

## 面板数据分位数回归模型的参数估计与变量选择

何晓霞, 徐 伟, 李 缓, 吴传菊  
(武汉科技大学理学院, 湖北 武汉 430065)

**摘 要:** 本文研究了基于面板数据的分位数回归模型的变量选择问题. 通过增加改进的自适应 Lasso 惩罚项, 同时实现了固定效应面板数据的分位数回归和变量选择, 得到了模型中参数的选择相容性和渐近正态性. 随机模拟验证了该方法的有效性. 推广了文献 [14] 的结论.

**关键词:** 面板数据; 分位数回归; 自适应 Lasso; 变量选择; 渐近正态性

MR(2010) 主题分类号: 62F12; 62J05

中图分类号: O211.7

文 献 标 识 码: A

文章编号: 0255-7797(2017)05-1101-10

### 1 引言

近年来, 由于计算机技术的日益成熟, 分位数回归在理论和方法上都得到了广泛的应用. Koenker<sup>[1]</sup> 首次提出了分位数回归, 如今分位数回归作为均值回归分析的稳健替代, 被广泛地用于探索响应变量与协变量之间的潜在关系. 在实际应用中, 分位数回归可以刻画响应变量更多的分布特征. Koenker<sup>[2]</sup> 发现分位数回归的结果可以提供比普通条件均值回归更丰富, 更有针对性. 特别是, 它提供了探索异质性的来源与合作的响应变量一种方法, 并深入研究了分位回归模型及其估计. 王新宇<sup>[3]</sup> 系统地介绍了分位数的基本模型及其扩展、分位数回归模型的经典统计推断. Tang 等<sup>[4]</sup> 研究了加权复合分位数 (WCQ) 与随机截尾线性回归模型. 在这个模型中, 提出了可变选择的自适应惩罚程序, 并证明了一致性和渐近正态性. Wang 和 Yin<sup>[5]</sup> 研究了无界意义下的在线变化分位数回归算法.

分位数回归模型中的变量选择问题一直受到广泛的关注. Shows 等<sup>[6]</sup> 针对一种多元线性模型, 提出了对随机删失数据的自适应 Lasso 加权 LAD (AWLAD) 变量选择方法. Wang 等<sup>[7]</sup> 提出了 BIC 调整参数选择方法, 证明了这种方法能够辨别出真模型, 并在模拟中验证了理论的有效性. Wu 等<sup>[8]</sup> 研究了惩罚分位数回归, 在一些较弱的条件下得到了 SCAD 和自适应 Lasso 惩罚分位数回归的 Oracle 性质. Zou<sup>[9]</sup> 提出了分位数回归模型的自适应 Lasso 的变量选择方法, 也得到了其 Oracle 性质. 吕亚召等<sup>[10]</sup> 研究部分线性单指标复合分位数回归模型, 提出了用自适应 Lasso 的变量选择方法, 该方法用 BIC 选择最优调整参数, 在随机模拟中验证了所提方法的优良性.

相对于横截面或是时间序列数据来说, 面板数据含有更多的信息, 因此, 面板数据回归模型的研究越来越受关注. 巴尔塔基<sup>[11]</sup> 提出了面板数据模型及其参数的估计方法, 并给出了实际应用. 李扬等<sup>[12]</sup> 提出了惩罚似然变量选择问题, 证明了面板数据的自适 Lasso 具有 Oracle 性质. 在选择最优调整参数时, 模拟显示 BIC 和 GCV 的选择结果一般比 AIC 有优

\*收稿日期: 2015-09-26

接收日期: 2016-02-25

基金项目: 国家自然科学基金资助 (11201356).

作者简介: 何晓霞 (1979-), 女, 湖北大悟, 副教授, 主要研究方向: 数理统计.

势. 曲婷等<sup>[13]</sup>对平衡纵向数据模型, 通过 Lasso 方法可将模型的系数压缩到 0, 采用 AIC 和 BIC 准则选取最优参数, 从而达到变量选择的目的. Koenker<sup>[14]</sup>首次提出了面板数据分位数回归模型, 用加权的形式控制分位数对效应的影响, 并加入  $l_1$  惩罚项, 既保持了线性规划形式, 又保持了结果设计矩阵的稀疏性. 李翰芳等<sup>[15]</sup>对随机效应面板数据, 通过引入条件 Laplace 先验, 构造了一种新的贝叶斯 Lasso 分位数回归法, 与一般贝叶斯分位回归法相比更有效的将异质变量的系数压缩到 0, 从而起到变量选择的作用.

分位数回归对误差项的分布没有具体的限制, 对异质点或者是非正态分布的参数估计具有一定的稳健性, 将分位数回归和面板数据模型两者结合起来, 在控制个体差异的同时, 可以分析各种变量在不同分位点之间的关系. 基于面板数据的分位数回归模型, 本文提出了一种在改进的自适应 Lasso 的罚函数下对变量进行选择的方法, 对系数变量的值进行压缩, 使得异质变量的系数为 0, 从而达到变量选择的效果, 并证明了相合性和渐近正态性, 在模拟中验证了选择的有效性.

