Biometrika (2011), **98**, 4, *pp.* 995–999
doi: 10.1093/biomet/asr052
Advance Access publication 3 November 2011

Printed in Great Britain

Wild bootstrap for quantile regression

BY XINGDONG FENG

School of Statistics and Management, Shanghai University of Finance and Economics, 777 Guoding Road, Shanghai 200433, China

xd.feng@mail.shufe.edu.cn

XUMING HE

Department of Statistics, University of Michigan, 439 West Hall, 1085 South University Avenue, Ann Arbor, Michigan 48109, U.S.A.

xmhe@umich.edu

AND JIANHUA HU

Department of Biostatistics, University of Texas M.D. Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, Texas 77030, U.S.A.

JHu@mdanderson.org

SUMMARY

The existing theory of the wild bootstrap has focused on linear estimators. In this note, we broaden its validity by providing a class of weight distributions that is asymptotically valid for quantile regression estimators. As most weight distributions in the literature lead to biased variance estimates for nonlinear estimators of linear regression, we propose a modification of the wild bootstrap that admits a broader class of weight distributions for quantile regression. A simulation study on median regression is carried out to compare various bootstrap methods. With a simple finite-sample correction, the wild bootstrap is shown to account for general forms of heteroscedasticity in a regression model with fixed design points.

Some key words: Bahadur representation; Heteroscedastic error; Quantile regression; Wild bootstrap.

1. INTRODUCTION

The bootstrap is a well-established method of inference in regression models. Common bootstrap methods, such as the residual bootstrap and the paired bootstrap, are described in Efron & Tibshirani (1994), and their asymptotic properties can be found in Shao & Tu (1995) and Mammen (1991) among others. For bootstrapping M estimators, Lahiri (1992) considered a modified version of the residual bootstrap, Rao & Zhao (1992) proposed a resampling method using random weights on the loss functions, and Knight (1999) established the validity of the paired bootstrap method. To account for heteroscedasticity, Wu (1986) and Liu (1988) proposed the wild bootstrap, randomly weighting the residuals. Other researchers (Davidson & Flachaire, 2008; Mammen, 1993) have considered the properties of the wild bootstrap, but the existing theory has focused on linear estimators.

In this note, we consider the wild bootstrap for quantile regression estimators (Koenker & Bassett, 1978). We find that many classical choices of the weight distribution in the wild bootstrap are invalid for these estimators with nonlinear score functions. We suggest a simple modification of the wild bootstrap to suit asymmetric loss functions in quantile estimation, and identify a class of weight distributions under which our method is asymptotically valid. We also report a simulation study on median regression to demonstrate the relevance of our results in finite-sample problems.

2. THEORETICAL DEVELOPMENT

Consider a linear model

$$y_i = x_i^{\mathrm{T}}\beta_0 + e_i \quad (i = 1, \ldots, n), \tag{1}$$

where y_i is the ith observation, x_i is the ith nonstochastic design point in $\mathbb{R}^m$ and e_i is an independent error variable with probability density f_i. For identifiability, we assume that, for a quantile level $\tau \in (0, 1)$ of interest, the conditional τth quantile of e_i given x_i is zero. The quantile regression estimator of β_0 minimizes the objective function

$$\sum_{i=1}^{n} \rho_\tau(y_i - x_i^{\mathrm{T}}\beta),$$

where $\rho_\tau(u) = u\{\tau - I(u < 0)\}$ is the quantile loss function, and $\psi_\tau(u) = \tau - I(u < 0)$ is the score function. We use the wild bootstrap, as described in the following steps:

Step 1. Fit (1) to the data, and denote the estimate of the parameter vector by $\hat{\beta}$ and use $\hat{e}_i$ $(i = 1, \ldots, n)$ to represent the residuals.

Step 2. Generate the weights w_i from an appropriate distribution satisfying the conditions stated later, and let $e_i^* = w_i|\hat{e}_i|$.

Step 3. Calculate the bootstrapped sample as $y_i^* = x_i^{\mathrm{T}}\hat{\beta} + e_i^*$.

Step 4. Refit (1) to the bootstrapped sample and denote the bootstrap estimate by $\hat{\beta}^*$.

Step 5. Repeat Steps 2–4 B times, and estimate the variance of $\hat{\beta}$ by the sample variance of the B copies of $\hat{\beta}^*$.

The proposed wild bootstrap is based on Liu (1988), but the important differences lie in the choice of the weight distribution and the generation of the residuals in Step 2. The original wild bootstrap method uses the residuals $\hat{e}_i$ in Step 2, but we use the absolute residuals, because it is easier to find valid weight distributions for asymmetric loss function ρ_τ whenever $\tau \neq 0{\cdot}5$.

