

右删失数据下分位数回归的光滑经验似然检验 *

李忠桂* 何书元

(首都师范大学数学科学学院, 北京, 100048)

摘要: 关于线性分位数回归模型的参数检验问题, 对完全观测数据, 已有文献用经验似然 (EL) 法和光滑经验似然 (SEL) 法构造的检验统计量在原假设下均以卡方分布 χ_M^2 为渐近分布. 对右删失数据, 已有文献用 EL 法构造的检验统计量以加权卡方分布为渐近分布, 而权重是待估的. 对右删失数据, 本文用 EL 法和 SEL 法构造的检验统计量在原假设下均依分布收敛到 χ_M^2 , 因此无需估计权重. 由于 SEL 法的估计函数是光滑的, 故可以进行 Bartlett 纠偏. 随机模拟结果表明与已有的方法相比, SEL 法经过 Bartlett 纠偏后有更高的精度.

关键词: 分位数回归; 右删失; 光滑经验似然

中图分类号: O212.1

英文引用格式: LI Z G, HE S Y. Smoothed empirical likelihood testing for quantile regression models under right censorship [J]. Chinese J Appl Probab Statist, 2019, 35(2): 153–164. (in Chinese)

§1. 引言

分位数回归模型由 Koenker 和 Bassett^[1] 在 1978 年提出, 在之后的几十年里, 人们对模型参数的估计和检验做了许多研究. 对 $p \in (0, 1)$, 线性 p 分位数回归模型是

$$Y = X^\top \beta + U, \quad (1)$$

其中, β 是 M 维列向量, $P(U \leq 0 | X) = p$, $Y \in R$, $X \in R^M$ 的第一个分量是 1. 本文将研究 $H_0: \beta = \beta_0$ vs. $H_1: \beta \neq \beta_0$ 的检验, 其中 β_0 是待检验值.

Otsu^[2] 用经验似然 (EL) 法和光滑经验似然 (SEL) 法构造 β 的置信域, Whang^[3] 详细介绍了 SEL 法在模型 (1) 中的应用, 但是他们的工作都是在完全观测数据下进行的.

当 Y 被随机变量 C 右删失时, 观测到的是 $Z \triangleq Y \wedge C$ 的样本, Otsu^[2] 和 Whang^[3] 的结论不再适用. 对右删失数据, Ying 等^[4] 构造了一个估计方程, 用以估计中位数回归模型的参数. Leng 和 Tong^[5] 推广了 Ying 等^[4] 的方法. Wang 和 Wang^[6] 用局部加权最小绝对距离法估计 β . 在这三篇文献中, β 的估计都是渐近正态的, 但渐近方差的估计十分复杂.

*国家自然科学基金项目 (批准号: 11671274、11231010) 资助.

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

本文 2017 年 5 月 15 日收到, 2018 年 3 月 2 日收到修改稿.

Qin 和 Tsao^[7] 与 Zhao 和 Chen^[8] 用 EL 法为中位数回归模型的参数构造置信域, 但是他们的经验似然比统计量以加权卡方分布为渐近分布, 而权重是待估的.

对右删失数据, 本文根据文献 [9] 的影响函数方法得到 EL 法的估计函数 W_{ni} (见 (9)), 据此构造的检验统计量在原假设下依分布收敛到 χ_M^2 , χ_M^2 是自由度为 M 的卡方分布, 但 W_{ni} 不是 β 的光滑函数, 因此无法对该统计量进行 Bartlett 纠偏. 参照文献 [3], 本文用 SEL 法解决这个问题, 构造的检验统计量在原假设下仍依分布收敛到 χ_M^2 . 随机模拟表明经过 Bartlett 纠偏后, SEL 法的精度高于正态渐近法和文献 [7] 的 EL 法.

第 2 节是准备工作. 第 3 节用 EL 法构造以 χ_M^2 为渐近分布的检验统计量. 第 4 节推广文献 [9] 中右删失数据下的中心极限定理. 第 5 节依据推广的中心极限定理对模型 (1) 的 SEL 检验做统计推断. 第 6 节是随机模拟和实例分析.

§2. 准备和假设

为了方便, 本文引用文献 [10] 的符号和条件.

对任意单调函数 $h(x)$, 设 $h(x-)$ 是 $h(x)$ 的左极限, $h\{x\} \triangleq h(x) - h(x-)$. 对分布函数 F , 设 $\bar{F} \triangleq 1 - F$, $b_F \triangleq \sup\{y : F(y) < 1\}$. $\forall a, b \in R$, $a \wedge b \triangleq \min(a, b)$. $\forall a = (a_1, a_2, \dots, a_M)^\top$, $b = (b_1, b_2, \dots, b_M)^\top$, 当 $a_j \leq b_j$, $j = 1, 2, \dots, M$ 时, 记 $a \leq b$.

关于模型 (1), 在右删失下, 观测到的不再是 (Y, X) 的样本, 而是 (Z, X, δ) 的样本

$$(Z_i, X_i, \delta_i), \quad i = 1, 2, \dots, n,$$

其中, $Z = Y \wedge C$, $\delta = I[Y \leq C]$ 为 $[Y \leq C]$ 的示性函数, $Z_i = Y_i \wedge C_i$, $\delta_i = I[Y_i \leq C_i]$.