## 2 模型与方法

### 2.1 经典高斯随机效应模型

考虑一般的随机效应面板数据模型

$$y_{ij} = x_{ij}^T \beta + \alpha_i + u_{ij}, \quad (2.1)$$

其中  $y_{ij}$  是因变量,  $x_{ij}$  是自变量,  $\alpha_i$  是不可观测的时间不变效应,  $u_{ij}$  是误差项. 写成矩阵的形式如下  $y = X^T \beta + Z\alpha + u$ , 其中  $y$  是  $n \times 1$  维,  $X$  是  $nm \times p$  维,  $Z$  是  $nm \times n$  维的虚拟变量的关联矩阵,  $\alpha$  和  $u$  是独立的随机向量.

### 2.2 加权分位数面板数据模型

令  $\rho_{\tau_k}(u) = u(\tau_k - I(u \leq 0))$ ,  $y_{ij}$  的分位数函数为

$$Q_{y_{ij}}(\tau_k | x_{ij}, \alpha_i) = x_{ij}^T \beta(\tau_k) + \alpha_i, \quad j = 1, \dots, m; i = 1, \dots, n. \quad (2.2)$$

为了更好的估计参数, 对 (2.1) 式提出加权分位数估计方法,

$$\min_{\alpha, \beta} \sum_{k=1}^K \sum_{j=1}^m \sum_{i=1}^n \omega_k \rho_{\tau_k}(y_{ij} - x_{ij}^T \beta(\tau_k) - \alpha_i), \quad (2.3)$$

最小化 (2.3) 是一个凸规划问题, 加权分位数回归估计方法可以凸优化来实现. 在分位数函数 (2.2) 中,  $\alpha$  与因变量的条件分位数相对应, 为了更好的估计截面的分位数方程, Koenker<sup>[14]</sup>引入了惩罚项  $\lambda_1 \sum_{i=1}^n |\alpha_i|$ , 代替高斯惩罚项,

$$(\tilde{\alpha}, \tilde{\beta}) = \operatorname{argmin}_{\alpha, \beta} \sum_{k=1}^K \sum_{j=1}^m \sum_{i=1}^n \omega_k \rho_{\tau_k}(y_{ij} - x_{ij}^T \beta(\tau_k) - \alpha_i) + \lambda_1 \sum_{i=1}^n |\alpha_i|. \quad (2.4)$$

惩罚项  $\lambda_1 \sum_{i=1}^n |\alpha_i|$  与高斯惩罚项相比, 保持了结果设计矩阵的稀疏性的统计优势和线性规划计算优势.

### 3 渐近性质

在 Koenker<sup>[14]</sup> 的定理 1 中给出了  $\beta(\tau_1)$  ( $\tau$  的个数为 1) 正则化后,  $\tilde{\delta}_1$  的渐近正态性, 本节用另一方法, 将  $\beta(\tau_k)$  中  $\tau$  的个数推广到  $q$  个, 证明正则化后  $\tilde{u}$  的渐近正态性. 为了得到  $(\alpha, \beta)$  估计的渐近性质, 下面给出正则化条件

(A1)  $y_{ij}, i = 1, \dots, n, j = 1, \dots, m$  是独立的, 其条件分布为  $F_{ij}$  (在  $x_{ij}, \alpha_i$  条件下), 连续密度函数为  $f_{ij}, 0 < f_{ij} < \infty$  处处具有有界的微分  $f'_{ij}$ .

(A2) 令  $\xi_{ij}(\tau_k) = Q_{y_{ij}}(\tau_k | x_{ij}, \alpha_i) = x_{ij}^T \beta(\tau_k) + \alpha_i$ , 当  $mn \rightarrow \infty$ , 式子

$$\begin{aligned} & \frac{1}{mn} \sum_{j=1}^m \sum_{i=1}^n \omega_k E[f(\xi_{ij}(\tau_k) | X) z_{ij}^T z_{ij}], \quad \frac{1}{mn} \sum_{j=1}^m \sum_{i=1}^n \omega_k E[f(\xi_{ij}(\tau_k) | X) z_{ij}^T x_{ij}], \\ & \frac{1}{mn} \sum_{j=1}^m \sum_{i=1}^n \omega_k E_X[f(\xi_{ij}(\tau_k) | X) x_{ij}^T x_{ij}] \end{aligned}$$

均收敛到正定矩阵.

(A3)  $\max_{1 \leq i \leq n, 1 \leq j \leq m} \|x_{ij}\| / \sqrt{mn} \rightarrow 0$ .

(A4) 随机变量  $\alpha_i$  与  $x_{ij}$  独立.

考虑下面的目标函数

$$V_{mn}(\delta) = \sum_{k=1}^K \sum_{j=1}^m \sum_{i=1}^n \omega_k [\rho_{\tau_k}(y_{ij} - \xi_{ij}(\tau_k) - v_{ijk}) - \rho_{\tau_k}(y_{ij} - \xi_{ij}(\tau_k))], \quad (2.5)$$

其中  $v_{ijk} = z_{ij}^T \delta_0 / \sqrt{mn} + x_{ij}^T \delta_k / \sqrt{mn}$ , 其中  $\tilde{\delta} = (\tilde{\delta}_0, \tilde{u})^T = (\sqrt{mn}(\tilde{\alpha} - \alpha), \tilde{\delta}_1, \dots, \tilde{\delta}_K)^T$ ,  $\tilde{\delta}_k = \sqrt{mn}(\tilde{\beta}(\tau_k) - \beta(\tau_k))^T, k = 1, \dots, K$ ,  $\tilde{\delta}$  是 (2.5) 式的最小值点. 则有如下定理.

