

含测量误差的复合分位数回归的 SIMEX 估计 *

杨宜平^{*} 余 鲁 吴东晟

(重庆工商大学数学与统计学院, 重庆, 400067)

摘要: 考虑含测量误差的线性回归模型, 采用模拟外推 (SIMEX) 方法并结合复合分位数回归构造了回归系数的估计. 所得回归系数估计不仅消除了测量误差对估计造成的偏差, 而且保留了复合分位数回归估计的优点. 在一些正则条件下, 证明了估计的渐近性质. 模拟研究了所提出方法的有限样本性质, 并进行了实例分析.

关键词: 复合分位数回归; 测量误差; 渐近正态; SIMEX

中图分类号: O212.7

英文引用格式: YANG Y P, YU L, WU D S. SIMEX estimation for composite quantile regression model with measurement error [J]. Chinese J Appl Probab Statist, 2020, 36(2): 111–122. (in Chinese)

§1. 引言

最小二乘法是估计线性回归模型中回归系数的最基本的方法, 但需对误差分布作特定假定, 如均值为零、方差相同等. 在分析实际问题时, 往往不满足这些假定, 如数据存在异方差、尖峰、厚尾或异常点等情况. 为了弥补最小二乘法的不足, Koenker 和 Bassett^[1] 提出了分位数回归估计. 该方法不需要对误差分布作特定限制, 具有较好稳健性, 因而被广泛应用. Zou 和 Yuan^[2] 综合多个分位点的信息提出了复合分位数 (CQR) 估计, 提高了估计的效率. 由于 CQR 方法具有一些好的性质, 该方法引起诸多学者的兴趣. Kai 等^[3] 将复合分位数回归方法推广到了非参数模型, 结合局部多项式方法提出了非参数函数的复合分位数回归估计. 吕亚召等^[4] 讨论了部分线性单指标模型的复合分位数回归估计及变量选择问题. 刘慧蓝^[5] 对复合分位数回归的参数与半参数模型进行了系统的讨论.

当前关于复合分位数回归的研究, 大多假定协变量能直接观测. 在实际应用中, 常遇到测量误差数据. 大多是在均值回归的框架下解决测量误差带来的偏差问题^[6–9]. 对协变量含测量误差的复合分位数回归研究较少, 主要由于无法直接校正测量误差引起的偏差.

*重庆市社科规划委托项目 (批准号: 2019WT58)、重庆工商大学校内预研项目 (批准号: 2019ZKYYA119)、重庆高校创新团队建设计划资助项目 (批准号: CXTDX201601026)、2018 年重庆市《统计学》研究生导师团队 (批准号: yds183002)、第五批重庆市高等学校优秀人才支持计划 (批准号: 68021900601) 和经济社会应用统计重庆市重点实验室资助.

*通讯作者, E-mail: yeepingyang@foxmail.com.

本文 2018 年 6 月 22 日收到, 2018 年 9 月 13 日收到修改稿.

本文讨论含测量误差的复合分位数回归模型中回归系数的估计问题. 尽管 Jiang^[10] 通过最小化正交残差的复合分位数损失函数构造了回归系数的估计, 但该方法要求回归误差和测量误差服从球形对称分布. 为了避免该假定, 本文采用模拟外推 (SIMEX) 方法解决复合分位数含测量误差的问题. SIMEX 方法的思想是通过模拟增加额外的方差, 以便考察随着额外方差的增加参数估计的变化情况, 然后外推到无测量误差的情形. 由于该方法具有广泛适应性, 大量文献采用该方法处理了不同模型含测量误差的情形. Liang 和 Ren^[11] 利用 SIMEX 方法讨论了含测量误差的广义部分线性模型的估计问题. Apanasovich 等^[12] 研究了部分线性模型参数部分和非参数部分含测量误差的 SIMEX 估计. Yang 等^[13] 考虑了含测量误差的单指标模型的 SIMEX 估计. 关于测量误差模型的更多相关研究可参见专著 [14, 15].

本文利用 SIMEX 方法给出了含测量误差的复合分位数回归中回归系数的估计过程, 在一些正则条件下, 研究了估计的大样本性质. 然后, 通过模拟研究, 比较了在不同情形下 SIMEX 估计、偏差校正的最小二乘估计、正交校正估计^[10] 以及 Naive 估计的优劣. 最后, 利用本文提出的方法分析了心脏病数据.

§2. 方法与主要结果

考虑含测量误差的回归模型

$$\begin{cases} Y_i = X_i^\top \beta + \varepsilon_i, \\ W_i = X_i + U_i, \quad i = 1, 2, \dots, n, \end{cases} \quad (1)$$

其中 Y_i 是响应变量, β 是 p 维回归系数, X_i 是 p 维不可观测的协变量, 而实际观察的是带有测量误差的观测变量 W_i , ε_i 是模型误差且满足 $P(\varepsilon_i \leq a_k | X_i) = \tau_k$, 即 a_k 是 ε 的 τ_k 分位点, U_i 是测量误差. 类似 Liang 和 Ren^[11] 及 Carroll 等^[16], 假定 U_i 与 (X_i, Y_i) 相互独立, 且 $U_i \sim N(0, \Sigma_{uu})$. 在这里, 假定 Σ_{uu} 已知. 当 Σ_{uu} 未知时, 可以采用 Carroll 等^[17] 提出的方法获得 Σ_{uu} 的相合估计. 对模型 (1), 如果采用经典的最小二乘估计, 当模型误差服从正态分布时, 具有较好的性质. 然而对尖峰、厚尾及异常点等情况敏感, 不具有稳健性. 由于最小二乘估计的缺陷, 促使本文提出分位数回归估计方法.

下面讨论模型 (1) 中参数 β 的估计. 参数 β 的 SIMEX 估计过程分以下模拟、估计以及外推三个过程.

