Wild Bootstrap Variance Estimation for Quantile Regression Estimator

Russel Sui, Steven Xu, Zun Yin, Miao Yu
Department of Statistics
North Carolina State University

Contents

List of Figures

List of Tables

Wild Bootstrap Variance Estimation for Quantile Regression Estimator

***Abstract*—This report examines the performance of wild bootstrap quantile regression estimator, which provides a variance estimation for the usual quantile regression estimator. We introduce the basic concepts of quantile regression and wild bootstrap to facilitate understanding. A simulation study is conducted to compare wild bootstrap with other variance estimation methods in a finite sample setting.**

I. Introduction: Quantile Regression

Quantile regression (QR) is usually utilized as an alternative to traditional mean regression when research of interest lies on non-central part of the conditional distribution of the response. It is particularly useful if the data express clear heteroscedasticity, when inference on mean might not provide a comprehensive depiction of the relationship between covariates and response. Moreover, its robustness to outliers and heavy-tailness of the error distribution provides more faithful estimates in real application over mean regression. While acknowledging the fact that countless modern parameteric and nonparametric quantile regression models have been proposed in the literature, we restrict our discussion in this report to the simplest case, linear quantile regression (LQR) [4].

Consider an ordinary least square regression (OLS) setting:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \sigma \boldsymbol{e} \ , \boldsymbol{e} \overset{iid}{\sim} \mathcal{N}(0,1) \tag{1}$$

where $\boldsymbol{y}$ is an n dimensional response variable, $\boldsymbol{X}$ is an $n \times p$ covariate matrix and $\boldsymbol{\beta}$ is a p dimensional parameter. Analogously, LQR assumes that the covariates impact the τ-th conditional quantile of $\mathbf{y}$ linearly through $\beta(\tau)$, i.e.

$$q_\tau(\mathbf{y}|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}(\tau) \tag{2}$$

It is straightforward to see that $\boldsymbol{\beta}(\tau)$ inherits the rate of change interpretation from OLS, and with variances for QR estimators one could perform hypothesis tests or construct confidence intervals to detect significant covariate effect on the τ-th quantile of response. As mentioned above, the normality assumption on the error distribution in (1) is usually skeptical, and to demonstrate the superiority of (2) we relax (1) to be

$$y_i = \mathbf{x}_i^T \beta + \sigma(\mathbf{x}_i) e_i \ , e_i \overset{ind}{\sim} F_i \tag{3}$$

for the ith element of $\boldsymbol{y}$, $i = 1, \ldots, n$, where variance is allowed to depend on $\mathbf{x}$. Notice that in order for β in (3) to be identifiable we must have $q_\tau(e_i|\mathbf{x}_i) = 0$.

II. Variance Estimation of QR Estimators

Estimating $\boldsymbol{\beta}(\tau)$ is not a difficult task in general. In theory, finding this estimator subjects to minimizing the sample version of pinball loss, i.e.

$$\hat{\boldsymbol{\beta}}_\tau = \underset{\beta \in \mathbb{R}^p}{\arg\min} \sum_{i=1}^{n} \rho_\tau(y_i - \boldsymbol{x_i^T\beta}) \tag{4}$$

where

$$\rho_\tau(\mu) = \mu(\tau - I(\mu < 0))$$

which is usually handled by linear programming. Direct estimation of $\text{Var}\big[\hat{\boldsymbol{\beta}}(\tau)\big]$, however, can be troublesome. Although large sample asymptotic theory exists for QR estimators, in practice it is shown to be unsatisfactory since only finite sample is available. Moreover, it would be contradictory if the estimation methods cannot accommodate the absence of informative assumption on error distribution. Nonparametric bootstrap is a class of methods that can handle the above two conditions. To date, various nonparametric bootstrapping methods have been proposed to solve linear regression problems, such as paired bootstrap [1] [3], residual bootstrap, random weight bootstrap [10] [7]. However, all of them have been shown to provide either inflated or deflated estimate of $\text{Var}\big[\hat{\boldsymbol{\beta}}(\tau)\big]$ when involving heteroscedasticity and clustering of data.

