DOI:10.15923/j.cnki.cn22-1382/t.2017.2.04

分位数回归区间估计方法比较分析

袁晓惠， 鞠婷婷， 陈 晶

（长春工业大学 基础科学学院，吉林 长春 130012）

摘 要： 介绍了分位数回归模型参数的 3 类区间估计方法，分别为直接法、自助法、诱导光滑法，通过模拟比较他们在覆盖率与置信区间长度方面的表现。

关键词： 分位数回归；诱导光滑；自助法

中图分类号： O 212.1 **文献标志码：** A **文章编号：** 1674-1374(2017)02-0122-05

Comparison analysis of quantile regression interval estimation

YUAN Xiaohui， JU Tingting， CHEN Jing

(School of Basic Science, Changchun University of Technology, Changchun 130012, China)

Abstract: Three confidence interval estimationmethod for quantile regression model are introduced, which are direct method, bootstrap and induced smoothing method. The performance and the features of these methods for the confidence interval estimation are compared by simulation.

Key words: quantile regression; induced smoothing; bootstrap.

0 引 言

线性回归模型是统计学中最经典的模型。传统的线性回归研究因变量的条件均值随自变量的变化趋势。此类模型对随机误差的分布有较强的假定。Koenker 和 Bassett[1] 于 1978 年提出线性分位数回归，考虑因变量的条件分位数对自变量的影响，可以根据不同的条件分位数更全面地认识因变量的条件分布。与传统的线性回归相比，分位数回归模型使用范围更广，估计效果更准确。随着计算机技术的发展，分位数回归模型在经济、金融、生物医学、数据挖掘、环境科学等方面得到广泛应用[2-3]。

分位数回归模型的目标函数是非光滑的，其参数的估计存在一定的困难。针对分位数回归模型参数的区间估计问题，比较流行的有 4 类方法：

1）直接法。根据参数估计的渐近正态性，运用样本信息直接估计渐近方差中的未知量并构造置信区间。

2）秩得分法。根据秩检验统计量的反演运算构造置信区间。此方法易于理解，计算简单，但是计算速度较慢，尤其在处理大型多维数据时，此算法运行缓慢。

3）自助法[4]。基于重复抽样技术构造回归参数的置信区间。

4）诱导光滑法[5-6]。此方法给参数添加一个

收稿日期： 2016-11-21

基金项目： 吉林省科技厅青年科研基金资助项目(20150520055JH)

作者简介： 袁晓惠(1983－)，女，汉族，四川广元人，长春工业大学讲师，博士，主要从事缺失数据方向研究，E-mail:yuanxh@ccut.edu.cn.

正态随机扰动，对不光滑的估计函数在这个扰动下求期望，得到一个新的光滑估计函数，然后基于这个新的光滑估计函数得到回归参数的估计。

经过迭代，诱导光滑方法可以同时得到参数的点估计及其协方差估计，进而得到回归参数的区间估计。由于此方法不需要额外确定调谐参数（如核估计的窗宽），此估计方法得到广泛应用[7-9]。

文中主要介绍直接法、自助法、诱导光滑法构造分位数回归模型区间估计的算法步骤，并通过模拟比较这 3 种方法构造的置信区间的覆盖率和平均置信区间长度。

1　分位数回归模型及其区间估计

假定得到观测数据为$(\boldsymbol{x}_i, y_i)(1\leqslant i\leqslant n)$，$y_i$ 是响应变量，$\boldsymbol{x}_i$ 是 p 维协向量，分位数回归模型如下：

$$y_i=\boldsymbol{x}_i^{\mathrm{T}}\beta+\varepsilon_i$$

ε_i 是连续的相互独立的但可能不是同分布随机变量，满足 $p(\varepsilon_i\leqslant 0)=\tau$。假设 $f_i(\cdot)$ 是 ε_i 的密度函数且 $f_i(\cdot)>0$，回归参数 β 的点估计 $\hat{\beta}$ 可以通过最小化目标函数 $L(\beta)=n^{-1}\sum_{i=1}^{n}\rho(y_i-\boldsymbol{x}_i^{\mathrm{T}}\beta)$ 得到，其中 $\rho_\tau(u)=u(I(u<0)-\tau)$。运用 R 软件编程时，此估计利用 R 程序包 quantreg 中的函数 $rq()$ 进行计算。