1. 模拟过程. 对 $i = 1, 2, \dots, n$, 产生数据

$$W_{is}(\lambda) = W_i + (\lambda \Sigma_{uu})^{1/2} U_{is}, \quad s = 1, 2, \dots, \mathcal{S},$$

其中 $U_{is} \sim N(0, I_p)$, I_p 是 $p \times p$ 单位阵, \mathcal{S} 是给定的整数, $\lambda \geq 0$ 且 $\lambda \in \Lambda = \{\lambda_1, \lambda_2, \dots, \lambda_M\}$. 在实际应用中, λ 通常取 $[0, 2]$ 区间上均匀的格子点. λ 的作用主要是控制有多少额

外测量误差加入到 W_i 中.

根据假定, 可以得到 $E(W_{is}(\lambda) | X_i) = X_i$ 且 $\text{Var}(W_{is}(\lambda) | X_i) = \text{Var}\{(W_{is}(\lambda) - X_i)^2 | X_i\} = (1 + \lambda)\text{Var}(W_i | X_i)$. 因此, 当 $\lambda = -1$ 时, $\text{Var}(W_{is}(\lambda) | X_i) = 0$.

2. 估计过程. 基于模拟数据 $W_{is}(\lambda)$, 可以构造 β 的估计, $\hat{\beta}_s(\lambda)$, 即

$$(\hat{b}_{\tau_1}, \hat{b}_{\tau_2}, \dots, \hat{b}_{\tau_K}, \hat{\beta}_s(\lambda)) = \arg \min_{b_{\tau_1}, b_{\tau_2}, \dots, b_{\tau_K}, \beta} \sum_{k=1}^K \left[\sum_{i=1}^n \rho_{\tau_k}(Y_i - W_{is}^T(\lambda)\beta - b_{\tau_k}) \right], \quad (2)$$

其中 $\rho_{\tau_k}(r) = \tau_k r - rI(r < 0)$, $k = 1, 2, \dots, K$, 是分位数损失函数, $0 < \tau_1 < \tau_2 < \dots < \tau_K < 1$, b_{τ_k} 是 $W_{is}(\lambda)$ 关于 Y_i 回归模型中模型误差的 τ_k 分位点. 通常地, 取等间隔分位点, 即 $\tau_k = k/(K+1)$, $k = 1, 2, \dots, K$. Zou 和 Yuan^[2] 建议 $K = 19$ 是个较好的选择. 在本文模拟研究中, 取 $K = 19$.

重复 (2) 估计过程 \mathcal{S} 次, 取其平均值, 即

$$\hat{\beta}(\lambda) = \frac{1}{\mathcal{S}} \sum_{s=1}^{\mathcal{S}} \hat{\beta}_s(\lambda).$$

通过对 λ 一系列的取点, 可以得到 $\{\lambda_i, \hat{\beta}(\lambda_i), i = 1, 2, \dots, M\}$.

3. 外推过程. 用外推函数 $\mathcal{G}(\lambda, \Gamma)$ 来拟合估计过程中得到的 $\{\lambda_i, \hat{\beta}(\lambda_i), i = 1, 2, \dots, M\}$, 其中 Γ 是未知参数向量. Γ 的估计 $\hat{\Gamma}$ 可以通过最小化下面二次目标函数获得

$$\sum_{i=1}^M (\hat{\beta}(\lambda_i) - \mathcal{G}(\lambda_i, \Gamma))^2.$$

然后外推到 $\lambda = -1$, 得到 β 的 SIMEX 估计

$$\hat{\beta}_{\text{SIMEX}} = \mathcal{G}(-1, \hat{\Gamma}).$$

注记 1 当 $\lambda = 0$ 时, 对应的是 Naive 估计, $\hat{\beta}_{\text{Naive}} = \mathcal{G}(0, \hat{\Gamma})$, 即忽略测量误差直接用 W 取代 X 进行估计.

注记 2 外推函数 $\mathcal{G}(\lambda, \Gamma)$ 往往是未知的, 在实际应用中, 常用的外推函数为二次函数 $\mathcal{G}(\lambda, \Gamma) = \Gamma_1 + \Gamma_2\lambda + \Gamma_3\lambda^2$, 其中 $\Gamma = (\Gamma_1, \Gamma_2, \Gamma_3)^T$. Lin 和 Carroll^[18], Liang 和 Ren^[11] 指出该外推函数得到的估计可以达到很好的效果.

讨论 $\hat{\beta}_{\text{SIMEX}}$ 的渐近性质之前, 需要以下条件:

- C1 参数 $\beta(\lambda)$ 的定义域 Θ_λ 是 R^P 空间上的一个紧集, $\beta(\lambda)$ 是 Θ_λ 的一个内点;
- C2 $\Omega(\beta(\lambda), \lambda)$ 是关于 $\lambda \in \Lambda$ 的正定矩阵, 其中 $\Omega(\beta(\lambda), \lambda) = E\{W_{is}(\lambda)W_{is}^T(\lambda)\}$.
- C3 ε_s^* 的密度函数 $f(\cdot)$ 有大于零的下界, 一阶导函数连续且一致有界, 其中 ε_s^* 是 $W_s(\lambda)$ 关于 Y 回归模型的模型误差.