Recently, Feng et. al [2] successfully adapted wild bootstrap into QR estimation by modifying the original weight distribution [9] [5]. They have also shown that their estimator provides reliable estimate of $\text{Var}\big[\hat{\boldsymbol{\beta}}(\tau)\big]$ even when the sample size is small and heteroscedasticity is allowed for a clustered data. This report serves as a detailed illustration of the finite sample performance of wild bootstrap QR estimator when heteroscedasticity is present, by performing extensive simulation with comparison to competitive methods. Readers who are interested in the theoretical details are referred to the original paper.

III. Wild Bootstrap

Wild Bootstrap is a technique to generate new response vectors by assigning random weights to the absolute residuals obtained from the original fit. The new responses are then re-fitted on the original covariates to obtain a new QR estimator $\hat{\boldsymbol{\beta}}^*(\tau)$. [2] proved that under certain conditions, $\hat{\boldsymbol{\beta}}^*(\tau)$ converges in probability to $\hat{\boldsymbol{\beta}}(\tau)$, the quantile regression estimator conditioned on the observed sample, as the sample size $n \longrightarrow \infty$. Furthermore, the asymptotic variance of $\hat{\boldsymbol{\beta}}^*(\tau)$ is in fact equivalent to that of $\hat{\boldsymbol{\beta}}$. As a result, one could repeat the above process as many times as possible and then collect $\hat{\boldsymbol{\beta}}^*(\tau)$ from each iteration. The resulting sample variances would be a reliable estimate of $\text{Var}\big[\hat{\boldsymbol{\beta}}(\tau)\big]$.

An algorithm to describe the procedure is presented below:

Algorithm 1 Wild Bootstrap

```
1: Fit the linear model $y_i = \mathbf{x}_i^T\beta + \sigma(\mathbf{x}_i)e_i$ to obtain
   the quantile regression estimators $\hat{\boldsymbol{\beta}}(\tau)$. Denote the $i$th
   residual as $\hat{e}_i$;
2: Generate random weights $w_i$ for every $i$ by some weight
   distribution that satisfies the conditions of Theorem 1 in
   [2]. Compute new errors $e_i^* = w_i|\hat{e}_i|$;
3: Calculate the bootstrapped response $y_i^* = \mathbf{x}_i^T\beta + \sigma(\mathbf{x}_i)e_i^*$;
4: Refit the linear model with $\boldsymbol{y}^*$ and $\boldsymbol{X}$ to obtain QR
   estimators $\hat{\boldsymbol{\beta}}^*(\tau)$;
5: Repeat step 2-4 for B times and obtain the sample variances
   of B copies of $\hat{\boldsymbol{\beta}}^*(\tau)$ as an estimate of $\text{Var}\big[\hat{\boldsymbol{\beta}}(\tau)\big]$.
```

As mentioned in section II, there exist other methods for variance estimation of quantile regression estimators. However, it is believed that wild bootstrap outperforms its competitors when the data is scarce, heteroscedastic and clustering present. Therefore, for comparison purpose, we also implement several other methods in the simulation study. Brief introductions to those methods are listed below:

- Paired Bootstrap: Resample $(y_i, \boldsymbol{x}_i^T)$ pairs from the original data set with repetition to obtain new data sets. Fit the resampled data sets to obtain bootstrapped estimators and calculate sample variances;
- Random Weight Bootstrap: Similar as wild bootstrap by generating random weights from certain distributions, but apply weights to modify loss functions;
- Rank Score Method: Utilize inversion rank score test to approximate variance estimations without bootstrapping;
- Normal Approximation: Construct quasi-Bayesian credibility intervals based on normal approximation ”posterior”, as computed using the Powell kernel estimate.

IV. Simulation Study

In this section, we present the designs and results of our simulation study. We also analyze the indication of those results.