**定理 3.1** 假设在  $(A_1 - A_4)$  成立的条件下,  $\tilde{u} \rightarrow_d N(0, D^{-1} \Sigma D^{-1})$ , 其中  $D_0$  是  $B$  的协方差矩阵,  $B = (W_1, Z)^T$ ,

$$D^{-1} \Sigma D^{-1} = \begin{pmatrix} (\tilde{D}^{-1} D_0 \tilde{D}^{-1})_{22} & 0 & \dots & 0 \\ 0 & (\tilde{D}^{-1} D_0 \tilde{D}^{-1})_{33} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & (\tilde{D}^{-1} D_0 \tilde{D}^{-1})_{KK} \end{pmatrix}.$$

**证** 由 Knight<sup>[16]</sup> 的结果, 对任意  $x \neq 0$ , 有

$$\rho_{\tau}(x - y) = y[I(x < 0) - \tau] + \int_0^y [I(x \leq t) - I(x \leq 0)] dt.$$

将  $V_{mn}(\delta)$  分解成下面的形式

$$\begin{aligned} V_{mn}(\delta) &= \sum_{k=1}^K \sum_{j=1}^m \sum_{i=1}^n \omega_k [v_{ijk}(I(y_{ij} - \xi_{ij}(\tau_k) < 0) - \tau_k)] \\ &+ \sum_{k=1}^K \sum_{j=1}^m \sum_{i=1}^n \omega_k \int_0^{v_{ijk}} [I(y_{ij} - \xi_{ij}(\tau_k) \leq t) - I(y_{ij} - \xi_{ij}(\tau_k) \leq 0)] dt \end{aligned}$$

$$= \sum_{k=1}^K Z_{n,m,k} \delta_k + W_{n,m,k}^T \delta_0 + \sum_{k=1}^K B_{n,m,k},$$

其中

$$\begin{aligned} Z_{n,m,k} &= \frac{1}{\sqrt{mn}} \sum_{j=1}^m \sum_{i=1}^n \omega_k x_{ij}^T [I(y_{ij} - \xi_{ij}(\tau_k) < 0) - \tau_k], \\ W_{n,m,k} &= \frac{1}{\sqrt{mn}} \sum_{j=1}^m \sum_{i=1}^n z_{ij}^T \sum_{k=1}^K \omega_k [I(y_{ij} - \xi_{ij}(\tau_k) < 0) - \tau_k], \\ B_{n,m,k} &= \sum_{j=1}^m \sum_{i=1}^n \omega_k \int_0^{v_{ijk}} [I(y_{ij} - \xi_{ij}(\tau_k) \leq t) - I(y_{ij} - \xi_{ij}(\tau_k) \leq 0)] dt. \end{aligned}$$

由于  $E[I(y_{ij} - \xi_{ij}(\tau_k) < 0) - \tau_k] = 0$ , 结合中心极限定理和 Cramér-Wold 定理,  $Z_{n,m,k}$  和  $W_{n,m,k}$  依分布收敛到  $Z_k$  和  $W_1$ , 其中  $Z_k$  是一个正态随机变量, 均值为 0,  $W_1$  是一个  $n$  维正态向量, 均值为 0. 因此可以得到

$$\sum_{k=1}^K Z_{n,m,k} \delta_k + W_{n,m,k}^T \delta_0 \rightarrow_d \sum_{k=1}^K Z_k \delta_k + W_1^T \delta_0.$$

令

$$B_{n_i, m_j, k} = \omega_k \int_0^{v_{ijk}} [I(y_{ij} - \xi_{ij}(\tau_k) \leq t) - I(y_{ij} - \xi_{ij}(\tau_k) \leq 0)] dt,$$

则对任意  $\epsilon > 0$ , 有  $[B_{n_i, m_j, k}]^2 = [B_{n_i, m_j, k}]^2 I(v_{ijk} \geq \epsilon) + [B_{n_i, m_j, k}]^2 I(v_{ijk} < \epsilon)$ .

