Quantile regression and heteroskedasticity*

José A. F. Machado† J.M.C. Santos Silva‡

June 18, 2013

Abstract

This note introduces a wrapper for `qreg` which reports standard errors and t statistics that are asymptotically valid under heteroskedasticity and misspecification of the quantile regression function. Moreover, the result of an heteroskedasticity test are also presented to guide the researcher in the choice of the appropriate covariance matrix estimator to use.

Key words: Bootstrap, Covariance matrix, Robust standard errors.

1. INTRODUCTION

Quantile regression (Koenker and Bassett, 1978, Koenker, 2005) is a useful tool which is widely used in empirical work. Although the computation of the quantile regression estimates is relatively straightforward, obtaining the corresponding standard errors is often perceived as being more problematic.

Currently, Stata offers two ways of computing the covariance matrix of quantile regression estimates.[1] The `qreg` command computes standard errors that are valid when the

*Machado is consultant for the Research Department of Banco de Portugal. The authors gratefully acknowledge the partial financial support from Fundação para a Ciência e Tecnologia, program POCTI, partially funded by FEDER (FEDER/POCI 2010).

†Faculdade de Economia, Universidade NOVA de Lisboa. E-mail: jafm@fe.unl.pt.

‡University of Essex and CEMAPRE. E-mail: jmcss@essex.ac.uk.

[1] The recently announced Stata 13 allows the computation of heteroskedasticity-robust standard errors; the evaluation of this new estimator is left for future research.

errors are identically distributed, a case in which quantile regression is not particularly interesting. Alternatively, the commands `bsqreg` and `sqreg` compute the standard errors of the quantile regression estimates using the pairs-bootstrap, a procedure recommended by Buchinsky (1995).

Although there are doubts about its asymptotic validity in the case of the quantile regression (Machado and Parente, 2005), the pairs-bootstrap estimator is widely used and the simulation results reported by Buchinsky (1995) suggest that it is likely to perform well in practice. The problem with the bootstrap estimator is that it is somewhat impractical when the problem involves very large samples and many regressors because in this case the computation of the bootstrap covariance matrix using a reasonable number of bootstraps is still very time consuming.

This note introduces an ado file to estimate quantile regression which extends the `qreg` command by reporting standard errors and t statistics that are asymptotically valid under heteroskedasticity and misspecification of the quantile regression function.[2] Moreover, the command also presents the result of an heteroskedasticity test which the researcher can use as a guide in the choice of the appropriate covariance matrix estimator to use. A second ado file is provided which permits the computation of the same heteroskedasticity test after the standard Stata commands for quantile regression (`qreg`, `bsqreg` and `sqreg`) or after least squares regression (`reg`).

The reminder of this note is organized as follows. Section 2 briefly discusses the different estimators of the covariance matrix of quantile regression estimates. Section 3 introduces the heteroskedasticity test, and Section 4 briefly describes the ado files and illustrates the performance of the proposed methods using a small simulation study. Section 5 concludes.

[2] From version 3.1, `qreg2` also allows the computation of clustered standard errors; see Parente and Santos Silva (2013).

2. THE COVARIANCE OF THE QUANTILE REGRESSION ESTIMATOR

2.1. Asymptotic results

Consider the following linear quantile regression

$$y_i = x_i'\beta(\tau) + u_i(\tau)$$

where $\mathrm{Q}_{u(\tau)}(\tau|x_i) = 0$.[3] In their seminal paper, Koenker and Bassett (1978) (see also Koenker, 2005) have shown that the parameters of interest can be estimated by

$$\hat{\beta}(\tau) = \arg\min_b \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(u_i(\tau))$$

where $\rho_\tau(a) = a(\tau - \mathbf{1}(a < 0))$ is the so-called ✓*-function*, and that under suitably regularity conditions, including the assumption that the errors $u_i(\tau)$ are i.i.d.,

$$\sqrt{n}\left(\hat{\beta}(\tau) - \beta(\tau)\right) \xrightarrow{d} \mathcal{N}(0, V),$$

with

$$V = \left(\frac{\tau(1-\tau)}{\left[f_{u(\tau)}(0)\right]^2}\right)[E(xx')]^{-1},$$

where $f_{u(\tau)}(0)$ denotes the density of $u_i(\tau)$ at zero.

The assumption that the errors are i.i.d. was typical of the literature on robust regression of the 1970s (e.g., Huber, 1973, Hogg, 1979), but it was quickly realized that quantile regression offered much more than a robust estimator of a measure of central tendency. Indeed, quantile regression is particularly useful when the conditional distribution of y_i depends on the regressors in complex ways, in which case $u_i(\tau)$ will not be independent of x_i.