A. Median Regression Setting

We generate data from the following distribution:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \frac{\{2 + [1 + (x_{1i} - 8)^2 + x_{2i}]/10\}}{\sqrt{3}} e_i \tag{5}$$

where the true values for $\beta_0, \beta_1, \beta_2$ are all 1, errors are generated from independent t distribution with degrees of freedom 3, $\boldsymbol{x_1}$ are generated from independent standard log-normal distribution, and $\boldsymbol{x_2}$ are 1 for the first 80% of the entries, 0 for the rest. The symmetry of t_3 ensures that $q_{0.5}(e_i) = 0$, therefore preventing non-identifiability under a median regression.

Quantile regression was carried out using built-in function `rq` within the R package `quantreg` with various sample sizes $n = 100, 200, 400, 800, 1600, 3200, 5000$. The purpose for the changing sample size is to find a threshold sample size, below which wild bootstrap method provides the best variance estimations.

We specify certain weight distributions for wild bootstrap and random weight bootstrap. For wild bootstrap, we follow the proposal that satisfies the Condition 3,4,5 in [2] and define our first weight distribution as:

$$g(w) = \begin{cases} -w, & \text{if } -2\tau - \frac{1}{4} \leq w \leq -2\tau + \frac{1}{4} \\ w, & \text{if } 2(1-\tau) - \frac{1}{4} \leq w \leq 2(1-\tau) + \frac{1}{4}; \end{cases} \tag{6}$$

with the pre-condition that $\frac{1}{8} \leq \tau \leq \frac{7}{8}$. The second distribution outputs point-mass weights at $2(1-\tau)$ and -2τ with probabilities τ and $1-\tau$. Both distributions are proven to be valid weight distributions in the original paper. Notice that when the median is of interest this becomes the Rademacher distribution. For random weight bootstrap, the weight functions is exponential distribution with mean 1.

To account for residuals being less dispersed than reality in wild bootstrap method, we utilize the Bahadur representation of the original residual estimate and modify it by $\hat{e}_i = \hat{e}_i + \{\hat{f}(0)\}^{-1}h_i\psi_{0.5}(\hat{e}_i)$ prior to weighting, where $\hat{f}$ is the kernel density[1] estimated from the residuals, $h_i = x_i^T(\sum_k x_k x_k^T)^{-1}x_i$, and $\psi_{0.5}(\mu) = 0.5 - I(\mu < 0)$ is the score function of the loss function we try to minimize.

Before carrying out the large-scale simulation, we perform regular quantile regression for 5000 simulated data sets and obtain the sample standard errors of quantile regression estimators $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$. The values we obtain from this procedure serve as benchmark standard errors. Afterwards, we run each proposed method under sample size $n = 100$ for 100 Monte Carlo iterations, and record the sample standard errors for each iteration. We take the ratio between each sample standard errors and the benchmark values in order to evaluate the accuracy of variance estimations of quantile regression estimators.

Finally, we perform formal simulations under all methods introduced with all 7 different sample sizes. For each bootstrap-based method, we obtain 999 bootstrap samples within each iteration and construct a 90% confidence interval using the percentiles and standard errors calculated from bootstrap estimators. We then record the average interval length and coverage probability, together with the standard error of interval length. To control the standard errors of coverage probability within 0.01, 2500 MC iterations were performed under each simulation setting.

For non-bootstrap-based method, we similarly perform 2500 iterations and obtain interval lengths and coverage probabilities directly from R Outputs. In addition to the quantile methods, we also included the performance of Ordinary Least Squares,

[1] There are various and different packages in R that perform KDE, a discussion on what methods to use is included in Section V.

since median and mean are identical under symmetric error distribution.

