano 和 Bond，1991; Blundell 和 Bond，1998，该方法可以考察面板数据模型中因变量的动态效应，也即因变量滞后项对于当期的影响，因此在经验文献中的应有较为广泛(Millet 和 MoDonough，2013)。Arellano 和 Bond(1991)提出用更高阶的因变量滞后项作差分后方程的工具变量的“差分 GMM”法，而 Blundell 和 Bond(1998)为了克服弱工具变量和残差自相关的问题，结合差分 GMM 法和使用差分后的因变量滞后项作水平方程中因变量滞后项工具变量的“水平 GMM”法，提出了“系统 GMM”法，自此动态面板数据模型估计方法的框架基本成型并且被广泛应用。此后的文献从其他的角度对系统 GMM 进行了补充，如 Everaert(2013)给出的“正交滞后均值变换”(简称 OBMT 法)，提出了估计水平方程组的新的工具变量集合。

无论是 Aerllano 和 Bond(1991)、Blundell 和 Bond(1998)，还是最新的 Everaert(2013)，提出的估计方法都只针对标准的面板数据，然而在现实当中，有相当一部分面板数据是调查而来的，这些数据相邻两期的时间间隔可能并不一致，比如中国健康和营养调查数据(已公布的调查年份为 1989、1991、1993、1997、2000、2004、2006 和 2009 年)。如果真正的数据生成机制会产生相等时间间隔的面板数据，但是现实中的数据却是非等间隔的，那么传统外数据越来越多的情况下，需要对这种数据的估计方法进行更加系统的研究，提出精度更高的估计量。

研究数据非等间隔问题的文献主要关注时间序列模型(Millet 和 McDonough,2013) ,然而对非等间隔动态面板数据模型的研究并不多见。考虑不随时间变化、不可观测的个体效应影响的非等间隔动态面板数据模型的文献仅有 Mckenzie(2001) 、Millet 和 McDonough 外的影响纳入考察的范围,但是他研究的是非等间隔伪动态面板数据模型(dynamic pseudo-panel-models) ①。Qiao(2013a) 首次在文献中提出"非等间隔动态面板数据"的概念,但模型中并没有包含控制变量,这一忽略使得模型的适用性受到了限制: 控制变量的内生性和自相关性需要妥善的解决,否则估计量还是有偏的和不一致的。Millet 和 McDonough (2013) 和 Qiao(2013b) 从不同的角度探讨了非等间隔动态面板数据模型的估计方法,前者从不同的估计量角度出发说明传统的估计方法是有偏的和不一致的,并且基于差分、固定效应等方法给出了多个一致的估计量,而 Qiao(2013b) 则在 Qiao(2013a) 的基础上拓展了非线性最小二乘估计量和最短距度估计量,给出了更加一般的包含控制变量的模型和严格的证明。然而,这些估计量还远远不够、需要为应用研究者提供更多的、精度更高的一致估计量。另外,缺失观测期②中的遗漏变量问题会使已有的估计量依旧有偏不一致(Qiao,2013b) 。

本文将基于半差分法提出非等间隔动态面板数据模型的新的估计量,并且借助蒙特卡洛模拟的方法检验这些估计量在有限样本中的估计精度。研究发现,已有的动态面板数据模型的估计量在有限样本中的估计精度不高,而新提出的半差分估计量提升了估计精度,尤其是在考虑到缺失观测期中的控制变量可能会进入残差项的时候,基于半差分方法的估计量精度最高,受到的影响最小,而 Millet、McDonough(2013) 和 Qiao(2013a、2013b) 给出的估计量因为没有完全消除个体效应而依旧会产生一定的误差。新得到的半差分估计量被用来考察中美两国劳动收入过程,实证结果表明这一过程需要充分地考虑到个体的异质性(Guvenen,2007) ,中国居民的劳动收入差距有拉大的趋势,而美国的情况却相反,并且中国居民的劳动收入对收入冲击更加敏感。

与以往的文献相比,本研究: ①完整地总结了以往文献中讨论非等间隔动态面板数据模型的估计量,并且使用蒙特卡洛模拟的方法对各个渐进一致或者偏差较小的估计量进行了系统的对比考察; ②提出了基于半差分方法的一致估计量,丰富了非等间隔动态面板数据模型的估计方法,并且提高了估计的精度; ③基于新的估计方法,系统考察了中国居民劳动收入过程,并且通过和美国的情况进行对比,可以更好地了解中国居民劳动收入的状况。经验研究者在讨论一些需要借助非等间隔动态面板数据模型的问题时,可以根据估计精度和计算方便在这些估计量中加以选择。

本文接下来的内容安排如下。第二部分总结已有的非等间隔动态面板数据模型的估计量,提出半差分估计量,并且考虑缺失观测期中控制变量可能进入残差项的问题,从而给出这些潜在的估计方法。第三部分使用蒙特卡洛模拟的方法对比了这些一致估计在有限样本中的估计精度。最后一部分给出结论和未来研究的方向。

二、估计方法

(一) 模型设定

考虑如下的数据生成机制:

$$y_{i,t_j} = \alpha + x'_{i,t_j}\beta + \gamma y_{i,t_{j-1}} + \mu_i + \epsilon_{i,t_j}$$

