应 用 数 学
MATHEMATICA APPLICATA
2019, 32(4): 930-934

面板数据分位数回归模型的工具变量估计

邹灵, 吴东晟, 杨宜平
(重庆工商大学数学与统计学院, 经济社会应用统计重庆市重点实验室, 重庆 400067)

摘要: 针对含有内生变量的面板数据回归模型, 提出基于工具变量的分位数回归估计方法. 首先, 通过引入工具变量解决协变量的内生性问题, 然后利用分位数回归的方法对回归系数进行估计. 在一些正则条件下, 证明所提出估计的大样本性质, 通过模拟研究证实该方法的有限样本性质.

关键词: 面板数据; 分位数回归; 工具变量; 渐近正态
中图分类号: O212.1 **AMS(2000)主题分类:** 62G05; 62G20
文献标识码: A **文章编号:** 1001-9847(2019)04-0930-05

1. 引言

面板数据模型越来越广泛应用于经济、环境、生物等领域, 是近20年计量经济学模型的重要模型之一. 关于模型的参数估计, 大多数采用最小二乘法对兴趣参数进行估计. 当数据出现尖峰、厚尾、异方差、异常点等情况时, 最小二乘法不再适用. 为了解决该问题, Koenker和Bassett[1]首次提出分位数回归的方法, 分位数回归以其稳健的性质已经在医学和经济领域得到广泛的应用. 目前已有一些学者对面板数据分位数回归进行相关研究. Koenker[2] 通过正则化的方法将分位数回归用于面板数据当中; 罗幼喜和田茂再[3]讨论了固定效应的面板数据模型的三种分位数回归方法并使用蒙特卡洛进行模拟; Canay[4]通过数据转化消除固定效应, 提出一种简单的面板数据分位数回归方法; Billger和Lamarche[5]使用面板数据分位数回归的方法对英国和美国的移民收入分配进行研究; Galvao和Kato[6]研究了面板数据分位数回归的固定效应问题.

从上述可以看出, 面板数据分位数回归的使用越来越广泛, 以上研究都是基于协变量是外生变量的情况下讨论的面板数据分位数回归. 但在实际应用中, 很多变量都具有内生性. 如: 在研究外商直接投资对环境污染的影响时, 外商直接投资作为解释变量具有内生性[7]; 在讨论城镇居民人均消费支出和人均可支配收入时, 通过豪斯曼检验证实了城镇居民人均可支配收入是具有内生性的[8]. 然而当前关于含内生变量的面板数据的分位数回归的研究却很少. 因此促使本文讨论面板数据分位数回归模型的工具变量估计.

本文讨论含有内生变量的面板数据模型, 为了解决协变量的内生性问题, 引入工具变量消除协变量的内生性, 再通过组内中心化消除面板数据个体效应项, 采用分位数回归的方法估计回归系数, 并证明其渐近性质. 最后对Naive最小二乘估计、Naive分位数回归、两阶段最小二乘估计和分位数回归的工具变量估计进行模拟研究, 比较四种方法在不同分布下的估计效果.

2. 面板数据分位数回归模型的工具变量估计

讨论如下面板数据模型

* **收稿日期:** 2018-11-09
基金项目: 国家社会科学基金项目 (18BTJ035)
作者简介: 杨宜平, 女, 汉族, 湖北人, 教授, 研究方向: 非参数统计及数据分析.

$$Y_{it} = X_{it}^{\mathrm{T}}\theta + Z_{it}^{\mathrm{T}}\beta + \alpha_i + \varepsilon_{it}, i = 1, \cdots, N,\ t = 1, \cdots, T, \tag{2.1}$$

其中Y_{it}是响应变量, X_{it}是p维内生协变量, Z_{it}是q维外生协变量, (θ, β)是未知参数, α_i 是不可测量的个体固定效应, ε_{it}表示随机误差项. 为了模型可识别, 假定$\sum\limits_{i=1}^{N}\alpha_i = 0$.