为了得到定理 3, 引入一些记号. 记 $\widehat{\beta}(\Lambda) = (\widehat{\beta}^\top(\lambda_1), \widehat{\beta}^\top(\lambda_2), \dots, \widehat{\beta}^\top(\lambda_M))^\top$, $\boldsymbol{\Gamma} = (\Gamma_1^\top, \Gamma_2^\top, \dots, \Gamma_p^\top)^\top$, 其中 Γ_j 是外推过程中用来估计 β 的第 j 个分量的参数. $\mathcal{G}(\Lambda, \boldsymbol{\Gamma}) = \text{vec}\{\mathcal{G}(\lambda_m, \Gamma_j), j = 1, 2, \dots, p, m = 1, 2, \dots, M\}$, $\text{Res}(\boldsymbol{\Gamma}) = \widehat{\beta}(\Lambda) - \mathcal{G}(\Lambda, \boldsymbol{\Gamma})$, $s(\boldsymbol{\Gamma}) = [\partial/\partial(\boldsymbol{\Gamma})]\text{Res}(\boldsymbol{\Gamma})$, $D(\boldsymbol{\Gamma}) = s(\boldsymbol{\Gamma})s^\top(\boldsymbol{\Gamma})$, $\varepsilon_{is}^* = Y_i - W_{is}^\top(\lambda)\beta(\lambda)$,

$$\eta_{i\mathcal{S}}(\beta(\lambda), \lambda) = \frac{1}{\sum_{k=1}^K f(b_{\tau_k})} \frac{1}{\mathcal{S}} \sum_{s=1}^{\mathcal{S}} \sum_{k=1}^K W_{is}(\lambda)[I(\varepsilon_{is}^* < b_{\tau_k}) - \tau_k],$$

$$\begin{aligned}\Psi_{i\mathcal{S}}\{\beta(\Lambda), \Lambda\} &= \text{vec}\{\eta_{i\mathcal{S}}(\beta(\lambda), \lambda), \lambda \in \Lambda\}, \\ \mathcal{A}_{11}\{\beta(\Lambda), \Lambda\} &= \text{diag}\{\Omega(\beta(\lambda), \lambda), \lambda \in \Lambda\},\end{aligned}$$

且

$$\Sigma = \mathcal{A}_{11}^{-1}\{\beta(\Lambda), \Lambda\} C_{11}\{\beta(\Lambda), \Lambda\} [\mathcal{A}_{11}^{-1}\{\beta(\Lambda), \Lambda\}]^\top,$$

其中

$$C_{11}\{\beta(\Lambda), \Lambda\} = \text{Cov}[\Psi_{i\mathcal{S}}\{\beta(\Lambda), \Lambda\}].$$

定理 3 如果条件 C1–C3 成立, 则

$$\sqrt{n}(\widehat{\beta}_{\text{SIMEX}} - \beta) \xrightarrow{\mathcal{L}} N\{0, \mathcal{G}_{\boldsymbol{\Gamma}}(-1, \boldsymbol{\Gamma})\Sigma(\boldsymbol{\Gamma})[\mathcal{G}_{\boldsymbol{\Gamma}}(-1, \boldsymbol{\Gamma})]^\top\},$$

其中 $\xrightarrow{\mathcal{L}}$ 表示依分布收敛, $\mathcal{G}_{\boldsymbol{\Gamma}}(\lambda, \boldsymbol{\Gamma}) = \{\partial/\partial(\boldsymbol{\Gamma})\}\mathcal{G}(\lambda, \boldsymbol{\Gamma})$, $\Sigma(\boldsymbol{\Gamma}) = D^{-1}(\boldsymbol{\Gamma})s(\boldsymbol{\Gamma})\Sigma s^\top(\boldsymbol{\Gamma}) \cdot D^{-1}(\boldsymbol{\Gamma})$.

§3. 模拟研究

本节通过数值模拟研究所提出方法的有限样本性质. 在 SIMEX 估计过程中, 取 $\lambda = 0, 0.2, \dots, 2$ 和 $\mathcal{S} = 100$. 在不同情形下, 比较了五种不同的估计方法的估计效果, 即本文提出的方法 (SIMEX-CQR), Jiang^[10] 提出的正交校正估计 (OR-CQR), 基于复合分位数回归的 Naive 估计 (Naive-CQR), 基于最小二乘法的 Naive 估计 (Naive-LS) 以及偏差校正的最小二乘估计 (BC-LS), 其估计形式为

$$\widehat{\beta}_{\text{BC-LS}} = \left(\sum_{i=1}^n W_i W_i^\top - n \Sigma_{uu} \right)^{-1} \left(\sum_{i=1}^n W_i Y_i \right).$$

数据由模型 (1) 产生. 参数取值为 $\beta_1 = 1$, $\beta_2 = 2$. 样本数据生成如下: $X_{i1}, X_{i2} \sim N(0, 1)$, $U_i \sim N(0, 0.2^2 I_2)$, 模型误差考虑正态分布、t 分布和柯西分布三种情况, 即 $\varepsilon_i \sim 0.2^2 N(0, 1)$, $\varepsilon_i \sim 0.2 t(1)$ 及 $\varepsilon_i \sim 0.2 C(0, 1)$. 样本量分别取 $n = 50, 100$ 和 200 . 对以上每种情况, 重复运算 500 次.

首先, 为了验证本文提出的 SIMEX 估计是否具有渐近正态性, 给出了样本量 $n = 100$, 三种不同模型误差分布情形下 β_1 和 β_2 重复运行 500 次所得的 SIMEX 估计的 Q-Q 图. 对样本量 $n = 50$ 和 $n = 200$ 结果类似. 模拟结果见图 1. 从图 1 给出的 500 次计算出的 β 的 SIMEX 估计 Q-Q 图可以看出, Q-Q 图上的点近似地在一条直线附近, 因此所提出的 SIMEX 估计具有渐近正态性.

其次, 展示了外推过程中外推函数的拟合效果. 模拟中给出了样本量 $n = 100$, 三种不同模型误差分布情形下, 二次外推函数拟合点 $(\lambda_i, \hat{\beta}_1(\lambda_i))$ 和 $(\lambda_i, \hat{\beta}_2(\lambda_i))$ 的拟合效果图. 模拟结果见图 2. 图 2 仅展示了 500 次运行中其中 1 次的结果, 其它情形结果类似. 从图 2 可以看出二次外推函数很好地拟合了数据 $\{\lambda_i, \hat{\beta}(\lambda_i), i = 1, 2, \dots, 11\}$. 当 $\lambda = 0$ 时, 对应的是 Naive 估计. 然后外推到 $\lambda = -1$, 得到 SIMEX 估计.