假设 C 与 Y 独立且满足 $\mathbb{P}(Y \leq C | Y, X) = \mathbb{P}(Y \leq C | Y)$, 定义 $F(y) = \mathbb{P}(Y \leq y)$, $G(y) = \mathbb{P}(C \leq y)$, $F(y, x) = \mathbb{P}(Y \leq y, X \leq x)$, $F^1(y, x) = \mathbb{P}(Y \leq y, X \leq x, \delta = 1)$, 则

$$F^1(y, x) = \int_{u \leq y} \int_{w \leq x} \bar{G}(u-) F(du, dw). \quad (2)$$

定义 $H(y) = \mathbb{P}(Z \leq y)$, $H^0(y) = \mathbb{P}(Z \leq y, \delta = 0)$, $H^1(y) = \mathbb{P}(Z \leq y, \delta = 1)$, 则

$$H^0(y) = \int_{-\infty}^y \bar{F}(s) dG(s), \quad H^1(y) = \int_{-\infty}^y \bar{G}(s-) dF(s), \quad \bar{H}(y) = \bar{F}(y) \bar{G}(y). \quad (3)$$

用下面经验分布函数分别估计 $H(y)$, $H^0(y)$, $H^1(y)$ 和 $F^1(y, x)$,

$$\begin{aligned} H_n(y) &= \frac{1}{n} \sum_{i=1}^n I[Z_i \leq y], & H_n^0(y) &= \frac{1}{n} \sum_{i=1}^n I[Z_i \leq y, \delta_i = 0], \\ H_n^1(y) &= \frac{1}{n} \sum_{i=1}^n I[Z_i \leq y, \delta_i = 1], & F_n^1(y, x) &= \frac{1}{n} \sum_{i=1}^n I[Z_i \leq y, X_i \leq x, \delta_i = 1]. \end{aligned}$$

为了保证模型可识别, 假设 $b_F \leq b_G$, 则 $F(y)$ 和 $G(y)$ 的 Kaplan-Meier 估计为

$$F_n(y) = 1 - \prod_{s \leq y} \left[1 - \frac{H_n^1\{s\}}{\bar{H}_n(s-)} \right] \quad \text{和} \quad G_n(y) = 1 - \prod_{s \leq y} \left[1 - \frac{H_n^0\{s\}}{\bar{H}_n(s-)} \right]. \quad (4)$$

$F_n(y)$, $G_n(y)$, $H_n^1(y)$ 和 $H_n(y)$ 满足

$$\bar{H}_n(y) = \bar{F}_n(y)\bar{G}_n(y) \quad \text{和} \quad dH_n^1(y) = \bar{G}_n(y-)dF_n(y). \quad (5)$$

按文献 [9], 分布函数 $F(y, x)$ 的估计是

$$F_n(y, x) = \int_{s \leq y} \int_{w \leq x} \frac{1}{\bar{G}_n(s-)} F_n^1(ds, dw). \quad (6)$$

在本文中, \xrightarrow{L} 表示依分布收敛, $\xrightarrow{\text{a.s.}}$ 表示几乎处处收敛, 积分号 \int_a^b 表示 $\int_{(a,b]}$, \int 表示 \int_R 或 $\int_{R^{M+1}}$. 如无特殊说明, 下文出现的相同符号以本节定义为准.

§3. 右删失数据下的经验似然

对模型 (1), 考虑用 EL 法检验 H_0 vs. H_1 , 对完全观测数据, 文献 [2] 的估计函数是

$$g(Y, X, \beta) = [G_0(X^\top \beta - Y) - p]X, \quad (7)$$

其中 $G_0(u) \triangleq I[u \geq 0]$. 在 H_0 下, 易得 $E[g(Y, X, \beta_0)] = 0$, 文献 [2] 证明了 $n \rightarrow \infty$ 时,

$$-2 \sup \left\{ \sum_{i=1}^n \ln(np_i) : p_i \geq 0, \sum_{i=1}^n p_i = 1, \sum_{i=1}^n p_i g(Y_i, X_i, \beta_0) = 0 \right\} \xrightarrow{L} \chi_M^2. \quad (8)$$

对右删失数据需要构造新的检验统计量, 设 $\psi_n(s, \beta_0) = \int_{y \geq s} g(y, x, \beta_0) F_n(dy, dx)$, $\bar{\delta}_i = 1 - \delta_i$, 参照文献 [10] 用

$$W_{ni} = \frac{g(Z_i, X_i, \beta_0)\delta_i}{\bar{G}_n(Z_i-)} + \frac{\psi_n(Z_i, \beta_0)}{\bar{H}_n(Z_i-)} \bar{\delta}_i - \int_{-\infty}^{Z_i} \frac{\psi_n(s, \beta_0)}{\bar{H}_n^2(s-)} dH_n^0(s) \quad (9)$$

构造经验似然比检验统计量

$$l_{\text{EL}}(\beta_0) = -2 \sup \left\{ \sum_{i=1}^n \ln(np_i) : p_i \geq 0, \sum_{i=1}^n p_i = 1, \sum_{i=1}^n p_i W_{ni} = 0 \right\}. \quad (10)$$

定理 1 若 $S_0 \triangleq E(X_i X_i^\top)$ 正定, F 和 G 连续, 且 $\exists \lambda > 1/2$ 使得

$$\int \frac{\|x\|}{\bar{G}^\lambda(y)\bar{F}^{1/2}(y)} F(dy, dx) < \infty, \quad E \frac{X^\top X}{\bar{G}(Y)} < \infty, \quad (11)$$

则在 $H_0 : \beta = \beta_0$ 下, $l_{\text{EL}}(\beta_0) \xrightarrow{L} \chi_M^2$, $n \rightarrow \infty$.