由于解释变量X_{it}具有内生性, 已有估计不再适用. 为了消除内生变量对参数θ估计的影响, 假定存在一个工具变量ω_{it}, 且满足

$$X_{it} = \Gamma\omega_{it} + e_{it}, \tag{2.2}$$

其中Γ是$p \times k$的未知参数矩阵, ω_{it}是$k \times 1$的工具变量, 且与随机误差项ε_{it}不相关, 并满足$\mathrm{E}(e_{it}|\omega_{it}) = 0$.

下面讨论(θ, β)的分位数回归估计. 首先, 由(2.2)可以得到Γ的估计, 即

$$\hat{\Gamma} = \left(\sum_{i=1}^{N} X_i\omega_i^{\mathrm{T}}\right)\left(\sum_{i=1}^{N}\omega_i\omega_i^{\mathrm{T}}\right)^{-1},$$

其中$X_i = (X_{i1}, X_{i2}, \ldots, X_{iT})$, $\omega_i = (\omega_{i1}, \omega_{i2}, \ldots, \omega_{iT})$, 则$\hat{X}_{it} = \hat{\Gamma}\omega_{it}$, 那么模型(2.1)转化为:

$$Y_{it} = \hat{X}_{it}^{\mathrm{T}}\theta + Z_{it}^{\mathrm{T}}\beta + \alpha_i + \varepsilon_{it}, i = 1, \cdots, N,\ t = 1, \cdots, T.$$

由于α_i未知, 通过组内变化消除α_i的影响, 即

$$Y_{it}^{\mathrm{T}} - \bar{Y}_i^{\mathrm{T}} = (\hat{X}_{it}^{\mathrm{T}} - \bar{\hat{X}}_i^{\mathrm{T}})\theta + (Z_{it}^{\mathrm{T}} - \bar{Z}_i^{\mathrm{T}})^{\mathrm{T}}\beta + (\varepsilon_{it} - \bar{\varepsilon}_i) \Rightarrow \ddot{Y}_{it} = \ddot{\hat{X}}_{it}^{\mathrm{T}}\theta + \ddot{Z}_{it}^{\mathrm{T}}\beta + \ddot{\varepsilon}_{it},$$

其中$\bar{Y}_i = \frac{1}{T}\sum\limits_{t=1}^{T} Y_{it}$, $\bar{\hat{X}}_i = \frac{1}{T}\sum\limits_{t=1}^{T}\hat{X}_{it}$, $\bar{Z}_i = \frac{1}{T}\sum\limits_{t=1}^{T} Z_{it}$, $\bar{\varepsilon}_i = \frac{1}{T}\sum\limits_{t=1}^{T}\varepsilon_{it}$. 则$(\theta, \beta)$ 的分位数回归估计可通过求解以下目标函数得到

$$(\hat{\theta}_{(\tau)}, \hat{\beta}_{(\tau)}) = \arg\min\sum_{i=1}^{N}\sum_{t=1}^{T}\rho_\tau\left(\ddot{Y}_{it} - \ddot{\hat{X}}_{it}^{\mathrm{T}}\theta_{(\tau)} - \ddot{Z}_{it}^{\mathrm{T}}\beta_{(\tau)}\right),$$

其中$\rho_\tau(s) = \tau s - sI(s < 0)$.

3. 渐近性质

下面讨论(θ, β)的估计$(\hat{\theta}_{(\tau)}, \hat{\beta}_{(\tau)})$的渐近性质, 需如下正则条件:

(C1)Ω是正定矩阵, 其中

$$\Omega = \lim_{N\to\infty}\tfrac{1}{N}\sum_{i=1}^{N} V_i^{\mathrm{T}}V_i,\ V_i = \left\{(\Gamma\ddot{\omega}_i)^{\mathrm{T}}, \ddot{Z}_i^{\mathrm{T}}\right\},\ \ddot{\omega}_{it} = \omega_{it} - \tfrac{1}{T}\sum_{t=1}^{T}\omega_{it},$$

$$\ddot{\omega}_i = (\ddot{\omega}_{i1}, \cdots, \ddot{\omega}_{iT}),\ \ddot{Z}_i = (\ddot{Z}_{i1}, \cdots, \ddot{Z}_{iT}).$$