进一步, 比较了不同样本容量不同误差分布情形下的五种估计方法的偏差 (Bias) 和标准差 (SD). 表 1 和表 2 分别给出了 β_1 和 β_2 的计算结果. 从表 1 和表 2 可以看出:

1. Naive-CQR 和 Naive-LS 总是有偏的, 这是因为两种方法都是直接用带测量误差的 W 取代 X , 造成了所得估计是有偏估计.
2. 随着样本容量的增加, SIMEX-CQR 的偏差和标准差相应的变小. 可见样本量的大小会影响估计的精度.
3. 当模型误差分布是正态分布时 (模型误差和测量误差分布相同), SIMEX-CQR、OR-CQR 和 BC-LS 给出了较好的结果, 有效地消除了测量误差引起的偏差; 但当模型误差分布是 t 分布和柯西分布时, BC-LS 和 OR-CQR 效果很差, 偏差和标准差的值都较大, 而 SIMEX-CQR 估计效果仍旧很好. 由此可见, 当模型误差非正态分布时, BC-LS 效果很差; 当模型误差和测量误差分布不相同时, OR-CQR 效果很差.

最后, 比较了当 Σ_{uu} 未知时, 用估计得到的协方差阵带入后的估计效果. 为了估计 Σ_{uu} , 我们需要产生重复测量的样本, 即 $W'_i = X_i + U'_i$, 其中 $U'_i \sim N(0, 0.2^2 I_2)$. 表 3 给出了样本 $n = 100$, Σ_{uu} 已知与 Σ_{uu} 未知情形下参数 β 估计结果. 对样本量 $n = 50$ 和 $n = 200$ 结果类似. 从表 3 可以看出, 用估计得到的协方差阵 $\hat{\Sigma}_{uu}$ 带入后的估计与 Σ_{uu} 已知情形下的估计差别不大.

因此, 本文提出的 SIMEX 估计方法不需对模型误差分布作特定假定, 无论模型误差分布是何种形式, 都具有较好的性质. SIMEX-CQR 估计有效地减少了测量误差引起的偏差, 同时保留了复合分位数回归的优点.

§4. 实例分析

本节分析了一组来自弗明汉心脏病研究 (Framingham Heart Study) 的数据. 该数据收集了 1615 名男性的 5 个指标值. 对该数据集 Liang 等^[19] 建立部分线性模型分析了年龄、血清胆固醇水平以及血压之间的关系. 我们主要感兴趣的是血压与血清胆固醇水平的