其中,β 是一个关于控制变量的 $k \times 1$ 系数向量,假设数据中有 T 年、τ 期和 S 个不同的时间间隔,即 $t_j, j = 1,2,\cdots,\tau$,同时有 $t_j > 0, \forall j$。ϵ_{i,t_j} 序列不相关,也即其方差矩阵 V 是一个对角矩阵,但可以允许异方差的存在,使用 FGLS 进行修正(Qiao,2013b) 。

定义 $1 = t_1 < t_2 < \cdots < t_\tau = T$,具体地,$s_{t_j} = t_j - t_{j-1}$ 且 $\forall s_{t_j}$,有 $s_{t_j} = \{s_1, s_2, \cdots, s_S\}$。

$$\begin{aligned} y_{i,t_j} &= \gamma^{t_j - t_{j-1}} y_{i,t_{j-1}} + \sum_{h=0}^{t_j - t_{j-1} - 1} \gamma^h \left(\alpha + x'_{i,t_{j-h}}\beta + \mu_i + \epsilon_{i,t_j-h}\right) \\ &= \gamma^{s_{t_j}} y_{i,t_j - s_{t_j}} + x'_{i,t_j}\beta + \left(\frac{1-\gamma^{s_{t_j}}}{1-\gamma}\right)(\alpha + \mu_i) + \tilde{\epsilon}_{i,t_j} \end{aligned}$$

$$\tilde{\epsilon}_{i,t_j} = \sum_{h=1}^{s_{t_j}-1} \gamma^h \left(\alpha + x'_{i,t_j-h}\beta + \mu_i + \epsilon_{i,t_j-h}\right) + \epsilon_{i,t_j}$$

如果相邻两期观测之间的时间间隔是相等的话,那么 $s_{t_j} = 1$,上述等式将退化为标准的动态面板数据模型。

---

① 伪面板数据中每一期的个体都不同,但是不同的个体根据人口特性属于同一组群,因此估计时需要通过取组群的平均值从而消除个体固定效应。

② 如包含 1990、1992 和 1996 年的面板数据中"1994 年的数据",即为"缺失观测期"。

(二)已有的非等间隔动态面板数据估计量

1. 面板估计量。

Millet 和 McDonough(2013)说明,直接使用截面或者传统的动态面板数据模型会得到较大渐进误差的有偏的,不一致的在计量,主要是因为不随时间变化、不可观测的个体效应并不能被消除,而它又与自变量(因变量的滞后项)相关。

Qiao(2013b)给出的观测误差修正的广义矩估计方法(GMMC),提出使用因变量滞后项工具变量可以解决这个难题,正如 Blundell 和 Bond(1998)提出的系统 GMM 法,对于差分方程组,可以使用(所有)水平高阶滞后项作工具变量,而对于水平方程组可以使用(所有)差分高阶滞后项作工具变量。这样得到的估计量(FD-GMM 和 FE-GMM)依旧是有偏不一致的,但是偏差相对较小。具体证明的思路可以参照 Arellano 和 Bond(1991)以及 Blundell 和 Bond(1998)。另外,Millet 和 McDonough(2013)提出还可以使用 Mundlak(1978)给出的相关随机效应方法首先估计不随时间变化、不可观测的个体效应,使用 $\mu_i+\alpha$ 对 $\bar{x}'_i$ 进行回归,再使用 GMM 法得到相应估计(FD-R-GMM 和 FE-R-GMM)。

另外,Millet 和 McDonough(2013)说明 Everaert(2013)提出的 OBMT 法也可以为水平方程组提供新的工具变量集合。在控制变量 $x'_{i,t}$ 外生、没有序列自相关并且 T 趋于无穷的情况下,对水平方程组进行估计时可以获得一组新的工具变量,该工具变量由 $y_{i,t_{j-1}}$ 对滞后均值 $\frac{1}{j}\sum_{h=1}^{j} y_{i,t_{j-h}}$ 进行最小二乘估计得到的残差项给出,如果不满足上述情况(如观测期数 T 有限而观测个数 N 趋于无穷)得到的估计量有可以忽略的偏差。因此,可以在基于 GMM 方法的固定效应估计量的基础上再加上 Everaert(2013)给出的新的工具变量对上述方程组(模型)进行估计,得到两个不同的估计量(FE-OBMT 和 FE-R-OBMT)①。

需要说明的是,这几个估计量依旧要求"缺失观测期没有变量进入残差项"的假设,否则遗漏变量的问题会放大产生的偏差。降低估计的精度。

2. 非线性最小二乘估计量。

Qiao(2013a)在不包含控制变量的模型中给出了非线性最小二乘估计量以及证明,类似地,加入了控制变量的模型同样需要将数据各期进行堆叠,通过转换矩阵得到:

$$\begin{pmatrix}\bar{y}_3\\ \bar{y}_5\\ \bar{y}_9\\ \bar{y}_{12}\\ \bar{y}_{16}\\ \bar{y}_{18}\\ \bar{y}_{21}\end{pmatrix}=\begin{pmatrix}\bar{y}_1 & 0 & 0 & \bar{x}'_2 & \bar{x}'_3 & 0 & 0 & 100\\ \bar{y}_3 & 0 & 0 & \bar{x}'_4 & \bar{x}'_5 & 0 & 0 & 100\\ 0 & \bar{y}_5 & 0 & \bar{x}'_8 & \bar{x}'_9 & \bar{x}'_7 & \bar{x}'_6 & 111\\ 0 & 0 & \bar{y}_9 & \bar{x}'_{11} & \bar{x}'_{12} & \bar{x}'_{10} & 0 & 110\\ 0 & \bar{y}_{12} & 0 & \bar{x}'_{15} & \bar{x}'_{16} & \bar{x}'_{14} & \bar{x}'_{13} & 111\\ \bar{y}_{16} & 0 & 0 & \bar{x}'_{17} & \bar{x}'_{18} & 0 & 0 & 100\\ 0 & 0 & \bar{y}_{18} & \bar{x}'_{20} & \bar{x}'_{21} & \bar{x}'_{19} & 0 & 110\end{pmatrix}\begin{pmatrix}\gamma^2\\ \gamma^4\\ \gamma^3\\ \gamma\beta\\ \beta\\ \gamma^2\beta\\ \gamma^3\beta\\ \alpha(1+\gamma)\\ \alpha\gamma^2\\ \alpha\gamma^3\end{pmatrix}+\begin{pmatrix}\epsilon_3\\ \epsilon_5\\ \epsilon_9\\ \epsilon_{12}\\ \epsilon_{16}\\ \epsilon_{18}\\ \epsilon_{21}\end{pmatrix}$$

令要估计的参数为 $\pi=(\alpha,\beta,\gamma)$,用 $h(\pi)$ 来表示定义在 α、β 和 γ 之上的非线性系数向量,那么 $\forall S\geqslant 2$,可以将上述矩阵形式的方程写作:$\mathbf{y}=\widetilde{\mathrm{X}}h(\pi)+\epsilon$。如果识别条件 $h(\hat{\pi}_j)\neq h(\hat{\pi}_0)$,$\forall \pi_j\neq\pi_0$($\pi_0$ 是 π 的真实值)得到满足,则利用非线性最小二乘估计法可以得到 $h(\pi)$ 的一致估计量 $h(\hat{\pi})$,再由反函数的性质可以得到 $\hat{\pi}$(NLS)的一致性和渐进正态性,具体的证明和数据转换原理见 Qiao(2013b),这里不再赘述。

另外,Qiao(2013a)提出了更加容易计算的非线性最小二乘一步估计量,该估计量可以通过高斯—牛顿回归法得到,$\mathbf{y}-\widetilde{\mathrm{X}}h(\hat{\pi}_s)=\widetilde{\mathrm{X}}\frac{\partial h(\pi)}{\partial\pi}\Big|_{\pi=\hat{\pi}_s} * c+\hat{\epsilon}$,其中 $\hat{\pi}_s$ 是一致估计量,则一步估计量(OS-NLS)$\hat{\pi}_{osnls}=\hat{\pi}_s+\hat{c}$。Qiao(2013a,2013b)分别在简单模型和包含控制变量的一般模型中证明一步最小二乘估计量与非线性最小二乘估计渐进相等,而一步估量更加容易计算,对于误差组成的形式也没有限制。

① 需要注意的是,OBMT 方法并不适用于差分方程组,因此 FD 估计量以及后文的半差分估计量都不可以使用 OBMT 法新增添的工具变量集合。

这两个估计量也依赖于“缺失观测期没有变量进入残差项”的假设,也即遗漏变量的问题同样会影响到这两个估计量的精度。从转换矩阵的方式不难看出,如果存在遗漏变量的话,得到的转换矩阵 $\tilde{X}$ 也是内生的,因此得到的 $h(\hat{\pi})$ 和随后对于模型参数的估计量 $\hat{\pi}$ 也是有偏的和不一致的。

3. 最短距离估计量。

Qiao (2013a) 提出通过最短距离估计量 (OPMD) 来恢复 π 的 S 个初始值,即:

$$\hat{\pi}_{opmd} = \arg\min_{\pi}\{h(\hat{\pi}) - h(\pi)\}'\Omega^{-1}\{h(\hat{\pi}) - h(\pi)\}$$

其中,$\Omega = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\hat{V}\tilde{X}(\tilde{X}'\tilde{X})^{-1}$,$\hat{V}$ 是方差矩阵的估计。

与非线性最小二乘法一样,可以进一步得到最短距离的一步估计量(OSMD):

$$\hat{\pi}_{osmd} = \hat{\pi}_s + (\hat{G}'\Omega^{-1}\hat{G})^{-1}\hat{G}'\Omega^{-1}(h(\hat{\pi}) - h(\hat{\pi}_s))$$

其中,$\hat{G} = \frac{\partial h(\pi)}{\partial \pi}\Big|_{\pi=\hat{\pi}_s}$。

这两个估计量仍然依赖于“缺失观测期没有变量进入残差项”的假设,因此遗漏变量的问题也会影响到它们的估计精度,使得估计量有偏不一致。

(三) 半差分估计量

面板估计量没有消除不随时间变化、不可观测的个体效应,而非线性最小二乘估计量和最短距离估计量则通过转换矩阵避免了不随时间变化、不可观测的个体效应可能对估计带来的偏差,但是上述方法均没有消除不随时间变化、不可观测的个体效应。然而,半差分的方法可以做到。考虑如下的半差分模型:

$$y_{i,t_j} - \varphi_j y_{i,t_{j-1}} = \gamma^{s_{tj}} y_{i,t_j - s_{tj}} - \varphi_j \gamma^{s_{tj-1}} y_{i,t_j - s_{tj-1}} + (x_{i,t_j} - \varphi_j x_{i,t_{j-1}})\beta + (\alpha + \mu_i)\frac{\gamma^{s_{tj-1}} - \varphi_j \gamma_{s_{tj}}}{1-\gamma} + \tilde{\epsilon}_{i,t_j} - \varphi_j \tilde{\epsilon}_{i,t_{j-1}}$$