证明: 由于在 H_0 下 $E[g(Y, X, \beta_0)] = 0$, $E(X_i X_i^\top)$ 正定, 并且 (7) 和 (11) 表明

$$\int \frac{\|g(Y, X, \beta_0)\|}{\bar{G}^\lambda(y) \bar{F}^{1/2}(y)} F(dy, dx) < \infty, \quad \int \frac{\|g(Y, X, \beta_0)\|^2}{\bar{G}(y)} F(dy, dx) < \infty,$$

故根据文献 [10] 中定理 4.1 得 $l_{EL}(\beta_0) \xrightarrow{L} \chi_M^2$. \square

注记 2 上述定理将文献 [10] 中定理 4.1 的条件 $\int_{-\infty}^{b_F} \bar{F}^{-1} dG < \infty$ 替换为 $\int \|x\| / [\bar{G}^\lambda(y) \bar{F}^{1/2}(y)] F(dy, dx) < \infty$, 用文献 [11] 的定理 2.1 能证明这个替换是正确的.

§4. 右删失数据下中心极限定理的推广

为了本文的应用, 下面定理是对文献 [9] 中定理 3.1 的推广.

定理 3 设 $\xi_n(y, x)$, $\xi(y, x)$ 和 $\eta(y, x) \in R$ 可测, $\bar{\xi}_n(y, x) \triangleq \xi_n(y, x) - \xi(y, x)$ 满足 $|\xi_n(Y, X)| \leq \eta(Y, X)$ 和 $\lim_{n \rightarrow \infty} \bar{\xi}_n(Y, X) = 0$, a.s., F 和 G 连续, 且 $\exists \lambda > 1/2$ 使得

$$\int \frac{\eta(y, x)}{\bar{G}^\lambda(y) \bar{F}^{1/2}(y)} F(dy, dx) < \infty, \quad \int \frac{\eta^2(y, x)}{\bar{G}(y)} F(dy, dx) < \infty. \quad (12)$$

定义 $\mu = E[\xi(Y, X)]$, $\psi(s) = \int_{y \geq s} \xi(y, x) F(dy, dx)$, 则当 $n \rightarrow \infty$ 时,

$$\sqrt{n} \int \xi_n(y, x) [F_n(dy, dx) - F(dy, dx)] \xrightarrow{L} N(0, \sigma^2), \quad (13)$$

其中

$$\sigma^2 = \int \frac{\xi^2(y, x)}{\bar{G}(y)} F(dy, dx) - \mu^2 - \int \frac{\psi^2(s)}{\bar{H}^2(s-)} dH^0(s).$$

证明: 显然

$$\begin{aligned} & \sqrt{n} \int \xi_n(y, x) [F_n(dy, dx) - F(dy, dx)] \\ &= \sqrt{n} \int \xi(y, x) [F_n(dy, dx) - F(dy, dx)] + \sqrt{n} \int \bar{\xi}_n(y, x) [F_n(dy, dx) - F(dy, dx)], \end{aligned}$$

根据文献 [9], 上式等号右边第一部分依分布收敛到 $N(0, \sigma^2)$, 故只需证明

$$\sqrt{n} \int \bar{\xi}_n(y, x) [F_n(dy, dx) - F(dy, dx)] = o_P(1). \quad (14)$$

设 $Z_{(n)} = \max\{Z_1, Z_2, \dots, Z_n\}$, 根据 (2) 和 (6) 得到

$$\begin{aligned} & \sqrt{n} \left| \int \bar{\xi}_n(y, x) [F_n(dy, dx) - F(dy, dx)] \right| \\ & \leq \sqrt{n} \int \frac{|[G(y) - G_n(y)] \bar{\xi}_n(y, x)|}{\bar{G}(y)} F_n(dy, dx) + \sqrt{n} \left| \int \frac{\bar{\xi}_n(y, x)}{\bar{G}(y)} [F_n^1(dy, dx) - F^1(dy, dx)] \right| \end{aligned}$$

$$\triangleq \Delta_n^{(1)} + \Delta_n^{(2)}. \quad (15)$$

为了证明 (14), 下面证明 $\Delta_n^{(1)} = o_P(1)$, $\Delta_n^{(2)} = o_P(1)$.

根据 Egorov's 定理和文献 [11] 中定理 2.1 分别得到 $n \rightarrow \infty$ 时,

$$\int \frac{|\bar{\xi}_n(y, x)|}{\bar{G}^\lambda(y)\bar{F}^{1/2}(y)} F_n(dy, dx) \xrightarrow{\text{a.s.}} 0$$

和

$$\sqrt{n} \sup_{y \leq Z_{(n)}} \frac{|G(y) - G_n(y)|}{\bar{G}(y)} \bar{G}^\lambda(y) \bar{F}^{1/2}(y) = O_P(1),$$

随之推出 $\Delta_n^{(1)} = o_P(1)$. 由 Chebyshev 不等式和控制收敛定理知 $\forall \varepsilon > 0$,

$$\mathbb{P}(\Delta_n^{(2)} \geq \varepsilon) \leq \frac{1}{\varepsilon^2} \int \frac{\bar{\xi}_n^2(y, x)}{\bar{G}(y)} F(dy, dx) \rightarrow 0, \quad n \rightarrow \infty.$$

因此 $\Delta_n^{(2)} = o_P(1)$, 结合 $\Delta_n^{(1)} = o_P(1)$ 和 (15) 得 (14). \square

§5. 光滑经验似然和 Bartlett 纠偏

对模型 (1), 第 3 节构造了 H_0 vs. H_1 的检验统计量 l_{EL} , 但由于 g 不是 β 的光滑函数, 因此无法对 l_{EL} 进行 Bartlett 纠偏. 参照文献 [3], 下面运用 SEL 法解决这个问题.

