Journal of Data Science **6**(2008), 547-556

Quantile Regression: A Simplified Approach to a Goodness-of-fit Test

Rand R. Wilcox
University of Southern California

Abstract: Recently, He and Zhu (2003) derived an omnibus goodness-of-fit test for linear or nonlinear quantile regression models based on a CUSUM process of the gradient vector, and they suggested using a particular simulation method for determining critical values for their test statistic. But despite the speed of modern computers, execution time can be high. One goal in this note is to suggest a slight modification of their method that eliminates the need for simulations among a collection of important and commonly occurring situations. For a broader range of situations, the modification can be used to determine a critical value as a function of the sample size (n), the number of predictors (q), and the quantile of interest (γ). This is in contrast to the He and Zhu approach where the critical value is also a function of the observed values of the q predictors. As a partial check on the suggested modification in terms of controlling the Type I error probability, simulations were performed for the same situations considered by He and Zhu, and some additional simulations are reported for a much wider range of situations.

Key words: CUSUM processes, detecting curvature, robust methods.

1. Introduction

Consider the random variables $x_1, \ldots, x_p$, y having some unknown $(p+1)$-variate distribution. The usual regression setup is being considered where $x_1 = 1$ and there are $q = p - 1$ predictors $x_2, \ldots, x_p$. Numerous methods have been proposed for testing

$$H_0 : E(y|\mathbf{x}) = g(\mathbf{x}, \theta)$$

for some unknown parameter θ and some specified function $g(X, \theta)$, which include the classic Kolmogorov-Smirnov test, the Cramer-von Mises statistic, and likelihood ratio statistics. More recent proposals have been given by Härdle and Mammen (1993), González-Manteiga and Cao (1993), Hong and White (1995) and Hart (1997).

In recent years, there has been a growing interest in quantile regression, meaning that rather than focus on $E(y|\mathbf{x})$, the goal is to model the conditional γ quantile of y given $\mathbf{x}$ (e.g., Hahn, 1995; Koenker. & Xiao, 2002; He & Shao, 1996; Zheng, 1998). Such methods add a new and more detailed perspective regarding associations. As for testing the fit of a particular quantile regression model, Zheng (1998) derived a test that uses a kernel estimate of the conditional mean of $I(y \leq g(\mathbf{x}) - \gamma)$ and Fan, Zhang and Zhang (2001) used a generalized likelihood ratio statistic that uses a smoothing-based method that can depend critically on how the smoothing parameter is chosen. (For another approach based on kernel smoothing, see Horowitz & Spokoiny, 2002.) Yet another approach is to use a CUSUM process based on the residuals (e.g., Stute, 1997; also see Bierens and Ploberger, 1997; Stute, Thies & Zhu, 1998; Stute & Zhu, 2002; Zhu, 2003). More recently, He and Zhu (2003) derived an approach based on quantile regression that uses instead a vector-weighted cusum process of the gradient vector, the practical point being that it can be made more sensitive to departures from the model.

He and Zhu (2003) suggest a simulation method for approximating a p-value or α-level critical value when using their the test statistic. In effect, their estimate of the p-value is a function of the sample size n, the number of predictors q, the quantile of interest γ, and the observed values of the covariates, $\mathbf{x}$. They suggest how the number of iterations used in simulations might be reduced, but in practice a large number of replications might still be needed. (On the author's SUNBLADE 150 computer, even with $n = 20$, execution time can exceed 3 minutes, and execution time increases rapidly as n gets large. With $n = 200$ execution time can exceed 40 minutes.)

There are two goals in this note. The first is to suggest a simple modification of the He and Zhu method that is aimed at eliminating the need for any simulations when dealing with the commonly occurring situations where the goal is to test the fit of a linear model with $q \leq 6$, $n \leq 400$, $\gamma = .5$ and when the level of the test is chosen to be .10, .05, .025 or .01. When dealing with other values for q, n and γ, simulations are still needed to estimate a critical value, but based on the results reported here, it appears that this must be done only once for given values of n, q and γ. That is, the results given here suggest that generally, it is not necessary to assume that critical values are also a function of the observed covariates. The second goal in this paper is to report the results of a more extensive simulation study, compared to the simulation study by He and Zhu, regarding how well the method controls the probability of a Type I error. A related goal is to provide some indirect evidence that when the He and Zhu method is applied, the actual level of their approach is relatively insensitive to changes in the design space.