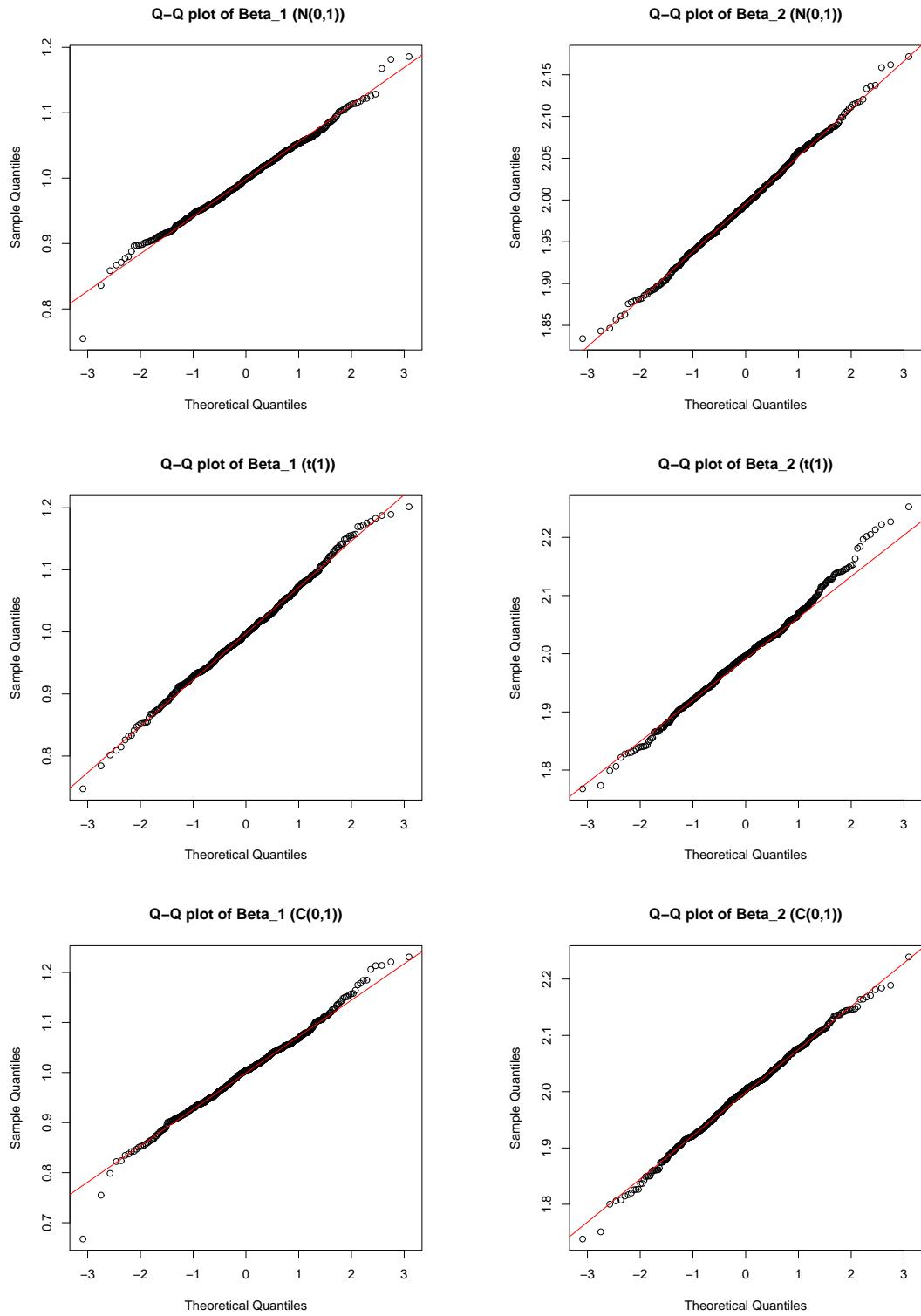


图 1 样本量 $n = 100$ 时, 三种误差分布情形下 SIMEX 估计的 Q-Q 图

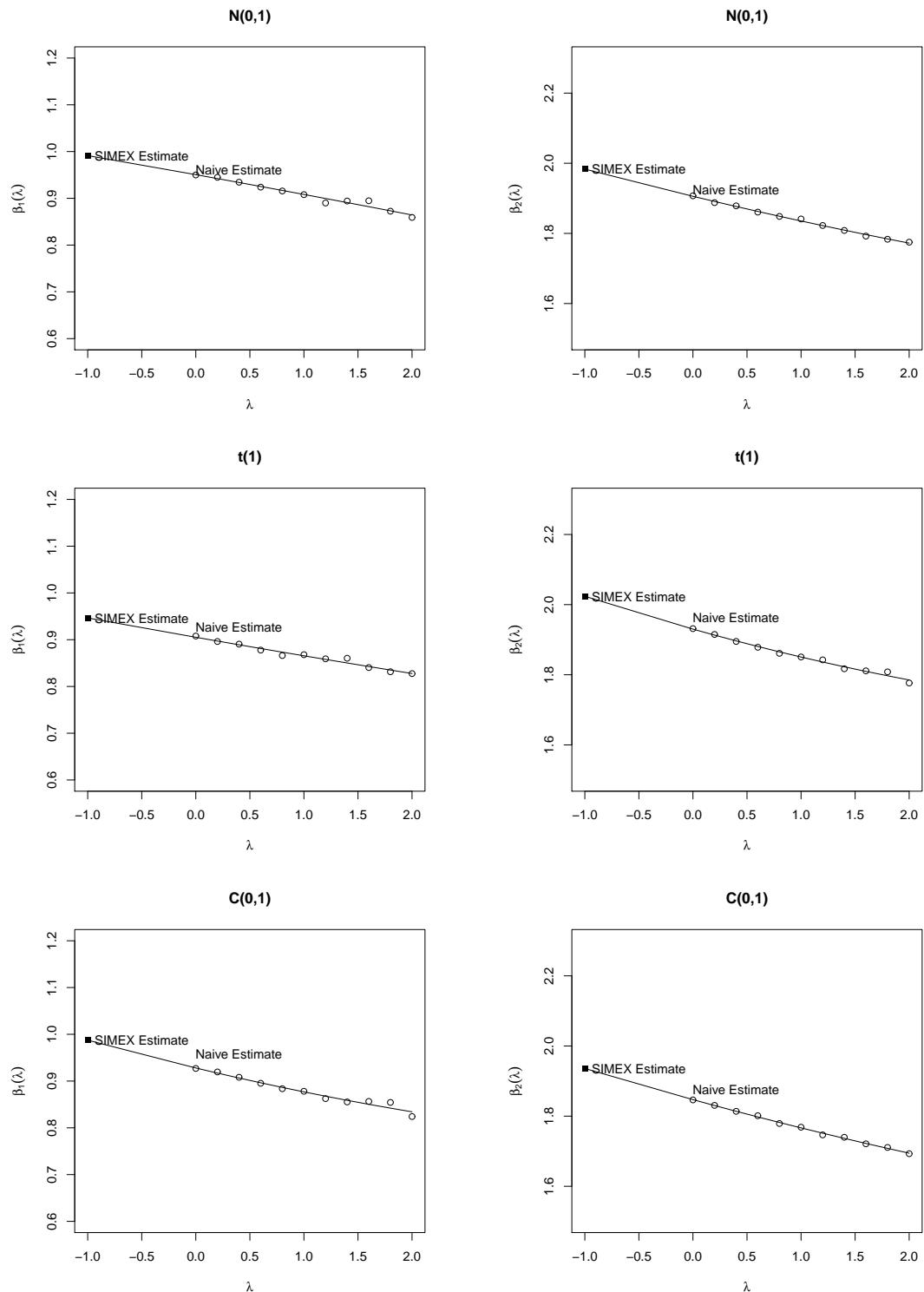


图 2 三种误差分布情形下, 外推过程拟合效果图. 圆点表示 $\{\lambda_i, \hat{\beta}(\lambda_i), i = 1, 2, \dots, 11\}$,
实线是二次拟合曲线, 外推到 $\lambda = -1$ 是 SIMEX 估计 (方形), $\lambda = 0$ 是 Naive 估计