设 $K(u)$ 是核函数, $G_h(w) = \int_{u < w/h} K(u) du$, 则 g 的光滑版本是

$$g_n(y, x, \beta) = [G_h(x^\top \beta - y) - p]x. \quad (16)$$

定义 $\psi(s, \beta) = \int_{y \geq s} g(y, x, \beta) F(dy, dx)$, $\hat{\psi}_n(s, \beta_0) = \int_{y \geq s} g_n(y, x, \beta_0) F_n(dy, dx)$,

$$\begin{cases} W_i = \frac{g(Z_i, X_i, \beta) \delta_i}{\bar{G}(Z_i-)} + \frac{\psi(Z_i, \beta)}{\bar{H}(Z_i-)} \bar{\delta}_i - \int_{-\infty}^{Z_i} \frac{\psi(s, \beta)}{\bar{H}^2(s-)} dH^0(s), \\ \hat{W}_{ni} = \frac{g_n(Z_i, X_i, \beta_0) \delta_i}{\bar{G}_n(Z_i-)} + \frac{\hat{\psi}_n(Z_i, \beta_0)}{\bar{H}_n(Z_i-)} \bar{\delta}_i - \int_{-\infty}^{Z_i} \frac{\hat{\psi}_n(s, \beta_0)}{\bar{H}_n^2(s-)} dH_n^0(s). \end{cases} \quad (17)$$

对右删失数据, 根据文献 [9] 的影响函数方法构造出 H_0 vs. H_1 的检验统计量如下

$$l_{SEL}(\beta_0) \triangleq -2 \sup \left\{ \sum_{i=1}^n \ln(np_i) : p_i \geq 0, \sum_{i=1}^n p_i = 1, \sum_{i=1}^n p_i \hat{W}_{ni} = 0 \right\}. \quad (18)$$

在给出 $l_{SEL}(\beta_0)$ 的渐近性质前, 先引入下面几个条件.

C1 核函数 $K(u)$ 满足

$$\int_R u^j K(u) du = \begin{cases} 1, & j = 0; \\ 0, & j = 1, 2, \dots, r-1; \\ c_r, & j = r, \end{cases}$$

其中, 偶数 $r \geq 2$, c_r 为正常数.

C2 设 $f(u|X)$ 是 X 已知时 U 的条件密度函数. 对 X 和 u 的值, $f(u|X)$ 有界, 且在 $u=0$ 的某个邻域内关于 u 存在 r 阶连续导数.

C3 $S_0 = E(X_i X_i^\top)$ 和 $D_0 = E[f(0|X_i) X_i X_i^\top]$ 正定.

C4 窗宽 h 满足 $\lim_{n \rightarrow \infty} nh^{2r} = 0$.

C5 分布函数 F 和 G 连续, 且 $\exists \lambda > 1/2$ 使得对 $\eta(y, x) \triangleq \|x\|$, (12) 成立.

定理 4 若条件 C1–C5 成立, 则在 $H_0 : \beta = \beta_0$ 下, $l_{\text{SEL}}(\beta_0) \xrightarrow{L} \chi_M^2$, $n \rightarrow \infty$.

由于 g_n 是 β 的光滑函数, 故可以对 l_{SEL} 进行 Bartlett 纠偏. 参照文献 [3], 用经 Bartlett 纠偏后的光滑经验似然 (BSEL) 法检验 H_0 vs. H_1 , 拒绝域为

$$R_{\text{BSEL}} = \{\beta : l_{\text{BSEL}}(\beta) > \chi_{M, 1-\alpha}^2\}, \quad (19)$$

其中, $l_{\text{BSEL}}(\beta) \triangleq (1 - n^{-1}b)l_{\text{SEL}}(\beta)$, $\chi_{M, 1-\alpha}^2$ 是 χ_M^2 的 $1 - \alpha$ 分位数, b 是纠偏因子, 其计算见 (24). 当 $\beta_0 \in R_{\text{BSEL}}$ 时, 拒绝原假设 H_0 .

为证明定理 4, 先介绍几个引理备用.

引理 5 若条件 C1–C5 成立, 则在 H_0 下 $n \rightarrow \infty$ 时,

- (i) $g_n(Y, X, \beta_0) \xrightarrow{\text{a.s.}} g(Y, X, \beta)$,
- (ii) $E[g_n(Y, X, \beta_0)] = o(n^{-1/2})$.

证明: (i) 由于 $\forall w \neq 0$, $\lim_{h \rightarrow 0} G_h(w) = G_0(w)$, 故 $\lim_{n \rightarrow \infty} g_n(Y, X, \beta_0) = g(Y, X, \beta)$, a.s.

(ii) 文献 [3] 引理 1 证明了 $E[g_n(Y, X, \beta_0)] = O(h^r)$, 结合条件 C4 得 (ii). \square

引理 6 设 $\widehat{V}_{ni} = \widehat{W}_{ni} \overline{H}_n(Z_i-) \overline{H}(Z_i)$, $V_i = W_i \overline{H}^2(Z_i)$. 若条件 C1–C5 成立, 则在 H_0 下, $n^{-1} \sum_{i=1}^n \|\widehat{W}_{ni} - W_i\|^2 = o_P(1)$, 且 $n^{-1} \sum_{i=1}^n \|\widehat{V}_{ni} - V_i\|^2 \xrightarrow{\text{a.s.}} 0$, $n \rightarrow \infty$.