The asymptotic distribution of the quantile regression estimator under more general conditions was considered by Koenker and Bassett (1982), Powell (1984), Chamberlain (1994), and Kim and White (2003) leading to the conclusion that, when the errors are

[3] In the interest of space, standard notation is used throughout so that the definition of the symbols used is kept to a minimum.

independent but not identically distributed and the quantile regression is possibly misspecified,

$$\sqrt{n}\left(\hat{\beta}(\tau) - \beta(\tau)\right) \xrightarrow{d} \mathcal{N}\left(0, D^{-1}AD^{-1}\right),$$

where

$$D = \mathrm{E}\left[f_{u(\tau)}(0|x_i)\, x_i x_i'\right], \qquad A = \mathrm{E}\left[(\tau - \mathbf{1}(y_i < x_i'\beta(\tau)))^2 x_i x_i'\right],$$

and $f_{u(\tau)}(0|x_i)$ denotes the conditional distribution of $u(\tau)$ evaluated at zero. It is clear that when the model is correctly specified and the errors are i.i.d., $D^{-1}AD^{-1} = V$.

2.2. Estimation of the covariance matrix

Stata's `qreg` command uses a covariance matrix estimator of the form $R_2^{-1}R_1R_2^{-1}$, where

$$R_1 = \frac{1}{n\hat{f}_0^2}\sum_{i=1}^{n}\left(\tau - \mathbf{1}\left(\hat{u}(\tau)_i < 0\right)\right)^2 x_i x_i', \qquad R_2 = \frac{1}{n}\sum_{i=1}^{n} x_i x_i',$$

and $\hat{f}_0$ is a nearest-neighbor-type estimator of $f_{u(\tau)}(0)$ (see Rogers, 1993, for details).

It is easy to see that the estimator used by `qreg` converges to $D^{-1}AD^{-1}$ when $f_{u(\tau)}(0|x_i) = f_{u(\tau)}(0)$, but it is generally inconsistent in the more interesting case in which the errors are not identically distributed.

It is, however, possible to construct consistent estimators of A and D and thereby obtain a consistent estimator of the asymptotic covariance matrix of the quantile regression estimator. Indeed, following Powell (1984), Chamberlain (1994), and Kim and White (2003), A can be consistently estimated by

$$\hat{A} = \frac{1}{n}\sum_{i=1}^{n}(\tau - \mathbf{1}\left(\hat{u}(\tau)_i < 0\right))^2 x_i x_i',$$

whereas, for an appropriately defined smoothing parameter δ_n,

$$\hat{D} = \frac{1}{2n\delta_n}\sum_{i=1}^{n}\mathbf{1}\left(-\delta_n \le \hat{u}(\tau)_i \le \delta_n\right) x_i x_i'$$

is a consistent estimator of D.[4]

[4] To accommodate censored data, Powell (1984) uses a one-sided rectangular kernel in the estimation of $\hat{D}$. For non-censored data, a more standard kernel centred at zero is preferable, see Buchinsky (1995).

In order to implement this estimator all that is needed is to define a practical method of choosing δ_n. The simple solution used in `qreg2`,[5] the ado file to be introduced in the next section, is to choose δ_n as described in Koenker (2005, p. 81). In particular, we define

$$\delta_n = \kappa \left[\Phi^{-1}\left(\tau + h_n\right) - \Phi^{-1}\left(\tau - h_n\right)\right],$$

where h_n is (see Koenker, 2005, p. 140)

$$h_n = n^{-1/3}\left(\Phi^{-1}\left(1 - \frac{0.05}{2}\right)\right)^{2/3}\left(\frac{1.5\left(\phi\left(\Phi^{-1}\left(\tau\right)\right)\right)^2}{2\left(\Phi^{-1}\left(\tau\right)\right)^2 + 1}\right)^{1/3},$$

and κ is a robust estimate of scale. After some experimentation, we decided to define κ as the MAD (median absolute deviation) of the τ-th quantile regression residuals.

As an alternative to the use of analytical covariance matrix estimators, the Stata commands `bsqreg` and `sqreg` estimate the covariance matrix using pairs-bootstrap, as recommended by Buchinsky (1995). Asymptotically, the bootstrap estimates have no advantage over the estimator constructed using $\hat{A}$ and $\hat{D}$, and are still too computer intensive to be practical in applications using large data sets. Moreover, although the results in Buchinsky (1995) suggest that the bootstrap covariance matrix may perform well in finite samples, it is unlikely that the 20 bootstrap draws used by default in Stata will be enough to achieve satisfactory results in more demanding applications, further increasing the computational cost of this approach.