Koenker[10] 的专著中给出了 $\hat{\beta}$ 的渐近正态性质。

$$\sqrt{n}\,(\hat{\beta}-\beta)\xrightarrow{d}N(0,\Sigma)$$

其中

$$\Sigma=\lim_{n\to\infty}A_n^{-1}B_nA_n^{-1}$$

$$B_n=\tau(1-\tau)n^{-1}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}$$

$$A_n=n^{-1}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}f_i(0)$$

1.1　直接法

直接法基于估计的渐近正态分布来构造参数的渐近置信区间。首先，利用样本数据估计出 $\hat{\Sigma}=\hat{A}_n^{-1}B_n\hat{A}_n^{-1}$，其中 $\hat{A}_n=n^{-1}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\hat{f}_i(0)$。令 $\hat{\sigma}=\sqrt{\mathrm{diag}(\hat{\Sigma})}$，那么可得 β 的置信水平为 $1-\alpha$ 的渐近区间估计为：

$$(\hat{\beta}-n^{-\frac{1}{2}}Z_{1-\frac{1}{2}\alpha}\hat{\sigma},\hat{\beta}+n^{-\frac{1}{2}}Z_{1-\frac{1}{2}\alpha}\hat{\sigma})$$

渐近协方差估计 $\hat{\Sigma}$ 涉及密度函数 $\hat{f}_i(0)$，文中采用 Koenker 和 Machado[11] 提出的方法估计为：

$$\hat{f}_i(0)=\max\left(0,\frac{2h_n}{\boldsymbol{x}_i^{\mathrm{T}}(\hat{\beta}_{\tau+h_n}-\hat{\beta}_{\tau-h_n})-10^{-4}}\right)$$

这里 h_n 是窗宽，当 $n\to\infty$，$h_n\to 0$，根据 Hall 和 Sheather[12] 方法选取

$$h_n=1.57n^{-\frac{1}{3}}\left(\frac{1.5\varphi^2\{\Phi^{-1}(\tau)\}}{2\{\Phi^{-1}(\tau)\}^2+1}\right)^{\frac{2}{3}}$$

1.2　自助法

自助法是 Efron[4] 于 1979 年提出的一种再抽样统计方法，通过不断地从原始数据集中有放回抽取新样本，组成新的数据集。渐近理论保证了基于新的数据集计算的估计量与基于原始数据集的估计量有相同的渐近分布。此方法适用于那些难以用常规方法（如极大似然法、矩估计法等）导出参数的区间估计、假设检验等问题。

文中主要介绍如下两种自助法。

1.2.1　成对数据自助法

Arcones 和 Gine[13] 提出成对自助法来构造 M-估计的置信区间。成对数据自助法的步骤如下：

1）令 $b=1$；

2）从数据 $\{(\boldsymbol{x}_i,y_i)\}_{i=1}^{n}$ 有放回的抽取 1 组成对样本 $\{(\boldsymbol{x}_i^b,y_i^b)\}_{i=1}^{n}$；

3）求解目标函数得到估计 $\hat{\beta}^b=\underset{\beta}{\mathrm{argmin}}n^{-1}\sum_{i=1}^{n}\rho(y_i^b-(\boldsymbol{x}_i^b)^{\mathrm{T}}\beta)$，令 $b=b+1$；

4）重复步骤 2）和 3），直到产生 B 个 β 的估计。

在一些正则条件下可知，$\hat{\beta}-\beta$ 与 $\hat{\beta}^{(b)}-\beta$ 有相同的渐近分布。

1.2.2　加权自助法

Jin[14] 等 2001 年提出一种通过扰动目标函数的重抽样方法。Tang 和 Leng[15] 运用此方法构造纵向数据分位数回归参数的置信区间。此方法应用于分位数回归区间估计的步骤如下：

1）令 $b=1$；

2）从参数为 1 的指数分布中产生随机数 $V_i\sim\exp(1)$，$i=1,2,\cdots,n$；

3）求解加权目标函数得到估计 $\hat{\beta}^{(b)}=$

$\text{argmin}\sum_{i=1}^{n}V_i\rho_\tau(y_i-x_i^{\mathrm{T}}\beta)$，令 $b=b+1$；

4)重复步骤2)和3)，直到产生 B 个 β 的估计。

Jin[14]等证明了 $\hat{\beta}^{(b)}-\beta$ 与 $\hat{\beta}-\beta$ 有相同的渐近分布。

对于上述两种自助法得到的参数估计，如果 $\hat{\beta}-\beta$ 与 $\hat{\beta}^{(b)}-\beta$ 有相同的渐近分布，则可以通过 $\hat{\beta}^{(b)}(b=1,2,\cdots,B)$ 得到 $\hat{\beta}$ 的渐近方差估计 $\hat{\Sigma}_B=\frac{n}{B}\sum_{b=1}^{B}(\hat{\beta}^{(b)}-\bar{\beta})(\hat{\beta}^{(b)}-\bar{\beta})^{\mathrm{T}}$，其中 $\bar{\beta}=\frac{1}{B}\sum_{b=1}^{B}\hat{\beta}^{(b)}$。令 $\hat{\sigma}_B=\sqrt{\text{diag}(\hat{\Sigma}_B)}$，那么 β 的置信水平为 $1-\alpha$ 的渐近区间估计为：

$$(\hat{\beta}-n^{-\frac{1}{2}}Z_{1-\frac{1}{2}\alpha}\hat{\sigma}_B,\hat{\beta}+n^{-\frac{1}{2}}Z_{1-\frac{1}{2}\alpha}\hat{\sigma}_B)$$