证明: 设 $\widehat{\psi}(s, \beta_0) = \int_{y \geq s} g_n(y, x, \beta_0) F(dy, dx)$,

$$\widehat{W}_i = \frac{g_n(Z_i, X_i, \beta_0) \delta_i}{\overline{G}(Z_i-)} + \frac{\widehat{\psi}(Z_i, \beta_0)}{\overline{H}(Z_i-)} \bar{\delta}_i - \int_{-\infty}^{Z_i} \frac{\widehat{\psi}(s, \beta_0)}{\overline{H}^2(s-)} dH^0(s),$$

则

$$n^{-1} \sum_{i=1}^n \|\widehat{W}_{ni} - W_i\|^2 \leq n^{-1} \sum_{i=1}^n \|\widehat{W}_{ni} - \widehat{W}_i\|^2 + n^{-1} \sum_{i=1}^n \|\widehat{W}_i - W_i\|^2.$$

同文献 [12] 中引理 4.2 可得 $n^{-1} \sum_{i=1}^n \|\widehat{W}_{ni} - \widehat{W}_i\|^2 = o_P(1)$ 和 $n^{-1} \sum_{i=1}^n \|\widehat{W}_i - W_i\|^2 = o_P(1)$,

故 $n^{-1} \sum_{i=1}^n \|\widehat{W}_{ni} - W_i\|^2 = o_P(1)$. 同理可得 $n^{-1} \sum_{i=1}^n \|\widehat{V}_{ni} - V_i\|^2 \xrightarrow{\text{a.s.}} 0$. \square

引理 7 设 $\Sigma = E(W_i W_i^\top)$, 若条件 C1–C5 成立, 则在 H_0 下 $n \rightarrow \infty$ 时,

- (i) $\max_{1 \leq i \leq n} \|\widehat{W}_{ni}\| = o_P(\sqrt{n})$,
- (ii) $S_n \triangleq n^{-1} \sum_{i=1}^n \widehat{W}_{ni} \widehat{W}_{ni}^\top = \Sigma + o_P(1)$,
- (iii) $n^{-1/2} \sum_{i=1}^n \widehat{W}_{ni} \xrightarrow{L} N(0, \Sigma)$.

证明: (i) 已知 W_1, W_2, \dots, W_n i.i.d., $E(W_i) = 0$, 且 $\text{Var}(W_i) = \Sigma$ 有限, 结合文献 [13] 的引理 3 便得 $\max_{1 \leq i \leq n} \|W_i\| = o_P(\sqrt{n})$. 联系引理 6 得

$$\max_{1 \leq i \leq n} \|\widehat{W}_{ni}\| \leq \sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n \|\widehat{W}_{ni} - W_i\|^2 \right)^{1/2} + \max_{1 \leq i \leq n} \|W_i\| = o_P(\sqrt{n}).$$

(ii) 注意到

$$\widehat{W}_{ni} \widehat{W}_{ni}^\top = W_i W_i^\top + (\widehat{W}_{ni} - W_i) W_i^\top + W_i (\widehat{W}_{ni} - W_i)^\top + (\widehat{W}_{ni} - W_i) (\widehat{W}_{ni} - W_i)^\top,$$

结合 Hölder's 不等式, 强大数定理和引理 6 推出 $S_n = \Sigma + o_P(1)$.

(iii) 根据 (6) 和引理 5 (ii) 得到

$$\begin{aligned} \frac{1}{\sqrt{n}} \sum_{i=1}^n \widehat{W}_{ni} &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \left[\frac{g_n(Z_i, X_i, \beta_0) \delta_i}{\bar{G}_n(Z_i^-)} + \frac{\widehat{\psi}_n(Z_i, \beta_0)}{\bar{H}_n(Z_i^-)} \bar{\delta}_i - \int \frac{\widehat{\psi}_n(s, \beta_0)}{\bar{H}_n^2(s^-)} I[Z_i \geq s] dH_n^0(s) \right] \\ &= \sqrt{n} \int g_n(y, x, \beta_0) [F_n(dy, dx) - F(dy, dx)] + o(1). \end{aligned}$$

取 $\xi(y, x) = g(y, x, \beta)$, $\xi_n(y, x) = g_n(y, x, \beta_0)$, 结合条件 C5, 引理 5 (i), 定理 3 和 Cramér-Wold's 定理即得 (iii). \square

记 $\Sigma = (\sigma_{ij})_{i,j=1}^M$, x_i 是 x 的第 i 个分量, $\varphi_i(s) \triangleq \int_{y \geq s} [I(x^\top \beta \geq y) - p] x_i F(dy, dx)$, 则

$$\sigma_{ij} = \int \frac{[I(x^\top \beta \geq y) - p]^2 x_i x_j}{\bar{G}(y)} F(dy, dx) - \int \frac{\varphi_i(s) \varphi_j(s)}{\bar{H}^2(s^-)} dH^0(s). \quad (20)$$

下面完成定理 4 的证明.

定理 4 的证明: $\forall c \in R^M$, $c^\top \Sigma c = \text{Var}(c^\top W_i) \geq \text{Var}[c^\top g(Y, X, \beta)] = p(1-p)c^\top S_0 c$, 故由 S_0 正定知 Σ 亦正定.

设 A 是 R^M 的全体单位向量, 结合 Σ 正定和引理 6, 参照文献 [10] 中定理 4.1 的证明可知 $\exists \delta > 0$ 使得

$$\liminf_{n \rightarrow \infty} \inf_{\alpha \in A} \frac{1}{n} \sum_{i=1}^n I[\alpha^\top \widehat{W}_{ni} > 0] \geq \delta, \quad \text{a.s.} \quad (21)$$

根据文献 [13] 中定理 1 的证明, 知道了 (21) 和引理 7 便能推出 $n \rightarrow \infty$ 时,

$$l_{\text{SEL}}(\beta_0) = \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \widehat{W}_{ni} \right)^\top S_n^{-1} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \widehat{W}_{ni} \right) + o_P(1) \xrightarrow{L} \chi_M^2.$$