其中,$\varphi_j = \frac{1-\gamma^{s_{tj}}}{1-\gamma_{s_{tj-1}}}$,因此不随时间变化、不可观商量的个体效应可以被消除。

然而,φ_j 不一定已知,因此需要分情况讨论。如果 φ_j 是已知的,那么可以通过面板估计量提到的工具变量法进行估计,得到已知 φ_j 的估计量(QD-K-GMM);如果 φ_j 并不是已知的,那么可以使用三种方法进行估计。第一,使用联合置信区间法先估计 γ 从而得到对 φ_j 的估计,进而得到估计量(QD-UCL-GMM)。第二,借鉴 Lee 等(2012) 的思路使用分位数回归法,对某一个工具变量向量 z_{i,t_j}(或使用工具变量进行一阶段估计,算出因变量滞后项的预测值),估计下式:

$$y_{i,t_j} - \left(\frac{1-\gamma^{s_{tj}+s_{tj-1}}}{1-\gamma^{s_{tj-1}}}\right)y_{i,t_{j-1}} + \left(\frac{\gamma_{s_{tj-1}} - \gamma^{s_{tj}+s_{tj-1}}}{1-\gamma^{s_{tj-1}}}\right)y_{i,t_{j-2}} = z_{i,t_j}\eta + (x_{i,t_j} - \varphi_j x_{i,t_{j-1}})'\beta + \tilde{\epsilon}_{i,t_j} - \varphi\tilde{\epsilon}_{i,t_{j-1}}$$

令 W 为正定的权重矩阵,可以通过 $\hat{\gamma} = \arg\min_{\gamma}\hat{\eta}(\gamma)'W\hat{\eta}(\gamma)$ 得到最优的 $\hat{\gamma}$(QD-QR-GMM)。第三,利用 Nauges 和 Thomas(2003) 给出的广义矩方法,矩条件为:

$$E[x_{i,t_{j-1}}(\tilde{\epsilon}_{i,t_j} - \varphi_j\tilde{\epsilon}_{i,t_{j-1}})] = 0,\ E[x_{i,t_j}(\tilde{\epsilon}_{i,t_j} - \varphi_j\tilde{\epsilon}_{i,t_{j-1}})] = 0,\ E[y_{i,t_{j-2}}(\tilde{\epsilon}_{i,t_j} - \varphi_j\tilde{\epsilon}_{i,t_{j-1}})] = 0$$

通过上述的矩条件可以得到对 γ 的估计进而估计出 φ_j。得到 φ_j 的估计值之后,可以得到一致估计量(QD-GMM-GMM)。

半差分虽然消除了不可观测、不随时间变化的个体固定效应,但是依旧需要注意缺失观测期中遗漏变量的问题。但是从直观上看,半差分是一致的,会好于面板估计量,而且没有像非线性最小二乘法和最短距离法那样通过调整理常数项数值使得个体固定效应的均值为 0,从而通过取均值的方法避开了个体固定效应的影响,从理论上看半差分估计量由于完全地消除了个体固定效应从而获得了更高的精度。

(四) 拓展: 缺失观测期数中的变量

需要注意的是,非等间隔面板数据带来最大的问题是缺失时期中可能会包含一些进入残差项的变量,这一遗漏变量的问题会使得工具变量无效; 上述的估计方法均会因此得不到对非等间隔动态面板数据模型一致的估计量。在现实当中,控制变量有可能由于自相关性进入残差项,因此需要对上述方法进行拓展,考虑遗漏变量的问题。

对于非等间隔动态板数据模型,需要估计的是残差项中的 $\sum_{h=0}^{s-1}\gamma^h x'_{i,t_{j-h}}\beta$。结合 Millet 和 McDonough (2013) 总结出的估计思路,一共有四种方法可供选择。

(1) 滞后一期值法:

$$\sum_{h=0}^{s-1}\gamma^h x'_{i,t_{j-h}}\beta = x'_{i,t_{j-1}}\left(\frac{\gamma-\gamma^s}{1-\gamma}\right)\beta$$

(2) 当期值法:

$$\sum_{h=0}^{s-1}\gamma^h x'_{i,t_{j-h}}\beta = x'_{i,t_j}\left(\frac{\gamma-\gamma^s}{1-\gamma}\right)\beta$$

(3) 平均值法:

$$\sum_{h=0}^{s-1}\gamma^h x'_{i,t_{j-h}}\beta = \frac{x'_{i,t_{j-1}} + x'_{i,t_j}}{2}\left(\frac{\gamma-\gamma^s}{1-\gamma}\right)\beta$$

(4) 参数为 ρ 的一阶自回归法:

$$\sum_{h=0}^{s-1}\gamma^h x'_{i,t_{j-h}}\beta = x'_{i,t_{j-1}}\left(\sum_{h=0}^{s-1}\rho^{s-h}\gamma^h\right)\beta$$