2. Review of the He and Zhu Method

In its more general form, He and Zhu (2003) test the hypothesis that $\mathbf{x}$ and y can be modeled by

$$y = g(\mathbf{x}, \theta) + s(\mathbf{x})\epsilon,$$

where $\theta \in \mathcal{R}^p$ is an unknown parameter, g is a given function except for the parameter θ, ϵ is a random variable having γth quantile 0, and $s(\mathbf{x})$ is a scale function that is consistently estimable. The focus here is on the important special case where the goal is to test the hypothesis that $\mathbf{x}$ and y can be modeled by

$$y = \mathbf{x}'\beta + (\mathbf{x}'\tau)\epsilon. \tag{2.1}$$

That is, the errors are iid except for a linear scale $\mathbf{x}'\tau$.

Given a set of observations $\{(\mathbf{x}_i, y), i = 1, \ldots n\}$, β is estimated by minimizing

$$\sum \rho_\gamma(y_i - \mathbf{x}_i'\beta),$$

where $\rho_\gamma(r) = \gamma r^+ + (1-\gamma)r^-$, with r^+ as the positive part and r^- the negative part of r. The solution can be found via linear programming (e.g., Koenker and d'Orey, 1987) and easy-to-use software is available within S-Plus and R. The resulting estimate of β, for given γ, is denoted by $\hat{\beta}$.

Following He and Zhu, for any $\mathbf{x}$, $\mathbf{t} \in \mathcal{R}^p$, $\mathbf{x} \leq \mathbf{t}$ if and only if each component of $\mathbf{x}$ is less than or equal to each component of $\mathbf{t}$. Let $\psi(r) = \gamma I(r > 0) + (\gamma - 1)I(r < 0)$ be the derivative of ρ_γ, $r_i = y_i - \mathbf{x}_i'\hat{\beta}$, and let

$$\mathbf{R}_n(\mathbf{t}) = n^{-1/2} \sum_{j=1}^{n} \psi(r_j)\mathbf{x}_j I(\mathbf{x}_j \leq \mathbf{t}).$$

Their test statistic is

$$T_n = \max_{\|a\|=1} n^{-1} \sum (\mathbf{a}'\mathbf{R}_n(\mathbf{x}_j))^2, \tag{2.2}$$

the largest eigenvalue of $n^{-1}\sum \mathbf{R}_n(\mathbf{x}_i)\mathbf{R}_n'(\mathbf{x}_i)$. What is required is an estimate of the null distribution of T_n, and He and Zhu describe a simulation method that, as previously indicated, is based in part on the $\mathbf{x}_i$ values. Their approximation, in the more general setting considered in their paper, is to use simulations based on

$$\mathbf{R}^*(\mathbf{t}) = n^{-1/2} \sum_{j=1}^{n} \omega_j \{I(\mathbf{x}_j \leq \mathbf{t})\dot{\mathbf{g}}(\mathbf{x}_j, \hat{\theta}) - \mathbf{S}_n(\mathbf{t})\dot{\mathbf{g}}(\mathbf{x}_j, \hat{\theta})\},$$

where $\dot{\mathbf{g}}(\mathbf{x}_j, \hat{\theta})$ is the partial derivative of $\mathbf{g}$ with respect to θ,

$$\mathbf{S}_n(\mathbf{t}) = n^{-1} \sum \dot{\mathbf{g}}(\mathbf{x}_j, \hat{\theta})\dot{\mathbf{g}}'(\mathbf{x}_j, \hat{\theta}) I(\mathbf{x}_j \leq \mathbf{t}),$$