表 1 不同样本容量不同模型误差下参数 β_1 估计的偏差 (Bias) 和标准差 (SD)

| n | Method | N(0, 1) | | t(1) | | C(0, 1) | |
|-----|-----------|---------|--------|---------|--------|---------|--------|
| | | Bias | SD | Bias | SD | Bias | SD |
| 50 | SIMEX-CQR | -0.0058 | 0.0785 | -0.0032 | 0.1064 | 0.0030 | 0.1100 |
| | OR-CQR | -0.0053 | 0.0781 | 0.3501 | 0.8237 | 0.3467 | 0.8217 |
| | BC-LS | -0.0039 | 0.0690 | -0.0243 | 1.5790 | 0.0483 | 2.9557 |
| | Naive-CQR | -0.0438 | 0.0683 | -0.0406 | 0.1021 | -0.0361 | 0.1074 |
| | Naive-LS | -0.0451 | 0.0671 | -0.0625 | 1.5241 | -0.0407 | 2.8251 |
| 100 | SIMEX-CQR | -0.0046 | 0.0575 | -0.0020 | 0.0742 | 0.0015 | 0.0753 |
| | OR-CQR | -0.0041 | 0.0579 | 0.3759 | 0.8044 | 0.4521 | 0.8636 |
| | BC-LS | -0.0029 | 0.0515 | -0.0395 | 4.5569 | 0.0357 | 1.9966 |
| | Naive-CQR | -0.0405 | 0.0489 | -0.0394 | 0.0706 | -0.0374 | 0.0715 |
| | Naive-LS | -0.0407 | 0.0485 | -0.0757 | 4.3366 | -0.0498 | 1.9206 |
| 200 | SIMEX-CQR | -0.0031 | 0.0423 | -0.0010 | 0.0488 | 0.0078 | 0.0479 |
| | OR-CQR | -0.0034 | 0.0492 | 0.4518 | 0.7637 | 0.4579 | 0.8614 |
| | BC-LS | -0.0017 | 0.0414 | 0.3478 | 6.6300 | 0.0595 | 3.3407 |
| | Naive-CQR | -0.0397 | 0.0378 | -0.0403 | 0.0478 | -0.0389 | 0.0468 |
| | Naive-LS | -0.0393 | 0.0383 | 0.2935 | 6.3314 | 0.1545 | 3.2119 |

表 2 不同样本容量不同模型误差下参数 β_2 估计的偏差 (Bias) 和标准差 (SD)

| n | Method | N(0, 1) | | t(1) | | C(0, 1) | |
|-----|-----------|---------|--------|---------|--------|---------|--------|
| | | Bias | SD | Bias | SD | Bias | SD |
| 50 | SIMEX-CQR | -0.0065 | 0.0708 | 0.0043 | 0.1101 | 0.0057 | 0.1124 |
| | OR-CQR | -0.0067 | 0.0757 | 0.4106 | 0.6015 | 0.4041 | 0.6379 |
| | BC-LS | 0.0054 | 0.0713 | 0.1184 | 2.1190 | 0.1520 | 1.9483 |
| | Naive-CQR | -0.0807 | 0.0681 | -0.0708 | 0.1078 | -0.0706 | 0.1051 |
| | Naive-LS | -0.0798 | 0.0653 | 0.0300 | 2.0171 | 0.0636 | 1.8524 |
| 100 | SIMEX-CQR | -0.0051 | 0.0461 | -0.0038 | 0.0752 | 0.0013 | 0.0775 |
| | OR-CQR | -0.0049 | 0.0466 | 0.4297 | 0.5774 | 0.4778 | 0.6566 |
| | BC-LS | -0.0036 | 0.0409 | -0.5030 | 6.9460 | -0.1278 | 3.1354 |
| | Naive-CQR | -0.0731 | 0.0399 | -0.0794 | 0.0739 | -0.0751 | 0.0735 |
| | Naive-LS | -0.0722 | 0.0386 | -0.5617 | 6.6418 | -0.2015 | 3.0212 |
| 200 | SIMEX-CQR | -0.0032 | 0.0449 | -0.0029 | 0.0528 | -0.0008 | 0.0515 |
| | OR-CQR | -0.0026 | 0.0496 | 0.4925 | 0.5423 | 0.5218 | 0.6489 |
| | BC-LS | -0.0023 | 0.0372 | 0.2125 | 4.7889 | -0.2012 | 3.0331 |
| | Naive-CQR | -0.0729 | 0.0353 | -0.0748 | 0.0501 | -0.0746 | 0.0495 |
| | Naive-LS | -0.0718 | 0.0327 | 0.1256 | 4.5910 | -1.3490 | 2.9126 |

表 3 Σ_{uu} 已知与 Σ_{uu} 未知情形下参数 β 估计的偏差 (Bias) 和标准差 (SD), $n = 100$

| β | Σ_{uu} | N(0, 1) | | t(1) | | C(0, 1) | |
|-----------|---------------------|---------|--------|---------|--------|---------|--------|
| | | Bias | SD | Bias | SD | Bias | SD |
| β_1 | Σ_{uu} | -0.0046 | 0.0575 | -0.0020 | 0.0742 | 0.0015 | 0.0753 |
| | $\hat{\Sigma}_{uu}$ | -0.0047 | 0.0566 | -0.0021 | 0.0746 | 0.0014 | 0.0755 |
| β_2 | Σ_{uu} | -0.0051 | 0.0461 | -0.0038 | 0.0752 | 0.0013 | 0.0775 |
| | $\hat{\Sigma}_{uu}$ | -0.0055 | 0.0468 | -0.0039 | 0.0751 | 0.0018 | 0.0757 |

关系, 建立如下模型:

$$\begin{cases} Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i; \\ W_i = X_i + U_i, \quad i = 1, 2, \dots, n, \end{cases}$$

其中响应变量 Y_i 是他们在固定两年内的平均血压 (BP), W_i 是血清胆固醇水平对数 ($\ln(SC)$) 的标准化变量. 类似文献 [19], W 的测量是带测量误差的, 由于该数据集 W 含有重复测量的观察值, 可以采用重复测量的方法来估计测量误差的方差. 图 3 给出了 Y 的直方图和密度图, 从图 3 可以看出, Y 的分布并非正态分布. 因此, 与均值回归相比, 用复合位数回归方法分析该数据可能更合理. 图 4 展示了 SIMEX-CQR 估计在外推步的轨迹, 从图 4 可以看出在外推步中采用二次外推函数很好地拟合了数据 $\{\lambda_i, \hat{\beta}(\lambda_i), i = 1, 2, \dots, 11\}$.