We shall use $|\cdot|$ to indicate the supremum norm of a vector. With the proposed bootstrap method, we have the following result for any fixed $\tau \in (0, 1)$. The proof can be found in the Supplementary Material.

THEOREM 1. *Assume model (1), with weights w_i independently drawn from a distribution with a bounded probability density function g. Let $\hat{\beta}$ be the quantile estimator of β_0. If Conditions* 1–6 *hold, then for almost all samples $\{(x_i^T, y_i) : i = 1, \ldots, n\}$, then the conditional distribution of $n^{1/2}(\hat{\beta}^* - \hat{\beta})$ given the data converges to $N\{0, \tau(1-\tau)M^{-1}QM^{-1}\}$ in distribution under resampling, as $n \to \infty$, where Q and M are specified below in Condition* 6.

Condition 1. The conditional densities f_i are bounded with $f_i(0) > 0$, and $f_i(y) - f_i(0) = O(|y|^{1/2})$ uniformly in i as $y \to 0$.

Condition 2. For $x_i \in \mathbb{R}^m$, $\sum_{i=1}^n |x_i|^3 = O(n)$ and $\max_{1 \leqslant i \leqslant n} |x_i| = O(n^{1/4})$.

Condition 3. For some positive constants c_1 and c_2, $\sup\{w \in \mathbb{G} : w \leqslant 0\} = -c_1$ and $\inf\{w \in \mathbb{G} : w \geqslant 0\} = c_2$, where $\mathbb{G}$ is the support of the weight distribution.

Condition 4. The weight distribution G satisfies $\int_0^{+\infty} w^{-1} dG(w) = -\int_{-\infty}^0 w^{-1} dG(w) = 1/2$, and $E_W(|w|) < \infty$, where the expectation E_W is taken under G.

Condition 5. The τth quantile of the weight w is zero, that is, $G(0) = \tau$.

Condition 6. The limits $Q = \lim_{n\to\infty} n^{-1} \sum_{i=1}^n x_i x_i^{\mathrm{T}}$, $M = \lim_{n\to\infty} n^{-1} \sum_{i=1}^n f_i(0) x_i x_i^{\mathrm{T}} > 0$ exist.

Conditions similar to 1, 2 and 6 are routinely used in the Bahadur representation of the quantile regression estimators. In fact, they imply that (Koenker, 2005, p. 122)

$$\hat{\beta} - \beta_0 = (nM)^{-1} \sum_{i=1}^{n} x_i \psi_\tau(e_i) + o_p(n^{-1/2}). \tag{2}$$

Conditions 3 and 4 are used to remove complications arising from using weights near zero. Condition 5 ensures that the conditional τth quantile of the bootstrapped residuals is zero. A simple weight distribution that satisfies those conditions is the two-point mass distribution with probabilities $1 - \tau$ and τ at $w = 2(1 - \tau)$ and -2τ, respectively.

Comparing the result in Theorem 1 and the asymptotic sampling distribution of $\hat{\beta}$ from (2), we have shown that the wild bootstrap distribution can provide asymptotically valid inference. For example, the wild bootstrap method consistently estimates the variance of $\hat{\beta}$.

In the special case of median regression with $\tau = 0{\cdot}5$, the loss function is symmetric. In that case, one can use residuals instead of absolute residuals in Step 2, and Condition 4 can be relaxed to $E_W(|w|^{-1}) = 1$.

3. Simulation study

A small-scale simulation study with fixed design points is reported here to demonstrate applicability of the proposed wild bootstrap in finite-sample problems. The data are generated from

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + 3^{-1/2}[2 + \{1 + (x_{1i} - 8)^2 + x_{2i}\}/10]\epsilon_i \quad (i = 1, \ldots, n),$$

where $(\beta_0, \beta_1, \beta_2) = (1, 1, 1)$, ϵ_i are drawn from the t distribution with 3 degrees of freedom and x_{1i} are generated from the standard log-normal distribution with fixed seed in R (R Development Core Team, 2011). We choose x_{2i} to be 1 for the first 80% of the observations and 0 for the rest, and consider the median regression with sample sizes ranging from $n = 20$ to 5000.