by assumption the design has been normalized so that $n^{-1}\sum \dot{\mathbf{g}}(\mathbf{x}_j,\hat{\theta})\dot{\mathbf{g}}'(\mathbf{x}_j,\hat{\theta}) - I = o(1)$, and where the random variable ω_j takes the values $(\gamma, -\gamma, 1-\gamma, \gamma-1)$ with probability $((1-\gamma)/2, (1-\gamma)/2, \gamma/2, \gamma/2)$. A reference distribution for T_n is simulated by generating $\mathbf{R}^*(\mathbf{t})$ and then computing the largest eigenvalue of $n^{-1}\sum \mathbf{R}^*_n(\mathbf{x}_i)\mathbf{R}^{*\prime}_n(\mathbf{x}_i)$.

3. The Proposed Simplification

Attention is focused on the model given by (1) and it is still assumed that the design has been normalized so that $n^{-1}\sum \mathbf{x}_j\mathbf{x}_j' - I = o(1)$. A simple strategy is to determine a critical value when both $\mathbf{x}$ (prior to being normalized) and y have normal distributions and then use this critical value when the normality assumption is violated, thus avoiding the need for simulations. (In essence, this is the same strategy used by Gosset to derive Student's T test.) However, in the simulations reported here, this approach was found to be unsatisfactory. When the marginal distributions have heavy-tails, the Type I error probability can exceed .085 when testing at the .05 level with $n = 20$ or 50. The main result here is that this problem was eliminated by modifying slightly the partial ordering among the design points used by He and Zhu.

Consider

$$\mathbf{R}_n(\mathbf{x}_i) = n^{-1/2}\sum_{k=1}^{n}\psi(r_k)\mathbf{x}_k I(\mathbf{x}_k \le \mathbf{x}_i).$$

For fixed j, let U_{ij} be the ranks of the n values in the jth column of $\mathbf{x}$, $j = 2,\ldots,q$. Let $F_i = \max U_{ij}$, the maximum being taken over $j = 2,\ldots,q$. If $\mathbf{x}_k \le \mathbf{x}_i$, then $F_k \le F_i$. To see this, let $\mathbf{U}_i = (U_{i2},\ldots,U_{iq})$ and note that $\mathbf{x}_k \le \mathbf{x}_i$ implies that $\mathbf{U}_k \le \mathbf{U}_i$. But $\mathbf{U}_k \le \mathbf{U}_i$ means in particular that $F_k \le F_i$. It follows that the sum used to compute $\mathbf{R}_n(\mathbf{x}_i)$ includes all terms for which $F_k \le F_i$. And if $F_k > F_i$, the term is not included. Let

$$\mathbf{W}_i = n^{-1/2}\sum_{k=1}^{n}\psi(r_k)\mathbf{x}_k I(F_k \le F_i),$$

and let C_n be the largest eigenvalue of

$$\mathbf{Z} = \frac{1}{n}\sum \mathbf{W}_i\mathbf{W}_i'.$$

Although very similar to the He and Zhu test statistic, T_n and C_n generally differ. It can be shown by example that it is possible to have $F_k \le F_i$ yet neither $\mathbf{x}_k \le \mathbf{x}_i$ or $\mathbf{x}_k > \mathbf{x}_i$. That is, the sum when computing W_i contains all of the terms used to compute T_n plus possibly some additional terms.

The first modification considered here consists of replacing the test statistic T_n with C_n, determining critical values under normality, and then using these critical values when sampling from non-normal distributions. This improved control over the probability of a Type I error, but with heavy-tailed distributions, estimated Type I error probabilities still exceeded .075 when testing at the .05 level, and so another modification was considered where C_n is computed as before, only now

$$\mathbf{W}_i = n^{-1/2} \sum_{k=1}^{n} \psi(r_k)\mathbf{x}_k I(F_k \geq F_i).$$

The resulting test statistic is labeled D_n. Now control over the Type I error probability was found to be good in the simulations reported here, and power compared well to the approach used by He and Zhu.