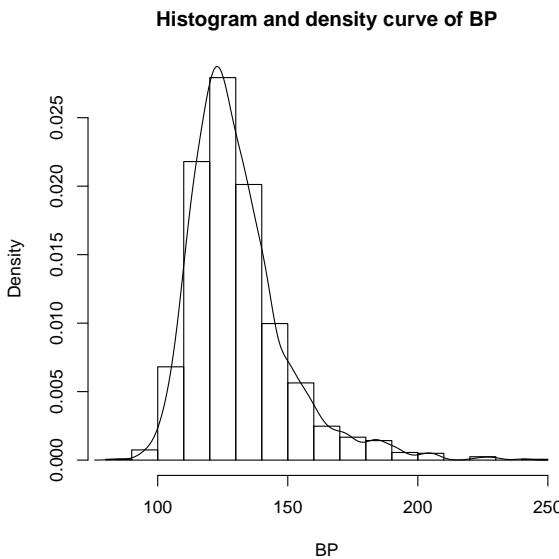


图 3 固定两年内的平均血压 (BP) 的直方图和密度图

为了比较, 类似第 3 节, 我们采用五种方法分析了该数据集, 计算结果见表 4. 从表 4 可以看出, 基于 SIMEX-CQR 方法、OR-CQR 方法以及 BC-LS 方法所得到的 β_2 的估计

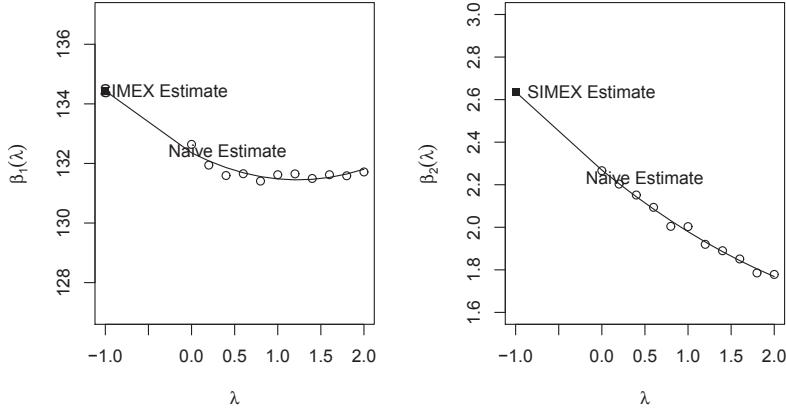


图 4 心脏病数据外推过程拟合效果图. 圆点表示 $\{\lambda_i, \hat{\beta}(\lambda_i), i = 1, 2, \dots, 11\}$, 实线是二次拟合曲线, 外推到 $\lambda = -1$ 是 SIMEX 估计 (方形), $\lambda = 0$ 是 Naive 估计

比 Naive-CQR 和 Naive-LS 估计要大. 该结果表明, 当考虑测量误差时, 血清胆固醇水平与血压具有更强的正相关, 该结果类似文献 [19] 的结果. 由于 Y 的分布非正态分布, 采用 SIMEX-CQR 方法和 OR-CQR 方法分析该数据比 BC-LS 方法更合理. 注意到, OR-CQR 方法要求回归误差和测量误差服从球形对称分布, 如何检验回归误差和测量误差服从球形对称分布, 是值得进一步讨论的问题. 在测量误差和回归误差分布都未知的情形下, 用本文提出的 SIMEX-CQR 方法分析实际数据可能更可靠一些.

表 4 不同估计方法下参数 β 估计

| | SIMEX-CQR | OR-CQR | BC-LS | Naive-CQR | Naive-LS |
|-----------|-----------|----------|----------|-----------|----------|
| β_1 | 134.4360 | 129.0969 | 130.7573 | 132.6395 | 130.7573 |
| β_2 | 2.6363 | 2.7682 | 2.5541 | 2.2660 | 2.1668 |

§5. 定理的证明

定理 3 的证明: 假定 $\beta(\lambda)$ 是基于模型 $Y = W_s^\top(\lambda)\beta(\lambda) + \varepsilon_s^*$ 的真值. 令 $Q_n = \sqrt{n} \cdot (\hat{\beta}_s(\lambda) - \beta)$ 且 $R_{nk} = \sqrt{n}(\hat{b}_{\tau_k} - b_{\tau_k})$. 那么 $(R_{n1}, R_{n2}, \dots, R_{nK}, Q_n)$ 是最小化下面的目标函数:

$$\mathcal{L}_n = \sum_{k=1}^K \sum_{i=1}^n \left[\rho_{\tau_k} \left(\varepsilon_{is}^* - b_{\tau_k} - \frac{R_{nk} + W_{is}^\top(\lambda)Q_n}{\sqrt{n}} \right) - \rho_{\tau_k}(\varepsilon_{is}^* - b_{\tau_k}) \right].$$

类似文献 [2] 中定理 2.1 的证明有

$$\mathcal{L}_n \xrightarrow{P} \sum_{k=1}^K \frac{R_{nk} + W_{is}^\top(\lambda)Q_n}{\sqrt{n}} [I(\varepsilon_{is}^* < b_{\tau_k}) - \tau_k] + \sum_{k=1}^K \frac{1}{2} f(b_{\tau_k}) R_{n,k}^2$$

$$+ \frac{1}{2} \sum_{k=1}^K f(b_{\tau_k}) Q_n^\top \Omega^{-1}(\beta(\lambda), \lambda) Q_n.$$

由于 \mathcal{L}_n 是凸函数, 则

$$\sqrt{n} (\hat{\beta}_s(\lambda) - \beta(\lambda)) = \frac{\Omega^{-1}(\beta(\lambda), \lambda)}{\sum_{k=1}^K f(b_{\tau_k})} \sum_{k=1}^K \sum_{i=1}^n n^{-1/2} W_{is}(\lambda) [I(\varepsilon_{is}^* < b_{\tau_k}) - \tau_k] + o_p(1).$$

根据 $\hat{\beta}(\lambda)$ 的定义有

$$\sqrt{n} (\hat{\beta}(\lambda) - \beta(\lambda)) = \Omega^{-1}(\beta(\lambda), \lambda) n^{-1/2} \sum_{i=1}^n \eta_{i\mathcal{S}}(\beta(\lambda), \lambda) + o_p(1). \quad (3)$$

利用 (3), $\sqrt{n} (\hat{\beta}(\Lambda) - \beta(\Lambda))$ 的极限分布为多元正态 $N(0, \Sigma)$.