一方面,

$$\begin{aligned} mnE[B_{n_i, m_j, k}]^2 I(v_{ijk} \geq \epsilon) &\leq mnE\omega_k^2 \left( \int_0^{v_{ijk}} 2dt \right)^2 I(v_{ijk} \geq \epsilon) \\ &\leq 4mn\omega_k^2 E[|v_{ijk}|^2 I(v_{ijk} \geq \epsilon)] \\ &\rightarrow 0, \text{ 当 } n, m \rightarrow \infty. \end{aligned}$$

另一方面,

$$\begin{aligned} &mnE[B_{n_i, m_j, k}]^2 I(v_{ijk} < \epsilon) \\ &\leq mnE(\omega_k^2 \int_0^{v_{ijk}} 2dt \int_0^{v_{ijk}} [I(y_{ij} - \xi_{ij}(\tau_k) \leq t) - I(y_{ij} - \xi_{ij}(\tau_k) \leq 0)] dt I(v_{ijk} < \epsilon)) \\ &\leq 2mn\omega_k^2 \epsilon E_X \left( \int_0^{v_{ijk}} [F(\xi_{ij}(\tau_k) + t|X) - F(\xi_{ij}(\tau_k)|X)] dt \right) \\ &\leq 2mn\omega_k^2 \epsilon E_X \left( \int_0^{v_{ijk}} f(\xi_{ij}(\tau_k)|X) t dt \right) \\ &\leq 2mn\omega_k^2 \epsilon f(\xi_{ij}(\tau_k)|X) E_X |z_{ij} \delta_0 + x_{ij}^T \delta_k^2| \\ &\rightarrow 0, \text{ 当 } \epsilon \rightarrow 0. \end{aligned}$$

因此当  $mn \rightarrow \infty$  时,

$$\text{Var}\left(\sum_{j=1}^m \sum_{i=1}^n B_{n_i, m_j, k}\right) = \sum_{j=1}^m \sum_{i=1}^n \text{Var}(B_{n_i, m_j, k}) \leq mn E(B_{n_i, m_j, k})^2 \rightarrow 0.$$

则有

$$\sum_{j=1}^m \sum_{i=1}^n [B_{n_i, m_j, k} - E(B_{n_i, m_j, k})] = o_p(1).$$

另外, 由于

$$\begin{aligned} E(B_{n, m, k}) &= \sum_{j=1}^m \sum_{i=1}^n E(B_{n_i, m_j, k}) \\ &= mn E\left(\omega_k \int_0^{v_{ijk}} [I(y_{ij} - \xi_{ij}(\tau_k) \leq t) - I(y_{ij} - \xi_{ij}(\tau_k) \leq 0)] dt\right) \\ &= mn \omega_k E_X \left( \int_0^{v_{ijk}} [F(\xi_{ij}(\tau_k) + t|X) - F(\xi_{ij}(\tau_k)|X)] dt \right) \\ &= mn \omega_k E_X \left( \int_0^{v_{ijk}} f(\xi_{ij}(\tau_k)|X) t dt \right) + o_p(1) \\ &= \frac{1}{2} \delta_0^T G_k \delta_0 + \frac{1}{2} \delta_k^T D_k \delta_k + \delta_0 H_k \delta_k + o_p(1), \end{aligned}$$

其中

$$\begin{aligned} G_k &= \omega_k E_X [f(\xi_{ij}(\tau_k)|X) z_{ij}^T z_{ij}], \quad H_k = \omega_k E_X [f(\xi_{ij}(\tau_k)|X) z_{ij}^T x_{ij}], \\ D_k &= \omega_k E_X [f(\xi_{ij}(\tau_k)|X) x_{ij}^T x_{ij}]. \end{aligned}$$

则

$$\begin{aligned} V_{mn}(\delta) &= \frac{1}{2} \delta_0^T G \delta_0 + \frac{1}{2} \sum_{k=1}^K \delta_k^T D \delta_k + \delta_0^T \sum_{k=1}^K H_k \delta_k + W_1^T \delta_0 + \sum_{k=1}^K Z_k \delta_k + o_p(1) \\ &= \frac{1}{2} \delta_0^T G \delta_0 + \frac{1}{2} u^T D' u + \delta_0^T H^T u + W_1^T \delta_0 + Z^T u + o_p(1) \\ &= \frac{1}{2} \delta^T \tilde{D} \delta + \delta^T B + o_p(1), \end{aligned}$$

其中  $H = (H_1, H_2 \cdots H_K)^T$ ,  $Z = (z_1, z_2 \cdots z_K)^T$ ,  $G = \sum_{k=1}^K G_k$ ,  $B = (W_1, Z)^T$ ,

$$D' = \begin{pmatrix} D_1 & 0 & \cdots & 0 \\ 0 & D_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & D_K \end{pmatrix}, \quad \tilde{D} = \begin{pmatrix} G & H^T \\ H & D' \end{pmatrix}.$$

即  $V_{mn}(\delta) \rightarrow \frac{1}{2} \delta^T \tilde{D} \delta + \delta^T B$ .

由 Koenker<sup>[14]</sup> 中引理 1, 可以得到  $\tilde{u} \rightarrow_d N(0, D^{-1}\Sigma D^{-1})$ .