A very small advantage of the test statistic D_n over T_n is that it is simpler and more efficient when writing software code in something like R or S-Plus. Nested loops are easily avoided and built-in functions for computing ranks can be used to reduce execution time. And as is evident, a major component of the test statistic is invariant under monotone transformations of the covariates; only the ranks of the marginal distributions of $\mathbf{x}$ are needed. However, the test statistic can be affected by monotone transformations because this can alter the $\psi(r_i)$ values.

4. Some Special Cases

Simulations were used to approximate critical values in the manner just described for $q = 1, \ldots, 6$ predictors; $n = 10$, 20, 30, 50, 100, 200 and 400; $\gamma = .5$; and $\alpha = .1$, .05, .025 and .01. For $n \leq 100$ it was found that a very good approximation of the α level critical, c_α, is given by

$$c_\alpha = \frac{d}{n^{1.5}},$$

where the values for d are given in Table 1. (When the null hypothesis is true, $T_n \to 0$ as $n \to \infty$.) By good approximation is meant that in simulations, the actual level of the test is reasonably close to the nominal level, details of which are given later in the paper.

Table 1: Values of d for Approximating c_α, $\gamma = .5$, $10 \leq n < 100$

	q					
α	1	2	3	4	5	6
.100	0.799	0.763	0.559	0.422	0.334	0.272
.050	1.050	0.955	0.635	0.477	0.384	0.306
.025	1.127	1.909	0.719	0.537	0.430	0.337
.010	1.382	1.241	0.818	0.599	0.472	0.360

However, for $n = 200$, the approximation based on Table 1 is not quite satisfactory, and so for $100 \leq n \leq 200$, it is suggested that the values for d in Table 2 be used instead. For $n < 200 \leq 400$, the values in Table 3 give better results.

Table 2: Values of d for Approximating c_α, $\gamma = .5$, $100 < n \leq 200$

	q					
α	1	2	3	4	5	6
.100	1.125	1.104	0.853	0.613	0.386	0.245
.050	1.360	1.558	1.056	0.746	0.461	0.286
.025	1.613	1.732	1.271	0.854	0.531	0.329
.010	1.998	2.226	1.665	1.063	0.632	0.394

Table 3: Values of d for Approximating c_α, $\gamma = .5$, $200 < n \leq 400$

	q					
α	1	2	3	4	5	6
.100	1.262	1.472	1.216	0.752	0.487	0.303
.050	1.704	1.858	1.524	0.917	0.582	0.354
.025	2.095	2.157	1.794	1.064	0.678	0.409
.010	2.647	2.541	2.118	1.332	0.768	0.487

5. A Simulation Study

Simulations were used to study the small-sample properties of the method just described. The initial set of simulations were based on the same situations considered by He and Zhu (2003). The first of these null cases is

$$y_i = \epsilon$$

with y_i and ϵ having independent standard normal distributions. The second was

$$y_i = 1 + x_{i1} + x_{i2} + \epsilon$$

with x_{i1} having a binomial distribution $b(5, .5)$, x_{2i} has a standard normal distribution and $\epsilon + 1$ has a standard lognormal distribution. And the third was

$$y_i = x_{i3} + (1 - x_{i1}/2)\epsilon,$$

where x_1, x_2 and x_3 have independent uniform distributions and now $\epsilon + 1$ has a gamma distribution with both shape and scale parameters set equal to 1. It is merely noted that the estimated probability of a Type I error was very similar to

the estimates reported by He and Zhu. (They differed from the nominal level by at most a few units in the third decimal place.) Here the focus is on a broader range of situations.