定理 4 证毕. \square

注记 8 当 $P(Y \leq C) = 1$ 时, 不存在右删失. 此时, $H^0(y) \equiv 0$, 并且 a.s. $\overline{G}(Y-) = 1$, $\delta = \delta_1 = \dots = \delta_n = 1$, $H_n^0(y) \equiv 0$, 由 Kaplan-Meier 估计 (4) 知 $G_n(y) \equiv 0$. 因此

$$W_i = [G_0(X_i^\top \beta - Y_i) - p]X_i, \quad \widehat{W}_{ni} = [G_h(X_i^\top \beta_0 - Y_i) - p]X_i, \quad \Sigma = p(1-p)S_0.$$

由 (18) 和定理 4 知, 对完全观测数据

$$\begin{aligned} l_{\text{SEL}}(\beta_0) &= -2 \sup \left\{ \sum_{i=1}^n \ln(np_i) : p_i \geq 0, \sum_{i=1}^n p_i = 1, \sum_{i=1}^n p_i [G_h(X_i^\top \beta_0 - Y_i) - p]X_i = 0 \right\} \\ &\xrightarrow{L} \chi_M^2, \quad n \rightarrow \infty. \end{aligned}$$

这与文献 [3] 的结论一致.

§6. 随机模拟和实例分析

对 $H_0 : \beta = \beta_0$ vs. $H_1 : \beta \neq \beta_0$, 本节将比较根据文献 [4] 的正态渐近法 (NA)、文献 [7] 的 EL 法 (QEL) 及本文的 BSEL 法构造的三个检验法则犯第一类错误的概率, 检验的显著水平为 α , 所有计算均通过 R 语言实现.

对 p 分位数回归模型 (1), 文中采用文献 [6] 的方法估计 β (设定窗宽为 $h_w = n^{-0.4}$), 记得到的估计为 $\widehat{\beta}$. 当 $p = 0.5$ 时, 参照文献 [7], 定义

$$\begin{aligned} U_{ni}(\beta) &\triangleq \left[\frac{I[Z_i - X_i^\top \beta \geq 0]}{1 - G_n(X_i^\top \beta)} - 0.5 \right] X_i, & \widehat{\Gamma}_1 &\triangleq \frac{1}{n} \sum_{i=1}^n U_{ni}(\widehat{\beta}) U_{ni}^\top(\widehat{\beta}), \\ \widehat{\Gamma}_2 &\triangleq \frac{1}{4n} \sum_{i=1}^n \left(\bar{\delta}_i \sum_{j=1}^n I[X_j^\top \widehat{\beta} \geq Z_i] X_j \Big/ \sum_{j=1}^n I[Z_j \geq Z_i] \right)^{\otimes 2}, & \widehat{\Gamma} &\triangleq \widehat{\Gamma}_1 - \widehat{\Gamma}_2, \end{aligned}$$

其中对 $\nu \in R^M$, $\nu^{\otimes 2} \triangleq \nu \nu^\top$.

NA 检验的拒绝域为

$$R_{\text{NA}} = \{\beta : l_{\text{NA}}(\beta) > \chi_{M, 1-\alpha}^2\}, \quad (22)$$

其中 $l_{\text{NA}}(\beta) \triangleq n^{-1} \left(\sum_{i=1}^n U_{ni}(\beta) \right)^{\top} \widehat{\Gamma}^{-1} \left(\sum_{i=1}^n U_{ni}(\beta) \right)$, 当 $\beta_0 \in R_{\text{NA}}$ 时, 拒绝原假设 H_0 .

QEL 检验的拒绝域为

$$R_{\text{QEL}} = \{ \beta : l_{\text{QEL}}(\beta) > c_{1-\alpha} \}, \quad (23)$$

$$l_{\text{QEL}}(\beta) \triangleq -2 \sup \left\{ \sum_{i=1}^n \ln(np_i) : p_i \geq 0, \sum_{i=1}^n p_i = 1, \sum_{i=1}^n p_i U_{ni}(\beta) = 0 \right\},$$

其中, $c_{1-\alpha}$ 是 $l_1\zeta_1 + l_2\zeta_2 + \dots + l_M\zeta_M$ 的 $1 - \alpha$ 分位数, l_1, l_2, \dots, l_M 是 $\widehat{\Gamma}_1^{-1}\widehat{\Gamma}$ 的特征值, $\zeta_1, \zeta_2, \dots, \zeta_M \stackrel{\text{i.i.d.}}{\sim} \chi_1^2$. 当 $\beta_0 \in R_{\text{QEL}}$ 时, 拒绝原假设 H_0 .

BSEL 检验以 (19) 定义的 R_{BSEL} 为拒绝域. 计算 l_{SEL} 时, 核函数和窗宽分别取为 $K(u) = \exp(-u^2/2)/\sqrt{2\pi}$ 和 $h = n^{-0.8}$, 纠偏因子 b 的计算如下

$$\begin{cases} b = \left(\frac{1}{2}t_1 - \frac{1}{3}t_2 \right) / M, & t_1 = \frac{1}{n} \sum_{i=1}^n [\widehat{W}_{ni}^{\top}(\widehat{\beta}) \widehat{\Sigma}^{-1} \widehat{W}_{ni}(\widehat{\beta})]^2, \\ t_2 = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n [\widehat{W}_{ni}^{\top}(\widehat{\beta}) \widehat{\Sigma}^{-1} \widehat{W}_{nj}(\widehat{\beta})]^3, & \widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^n \widehat{W}_{ni}(\widehat{\beta}) \widehat{W}_{ni}^{\top}(\widehat{\beta}), \\ \widehat{W}_{ni}(\widehat{\beta}) = \frac{g_n(Z_i, X_i, \widehat{\beta}) \delta_i}{G_n(Z_i-)} + \frac{\widehat{\psi}_n(Z_i, \widehat{\beta})}{H_n(Z_i-)} \bar{\delta}_i - \int_{-\infty}^{Z_i} \frac{\widehat{\psi}_n(s, \widehat{\beta})}{H_n^2(s-)} dH_n^0(s). \end{cases} \quad (24)$$

本文选用下面两个中位数回归模型进行模拟.