(C2)$\ddot{\varepsilon}_{it}^*$的密度函数$f(\cdot)$有大于零的下界, 一阶导函数连续且一直有界, 其中

$$\ddot{\varepsilon}_{it}^* = \ddot{\varepsilon}_{it} + \ddot{e}_{it}^{\mathrm{T}}\theta,\ \ddot{e}_{it}^{\mathrm{T}} = e_{it} - \tfrac{1}{T}\sum_{i=1}^{T} e_{it}.$$

定理 3.1　如果(C1)和(C2)成立, 则有

$$\sqrt{N}\begin{pmatrix}\hat{\theta}_{(\tau)} - \theta_{(\tau)} \\ \hat{\beta}_{(\tau)} - \beta_{(\tau)}\end{pmatrix} \xrightarrow{d} N\left(0, \frac{\psi(\tau)}{f^2(0)}\Omega^{-1}\right),$$

其中$\psi(\tau) = \mathrm{E}(\eta + f(0)e^{\mathrm{T}}\theta)^2$, $\eta = (I(\ddot{\varepsilon}^* \le 0) - \tau)$.

证明: 令$Q_{N_1} = \sqrt{N}(\hat{\theta}_{(\tau)} - \theta_{(\tau)})$, $Q_{N_2} = \sqrt{N}(\hat{\beta}_{(\tau)} - \beta_{(\tau)})$, 则$(\hat{\theta}_{(\tau)}, \hat{\beta}_{(\tau)})$是下式的最小化解

$$L_N = \sum_{i=1}^{N}\sum_{t=1}^{T}\left[\rho_\tau\left(\ddot{\varepsilon}_{it}^* - \gamma_{it} - \frac{(\Gamma\ddot{\omega}_{it})^{\mathrm{T}}Q_{N_1} + \ddot{Z}_{it}^{\mathrm{T}}Q_{N_2}}{\sqrt{N}}\right) - \rho_\tau(\ddot{\varepsilon}_{it}^* - \gamma_{it})\right],$$

其中$\gamma_{it}=\left[(\hat{\Gamma}-\Gamma)\ddot{\omega}_{it}\right]^{\mathrm{T}}\theta$. 由Knight[9]中等式(2-13), 我们有

$$L_N=\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{(\Gamma\ddot{\omega}_{it})^{\mathrm{T}}Q_{N_1}+\ddot{Z}_{it}^{\mathrm{T}}Q_{N_2}}{\sqrt{N}}\left(I(\ddot{\varepsilon}_{it}^{*}\le 0)-\tau\right)$$
$$+\sum_{i=1}^{N}\sum_{t=1}^{T}\int_{\gamma_{it}}^{\gamma_{it}+\left\{(\Gamma\ddot{\omega}_{it})^{\mathrm{T}}Q_{N_1}+\ddot{Z}_{it}^{\mathrm{T}}Q_{N_2}\right\}/\sqrt{N}}\left[I(\ddot{\varepsilon}_{it}^{*}\le v)-I(\ddot{\varepsilon}_{it}^{*}\le 0)\right]\mathrm{d}v=A_N+B_N.$$

首先, 考虑B_N,

$$\begin{aligned}\mathrm{E}(B_N|\omega_{it},Z_{it})=&\sum_{i=1}^{N}\sum_{t=1}^{T}\int_{\gamma_{it}}^{\gamma_{it}+\left[(\Gamma\ddot{\omega}_{it})^{\mathrm{T}}Q_{N_1}+\ddot{Z}_{it}^{\mathrm{T}}Q_{N_2}\right]/\sqrt{N}}vf(0)\left[1+o(1)\right]\mathrm{d}v\\=&\frac{1}{2}f(0)(Q_{N_1}^{\mathrm{T}},Q_{N_2}^{\mathrm{T}})\Omega_n(Q_{N_1}^{\mathrm{T}},Q_{N_2}^{\mathrm{T}})^{\mathrm{T}}\\&+\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\sum_{t=1}^{T}f(0)\gamma_{it}\left\{(\Gamma\ddot{\omega}_{it})^{\mathrm{T}}Q_{N_1}+\ddot{Z}_{it}^{\mathrm{T}}Q_{N_2}\right\}+o(1),\end{aligned}$$