The number of covariates considered was $q = 1$ and 4, similar results were obtained in both cases, so only the results for $q = 4$ are reported. The marginal distributions for $\mathbf{x}$ were generated from the family of g-and-h distributions, which contains normal distributions as a special case (Hoaglin, 1985). If Z has a standard normal distribution, then

$$X = \begin{cases} \frac{\exp(gZ)-1}{g}\exp(hZ^2/2), & \text{if } g > 0 \\ Z\exp(hZ^2/2), & \text{if } g = 0 \end{cases}$$

has a g-and-h distribution where g and h are parameters that determine the first four moments. The four distributions used here were the standard normal ($g = h = 0.0$), a symmetric heavy-tailed distribution ($g = 0.0$, $h = 0.2$), an asymmetric distribution with relatively light tails ($g = 0.2$, $h = 0.0$), and an asymmetric distribution with heavy tails ($g = h = 0.2$). Table 3 shows the skewness (κ_1) and kurtosis (κ_2) for each distribution considered. Additional properties of the g-and-h distribution are summarized by Hoaglin (1985). The error term ϵ was also taken to have one of the g-and-h distribution.

Table 4: Some properties of the g-and-h distribution

g	h	κ_1	κ_2
0.0	0.0	0.00	3.00
0.0	0.2	0.00	21.46
0.2	0.0	0.61	3.68
0.2	0.2	2.81	155.98

Observations were generated with $n = 20$ and $y_i = \lambda(x_{i1})\epsilon$, where three choices for λ were used to reflect three types of variance patterns: $\lambda(x_{i1}) \equiv 1$ (homoscedasticity), $\lambda(x_{i1}) = |x_{i1}|+1$ and $\lambda(x_{i1}) = 1/(|x_{i1}|+1)$. For convenience, these three choices for λ are denoted by VP1, VP2 and VP3.

Note that the distributions considered here for $\mathbf{x}$ include much heavier tailed distributions than those considered by He and Zhu. It is well known that with $\gamma = .5$ (L_1 regression), protection against the deleterious affects of outliers among the y values is achieved, but that outliers among the design space can result in a regression line that poorly reflects the association among the bulk of the points. So a practical issue is the extent to which heavy-tailed distributions cause problems when trying to control the probability of a Type I error.

Table 5 shows the estimated probability of a Type I error when testing at the .05 level. The estimates are based on 1,000 replications and range between .026 and .055.

Table 5: Estimated Probability, $\hat{\alpha}$, of a Type I error, $n = 20$, $p = 4$

x		ϵ		$\hat{\alpha}$		
g	h	g	h	VP1	VP2	VP3
0.0	0.0	0.0	0.0	.048	.033	.028
0.0	0.0	0.0	0.2	.046	.034	.033
0.0	0.0	0.2	0.0	.033	.039	.038
0.0	0.0	0.2	0.2	.039	.029	.026
0.0	0.2	0.0	0.0	.034	.044	.042
0.0	0.2	0.0	0.2	.043	.032	.044
0.0	0.2	0.2	0.0	.038	.029	.051
0.0	0.2	0.2	0.2	.042	.034	.046
0.2	0.0	0.0	0.0	.028	.034	.033
0.2	0.0	0.0	0.2	.032	.034	.035
0.2	0.0	0.2	0.0	.035	.045	.034
0.2	0.0	0.2	0.2	.030	.037	.036
0.2	0.2	0.0	0.0	.036	.038	.055
0.2	0.2	0.0	0.2	.040	.028	.051
0.2	0.2	0.2	0.0	.041	.034	.043
0.2	0.2	0.2	0.2	.032	.036	.041

Power comparisons, based on using T_n and D_n, were made as well for the same situations considered by He and Zhu (2003). It is merely noted that there appears little or no difference in terms of power.

6. Concluding Remarks

Of course, it is not possible to prove by simulation that the approach used here always maintains reasonable control over the probability of a Type I error. However, the distributions considered would seem to include fairly extreme departures from normality, and the types of heteroscedasticity would seem to be relatively extreme as well, suggesting that in general the method would be expected to perform in a satisfactory manner. In summary, all indications are that a very quick and relatively simple method can be applied for the important special cases considered here. For other situations, simulations must again be used to approximate accurate confidence intervals, but it is a simple matter to write software that stores critical values when these special cases come up and then to use these critical values in any future investigation where the same values for n, p and γ occur. An R and S-Plus function that accomplishes this goal, called qrchk, is available from the author upon request. (This function uses a more refined approximation of the critical values; it interpolates based critical values corresponding to the sample sizes 10, 20, 30, 50, 100, 200, and 400.)