In addition to the point-mass distribution mentioned in § 2, we use the following weight distribution, which satisfies the conditions of Theorem 1 for $1/8 < \tau < 7/8$:

$$g(w) = G'(w) = \begin{cases} -w \ (-2\tau - 1/4 \leqslant w \leqslant -2\tau + 1/4), \\ w \ (2(1 - \tau) - 1/4 \leqslant w \leqslant 2(1 - \tau) + 1/4). \end{cases} \tag{3}$$

We use the function rq of the R package quantreg to fit the median regression model, and compare the performance of the wild bootstrap method using the distribution given in (3) or the Bernoulli distribution with equal probabilities at -1 and 1, the wild bootstrap method with the weights w drawn from the standard normal distribution, the paired bootstrap method, the residual bootstrap method by resampling the residuals and the random weight method of Rao & Zhao (1992) with the weights w drawn from the exponential distribution exp(1) or the Poisson distribution with $\lambda = 1$.

A finite-sample correction factor is generally recommended for the wild bootstrap to reflect the fact that the residuals are slightly less dispersed than the model errors. We note from the Bahadur representation of the estimator in the independent and identically distributed error case that

$$\hat{e}_i = e_i - \{f(0)\}^{-1} h_i \psi_\tau(e_i) + o_p(n^{-1/2}),$$

where $h_i = x_i^T (\sum_k x_k x_k^T)^{-1} x_i$. This suggests that we replace the residual $\hat{e}_i$ in the wild bootstrap by $\hat{e}_i + \{\hat{f}(0)\}^{-1} h_i \psi_\tau(\hat{e}_i)$, where $\hat{f}(0)$ is estimated from the residuals. Specifically, we obtain $\hat{f}(0)$ via implementing the akj function in the R package quantreg, which is a univariate adaptive kernel density estimation method as used in Portnoy & Koenker (1989). Despite the fact that it is motivated by the independent and identically distributed error models, the proposed finite-sample correction is also useful in general error models.

To distinguish the methods, we first use 100 Monte Carlo samples with size n. Under each bootstrap method, 999 bootstrapped samples are used to estimate the standard errors of the parameter estimates for

Table 1. *Comparison of nominal* 90% *confidence intervals at* $n = 50$. *The bootstrap-t intervals are used for the following bootstrap-based methods*

	β_0		β_1		β_2	
	Coverage (%)	Length (SE)	Coverage (%)	Length (SE)	Coverage (%)	Length (SE)
CG	87·8	5·5 (0·02)	91·5	1·5 (0·01)	90·4	5·7 (0·02)
BN	87·1	5·5 (0·02)	91·1	1·5 (0·01)	89·2	5·7 (0·02)
PB	93·6	7·0 (0·03)	94·6	1·7 (0·01)	95·1	7·4 (0·02)
RW	93·6	7·0 (0·03)	94·2	1·6 (0·01)	95·3	7·6 (0·03)
ND	90·7	6·4 (0·02)	87·7	1·4 (0·01)	92·1	6·8 (0·02)
RK	87·7	5·9 (0·02)	90·0	1·7 (0·01)	90·8	∞

CG and BN refer to the proposed wild bootstrap methods using weights generated from (3) and the Bernoulli distribution with equal probabilities at -1 and 1, respectively; PB denotes the paired bootstrap; RW is the random weight bootstrap with weights generated from the exponential distribution with mean 1. Two other methods in the comparison are ND and RK as implemented in the R package quantreg, where ND refers to the normal approximation-based intervals allowing nonidentically distributed errors, and RK refers to the rank score method. Coverage is the estimated coverage probability of confidence intervals; Length (SE) gives the average lengths and their standard errors. Some intervals from RK are of infinite length.

each sample. Moreover, we use 5000 Monte Carlo samples to estimate the standard errors of the parameter estimates as the benchmark for comparison. Figure 1 in the Supplementary Material shows the clear biases of the wild bootstrap with weights drawn from the standard normal distribution and the regular residual bootstrap. The proposed method shows better performance over the paired bootstrap and the random weight bootstrap of Rao & Zhao (1992) in small samples, because the latter methods effectively introduce sampling variability to the design points.

We further compare the performance of confidence intervals for the parameters β_0, β_1 and β_2 among several competitive methods as shown in Table 1, using 10 000 Monte Carlo samples. For various resampling methods, we use confidence interval constructions based on the bootstrap-t method and the naive percentile method; we refer to Efron & Tibshirani (1994, § 12.5, § 13.2–13.3) and Davison & Hinkley (1997, Ch. 5) for details. We report only the results from the bootstrap-t method at $n = 50$ here, but additional results from different methods, different sample sizes and different confidence levels can be found in the Supplementary Material.

In addition to resampling methods, Table 1 also includes two commonly used inference methods in quantile regression, one based on large-sample approximation under nonidentically distributed errors, the other based on inversion of quantile regression rank score tests. We refer to Koenker (2005, Ch. 3) for details of these methods. Because the bootstrap method of Rao & Zhao (1992) with weights drawn from the Poisson distribution with mean 1 gives similar performance to the paired bootstrap, we do not give results from the former. In Table 1 and the Supplementary Material, the standard errors of the coverage estimates are no larger than 0·4%. The proposed wild bootstrap methods perform better overall than other methods, especially for the slope parameters β_1 and β_2. In particular, other resampling methods tend to be overly conservative.