模型 A

$$Y_i = X_i^{\top} \beta + U_i, \quad i = 1, 2, \dots, n, \quad (25)$$

其中, $\beta = (1, 2)^{\top}$, $X_i = (1, X_{2,i})^{\top}$, $X_{2,i} \sim U[0, 1]$, $U_i \sim N(0, 1)$, $C_i \sim \text{Exp}(\lambda)$, 且 $X_1, X_2, \dots, X_n, U_1, U_2, \dots, U_n, C_1, C_2, \dots, C_n$ 相互独立.

模型 B 对 (25), $\beta, X_1, X_2, \dots, X_n$ 和 C_1, C_2, \dots, C_n 的设定同模型 A. $U_i \sim \text{Exp}(2) - u_0$, 其中 $u_0 = (\ln 2)/2$ 是 $\text{Exp}(2)$ 的中位数.

在模拟时, 选取适当的 λ , 使得删失概率 $P(Y > C)$ 达到 15%、25% 和 40%. 样本量取 $n \in \{30, 60, 90\}$. 模拟重复次数为 $N = 25000$ 次, 表 1 输出的是 NA、QEL 和 BSEL 三个检验犯第一类错误的经验概率. 检验的功效是当 H_1 为真时检验拒绝 H_0 的概率, 为了研究检验的功效, 取 $\beta_0 = (1.2, 2.2)^{\top}$, 表 2 输出的是三个检验的经验功效. $\beta_0 = (0.8, 1.8)^{\top}$ 时模拟结果与表 2 类似, 这里就不再列出了. 观察表 1 和表 2 得到下面结论.

- 1) 在表 1 中, 当删失概率较大且样本量较小时, NA 检验优于其余两个检验. 例如对模型 A, 当 $P(Y > C) = 40\%$, $n = 30$ 时 NA 检验犯第一类错误的概率约为 0.091 ($\alpha = 0.1$) 和 0.044 ($\alpha = 0.05$), 较其余两个检验更接近 α .
- 2) 当删失概率较小或样本量较大时, BSEL 检验优于其余两个检验, 如当 $P(Y > C) = 15\%, 25\%$, 或 $n = 60, 90$ 时 BSEL 检验犯第一类错误的概率更接近 α .

表 1 NA、QEL 和 BSEL 三个检验犯第一类错误的经验概率

模型	$P(Y > C)$	n	$\alpha = 0.1$			$\alpha = 0.05$		
			NA	QEL	BSEL	NA	QEL	BSEL
A	15%	30	0.091	0.105	0.103	0.043	0.057	0.055
		60	0.097	0.107	0.102	0.047	0.054	0.050
		90	0.098	0.104	0.100	0.050	0.054	0.050
	25%	30	0.086	0.115	0.104	0.041	0.063	0.054
		60	0.093	0.105	0.097	0.046	0.054	0.048
		90	0.096	0.104	0.098	0.046	0.054	0.049
	40%	30	0.091	0.152	0.120	0.044	0.093	0.072
		60	0.094	0.111	0.099	0.044	0.057	0.051
		90	0.096	0.107	0.101	0.046	0.055	0.049
B	15%	30	0.088	0.112	0.103	0.041	0.059	0.055
		60	0.095	0.108	0.098	0.047	0.056	0.051
		90	0.098	0.104	0.099	0.050	0.054	0.050
	25%	30	0.085	0.116	0.106	0.040	0.064	0.058
		60	0.093	0.110	0.098	0.044	0.057	0.048
		90	0.094	0.107	0.099	0.048	0.056	0.049
	40%	30	0.087	0.132	0.117	0.040	0.076	0.067
		60	0.090	0.112	0.097	0.044	0.058	0.047
		90	0.092	0.107	0.097	0.044	0.056	0.049

3) NA 检验犯第一类错误的概率总是小于 α , QEL 检验犯第一类错误的概率总是大于 α , BSEL 检验犯第一类错误的概率在 α 左右波动, 但它们都随 $P(Y > C)$ 增大远离 α .

4) 随 n 增大, 虽然三个检验犯第一类错误的概率都在逼近 α , 但 BSEL 的逼近速度明显快于其余两个检验, 表明 BSEL 的精度更高.

5) 在表 2 中, 三个检验的经验功效都随 n 增大而增大, 随 α 增大而增大, 随 $P(Y > C)$ 增大而减小. 比较而言, BSEL 的功效略高于其余两个检验.

综上所述, 与正态渐近法和文献 [7] 的 EL 法相比, SEL 法经过 Bartlett 纠偏后有更高的精度.