1.3 诱导光滑法

诱导光滑法最初是由 Brown 和 Wang[5]于2005年提出，用于估计秩估计的渐近方差。Wang[6]等将之用于构造分位数回归区间估计。由于此光滑方法不像核估计等需要额外估计窗宽，使之得到许多统计学家的青睐。

诱导光滑算法步骤如下：

1)设定 Γ 的初始值：$\Gamma^{(0)}=n^{-1}I_p$；

2)在第 j 步迭代中($j=1,2,\cdots$)，令 $\sigma_i^2=x_i^{\mathrm{T}}\Gamma^{(j-1)}x_i$，$b_i=(y_i-x_i^{\mathrm{T}}\beta^{(j-1)})/\sigma_i$，$\Phi(u)$ 为标准正态分布的分布函数。记 $\tilde{U}(\beta)=\frac{1}{n}\sum_{i=1}^{n}x_i\{\Phi(b_i)-1+\tau\}$，通过求解 $\tilde{U}(\beta)=0$ 得到 $\beta^{(j)}$；

3)通过 $\beta^{(j)}$ 和 $\Gamma^{(j-1)}$ 来计算出 $\tilde{A}_n=n^{-1}\sum_{i=1}^{n}\frac{\varphi(b_i)}{\sigma_i}x_ix_i^{\mathrm{T}}$，继而 $\Gamma^{(j)}=n^{-1}\tilde{A}_n^{-1}B_n(\tilde{A}_n^{-1})^{\mathrm{T}}$；

4)令 $j=j+1$，重复步骤2)和3)的迭代，直至 $\max|\Gamma^{(j+1)}-\Gamma^{(j)}|<10^{-4}$ 停止迭代。

令 $\hat{\sigma}_I=\sqrt{\text{diag}(\Gamma^{(j)})}$，那么可得 β 的置信水平为 $1-\alpha$ 的渐近区间估计为：

$$(\hat{\beta}-Z_{1-1/2\alpha}\hat{\sigma}_I,\hat{\beta}+Z_{1-1/2\alpha}\hat{\sigma}_I)$$

2 模拟比较

通过模拟研究从置信区间长度和覆盖率两个角度来比较上述3类方法在构造分位数回归参数的置信区间上的表现。从如下分位数回归模型产生数据$(x_i,y_i)(1\leqslant i\leqslant n)$：

$$y_i=\beta_0+x_i\beta_1+\sigma(x_i)(\varepsilon_i-Q_\tau(\varepsilon_i))$$
$$i=1,2,\cdots,n$$

其中，$\beta=(\beta_0,\beta_1)^{\mathrm{T}}=(1,1)^{\mathrm{T}}$，$x_i\sim N(0,1)$，$\varepsilon_i\sim N(0,1)$，令 $\sigma(x_i)=1$ 及 $\sigma(x_i)=\sqrt{1+|x_i|}$ 分别表示误差同方差和异方差性。在模拟过程中，$\tau=0.5$，重复试验1 000次。

β 的置信水平为95%的置信区间的平均长度和覆盖率($\sigma(x_i)=1$)见表1。

表1 β 的置信水平为95%的置信区间的平均长度和覆盖率($\sigma(x_i)=1$)

n	估计方法	置信区间平均长度		覆盖率	
		β_0	β_1	β_0	β_1
20	直接法	1.205 2	1.194 1	0.946	0.924
	成对数据自助法	1.236 3	1.391 2	0.953	0.962
	加权自助法	1.225 6	1.313 9	0.951	0.954
	诱导光滑法	1.136 8	1.229 8	0.930	0.926
50	直接法	0.731 6	0.680 1	0.948	0.905
	成对数据自助法	0.744 7	0.790 7	0.950	0.955
	加权自助法	0.745 4	0.772 0	0.951	0.948
	诱导光滑法	0.693 5	0.713 8	0.930	0.923
100	直接法	0.508 6	0.483 0	0.945	0.922
	成对数据自助法	0.518 4	0.537 0	0.942	0.949
	加权自助法	0.517 0	0.528 4	0.948	0.947
	诱导光滑法	0.487 5	0.492 2	0.936	0.935

从表 1 可以看出，直接法和诱导光滑法的置信区间平均长度比自助法估计的置信区间长度短。当样本量为 20 时，直接法和诱导光滑法的覆盖率较低，但是当样本量增至 50 和 100 时，他们的覆盖率都有所增加。