3. QUANTILE REGRESSION TESTS FOR HETEROSKEDASTICITY

Although Stata does not provide any specific command to perform a quantile regression-based heteroskedasticity test, versions of the test suggested by Koenker and Bassett (1982) can easily be implemented by using `iqreg` to estimate an inter-quantile regression and then using `test` to check the significance of the estimated slopes, or of a sub-set of them.

Although this test is conceptually attractive and easy to implement, it is cumbersome in that the results of the inter-quantile often are not of independent interest and their

[5] A different bandwidth was used before version 3.1.

bootstrap covariance matrix estimated by Stata is computationally expensive in realistic applications.

An alternative way to perform a quantile regression-based heteroskedasticity test was proposed by Machado and Santos Silva (2000). Their test statistic can be easily computed as n times the R^2 of the auxiliary regression of $\rho_\tau(\hat{u}_i(\tau))$ on a constant and on appropriate functions of x. The test can then be performed by comparing the test statistic with the critical value from the $\chi^2_{(J-1)}$ distribution, where J is the number of parameters in the auxiliary regression.

The Machado-Santos Silva (MSS) test is simple enough to be routinely performed after quantile regression, thereby providing the researcher with information not only about the kind of covariance matrix that is more appropriate but also about the relevance of estimating multiple quantiles.[6]

4. THE `QREG2` AND `MSS` ADO FILES

4.1. `qreg2`

`qreg2` is a wrapper for `qreg` which estimates quantile regression and reports standard errors and t-statistics that are asymptotically valid under heteroskedasticity and misspecification. The robust covariance matrix is computed using $\hat{A}$ and $\hat{D}$, as described in Section 2. The results based on the non-robust covariance matrix can be displayed with the option `norobust`.

Additionally, `qreg2` reports the value of the objective function, defined as the average of the check function,[7] and the R^2, defined as the square of the correlation between the fitted values and the dependent variable. Unlike the R^2 reported by `qreg`, the one reported by

[6] It should be noted that, strictly speaking, both the MSS test and the test proposed by Koenker and Bassett (1982) will check not only for heteroskedasticity but also for other departures from the assumption that the errors are identically distributed.

[7] The "Min sum of deviations" reported by `qreg` corresponds to $2n$ times the value of the objective function reported by `qreg2`.

`qreg2` is not based on any assumption about the distribution of the errors. It should be noted, however, that in quantile regressions the R^2 is even less meaningful than usual.

Finally, `qreg2` reports the result of the Machado-Santos Silva (2000) test for heteroskedasticity. By default the test variables are the fitted values of the dependent variable and its squares as in the "Special case of the White test" described by Wooldridge (2009, p. 276), but alternative sets of test variables can be specified with the option `mss(varlist)`. Obviously, the performance of the test will depend on the quantile being estimated.

4.2. `mss`

The MSS test may be of interest even when `qreg2` is not appropriate, for example because the sample size is too small for the its covariance matrix estimator to be reliable. Moreover, the MSS test can also be performed after OLS regressions (see Im, 2000, and Machado and Santos Silva, 2000). The `mss` ado file implements the MSS test for `reg`, `qreg`, `bsqreg` and `sqreg` (in this case, the results of the first quantile are considered).

By default the test variables are the fitted values of the dependent variable and its squares as when the rest is performed after `qreg2`. Alternative sets of test variables can be specified simply by listing the test variables after `mss`.

4.3. Simulation results

In this section we present the results of a small simulation study on the performance of the covariance matrix estimators computed by `qreg`, `bsqreg`, and `qreg2`. The default options were used for all estimators; therefore median regression was used in all cases. Additionally, the performances of the MSS test for heteroskedasticity computed by `qreg2` (with the default options) and of a version of the Koenker and Bassett (1982) heteroskedasticity test are also evaluated.

The simulated data were generated by

$$y_i = 1 + \beta x_i + \exp(\omega x_i)\, \varepsilon_i, \quad i = 1, \ldots, n,$$

where $x_i \sim \chi^2_{(3)}$, $\varepsilon_i \sim \mathcal{N}(0,1)$, and $\beta = 1$. The performance of the covariance estimators are evaluated by comparing the rejection frequencies of $H_0 : \beta = 1$. The Koenker and Bassett (1982) test was performed by testing that x_i has a coefficient equal to zero in the inter-quartile regression of y_i on a constant and on x_i, performed using the `iqreg` command with the default options. For each of the designs, y_i, x_i and ε_i were newly generated for each of the 10000 replications used in the experiment.