#### 4 变量选择

在对数据进行统计分析时, 人们一般会借助一些相关变量对所关心的变量进行分析, 建模, 以便得到理想的结果, 一般称这些相关的变量为协变量, 而所关心的变量为因变量. 在开始建模的时候, 希望加入更多相关变量, 来得到更真实的结果, 然而, 随着协变量的增多, 异质变量存在的可能性就越大, 于是, 希望寻找一个有效方法来选出对响应变量有显著影响的协变量. 因此变量选择就是统计学中一个重要的问题. 本节对上述面板数据分位数模型的变量选择进行分析, 在 (4.1) 式中需要指定调节参数  $\lambda_2$ , 本文最优的调整参数  $\lambda_2$  可以通过 BIC (Bayesian information criterion) 准则选取. 在加权分位数估计的同时, 同时希望对变量做选择, 本节选的罚函数是自适应 Lasso 罚函数. 令

$$(\hat{\alpha}, \hat{\beta}) = \operatorname{argmin}_{\alpha, \beta} \sum_{k=1}^K \sum_{j=1}^m \sum_{i=1}^n \omega_k [\rho_{\tau_k}(y_{ij} - x_{ij}^T \beta(\tau_k) - \alpha_i)] \\ + \lambda_1 \sum_{i=1}^n |\alpha_i| + \lambda_2 \sum_{k=1}^K \omega_k \sum_{j=1}^p \frac{|\beta_j(\tau_k)|}{|\hat{\beta}_j(\tau_k)|}, \quad (4.1)$$

令  $\operatorname{BIC}(\lambda) = \log P_\lambda + df_\lambda \cdot \log(mn)/mn$ , 其中

$$P_\lambda = \sum_{k=1}^K \sum_{j=1}^m \sum_{i=1}^n \omega_k \rho_{\tau_k}(y_{ij} - x_{ij}^T \beta(\tau_k) - \alpha_i).$$

$df_\lambda$  为  $\hat{\beta}$  非零分量的个数. 这样可以得到调节参数的一个适合的估计  $\hat{\lambda} = \arg \min_\lambda \operatorname{BIC}(\lambda)$ . 令  $\beta^* = (\beta_1^{*T}, \beta_2^{*T})^T$  和  $\hat{\beta} = (\hat{\beta}_1^T, \hat{\beta}_2^T)^T$  分别是参数向量  $\beta$  的真实值和加权分位数自适应 Lasso 惩罚估计值. 不失一般性, 可以假设  $\beta_1^*$  包含  $\beta^*$  中的前  $q$  个非零部分, 同时  $\beta_2^*$  是剩下  $p - q$  维零向量, 则下面的定理.

**定理 4.1** 假设在 (A1)–(A5) 成立的条件下, 如果  $\sqrt{mn}\lambda_2 \rightarrow 0$ ,  $mn\lambda_2 \rightarrow \infty$ , 则当  $mn \rightarrow \infty$ ,  $\hat{\beta}$  满足

- (i) 选择相合性:  $P(\hat{\beta}_2 = 0) \rightarrow 0$ ;
- (ii) 渐近正态性:  $\sqrt{mn}(\hat{\beta}_1 - \beta_1^*) \rightarrow_d N(0, (D_1)^{-1}\Sigma(D_1)^{-1})$ .

**证** 令  $\hat{\delta} = (\hat{\delta}_0, \hat{u})^T = (\sqrt{mn}(\hat{\alpha} - \alpha), \hat{\delta}_1, \dots, \hat{\delta}_K)^T$ ,  $\hat{\delta}_k = \sqrt{mn}(\hat{\beta}(\tau_k) - \beta(\tau_k))^T$ ,  $k = 1, \dots, K$ ,  $\hat{\delta}$  是下面式子的最小值点.

$$L_{mn}(\delta) = \sum_{k=1}^K \sum_{j=1}^m \sum_{i=1}^n \omega_k [\{\rho_{\tau_k}(y_{ij} - \xi_{ij}(\tau_k) - v_{ij}) - \rho_{\tau_k}(y_{ij} - \xi(\tau_k))\}] \\ + \lambda_1 \sum_{i=1}^n |\alpha_i - \delta_{0i}/\sqrt{mn}| - |\alpha_i| \\ + \lambda_2 / |\hat{\beta}_j(\tau_k)| \sum_{k=1}^K \omega_k \sum_{j=1}^p |\beta_j(\tau_k) - \delta_{kj}/\sqrt{mn}| - |\beta_j(\tau_k)|.$$

(i) 因为  $L_{mn}(\delta)$  是对  $\delta$  的分段线性函数, 在每个可微的点, 对  $k = 1, 2, \dots, K, j = q + 1, \dots, p$  取  $L_{mn}(\delta)$  对  $\delta_{kj}$  的偏导, 有

$$\begin{aligned} & \frac{1}{\sqrt{mn}} \frac{\partial L_{mn}}{\partial \delta_{kj}} \\ &= -\frac{1}{\sqrt{mn}} \sum_{k=1}^K \sum_{j=1}^m \sum_{i=1}^n \omega_k [I(y_{ij} - \xi_{ij}(\tau_k) - v_{ijk} \leq 0) - \tau_k] + \lambda_2 mn \omega_k \frac{\text{sign}(\delta_{kj})}{|\tilde{\beta}(\tau_k)|} + o_p(1) \\ &= \sqrt{mn} \omega_k \frac{\text{sign}(\delta_{kj})}{|\tilde{\beta}(\tau_k)|} - \sum_{k=1}^K (\Psi_{mn}(v_{ijk})) + o_p(1), \end{aligned}$$