文章最后介绍一个实例, 数据来自 121 位肺癌病人的临床记录 (见文献 [4] 的表 1). 医生为病人提供 A 和 B 两种治疗方法, 每个病人接受一种疗法. T_i 是病人接受治疗后的存活时间 (单位: 天), δ_i 是右删失示性变量, $X_{2,i}$ 取 0 和 1 两个值, $X_{2,i}$ 取 0 表示病人接受疗法 A, 否则接受疗法 B, $X_{3,i}$ 是病人接受治疗时的年龄. 取 $Y_i = \log_{10}(T_i)$, 为了研究两种疗法对病人存活时间的影响, Ying 等^[4] 建立了中位数回归模型

$$Y_i = X_i^\top \beta + U_i, \quad i = 1, 2, \dots, 121,$$

表2 $\beta_0 = (1.2, 2.2)^\top$ 时 NA、QEL 和 BSEL 三个检验的经验功效

模型	$P(Y > C)$	n	$\alpha = 0.1$			$\alpha = 0.05$		
			NA	QEL	BSEL	NA	QEL	BSEL
A	15%	30	0.656	0.678	0.705	0.515	0.553	0.574
		60	0.914	0.906	0.929	0.850	0.841	0.871
		90	0.983	0.980	0.986	0.963	0.957	0.971
	25%	30	0.579	0.618	0.628	0.435	0.492	0.495
		60	0.868	0.849	0.885	0.780	0.763	0.804
		90	0.964	0.956	0.970	0.929	0.916	0.941
	40%	30	0.480	0.567	0.535	0.343	0.461	0.414
		60	0.764	0.737	0.776	0.652	0.623	0.657
		90	0.904	0.878	0.912	0.838	0.799	0.842
B	15%	30	0.659	0.684	0.698	0.561	0.598	0.607
		60	0.920	0.918	0.932	0.876	0.873	0.892
		90	0.985	0.984	0.988	0.975	0.974	0.979
	25%	30	0.599	0.635	0.640	0.491	0.539	0.537
		60	0.884	0.872	0.897	0.826	0.813	0.848
		90	0.971	0.966	0.977	0.953	0.945	0.958
	40%	30	0.529	0.570	0.567	0.421	0.481	0.470
		60	0.827	0.814	0.843	0.756	0.743	0.773
		90	0.945	0.935	0.952	0.913	0.900	0.923

其中 $X_i = (1, X_{2,i}, X_{3,i})^\top$. 按 Ying 等^[4] 的结论, β 的取值应该是 $\beta_0 = (3.028, -0.163, -0.004)^\top$, 用本文的方法检验 $H_0 : \beta = \beta_0$, 结论如下. 检验统计量 $l_{\text{BSEL}}(\beta_0) = 4.264 < \chi^2_{3,0.95}$, 检验的 p 值为 0.234, 所以检验不显著, 表明取 $\beta_0 = (3.028, -0.163, -0.004)^\top$ 是合理的.

致谢 审稿人提出的宝贵意见使文章的质量有较大提高, 作者对他们表示衷心感谢.

参 考 文 献

- [1] KOENKER R, BASSETT JR G. Regression quantiles [J]. *Econometrica*, 1978, **46**(1): 33–50.
- [2] OTSU T. Empirical likelihood for quantile regression [OL]. 2003 [2003-11-01]. <http://www.cirje.e.u-tokyo.ac.jp/research/workshops/stateng/statpaper2003/otsu.pdf>.
- [3] WHANG Y J. Smoothed empirical likelihood methods for quantile regression models [J]. *Econometric Theory*, 2006, **22**(2): 173–205.
- [4] YING Z, JUNG S H, WEI L J. Survival analysis with median regression models [J]. *J Amer Statist Assoc*, 1995, **90**(429): 178–184.

- [5] LENG C L, TONG X W. A quantile regression estimator for censored data [J]. *Bernoulli*, 2013, **19**(1): 344–361.
- [6] WANG H J, WANG L. Locally weighted censored quantile regression [J]. *J Amer Statist Assoc*, 2009, **104**(487): 1117–1128.
- [7] QIN G S, TSAO M. Empirical likelihood inference for median regression models for censored survival data [J]. *J Multivariate Anal*, 2003, **85**(2): 416–430.
- [8] ZHAO Y C, CHEN F M. Empirical likelihood inference for censored median regression model via nonparametric kernel estimation [J]. *J Multivariate Anal*, 2008, **99**(2): 215–231.
- [9] HE S Y, HUANG X. Central limit theorem of linear regression model under right censorship [J]. *Sci China Ser A*, 2003, **46**(5): 600–610.
- [10] HE S Y, LIANG W. Empirical likelihood for right censored data with covariables [J]. *Sci China Ser A*, 2014, **57**(6): 1275–1286.
- [11] GILL R. Large sample behaviour of the product-limit estimator on the whole line [J]. *Ann statist*, 1983, **11**(1): 49–58.
- [12] HE S Y, LIANG W, SHEN J S, et al. Empirical likelihood for right censored lifetime data [J]. *J Amer Statist Assoc*, 2016, **111**(514): 646–655.
- [13] OWEN A. Empirical likelihood ratio confidence regions [J]. *Ann statist*, 1990, **18**(1): 90–120.

Smoothed Empirical Likelihood Testing for Quantile Regression Models under Right Censorship

LI Zhonggui HE Shuyuan

(School of Mathematical Sciences, Capital Normal University, Beijing, 100048, China)

Abstract: This paper is focused on testing the parameters of the quantile regression models. For complete observation, it is shown in literature that the test statistics, based on empirical likelihood (EL) method and smoothed empirical likelihood (SEL) method, both converge weakly to the standard Chi-square distribution χ_M^2 under the null hypothesis. For right censored data, the statistics in literature, by the EL method, have a weighted Chi-square limiting distribution, but the weights are unknown. In this paper, we show that the statistics based on the EL method and the SEL method also converge weakly to χ_M^2 under the null hypothesis, so there is no need to estimate any weights. As its estimating function is smoothed, the SEL method can be Bartlett corrected. Numerical results show that the SEL method, via Bartlett correction, outperforms some recent methods.

Keywords: quantile regression; right censoring; smoothed empirical likelihood

2010 Mathematics Subject Classification: 62G10; 62N01