Finally, the results reported here indicate that the partial ordering applied to

the design points can make a practical difference. The reason for this is unknown. And it also raises the issue of whether alternative partial orderings have practical value.

References

Biereñs, H. J. and Ploberger, W. (1997). Asymptotic theory of integrated conditional moment tests. *Econometrika* **65**, 1129-1151.

Fan, J., Zhang, C. M. and Zhang, J. (2001). Generalized likelihood ratio statistics and Wilks Phenomenon. *Annals of Statistics* **29**, 153-193.

Gonzalez-Manteiga, W. and Cao, R. (1993). Testing the hypothesis of a general linear model using nonparametric regression estimation. *Test* **2**, 161-188.

Hahn, J. (1995). Bootstrapping quantile regression estimators. *Econometric Theory* **11**, 105-121.

Härdle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric

regression fits. *Annals of Statistics* **21**, 1926-1947.

Hart, J. D. (1997). *Nonparametric Smoothing and Lack-of-Fit Tests.* Springer-Verlag.

He, X. and Ng, P. (1999). Quantile splines with several covariates. *Journal of Statistical Planning and Inference* **75**, 343-352.

He, X., Ng, P. and Portnoy, S. (1998). Bivariate quantile smoothing splines. *Journal of the Royal Statistical Society, Ser. B* **60**, 537-550.

He, X., and Zhu, L.-X. (2003). A lack of fit test for quantile regression. *Journal of the American Statistical Association* **98**, 1013-1022.

Horowitz,J. L. and Spokoiny, V. G. (2002). An adaptive, rate-optimal test of linearity for median regression models. *Journal of the American Statistical Association* **97**, 822-835.

Hoaglin, D. C. (1985) Summarizing shape numerically: The g-and-h distributions. In *Exploring data tables, trends, and shapes.* (Edited by D. Hoaglin, F. Mosteller and J. Tukey), 461-515. Wiley.

Hong, T. and White, H. (1995). Consistent specification testing via nonparametric series regression. *Econometrika* **63**, 1133-1159.

Koenker, R. and d'Orey, V. (1987). Computing regression quantiles. *Applied Statistics* **36**, 383-393.

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

Koenker, R. and Xiao, Z. J. (2002). Inference on the quantile regression process. *Econometrica* **70**, 1583-1612.

LaRiccia, V. N. (1991). Smooth goodness-of-Fit tests: A quantile function approach. *Journal of the American Statistical Association* **86**, 427-431.

Stute, W. (1997). Nonparametric model checks for regression. *Annals of Statistics* **25**, 613-641.

Stute, W., Gonzalez-Manteiga, W. G. and Presedo-Quindimil, M. P. (1998). Bootstrap approximations in model checks for regression. *Journal of the American Statistical Association* **93**, 141-149.

Stute, W., Thies, S. and Zhu, L. X. (1998). Model checks for regression: An innovation process approach. *Annals of Statistics* **26**, 1916-1934.

Stute, W. and Zhu, L. X. (2002). Model checks for generalized linear models. *Scandinavian Journal of Statistics* **29**, 535-546.

Wooldridge, J. (1992). A test for functional form against nonparametric alternatives. *Econometric Theory* **8**, 452-475.

Yatchew, A. J. (1992). Nonparametric regression tests based on an infinite dimensional least squares procedure. *Econometric Theory* **8**, 435-451.

Zheng, J. X. (1998). A consistent nonparametric test of parametric models under conditional quantile regressions. *Econometric Theory* **14**, 223-238.

Zhu, L.-X. (1993). Model checking in dimension-reduction type for regression. *Statistical Sinica* **13**, 283-296.

Received May 17, 2007; accepted October 23, 2007.

Rand R. Wilcox
Department of Psychology
University of Southern California
Los Angeles, CA 90089-1061, USA
rwilciox@usc.edu