其中

$$\Psi_{mn}(v_{ijk}) = \sum_{j=1}^m \sum_{i=1}^n \omega_k [I(y_{ij} - \xi_{ij}(\tau_k) - v_{ijk} \leq 0) - \tau_k].$$

由定理 3.1, 可以得到  $\Psi_{mn}(0)$  收敛到均值为 0, 方差为  $\Xi$  的正态分布,  $\max_{1 < i < n, 1 < j < m} v_{ijk} \rightarrow 0$ . 类似 Tang<sup>[4]</sup> 的证明可以得到  $\sup_{\|\delta(k)\|} |\Psi_{mn}(v_{ijk}) - \Psi_{mn}(0) + \Xi v_{ijk}| = o_p(1)$ , 注意到  $\Psi_{mn}(0)$  依分布收敛到正态随机向量,  $\Xi$  是有限的, 当  $n \rightarrow \infty$  时  $v_{ijk} \rightarrow 0$ , 因此对  $k = 1, 2, \dots, K; j = q + 1, \dots, p$ , 有

$$\frac{1}{\sqrt{mn}} \frac{\partial L_{mn}(\delta)}{\partial \delta_{kj}} = \frac{\lambda_2 \omega_k \sqrt{mn} \text{sign}(\delta_{kj})}{|\tilde{\beta}_j(\tau_k)|} = \frac{\lambda_2 \omega_k mn \text{sign}(\beta_j(\tau_k))}{|\tilde{\beta}_j(\tau_k)|}.$$

当  $n$  充分大时,  $mn\lambda_2 \rightarrow \infty, \omega_k > 0$ , 则  $\partial L_{mn}(\delta)/\partial \delta_{kj}$  的符号由  $\beta_j(\tau_k)$  的符号决定. 因此当  $n$  充分大时, 对于  $k = 1, 2, \dots, K, j = q + 1, \dots, p$ , 如果  $\beta_j(\tau_k) > 0$ ,  $\partial L_{mn}(\delta)/\partial \delta_{kj}$  为正; 如果  $\beta_j(\tau_k) < 0$ ,  $\partial L_{mn}(\delta)/\partial \delta_{kj}$  为负. 可以得到  $P(\delta_{kj} = 0) \rightarrow 1, k = 1, 2, \dots, K, j = q + 1, \dots, p$ , 也就是说当  $mn \rightarrow \infty, P(\hat{\beta}_2 = 0) \rightarrow 1$ .

(ii) 正如定理 3.1 和 (i), 只需建立  $L_{mn}[\delta_0, (\beta_1^T, \mathbf{0})]$  的渐近形式, 因为  $\sqrt{mn}\lambda_2 \rightarrow 0, \beta_j(\tau_k) \rightarrow \beta_j^*(\tau(k)), k = 1, 2, \dots, K, j = q + 1, \dots, p$ , 则  $L_{mn}[\delta_0, (\beta_1^T, \mathbf{0})]$  可以表示为

$$L_{mn}[\delta_0, (\beta_1^T, \mathbf{0})] = \sum_{k=1}^K \sum_{j=1}^m \sum_{i=1}^n \omega_k [\{\rho_{\tau_k}(y_{ij} - \xi_{ij}(\tau_k) - v_{ijk}^1) - \rho_{\tau_k}(y_{ij} - \xi(\tau_k))\}] + O_p(1),$$

其中  $v_{ijk}^1 = z_{ij}^T \delta_0 / \sqrt{mn} + x_{ij}^{1T} \delta_k / \sqrt{mn}, x_{ij}^1 = (x_{ij1}, x_{ij2}, \dots, x_{ijq}), i = 1, \dots, n, j = 1, \dots, m, \hat{u}^1 = \sqrt{mn}(\hat{\beta}_1 - \beta_1^*)$ . 类似定理 3.1 的证明, 有

$$\begin{aligned} L_{mn}[\delta_0, (\beta_1^T, \mathbf{0})] &= \sum_{k=1}^K \sum_{j=1}^m \sum_{i=1}^n \omega_k [v_{ijk}^1 (I(y_{ij} - \xi_{ij}(\tau_k) < 0) - \tau_k)] \\ &\quad + \sum_{k=1}^K \sum_{j=1}^m \sum_{i=1}^n \omega_k \int_0^{v_{ijk}^1} [I(y_{ij} - \xi_{ij}(\tau_k) \leq t) - I(y_{ij} - \xi_{ij}(\tau_k) \leq 0)] dt + o_p(1) \\ &= w_1^T \delta_0 + \frac{1}{2} \delta_0^T G \delta_0 + \delta_0^T \sum_{k=1}^K H_k^1 \delta_k^1 + \frac{1}{2} \sum_{k=1}^K \delta_k^{1T} D_k^1 \delta_k^1 + \sum_{k=1}^K z_k^{1T} \delta_k^1 + o_p(1) \\ &= w_1^T \delta_0 + \frac{1}{2} \delta_0^T G \delta_0 + \delta_0^T H^1 u^1 + \frac{1}{2} u^T D_1 u^1 + o_p(1), \end{aligned}$$