从理论上来看,这四种方法给出的估计值是渐进相等的,但平均值法和参数为 ρ 的一阶自回归法利用了更多的数据信息。

三、蒙特卡洛模拟

(一) 数据生成机制

与 Everaert(2013)、Millet 和 McDonough(2013) 类似,定义数据生成机制:

$$y_{i,t} = \alpha_i + x'_{i,t}\beta + \gamma y_{i,t-1} + \epsilon_{i,t}$$

$$x_{i,t} = \theta\alpha_i + \rho x'_{i,t-1} + \xi_{i,t}$$

$$y_{t,0} = \alpha_i\frac{1-\rho+\beta\theta}{(1-\gamma)(1-\rho)} + \zeta_{i,0}$$

$$x_{i,0} = \frac{\theta\alpha_i}{1-\rho} + \xi_{i,0}\sqrt{\frac{1}{1-\rho^2}}$$

其中,$\epsilon_{i,t} \sim^{i.i.d.} N(0,\sigma_\epsilon^2)$,$\alpha_i \sim^{i.i.d.} N(0,(1-\gamma)^2\sigma_\epsilon^2)$,$\xi_{i,t} \sim^{i.i.d.} N\left(0,\left(\sigma_s^2 - \frac{\gamma^2}{1-\gamma^2}\sigma_\epsilon^2\right)\frac{(1-\gamma\rho)(1-\gamma^2)(1-\rho^2)}{\beta^2(1+\rho\gamma)}\right)$,$\zeta_{i,0} \sim^{i.i.d.} N\left(0,\frac{\sigma_c^2}{1-\gamma^2} + \frac{\beta^2\sigma_\xi^2(1+\gamma\rho)}{(1-\gamma\rho)(1-\gamma^2)(1-\rho^2)}\right)$。

根据 Everaert(2013),σ_s^2 为 $y_{i,t}$ 中可以被 $y_{i,t-1}$ 和 $x'_{i,t}$ 解释而不能被残差和不随时间变化、不可观测的个体效应的方差部分,令这一数值为 2,且 $\epsilon_{i,t}$ 服从一个标准正态分布。

考察以下 4 个数据生成机制:

①$x'_{i,t}$ 外生且序列不相关:$Cov(x'_{i,t},\alpha_i)=0$,$Cov(x'_{i,t},x'_{i,t-1})=0$:$\theta=0$,$\rho=0$

②$x'_{i,t}$ 非外生且序列不相关:$Cov(x'_{i,t},\alpha_i)\neq0$,$Cov(x'_{i,t},x'_{i,t-1})=0$:$\theta=1$,$\rho=0$

③$x'_{i,t}$ 外生且序列相关:$Cov(x'_{i,t},\alpha_i)=0$,$Cov(x'_{i,t},x'_{i,t-1})\neq0$:$\theta=0$,$\rho=0.5$

④$x'_{i,t}$ 非外生且序列不相关:$Cov(x'_{i,t},\alpha_i)\neq0$,$Cov(x'_{i,t},x'_{i,t-1})\neq0$:$\theta=1$,$\rho=0.5$

这里放松了控制变量 $x'_{i,t}$ 的外生性和非序列相关的假设,尤其是允许控制变量有较强的序列相关性,进而比较不同的估计量在有限样本中的表现。

与大多数蒙特卡洛模拟的设定类似,模拟的次数定为 1000 次。为了仿照中国健康和营养调查中数据的结构,时期定为 1、3、5、9、12、16、18 和 21 共 8 期,对应于已公布数据的 1989、1991、1993、1997、2000、2004、2006 和 2009 年。

(二) 模拟结果

Millet 和 McDonough(2013) 模拟了传统的横截面、面板估计量发现,它们的估计精度较低,偏差很大,因此在这里仅模拟总结出的 14 个渐进误差较小的和一致的估计量。表 1 给出了不考虑缺失观测期中遗漏变量时使用这 14 个估计量对模拟产生的数据进行估计得到的平均绝对百分比误差。对比 Millet 和 McDonough(2013) 的结果,依旧可以看到这 14 个估计量的精度要远远高于传统横截面和面板估计量①。

表 1 各个估计量的估计精度:不考虑缺失观测期中的遗漏变量

估计量	$\gamma=0.75$				$\beta=0.75$			
数据生成机制	①	②	③	④	①	②	③	④
FD-GMM	4.874	5.829	6.481	10.286	2.240	2.103	3.641	4.489
FD-R-GMM	5.104	5.342	6.165	9.789	2.290	2.259	3.551	4.612
FE-GMM	4.651	5.309	6.255	10.495	1.996	2.255	3.377	4.857
FE-R-GMM	4.665	5.068	6.160	10.693	2.127	2.247	3.714	4.510
FE-OBMT	5.442	4.993	6.165	10.753	1.913	2.422	3.649	4.603
FE-R-OBMT	5.381	4.899	6.670	10.851	1.974	2.141	3.551	4.680
NLS	3.545	3.070	4.575	7.424	1.965	2.074	2.405	3.983
OSNLS	3.166	3.214	4.819	7.043	1.844	2.108	2.349	3.551
OPMD	3.222	3.378	4.690	6.487	1.847	2.017	2.549	3.625
OSMD	3.556	3.387	4.231	6.728	1.857	2.057	2.519	3.629
QD-K-GMM	1.797	2.025	2.261	3.319	1.399	1.567	2.004	2.283
QD-UCL-GMM	1.924	2.416	2.376	3.495	1.492	1.640	2.210	2.943
QD-QR-GMM	1.922	2.191	2.598	3.447	1.486	1.627	2.193	2.870
QD-GMM-GMM	1.918	2.105	2.298	3.433	1.463	1.619	2.184	2.565