B. Tail Regression Setting

As mentioned above, when strong heteroscedasticity is observed in the data. regression at the tail might be of interest. Therefore we also perform simulations for a second quantile candidate $\tau_2 = 0.9$. Data are again generated from (5), but the errors are taken as skewed normal (7) instead, i.e. $e_i \sim SN(\xi = 0, \omega = 1, \alpha = -3.077864)$. This specific value of the shape parameter α can be found by numerically solving the equation $Q_{0.9}[e_i|x_i] = 0$ using any statistical software.

$$SN(x|\xi,\omega,\alpha) = \frac{2}{\omega\sqrt{2\pi}} e^{-\frac{(x-\xi)^2}{2\omega^2}} \int_{-\infty}^{\alpha(\frac{x-\xi}{\omega})} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt \quad (7)$$

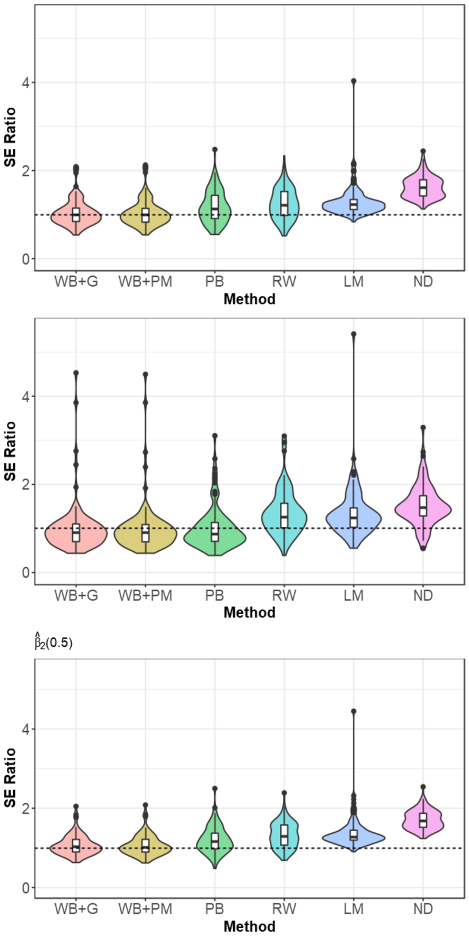

Fig. 1: Ratio Comparison of $\text{SE}\left[\hat{\beta}(0.5)\right]$

In this setting, we only perform small-scale simulations to obtain ratios between sample standard errors and benchmark errors.

C. Results

Violin plots for the ratios of standard errors for 0.5 and 0.9 QR estimators can be found in Figure 1 and 2 respectively. From these two plots we can clearly see that wild bootstrap method outperforms the others in both situations. For each of the parameters, the distribution of standard error ratio resulting from wild bootstrapping is centered around 1 which implies consistent estimation, out-performing other methods. It is noteworthy that the two weight functions in wild bootstrap do not make observable differences in terms of standard error estimation.