其中  $1/\sqrt{mn} \sum_{k=1}^K \sum_{j=1}^m \sum_{i=1}^n \omega_k [x_{ij}^{1T} (I(y_{ij} - \xi_{ij}(\tau_k) < 0) - \tau_k)] \rightarrow_d z_k^1$ ,  $D_k^1 = E_X(f(\xi_{ij}(\tau_k)|X)x_{ij}^{1T}x_{ij}^1)$ ,  $\delta_k^1 = (\sqrt{mn}(\hat{\beta}_1(\tau_k) - \beta_1(\tau_k)), \dots, \sqrt{mn}(\hat{\beta}_p(\tau_k) - \beta_p(\tau_k)))^T$ ,  $z_k^1$  是均值为 0 的  $q$  维正态随机向量. 则

$$L_{mn}[\delta_0, (\beta_1^T, \mathbf{0})] \rightarrow_d \frac{1}{2} \delta^{1T} \tilde{D}_1 \delta^1 + \delta^{1T} B^1.$$

同时可以得到

$$\hat{u}^1 \rightarrow_d N(0, (D_1)^{-1} \Sigma (D_1)^{-1}).$$

## 5 随机模拟

在本节给出两个例子, 比较不同的方法对参数估计值优势, 并验证自适应 Lasso 罚函数对变量选择的有效性.

**例 1** 考虑  $n = 50, m = 5, p = 1$ , 响应变量由下面的模型生成

$$y_{ij} = \alpha_i + x_{ij}\beta + u_{ij},$$

其中  $\beta = 1$ ,  $\alpha_i$  和  $u_{ij}$  服从标准正态分布,  $\omega = (0.25, 0.5, 0.25)$  在三个分位点  $\tau = (0.25, 0.5, 0.75)$ ,  $x_{ij}$  由高斯分布生成

$$x_{ij} = \gamma_i + v_{ij}, \quad (5.1)$$

$\gamma_i$  和  $v_{ij}$  独立同分布, 相应的组内相关系数,

$$\rho_x = \sigma_\gamma^2 / (\sigma_\gamma^2 + \sigma_u^2), \quad (5.2)$$

就是  $x_{ij}$  和  $x_{ik}$  之间的相关系数, 当  $j \neq k$  时, 在的模拟中, 都令  $\rho_x = 0.5$ . 而  $\lambda_1$  选择位置参数比  $\sigma_u/\sigma_\alpha$ ,  $\lambda_2$  的选择由上一节 BIC 得到,  $\alpha$  和  $u_{ij}$  分两种情况.

1. 都来自于标准正态;
2. 都来自于自由度为 3 的  $t$  分布.

这样可以得到分别在分位数回归的估计方法 (QR)、分位数效应罚函数估计 (PQR)、分位数回归自适应罚函数估计 (LPQR), 对  $\beta$  的估计, 如表 1, 可以看出在  $\alpha$  和  $u_{ij}$  的两种情况 PQR 和 LPQR 都比 QR 估计更优.

表 1: 例 1 中  $\beta$  参数估计

|       | 方法   | 平均值     | 偏差       | 标准差     |
|-------|------|---------|----------|---------|
| $N$   | QR   | 1. 0544 | 0. 0544  | 0. 0876 |
|       | PQR  | 1. 005  | 0. 0005  | 0. 0880 |
|       | LPQR | 0. 9851 | -0. 0149 | 0. 1013 |
| $t_3$ | QR   | 1. 0724 | 0. 0724  | 0. 2464 |
|       | PQR  | 0. 9865 | -0. 0135 | 0. 1502 |
|       | LPQR | 0. 9901 | -0. 0099 | 0. 1545 |

表 2: 例 2 中  $\beta$  罚估计

|       | 方法  | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_4$ | $\beta_5$ | $\beta_6$ | $\beta_7$ | $\beta_8$ |
|-------|-----|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| $N$   | 平均值 | 3. 0403   | 1. 6857   | -0. 0668  | -0. 1212  | 0. 1733   | 0. 0661   | 2. 1015   | 0. 0020   |
|       | 偏差  | 0. 0403   | 0. 1857   | -0. 0668  | -0. 1212  | 0. 1733   | 0. 0661   | 0. 1015   | 0. 0020   |
|       | 标准差 | 0. 0187   | 0. 0643   | 0. 1042   | 0. 1329   | 0. 0330   | 0. 0702   | 0. 0501   | 0. 1552   |
| $t_3$ | 平均值 | 3. 0650   | 1. 4789   | -0. n0151 | -0. 0481  | 0. 1574   | -0. 2183  | 2. 0239   | 0. 1626   |
|       | 偏差  | 0. 0650   | -0. 0211  | -0. 0151  | -0. 0481  | 0. 1574   | -0. 2183  | 0. 0239   | 0. 1626   |
|       | 标准差 | 0. 0591   | 0. 1803   | 0. 0383   | 0. 0369   | 0. 0477   | 0. 0931   | 0. 0914   | 0. 1106   |