注:结果通过 1000 次模拟,横截面观测数量为 500,一共包含第 1、3、5、9、12、16、18 和 21 共 8 期的数据。

在控制变量没有内生性并且没有序列相关的时候,面板估计量(FD-GMM、FD-R-GMM、FE-GMM、FE-R-GMM、FE-OBMT 和 FE-R-OBMT) 的估计误差与非线性最小二乘估计量(NLS 和 OS-NLS) 以及最短距离估计量(OPMD 和 OSMD) 的类似,其中面板估计量对 γ 的估计误差在 5% 左右,而对 β 的估计

① 在这一模型设定下,传统的横截面和面板估计量也可以通过模拟得到,模拟结果显示这两个估计量的偏差较大,其中横截面的估计量对 γ 的百分比误差在 30% ~90% 之间,面板估计量的百分比误差在 20% ~50% 之间,与汇报的相比可以看出已民用工业的估计量和半差分估计量大大提升了估计的精度。

误差在 2% 附近,非线性最小二乘估计量(NLS 和 OS-NLS) 和最短距离估计量(OPMD 和 OSMD) 对两个参数的估计误差稍低,对 γ 和 β 的估计误差可以控制在 3% 和 2% 左右。一个更加明显的现象是,半差分估计量(QD-K-GMM、QD-UCI-GMM、QD-QR-GMM 和 QD-GMM-GMM) 的精度最高,可以将 γ 和 β 的估计误差分别控制在 2% 和 1.5% 之内。考虑控制变量的内生性,也即表 1 中的数据生成机制②,可以看到相对精度与不考虑内生性之间的相似: 面板估计量因为存在偏差并且不是一致的,因此它们的估计偏差较大,对 γ 和 β 的估计误差分别在 5% 和 2.2% 左右,而非线性最小二乘和最短距离估计量的精度稍高,误差可以减小至 3.1% 和 2.1%,说明控制变量的内生性会增加基于传统 GMM 方法的面板估计量的渐进误差,而半差分估计量的精度依旧是最高的,使用它们对 γ 和 β 进行估计,误差仅有 2.5% 和 1.6%。仅考虑控制变量的自相关而不考虑内生性的结果与仅考虑内生性而没有序列相关的相对精度类似,面板估计量的误差接近 6% 和 3.5%,非线性最小二乘和最短距离估计量的误差为 5% 和 2.5%,而半差分估计量的误差可以降低到 2.5% 和 2.1% 左右。最后看控制变量既有内生性又有自相关的情况,面板估计量对 γ 和 β 的估计误差被进一步放大,误差达到了 10% 和 4.5%,而线性最小二乘和最短距离估计量的误差也有 7% 和 4%,表现最好的依旧是半差分估计量,对 γ 和 β 的估计误差可以控制在 4% 和 3% 之内。

对比估计量 FE-GMM 和 FE-OBMT、FE-R-GMM 和 FE-R-OBMT,可以发现使用不同的工具变量的组合对于估计精度的影响很小,几乎可以忽略不计。虽然后者的工具变量数量更多会使得估计更加有效,但是增加的工具变量个数引起的过度识别问题使得估计的渐进误差也同时增加,因此使用不同数目的工具变量组合得到的估计误差基本一致。另外,是否使用 Mundlak(1978) 给出的相关随机效应方法结果的影响也比较有限,无法看出 FD-GMM 和 FD-R-GMM、FE-GMM 和 FE-R-GMM 以及 FE-OBMT 和 FE-R-OBMT 之间的差别。非线性最小二乘和最短距离估计理也都与它们的一步估计量的精度类似,NLS 和 OS-NLS 以及 OPMD 和 OSMD 之间的平均绝对百分比误差的差别很小,与 Qiao(2013a) 的类似。

对于新提出的半差分估计量而言,如果已知 φ_j,那么估计的精度最高。然而现实往往需要对这一数值进行估计,使用联合置信区间(UCI) 、分位数回归(QR) 或者广义矩方法先估计 φ_j 再进行半差分估计,得到的估计精度也相差不大,其中估计精度最高的是广义矩方法,接下来分别是分位数回归法和联合置信区间法。表 1 的结果说明各个估计量的相对精度结果在 4 种不同的数据生成机制中基本一致,对模拟数据的估计结果是稳健的①。

考虑到缺失观测期中的变量可能会进入残差项进而造成遗漏变量的问题,可以使用滞后一期值、当期值、平均值和参数为 ρ 的一阶自回归法对“遗漏变量”进行估计。全部四种估计方法在模拟的数据中得到了类似的估计结果,这 14 个估计量的相对精度也比较接近,碍于篇幅和重要性,这里以参数为 ρ 的一阶自回归法得到的对遗漏变量的估计为主。