其中$\Omega_n=\frac{1}{n}\sum\limits_{i=1}^{N}V_i^{\mathrm{T}}V_i$, 定义$R_n=B_n-\mathrm{E}(B_n|\omega_{it},Z_{it})$, 容易证明$R_n=o_p(1)$, $\Omega_n=\mathrm{E}(\Omega_n)+o_p(1)=\Omega+o_p(1)$. 因此, 我们有

$$\begin{aligned}L_N=&\frac{1}{2}f(0)(Q_{N_1}^{\mathrm{T}},Q_{N_2}^{\mathrm{T}})\Omega(Q_{N_1}^{\mathrm{T}},Q_{N_2}^{\mathrm{T}})^{\mathrm{T}}\\&+\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\sum_{t=1}^{T}(\eta_{it}+f(0)\gamma_{it})\left\{(\Gamma\ddot{\omega}_{it})^{\mathrm{T}}Q_{N_1}+\ddot{Z}_{it}^{\mathrm{T}}Q_{N_2}\right\}+o_p(1),\end{aligned}$$

其中$\eta_{it}=(I(\ddot{\varepsilon}_{it}^{*}\le 0)-\tau)$. 则有

$$\sqrt{N}\begin{pmatrix}\hat{\theta}_{(\tau)}-\theta_{(\tau)}\\ \hat{\beta}_{(\tau)}-\beta_{(\tau)}\end{pmatrix}=-\frac{\Omega^{-1}}{f(0)}\sum_{i=1}^{N}\sum_{t=1}^{T}N^{-\frac{1}{2}}\left((\Gamma\ddot{\omega}_{it})^{\mathrm{T}},\ddot{Z}_{it}^{\mathrm{T}}\right)^{\mathrm{T}}\left\{\eta_{it}+f(0)e_{it}^{\mathrm{T}}\theta\right\}+o_p(1).$$

注意到$\mathrm{E}\left(\eta_{it}+f(0)e_{it}^{\mathrm{T}}\theta\right)=0$, $\mathrm{Var}\left(\eta_{it}+f(0)e_{it}^{\mathrm{T}}\theta\right)=\psi(\tau)$. 由中心极限定理可得定理3.1.

4. 模拟研究

本节通过模拟研究所提出方法的有限样本性质. 考虑如下含有内生变量的面板数据模型:

$$\begin{cases}Y_{it}=\theta X_{it}+\beta Z_{it}+\alpha_i+\varepsilon_{it},\\ X_{it}=\Gamma\omega_{it}+e_{it},\quad i=1,\cdots,N,\ t=1,\cdots,5.\end{cases}\tag{4.1}$$

该模型中$\theta=1.5$, $\beta=2$, $\Gamma=1$, 其中X_{it}是内生性变量, Z_{it}是外生性变量, $Z_{it}\sim\mathrm{N}(0,1)$, $\omega_{it}\sim\mathrm{N}(0,1)$, $e_{it}\sim\mathrm{N}(0,0.4^2)$, $\varepsilon_{it}=e_{it}+\delta_{it}$. 分别讨论服从以下分布, 具体如下: $\delta_{it}\sim 0.5\mathrm{N}(0,1)$, $\delta_{it}\sim 0.2\mathrm{t}(1)$, $\delta_{it}\sim 0.2\mathrm{C}(0,1)$. 我们计算了偏差(Bias)和标准差(SD), 分别取样本量为50、100和150. 模拟研究比较了四种方法: Naive 最小二乘估计(NLS)、Naive分位数回归估计(NQR)、两阶段最小二乘估计(2SLS)和分位数回归的工具变量估计(IVQR)的估计效果, 其中分位数回归给出的是分位点为0.5的估计. 重复试验次数1000次. 模拟研究结果如表4.1所示.

根据表4.1可以看出, 无论模型误差分布是何种情形, 对θ的估计, NLS和$\mathrm{NQR}_{0.5}$是有偏的. 由于Naive估计忽略了内生变量的影响, 直接用内生变量估计, 所得的估计是有偏. 对模型误差是正态分布时, 2SLS和$\mathrm{IVQR}_{0.5}$两者差别并不大, 但是, 当模型误差分布为非正态分布时, 2SLS方法的Bias和SD都很大, 效果不好, 而本文提出的$\mathrm{IVQR}_{0.5}$仍表现出良好的效果. 随着样本量的增加, 本文提出的$\mathrm{IVQR}_{0.5}$偏差和标准差变小. 因此, 本文提出的估计方法消除了内生变量对估计造成的偏差, 同时, 估计不受模型误差分布的影响.