表 3: 例 2 中  $\beta$  变量选择

| 方法    | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_4$ | $\beta_5$ | $\beta_6$ | $\beta_7$ | $\beta_8$ |
|-------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| $N$   | 平均值       | 2. 9784   | 1. 5839   | 0. 0000   | 0. 0000   | 0. 0000   | 2. 0569   | 0. 0000   |
|       | 偏差        | -0. 0216  | 0. 0839   | 0. 0000   | 0. 0000   | 0. 0000   | 0. 0569   | 0. 0000   |
|       | 标准差       | 0. 0221   | 0. 2415   | 0. 0000   | 0. 0000   | 0. 0000   | 0. 1540   | 0. 0000   |
| $t_3$ | 平均值       | 3. 0424   | 1. 5412   | 0. 0000   | 0. 0000   | 0. 0000   | 2. 0689   | 0. 0000   |
|       | 偏差        | 0. 0424   | 0. 0412   | 0. 0000   | 0. 0000   | 0. 0000   | 0. 0689   | 0. 0000   |
|       | 标准差       | 0. 0389   | 0. 3130   | 0. 0000   | 0. 0000   | 0. 0000   | 0. 1671   | 0. 0000   |

例 2 令  $m = 5, n = 50, p = 8$ , 响应变量来自下面的模型

$$y_{ij} = \alpha_i + x_{ij}^T \beta + u_{ij}, \quad (5.3)$$

$\beta = (3, 1.5, 0, 0, 0, 0, 2, 0)$ ,  $x_{ij}$  由 (5.1), (5.2) 式生成,  $\alpha_i$  和  $u_{ij}$  同样分两种情况.

1. 都来自于标准正态;
2. 都来自于自由度为 3 的  $t$  分布.

表 2 是分位数罚估计 (PQR) 分别对上面两种情形下  $\beta$  的估计, 表 3 是分位数自适应 Lasso 罚函数 (LPQR) 对参数的估计, 通过模拟可以看出 PQR 可以对参数做近似估计, 但对异质变量不能做选择, 而 LPQR 在参数估计的同时对变量做了选择, 0 参数都选择出来了, 不管是参数估计还是变量选择都比 PQR 有优势.

## 参 考 文 献

- [1] Koenker R. Bassett G. Regression quantiles[J]. Econo., 1978, 46: 33–50.
- [2] Koenker R. Quantile regression[M]. Cambridge: Cambridge University Press, 2005.
- [3] 王新宇. 分位数回归理论及其在金融风险测量中的应用 [M]. 北京: 经济科学出版社, 2010.
- [4] Tang L, Zhou Z, Wu C. Weighted composite quantile estimation and variable selection method for censored regression model[J]. Stat. Prob. Lett., 2012, 3: 653–663.
- [5] Wang B, Yin H. Varying quantile regression with online scheme and unbounded sampling[J]. J. Math., 2015, 34: 281–286.
- [6] Shows H, Lu W, Zhang H. Sparse estimation and inference for censored median regression[J]. Stat. Plan. Infer., 2010, 140: 1903–1917.
- [7] Wang H, Li R, Tsai C L. Tuning parameter selectors for the smoothly clipped absolute deviation method[J]. Biometrika, 2007, 94: 553–568.

- [8] Wu Y, Liu Y. Variable selection in quantile regression[J]. *Statist. Sinica*, 2009, 19: 801–817.
- [9] Zou H. The adaptive Lasso and its oracle properties[J]. *Amer. Stat. Assoc.*, 2006, 101: 1418–1429.
- [10] 吕亚召, 张日权等. 部分线性单指标模型的复合分位数回归及变量选择 [J]. *中国科学*, 2014, 12: 1299–1322.
- [11] 巴尔塔基. 面板数据计量经济分析 [M]. 北京: 中国人民大学出版社, 2010.
- [12] 李扬, 曾宪斌. 面板数据模型的惩罚似然变量选择方法研究 [J]. *统计研究*, 2014, 3: 83–89.
- [13] 曲婷, 王静. 基于 Lasso 方法的平衡纵向数据模型变量选择 [J]. *黑龙江大学自然科学学报*, 2012, 29: 715–722.
- [14] Koenker R. Quantile regression for longitudinal data[J]. *J. Multi. Anal.*, 2004, 91: 71–89.
- [15] 李翰芳, 罗幼喜等. 面板数据的贝叶斯 LASSO 分位回归方法 [J]. *数量经济技术经济研究*, 2013, 2: 138–149.
- [16] Knight K. Limiting distributions for  $L_1$  regression estimators under general conditions[J]. *Ann. Stat.*, 1998, 26: 755–770.

## PARAMETER ESTIMATION AND VARIABLE SELECTION IN THE QUANTILE REGRESSION MODEL FOR PANEL DATA

HE Xiao-xia, XU Wei, LI Huan, WU Chuan-ju

(*College of Science, Wuhan University of Science and Technology, Wuhan 430065, China*)

**Abstract:** In this paper, we consider the variable selection problem for the quantile regression model based on panel data. By adding an improved adaptive lasso penalty term, we realize the quantile regression and variable selection for the panel data with fixed effect simultaneously, and obtain the consistency and asymptotical normality for the selection of the parameters. Simulation studies show the validity of the proposed method, which extend that of [14].

**Keywords:** panel data; quantile regression; adaptive lasso; variable selection; asymptotic normality

**2010 MR Subject Classification:** 62F12; 62J05