表 2 给出了使用参数为 ρ 的一阶自回归法估计缺失观测期的遗漏变量时,各个估计量的平均绝对百分比误差。对比表 1,可以看到可能存在的遗漏变量问题会对估计量的精度产生重要的影响,误差均由于遗漏变量的存在而被放大,但是面板估计量、非线性最小二乘、最短距离估计量和半差分估计量的相对精度没有发生变化: 无论是否考虑控制变量的内生性、自相关性,这四组估计量的相对精度从高到低依次是半差分估计量、非线性最小二乘估计量、最短距离估计量和面板估计量。考虑内生性和序列相关均会放大误差,而后者对误差的影响更大,两者一并考虑时估计误差最大,但是半差分估计量依旧可以把对 γ 和 β 的估计误差控制在 5% 之内。是否使用 OBMT 法增加一组工具变量、Mundlak(1978) 给出的相关随机效应方法对于结果几乎没有影响,非线性最小二乘和最短距离也依然和它们的一步估计量的估计精度接近,半差分估计量中精度最高的依旧是假设 φ_j 已恬的 QD-K-GMM 估计量,接下来分别是广义矩方法、分位数回归法和联合置信区间法得到的估计量。

综合表 1 和表 2,可以看到相对估计精度并不

① 与 Millet 和 McDonough(2013) 的类似,本研究的工作论文版本 Qiao(2013b) 同时使用了均方根误差来衡量这些估计量在有限样本中的估计精度,得到了类似的相对精度考察结果,碍于篇幅和重要性,结果没有汇报出。

表 2　各个估计量的估计精度: 考虑缺失观测期中的变量(参数为 ρ 的一阶自回归法)

估计量	$\gamma=0.75$				$\beta=0.75$			
数据生成机制	①	②	③	④	①	②	③	④
FD-GMM	9.177	9.959	14.944	17.074	3.693	6.779	6.528	6.932
FD-R-GMM	8.737	9.419	15.338	18.267	3.470	5.591	6.811	6.902
FE-GMM	8.989	9.284	16.177	19.879	3.579	5.889	6.529	6.892
FE-R-GMM	8.999	8.829	16.905	19.886	3.586	5.926	6.644	6.998
FE-OBMT	9.155	8.113	15.460	19.934	3.463	5.599	6.641	7.003
FE-R-OBMT	9.277	8.390	15.018	19.563	3.452	5.432	6.493	7.178
NLS	6.145	6.668	6.858	8.469	3.536	4.832	4.911	5.655
OSNLS	5.799	6.006	6.569	8.044	3.164	4.733	5.173	6.036
OPMD	5.607	6.527	6.709	8.095	3.211	4.612	5.138	6.399
OSMD	5.816	6.695	6.805	8.152	3.335	4.745	5.299	6.245
QD-K-GMM	2.271	3.131	3.822	4.018	2.466	3.095	3.577	4.014
QD-UCL-GMM	2.942	3.932	4.524	4.699	2.840	3.662	3.479	4.646
QD-QR-GMM	2.712	3.543	4.431	4.229	2.665	3.554	3.659	4.567
QD-GMM-GMM	2.875	3.269	4.091	4.105	2.550	3.382	3.550	4.368

注: 结果通过 1000 次模拟,横截面观测数量为 500,一共包含第 1、3、5、9、12、16、18 和 21 共 8 期的数据。

会由于缺失观测期的遗漏变量而发生变化,这说明对于模拟数据的估计结果是稳健的,但是需要注意的是遗漏变量可能会对估计结果产生重要的影响。另外,半差分估计量的精度较高,是对已有估计量的一个改进,尤其是在缺失观测期有遗漏变量的情况下,半差分估计量依然能保持较高的估计精度。

四、应用

(一) 数据

由于计量方法的缺失,目前几乎没有文献涉及中国居民的劳动收入过程问题,然而对这一问题的研究是必要的,它有助于理解中国居民劳动收入的分布和差距情况,而使用美国的数据进行对比可以更好地了解这一状况。中国的数据来自于中国健康和营养调查(CHNS) ,该数据由中美合作调查而来,涵盖 9 个省 1989、1991、1993、1997、2000、2004、2006 和 2009 共 8 年 4400 户共 26000 人的随机样本数据(2011 年的数据正在勘误而没有公布) ,这一数据翔实地记录了家庭的收入情况和明细门类,详见 Qiao (2013a) 对这一数据的阐述。与 Qiao(2013a) 类似,样本选取了包含收入、户主年龄、性别、教育、职业、民族以及他们是在城市还是在农村等有效信息的住户进行研究,并且去除了增长过快或者过慢的观测值,最终得到了包含 450 个横截面观测的平衡面板数据。美国的数据使用综合社会调查数据(General Social Survey,GSS) ,样本包含代表性成年人口劳动收入的相关数据,调查年份为 1972 - 1993 年除了 1979、1981 和 1992 年每年一次,1994 年之后每两年一次。按照 Guvenen(2007) 给出的变量选择包含有效信息的个体,一共抽出了 1517 户。为了将中国和美国的情况进行对比,GSS 数据本参照 CHNS 的年份进行选取,得到了 1989、1991、1993、1996、2000、2004、2006 和 2008 年的数据。两套数据中各期劳动收入的均值显示,中国居民的劳动收入还是远低于美国,但是增长速度较快。

(二) 实证结果

根据 Guvenen(2007) ,劳动收入过程的考察主要在于上一期劳动收入对于当期的影响,同时控制暂时性和永久性收入冲击(用方差衡量) ,具体的技术细节详见 Gunvenen(2007) 和 Qiao(2013a) 。