为了验证本文提出的分位数回归的工具变量估计的渐近正态性, 图4.1和图4.2分别给出了样本量$N=100$, 不同误差分布情形下参数θ和β的Q-Q图.

表 4.1　四种估计方法下参数估计的偏差与标准差

			N(0,1)		t(1)		C(0,1)	
(θ,β)	N	估计方法	Bias	SD	Bias	SD	Bias	SD
θ	50	NLS	0.1366	0.0419	0.1777	4.7945	0.1853	3.4168
		$NQR_{0.5}$	0.1371	0.0509	0.1353	0.0554	0.1390	0.0559
		2SLS	-0.0025	0.0546	0.0913	5.3664	0.0408	3.7537
		$IVQR_{0.5}$	-0.0037	0.0774	-0.0032	0.0930	0.0017	0.0973
	100	NLS	0.1376	0.0282	0.2333	5.4741	-1.0718	25.5669
		$NQR_{0.5}$	0.1372	0.0349	0.1381	0.0380	0.1383	0.0382
		2SLS	-0.0006	0.0380	-0.1818	9.3313	-1.0643	18.2933
		$IVQR_{0.5}$	0.0016	0.0539	-0.0017	0.0664	0.0012	0.0672
	150	NLS	0.1382	0.0234	-0.2182	18.3360	0.1158	2.8675
		$NQR_{0.5}$	0.1373	0.0286	0.1381	0.0302	0.1360	0.0306
		2SLS	-0.0007	0.0308	-0.2769	17.2276	-0.0571	3.8050
		$IVQR_{0.5}$	-0.0015	0.0437	-0.0001	0.0535	-0.0003	0.0539
β	50	NLS	0.0011	0.0451	0.1523	4.8742	0.1075	3.6342
		$NQR_{0.5}$	0.0011	0.0542	-0.0014	0.0573	0.0022	0.0589
		2SLS	0.0018	0.0804	0.1578	4.9970	0.1123	3.6057
		$IVQR_{0.5}$	0.0020	0.0959	-0.0035	0.1053	0.0086	0.1145
	100	NLS	0.0005	0.0317	-0.0797	5.5831	0.9074	51.6534
		$NQR_{0.5}$	0.0005	0.0376	0.0005	0.0415	0.0002	0.0410
		2SLS	0.0004	0.0561	-0.0856	5.7102	0.9292	52.3910
		$IVQR_{0.5}$	0.0003	0.0664	0.0031	0.0799	-0.0016	0.0761
	150	NLS	-0.0002	0.0256	0.5632	17.2589	-0.0606	4.4265
		$NQR_{0.5}$	-0.0002	0.0312	0.0005	0.0333	-0.0003	0.0336
		2SLS	0.0002	0.0451	0.5692	17.5144	-0.0617	4.4415
		$IVQR_{0.5}$	0.0002	0.0535	0.0015	0.0623	-0.0002	0.0621

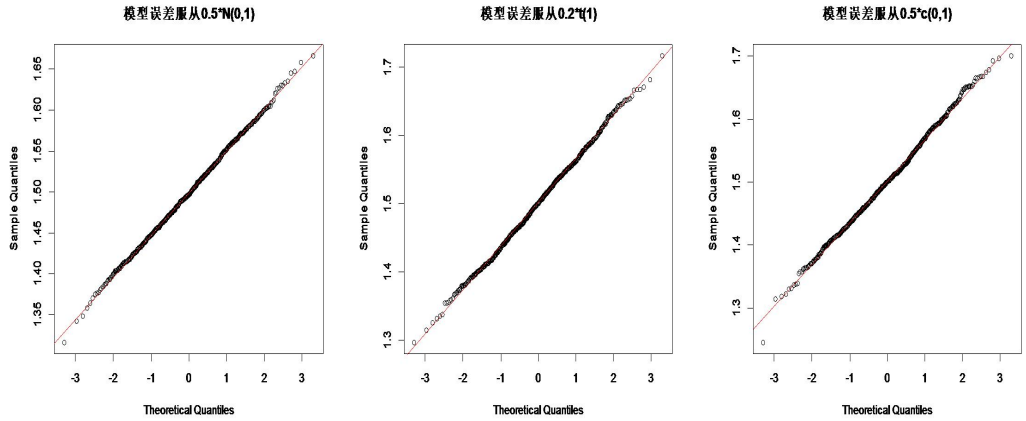

图4.1　参数θ的Q-Q图

从图4.1, 图4.2可以看出, 图上的点近似地在一条直线上. 因此, 所提出的IVQR估计具有渐近正态性.