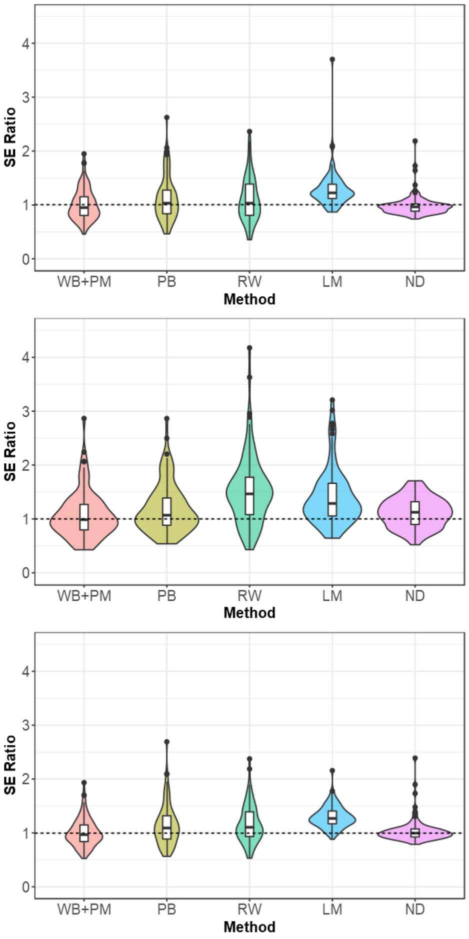

Fig. 2: Ratio Comparison of $\text{SE}\left[\hat{\beta}(0.9)\right]$

Since bootstrap methods are usually advantageous when the amount of observations is limited, coverage probability and length of bootstrap confidence interval is of highest interest in small-sample cases. Table I compares the nominal coverage probabilities and length of 90% naive percentile confidence intervals from different methods for the 0.5 QR estimator $\beta_1(0.5)$. The superiority of wild bootstrap QR estimator is obvious in small-sample cases. In fact, wild bootstrap method can perfectly maintain the nominal coverage from small-sample to large-sample simulation studies. Its demonstration

TABLE I: 90% Naive Percentile Confidence Interval for β_1

	n=100			200		
Methods	Coverage	Length	SE	Coverage	Length	SE
WB+G	89.0	0.78	0.01	90.4	0.52	0.00
WB+PM	88.4	0.76	0.01	90.0	0.52	0.00
PB	92.2	0.88	0.01	92.4	0.58	0.01
RW	88.3	1.20	0.01	89.9	0.79	0.01
ND	98.0	1.44	0.03	98.6	0.88	0.01
RK	86.4	0.79	0.01	87.0	0.53	0.01
LM	90.4	1.17	0.01	85.5	0.79	0.00

Notes: The coverage probabilities displayed are in percentage (%).

with varying sample sizes, as well as details of confidence intervals for $\beta_0(0.5)$ and $\beta_2(0.5)$, can be seen in Figure 3, 4, 5, Tables II and III respectively in Appendix B.

V. Conclusions and Future work

In this report we showed through extensive simulation study that wild bootstrap estimators can accurately estimate the variance of a single QR estimator for heteroscedastic and clustered data in small-sample cases. In application however, researchers might be interested in simultaneously estimating multiple quantiles to achieve a full description of the response distribution. Therefore a possible extension might be to study whether the wild bootstrap method can effectively estimate the covariance matrix of the QR estimators, since QR estimators for different quantiles of the same data are correlated.

In Section IV the method to estimate $\hat{f}(0)$ was left unspecified. [2] suggests `akj` (univariate adaptive kernel density estimation) [8] in the package `quantreg`, while we used `locfit` (local regression and likelihood model) [6] from its eponymous package. We found out that KDE methods that succeed in retaining the shape of the empirical distribution has a stronger finite-sample correlation effect than those succeed in recovering the shape of the generating distribution. Therefore in small-sample simulation study it might be appropriate to explore different combinations of KDE methods and tuning to compare the performance. For real data application, a clear guideline remains absent and future study is needed.

VI. Acknowledgement

The authors would like to thank Dr. Ana-maria Staicu for her insightful suggestions.

References

[1] Efron, B., and Tibshirani, R. J. *An introduction to the bootstrap*. CRC press, 1994.

[2] Feng, X., He, X., and Hu, J. Wild bootstrap for quantile regression. *Biometrika 98*, 4 (2011), 995–999.

[3] Knight, K. Asymptotics for l1-estimators of regression parameters under heteroscedasticity. *Canadian Journal of Statistics 27*, 3 (1999), 497–507.

[4] Koenker, R., and Bassett Jr, G. Regression quantiles. *Econometrica: journal of the Econometric Society* (1978), 33–50.

[5] Liu, R. Y., et al. Bootstrap procedures under some non-iid models. *The Annals of Statistics 16*, 4 (1988), 1696–1708.

[6] Loader, C. *Local Regression and Likelihood*. Springer, 1999.

[7] Rao, C. R., and Zhao, L. Approximation to the distribution of m-estimates in linear models by randomly weighted bootstrap. *Sankhyā: The Indian Journal of Statistics, Series A* (1992), 323–331.

[8] Silverman, B. W. *Density Estimation for Statistics and Data Analysis*. CRC press, 1986.

[9] Wu, C. F. J., et al. Jackknife, bootstrap and other resampling methods in regression analysis. *the Annals of Statistics 14*, 4 (1986), 1261–1295.