Guvenen(2007) 认为传统的观点由于忽略了个体的异质性而得出了错误的结论: 劳动收入并不服从一个随机游走的过程,而是收敛的,他在考虑个体的异质性之后得到滞后项的估计系数为 0.8,并且不确定的收入冲击的影响也不是非常重要。然而,受限于计量方法,对于劳动收入过程的研究只能使用等间隔面板数据(如 Guvenen(2007) 使用的 PSID 数据) ,这样可能使得结果比较依赖于不同的数据源,而如果对诸如 GSS 这样的非等间隔面板数据的某些年份进行删除又会丧失重要的数据信息。本研究总结并提出的非等间隔动态面板数据模型的估计量可以解决研究方法的问题,这里使用半差分估计量进行研究。

本文对中美两国劳动收入过程的考察显示,使用这四个半差分估计量无论是对中国的样本还是美国的样本进行估计,都得到了类似的结果。为了更加全面地考察劳动收入过程,本文同时使用受限的收入档案(Restricted Income Profiles,RIP) 模型和异质性的收入档案(Heterogeneous Income Profies,HIP) 模型。

对 CHNS 数据样本的估计结果显示,无论是否考虑个体的异质性(也即使用 RIP 模型或者 HIP 模型) ,中国居民劳动收入都呈现了明显的发散趋势,劳动收入差距在 1989 - 2009 年之间被进一步拉大,并且中国居民的劳动收入对收入的冲击相比于美国更加敏感,这些结果与 Qiao(2013a) 的类似。

对于 GSS 数据样本,使用 RIP 模型得到的结果与大多数研究美国劳动收入过程的文献类似,

劳动收入服从一个随机游走过程并且收入风险的冲击较大; 使用 HIP 模型、考虑了个人的异质性之后,可以看到收入风险的冲击迅速减小,而劳动收入本身也收敛了,这一发现与 Guvenen(2007) 的一致,说明研究劳动收入过程时需要充分地考虑个体的异质性。

五、结论

无论是发达国家还是发展中国家,调查数据形成的非等间隔面板数据越来越多,而已有的动态面板数据模型估计量的精度并不高,无论是面板估计量、非线性最小二乘估计量还是最短距离估计量。基于此,本文提出半差分估计量以提升估计效率,并且考虑了缺失观测期中的变量可能进入残差项的问题。蒙特卡洛模拟的结果验证了新提出的半差分估计量在有限样本当中的较高的估计精度,尤其是在缺失观测期包含遗漏变量的情况下。模拟说明新提出的半差分估计量对原来已有的估计量而言是一个重要的补充。与此同时,研究劳动收入过程时需要充分考虑个体异质性,中国居民劳动收入出现了发散的趋势,收入差距在被不断拉大。

未来关于计量方法的研究应该着力于经验研究当中更为常见的非平衡面板数据,一些个体难免在某些观测期数据缺失,直接删除使得样本磨损较大,因此需要讨论更加一般的"非等间隔非平衡动态面板数据模型"。另外,对于中国居民劳动收入过程还需要更加详实的研究,从而更好地理解中国居民劳动收入差距的演进过程。

参考文献

[1] Arellano,M. ,S. Bond. Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. The Review of Economic Studies,1991,58(2):277 – 297.

[2] Blundell,R. ,S. Bond. Initial conditions and moment restrictions in dynamic panel data models. Journal of Econometrics,1998,87(1): 115 – 143.

[3] Everaert,G. Orthogonal to backward mean transformation for dynamic panel data models. The Econometrics Journal,2013,16 (2):179 – 221.

[4] Guvenen,F. Learning Your Earning: Are Labor Income Shocks Really Very Persistent. The American Economic Review,2007,97 (3):687 – 712.

[5] Lee,N. ,H. Moon, M. Weidner. Analysis of Interactive Fixed Effects Dynamic Linear Panel Regression with Measurement Error. Economics Letters,2012,117(1):239 – 242.

[6] McKenzie,D. J. Estimation of AR (1) Models with Unequally Spaced Pseudo-panels. Econometrics Journal, 2001, 4 (1): 89 – 108.

[7] Milimet,D. L,I. K. McDonough. Dynamic Panel Data Models with Irregular Spacing: With Applications to Early Childhood Development. SSRN Working Paper 2260674,2013.

[8] Mundlak,Y. On the pooling of time series and cross section data. Econometrica,1978,46(1):69 – 85.

[9] Nauges,C. ,A. Thomas. Consistent estimation of dynamic panel data models with time-varying individual effects. Annales d' Economie et de Statisique,2003,70(2):53 – 75.

[10] Qiao,K. Consumption Inequality in China: Theory and Evidence form China Health and Nutrition Survey. Fronties of Economics in China,2013a,8(1):92 – 113.

[11] Qiao K. On the Estimation of Unequally Spaced Dynamic Panel Data Model. Peking Univesity Guanghua School of Management Discussion Page,2013b.

作者简介

乔坤元,男,1990 年生,新疆乌鲁木齐人,2012 年毕业于北京大学光华管理学院,获经济学、理学(统计学方向)双学士学位,现为北京大学光华管理学院硕博连读研究生,北京大学光华管理学院教学助理。研究方向为计量经济学、应用计量经济学、政治经济学。

(责任编辑: 曹　麦)