根据外推过程, 可得

$$\hat{\boldsymbol{\Gamma}} = \arg \min_{\boldsymbol{\Gamma}} \text{Res}^\top(\boldsymbol{\Gamma}) \text{Res}(\boldsymbol{\Gamma}).$$

$\hat{\boldsymbol{\Gamma}}$ 通过解如下估计方程得到

$$s(\hat{\boldsymbol{\Gamma}}) \text{Res}(\hat{\boldsymbol{\Gamma}}) = 0.$$

则

$$\sqrt{n} (\hat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}) \xrightarrow{\mathcal{L}} N(0, \Sigma(\boldsymbol{\Gamma})).$$

由于 $\hat{\beta}_{\text{SIMEX}} = \mathcal{G}(-1, \hat{\boldsymbol{\Gamma}})$, 那么根据 Δ 方法 SIMEX 估计的渐近方差为

$$\mathcal{G}_{\boldsymbol{\Gamma}}(-1, \boldsymbol{\Gamma}) \Sigma(\boldsymbol{\Gamma}) [\mathcal{G}_{\boldsymbol{\Gamma}}(-1, \boldsymbol{\Gamma})]^\top. \quad \square$$

参 考 文 献

- [1] KOENKER R, BASSETT JR G. Regression quantiles [J]. *Econometrica*, 1978, **46**(1): 33–50.
- [2] ZOU H, YUAN M. Composite quantile regression and the oracle model selection theory [J]. *Ann Statist*, 2008, **36**(3): 1108–1126.
- [3] KAI B, LI R Z, ZOU H. Local composite quantile regression smoothing: an efficient and safe alternative to local polynomial regression [J]. *J R Stat Soc Ser B Stat Methodol*, 2010, **72**(1): 49–69.
- [4] 吕亚召, 张日权, 赵为华, 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: 数学, 2014, **44**(12): 1299–1322.
- [5] 刘慧蓝. 基于复合分位数回归方法的统计模型的相关研究 [D]. 重庆: 重庆大学, 2016.
- [6] ZHAO P X, XUE L G. Variable selection for semiparametric varying coefficient partially linear errors-in-variables models [J]. *J Multivariate Anal*, 2010, **101**(8): 1872–1883.
- [7] STUTE W, XUE L G, ZHU L X. Empirical likelihood inference in nonlinear errors-in-covariates models with validation data [J]. *J Amer Statist Assoc*, 2007, **102**(477): 332–346.

- [8] YANG Y P, LI G R, TONG T J. Corrected empirical likelihood for a class of generalized linear measurement error models [J]. *Sci China Math*, 2015, **58**(7): 1523–1536.
- [9] CUI H J, CHEN S X. Empirical likelihood confidence region for parameter in the errors-in-variables models [J]. *J Multivariate Anal*, 2003, **84**(1): 101–115.
- [10] JIANG R. Composite quantile regression for linear errors-in-variables models [J]. *Hacet J Math Stat*, 2015, **44**(3): 707–713.
- [11] LIANG H, REN H B. Generalized partially linear measurement error models [J]. *J Comput Graph Statist*, 2005, **14**(1): 237–250.
- [12] APANASOVICH T V, CARROLL R J, MAITY A. SIMEX and standard error estimation in semi-parametric measurement error models [J]. *Electron J Stat*, 2009, **3**: 318–348.
- [13] YANG Y P, TONG T J, LI G R. SIMEX estimation for single-index model with covariate measurement error [J]. *AStA Adv Stat Anal*, 2019, **103**(1): 137–161.
- [14] 李高荣, 张君, 冯三营. 现代测量误差模型 [M]. 北京: 科学出版社, 2018.
- [15] FULLER W A. *Measurement Error Models* [M]. Hoboken, NJ: Wiley-Interscience, 2006.
- [16] CARROLL R J, MACA J D, RUPPERT D. Nonparametric regression in the presence of measurement error [J]. *Biometrika*, 1999, **86**(3): 541–554.
- [17] CARROLL R J, RUPPERT D, STEFANSKI L A, et al. *Measurement Error in Nonlinear Models: A Modern Perspective* [M]. 2nd ed. London: Chapman and Hall/CRC, 2006.
- [18] LIN X H, CARROLL R J. Nonparametric function estimation for clustered data when the predictor is measured without/with error [J]. *J Amer Statist Assoc*, 2000, **95**(450): 520–534.
- [19] LIANG H, HÄRDLE W, CARROLL R J. Estimation in a semiparametric partially linear errors-in-variables model [J]. *Ann Statist*, 1999, **27**(5): 1519–1535.

SIMEX Estimation for Composite Quantile Regression Model with Measurement Error

YANG Yiping YU Lu WU Dongsheng

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

Abstract: Composite quantile regression model with measurement error is considered. The SIMEX estimators of the unknown regression coefficients are proposed based on the composite quantile regression. The proposed estimators not only eliminate the bias caused by measurement error, but also retain the advantages of the composite quantile regression estimation. The asymptotic properties of the SIMEX estimation are proved under some regular conditions. The finite sample properties of the proposed method are studied by a simulation study, and a real example is analyzed.

Keywords: composite quantile regression; measurement error; asymptotic normality; SIMEX

2010 Mathematics Subject Classification: 62G08; 62G20