β 的置信水平为 95%的置信区间的平均长度和覆盖率($\sigma(x_i)=\sqrt{1+|x_i|}$)见表 2。

表 2　β 的置信水平为 95%的置信区间的平均长度和覆盖率($\sigma(x_i)=\sqrt{1+|x_i|}$)

n	估计方法	置信区间平均长度		覆盖率	
		β_0	β_1	β_0	β_1
20	直接法	1.534 8	1.514 8	0.944	0.853
	成对数据自助法	1.582 0	1.988 3	0.969	0.942
	加权自助法	1.553 6	1.880 3	0.959	0.932
	诱导光滑法	1.439 6	1.762 1	0.941	0.909
50	直接法	0.943 7	0.895 8	0.932	0.842
	成对数据自助法	0.961 4	1.195 0	0.949	0.935
	加权自助法	0.960 8	1.167 6	0.937	0.927
	诱导光滑法	0.890 5	1.079 7	0.923	0.910
100	直接法	0.657 7	0.621 4	0.942	0.862
	成对数据自助法	0.667 5	0.820 7	0.944	0.939
	加权自助法	0.668 1	0.812 7	0.943	0.939
	诱导光滑法	0.629 9	0.754 5	0.934	0.927

表 2 中，直接法的覆盖率较低。随着样本量增大，覆盖率也没有增加，说明直接法需要误差独立同分布的假定。当误差不是独立同分布时，构造的置信区间不是很好，而自助法和诱导光滑法的覆盖率都能接近 95%。虽然诱导光滑法的平均置信区间长度相比于自助法要短，但是当样本量较小时，诱导光滑法的覆盖率偏低。自助法中成对数据自助法的平均置信区间长度相对长一些，在覆盖率接近 95%时，加权自助法的平均置信区间长度相对短一些。加权自助法在小样本时表现较出色。

3　结　语

分别介绍了 3 类区间估计方法的算法，并通过模拟比较他们在覆盖率与置信区间长度方面的表现。从模拟结果可以看出，在直接法中，由于用核估计方法来估计渐近方差中未知的密度函数，依赖于误差独立同分布的假定。如果误差分布不是独立同分布时，此估计效果不是很理想。重复抽样法计算估计的算法虽然需要上百次的重新计算估计，计算量比较大，但是覆盖率较好。诱导光滑法计算方法简单，其估计的置信区间长度最小，但是在小样本时覆盖率较低。建议如果数据样本量比较小时，考虑用加权自助法估计参数的置信区间，当样本量较大时，用诱导光滑法构造参数的置信区间。

参考文献：

[1] Koenker R, Bassett G. The asymptotic distribution of the least absolute error estimator[J]. Journal of the American Statistical Association, 1978, 73: 618-622.

[2] 王纯杰，董小刚，陈嘉，等.基于分位数回归的长春市职工工资水平的分析[J].长春工业大学学报：自然科学版，2010，31(4)：367-373.

[3] 何大强，张海燕.吉林省农村居民消费水平分析[J].长春工业大学学报：自然科学版，2013，34(4)：452-456.

[4] Efron B. Bootstrap methods: another look at the Jackknife [J]. Annals of Statistics, 1979, 7(1): 1-26.

[5] Brown B M, Wang Y G. Standard errors and covariance matrices for smoothed rank estimators [J]. Bi-

ometrika,2005,92(1):149-158.

[6] Wang Y, Shao Q, Zhu M, et al. Quantile regression without the curse of unsmoothness [J]. Computational Statistics & Data Analysis, 2009, 53(10):3696-3705.

[7] Pang L, Lu W, Wang H. Variance estimation in censored quantile regression via induced smoothing [J]. Computational Statistics and Data Analysis, 2012,56(4):785-796.

[8] Leng C, Zhang W. Smoothing combined estimating equations in quantile regression for longitudinal data [J]. Statistics and Computing, 2014, 24(1): 123-136.

[9] Lu X, Fan Z. Weighted quantile regression for longitudinal data [J]. Computational Statistics, 2015, 30(2):569-592.

[10] Koenker R. Quantile regression [M]. Cambridge: Cambridge University Press,2005.

[11] Koenker R, Machado J A F. Goodness of fit and related inference processes for quantile regression [J]. Journal of the American Statistical Association,1999,94(448):1296-1310.

[12] Hall P, Sheather S J. On the distribution of a studentized quantile [J]. J. R. Stat. Soc. B.,1988,50: 381-391.

[13] Arcones M, Gine E. On the bootstrap of M-estimators and other statistical functionals[C]// In R. LePage & L. Billard (eds.), Exploring the Limits of Boorstrap,1992:13-47.

[14] Jin Z, Ying Z, Wei L J. A simple resampling method by perturbing the minimand[J]. Biometrika,2001,88(2):381-390.

[15] Tang C Y, Leng C. Empirical likelihood and quantile regression in longitudinal data analysis[J]. Biometrika,2011,98(4):1001-1006.