[10] ZHENG, Z. G. Random weighting method [j]. *ACTA Mathematicae applicatae sinica 2* (1987), 247–253.

Appendix A
Code

Please check *https://github.com/myu121/Bootstrap* for all simulation codes.

Appendix B
Figures & Tables

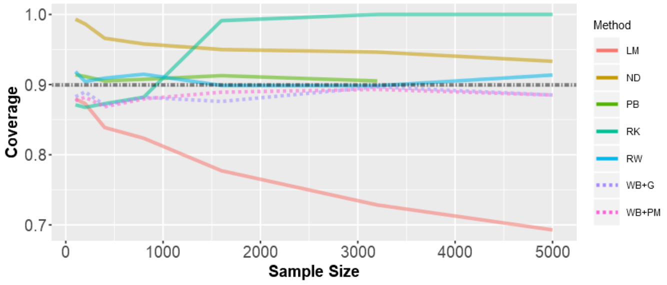

Fig. 3: 90% nominal coverage probabilities for β_0

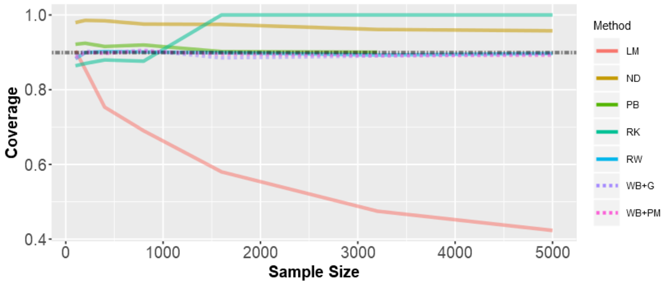

Fig. 4: 90% nominal coverage probabilities for β_1

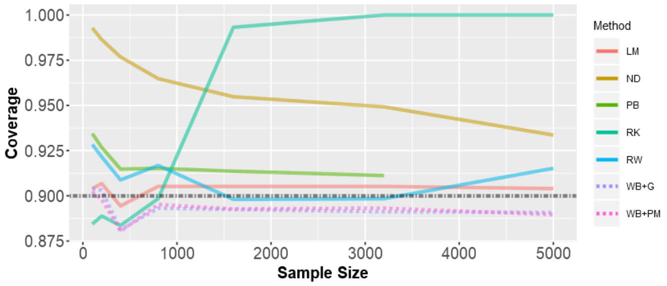

Fig. 5: 90% nominal coverage probabilities for β_2

TABLE II: 90% Naive Percentile Confidence Interval for β_0

	n=100			200		
Methods	Coverage	Length	SE	Coverage	Length	SE
WB+G	88.3	3.89	0.02	89.0	2.76	0.01
WB+PM	87.7	3.86	0.02	88.2	2.72	0.02
PB	91.4	4.35	0.03	91.2	2.97	0.02
RW	91.8	4.81	0.03	90.4	3.30	0.02
ND	99.3	6.73	0.04	98.6	4.20	0.01
RK	87.1	3.87	0.03	86.8	2.66	0.02
LM	88.0	5.24	0.03	87.3	3.75	0.02

Notes: The coverage probabilities displayed are in percentage (%).

TABLE III: 90% Naive Percentile Confidence Interval for β_2

	n=100			200		
Methods	Coverage	Length	SE	Coverage	Length	SE
WB+G	89.0	0.78	0.01	90.3	2.81	0.01
WB+PM	88.4	0.76	0.01	89.8	2.78	0.01
PB	92.2	0.88	0.01	92.7	3.00	0.01
RW	88.3	1.20	0.01	92.1	3.48	0.01
ND	98.0	1.44	0.03	98.6	4.28	0.01
RK	86.4	0.79	0.01	88.9	2.78	0.01
LM	90.4	1.17	0.01	90.7	3.94	0.02

Notes: The coverage probabilities displayed are in percentage (%).