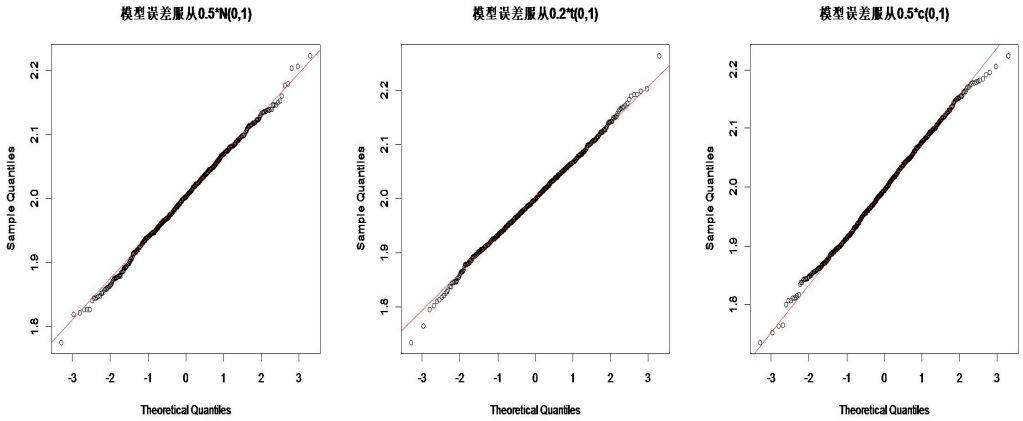

图4.2 参数β的Q-Q图

参考文献:

[1] KOENKER R, BASSETT G. Regression quantiles[J]. Econometrica, 1978, 46(1): 33-50.

[2] KOENKER R. Quantile regression for longitudinal data[J]. Journal of Multivariate Analysis, 2004, 91(1): 74-89.

[3] 罗幼喜, 田茂再. 面板数据的分位回归方法及其模拟研究[J]. 统计研究, 2010, 27(10): 81-87.

[4] CANAY I A. A simple approach to quantile regression for panel data[J]. Econometrics Journal, 2011, 14(3): 368-386.

[5] BILLGER S M, LAMARCHE C. A panel data quantile regression analysis of the immigrant earnings distribution in the United Kingdom and United States[J]. Empirical Economics, 2015, 49(2): 705-750.

[6] GALVAO A F, KATO K. Smoothed quantile regression for panel data[J]. Journal of Econometrics, 2016, 193(1): 92-112.

[7] 贺培, 刘叶. FDI对中国环境污染的影响效应——基于地理距离工具变量的研究[J]. 中央财经大学学报, 2016(06): 79-86.

[8] 李子奈, 潘文卿. 计量经济学[M]. 第三版. 北京: 高等教育出版社, 2010.

[9] KNIGHT K. Limiting distributions for L1 regression estimators under general conditions[J]. Annals of Statistics, 1998, 26(2):755-770.

Instrumental Variables Estimation of Quantile Regression Model with Panel Data

ZOU Ling, WU Dongsheng, YANG Yiping

(*College of Mathematics and Statistics, Chongqing Technology and Business University, Chongqing 400067,China*)

Abstract: In this paper, we consider the panel data regression model with endogenous variables. A quantile regression estimation is proposed based on the instrumental variables. To deal with the endogenous variables, we introduce some instrumental variables. The regression coefficient is estimated by the quantile regression. The large sample property is established for the proposed estimators under some regularity conditions. Simulation results illustrate the finite sample properties of the proposed method.

Key words: Panel data; Quantile regression; Instrumental variable; Asymptotic normality