If these bootstrap methods are used to test the hypothesis $\beta_k = 0$ $(k = 0, 1, 2)$ against the alternative hypothesis $\beta_k \neq 0$ based on the Wald test, we find that, by generating the data with $n = 50$ from the same model except that β_k is set to 0, our proposed wild bootstrap leads to Type I errors between 0·04 and 0·06, but the paired bootstrap method leads to Type I errors around 0·02, and the random weight method with weights drawn from exp(1) results in Type I errors as high as 0·08. The better performance of the wild bootstrap is evident.

4. DISCUSSION

In evaluating the relative performances of the wild bootstrap against other resampling methods, a differentiating factor is how much variability of the design points is to be considered. Both the wild bootstrap and

the resampling methods of Rao & Zhao (1992), Chatterjee & Bose (2005), and Hu & Zidek (1995) use the original design points x_i in the bootstrapped samples. However, the resampling methods of Rao & Zhao (1992) and Chatterjee & Bose (2005) apply weights to all points, effectively introducing variability in the design space. Alternatively, the paired bootstrap resamples x_i, so the difference between conditional and unconditional inference tends to appear more clearly in small sample cases. The wild bootstrap is most appropriate for conditional inference with fixed design points, especially in the presence of leverage points. We also find that simple finite-sample corrections derived from independent and identically distributed error models are often helpful for heteroscedastic error models. Further work is needed to adapt the method to data with dependent errors.

The asymptotic validity of the wide bootstrap given in the present paper is limited to first-order accuracy. Due to lack of sufficient smoothness in the quantile objective function, second-order accuracy is not expected from a resampling method. To achieve higher order accuracy, Hall et al. (1989) considered smoothing the quantile objective function.

Acknowledgement

This research work is partially supported by grants from the National Science Foundation, National Institute of Health, and the National Natural Science of China. The authors are grateful to anonymous reviewers for their comments and suggestions on an earlier draft of the paper.

Supplementary material

Supplementary material available at *Biometrika* online includes the proof of Theorem 1, Figure 1 and extended Tables 1–4.

References

Chatterjee, S. & Bose, A. (2005). Generalized bootstrap for estimating equations. *Ann. Statist.* **33**, 414–36.
Davidson, R. & Flachaire, E. (2008). The wild bootstrap, tamed at last. *J. Economet.* **146**, 162–9.
Davison, A. C. & Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press.
Efron, B. & Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*. New York: Chapman & Hall.
Hu, F. & Zidek, J. V. (1995). A bootstrap based on the estimating equations of the linear model. *Biometrika* **82**, 263–75.
Knight, K. (1999). Asymptotics for L_1-estimators of regression parameters under heteroscedasticity. *Can. J. Statist.* **27**, 497–507.
Koenker, P. (2005). *Quantile Regression*. Cambridge, UK: Cambridge University Press.
Koenker, R. & Bassett, G. (1978). Regression quantiles. *Econometrica* **46**, 33–50.
Hall, P., Diciccio, T.J., & Romano, J.P. (1989). On smoothing and the bootstrap. *Ann. Statist.* **17**, 692–704.
Lahiri, S. N. (1992). Bootstrapping M-estimators of a multiple linear regression parameter. *Ann. Statist.* **20**, 1548–70.
Liu, R. Y. (1988). Bootstrap procedures under some non-i.i.d. models. *Ann. Statist.* **16**, 1696–708.
Mammen, E. (1991). *When Does Bootstrap Work? Asymptotic Results and Simulations*. New York: Springer.
Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *Ann. Statist.* **21**, 255–85.
Portnoy, S. & Koenker, R. (1989). Adaptive l estimation of linear models. *Ann. Statist.* **17**, 362–81.
Rao, C. R. & Zhao, L. C. (1992). Approximation to the distribution of M-estimates in linear models by randomly weighted bootstrap. *Sankhya* **54**, 323–31.
R Development Core Team (2011). *R: A Language and Enviroment for Statistical Computing: Vienna, Austria: R Foundation for Statistical Computing*. ISBN 3-900051-07-0. URL: http://www.R-project.org.
Shao, J. & Tu, D. (1995). *The Jackknife and Bootstrap*. New York: Springer.
Wu, C. F. J. (1986). Jackknife bootstrap and other resampling methods in regression analysis. *Ann. Statist.* **14**, 1261–95.

[*Received September* 2010. *Revised June* 2011]