Table 1 gives the rejection frequencies at the 5% level for $H_0 : \beta = 1$ obtained using `qreg`, `bsqreg`, and `qreg2`. When $\omega = 0$ the estimators of the covariance matrix used by the different commands perform reasonably. As expected, when heteroskedasticity is present, the t statistic provided by `qreg` has an empirical size which is far from the desired significance level. In these cases, both `bsqreg`, and `qreg2` report t statistics whose empirical size is close to 5%. It is worth noting that in all cases considered the bootstrap estimator of the covariance matrix is actually outperformed by the analytical estimator implemented in `qreg2`.

Table 2 presents the rejections frequencies at the 5% level of the Koenker and Bassett (1982) heteroskedasticity test (labeled KB) and of the MSS test. Under the null ($\omega = 0$) both tests are reasonably well behaved even for $n = 100$. However, when heteroskedasticity is present, the MSS test is substantially more powerful than the KB test, at least in the conditions of this experiment.

Table 1: Rejection frequencies at the 5% level for $H_0 : \beta = 0$ using different commands

n	100			1000			10000		
ω	0.00	0.05	0.10	0.00	0.05	0.10	0.00	0.05	0.10
`qreg`	0.0791	0.1321	0.2050	0.0518	0.1121	0.1818	0.0522	0.1115	0.1850
`bsqreg`	0.0657	0.0722	0.0762	0.0697	0.0733	0.0707	0.0673	0.0695	0.0682
`qreg2`	0.0507	0.0706	0.0755	0.0533	0.0579	0.0588	0.0508	0.0513	0.0519

Table 2: Rejection frequencies at the 5% level for $H_0 : \omega = 0$

n	100			1000			10000		
ω	0.00	0.05	0.10	0.00	0.05	0.10	0.00	0.05	0.10
KB	0.0359	0.0966	0.3101	0.0558	0.8204	0.9997	0.0572	1.0000	1.0000
MSS	0.0490	0.2972	0.7780	0.0510	0.9970	1.0000	0.0467	1.0000	1.0000

5. CONCLUDING REMARKS

The results of the small simulation experiment reported in Section 4 confirm that, at least for realistic sample sizes, the asymptotically valid standard errors reported by `qreg2` perform as well as bootstrap standard errors computed by `bsqreg`, even when strong heteroskedasticity is present. Therefore, the `qreg2` command may prove to be useful for researchers estimating quantile regressions it that is avoids the need to use bootstrap, which can be very time consuming, especially in applications with many regressors and large samples.

REFERENCES

Buchinsky, M. (1995), "Estimating the Asymptotic Covariance Matrix for Quantile Regression Models a Monte Carlo Study," *Journal of Econometrics*, 68, 303-38.

Chamberlain, G. (1994), "Quantile Regression, Censoring and the Structure of Wages," in *Advances in Econometrics*, ed. C. A. Sims. Cambridge: Cambridge University Press, 171–209.

Hogg, R.V. (1979). "Statistical Robustness: One View of its Use in Applications Today," *The American Statistician*, 33, 108-115.

Huber, P.J. (1973). "Robust Regression: Asymptotics, Conjectures and Monte Carlo." *Annals of Statistics*, 1, 799-821.

Im, K.S. (2000). "Robustifying the Glejser test of Heteroskedasticity," *Journal of Econometrics*, 97, 179-188.

Kim, T.H. and White, H. (2003). “Estimation, Inference, and Specification Testing for Possibly Misspecified Quantile Regressions,” in T. Fomby and R.C. Hill, eds., *Maximum Likelihood Estimation of Misspecified Models: Twenty Years Later*, 107-132. New York (NY): Elsevier.

Koenker, R.W. (2005), *Quantile Regression*, New York: Cambridge University Press.

Koenker, R.W. and Bassett Jr., G.S. (1978), “Regression Quantiles,” *Econometrica*, 46, 33-50.

Koenker, R.W. and Bassett Jr., G.S. (1982), “Robust Tests for Heteroscedasticity Based on Regression Quantiles,” *Econometrica*, 50, 43-61.

Machado, J.A.F. and Parente, P. (2005), “Bootstrap Estimation of Covariance Matrices via the Percentile Method,” *The Econometrics Journal*, 8, 70-78.

Machado, J.A.F. and Santos Silva, J.M.C. (2000), “Glejser’s Test Revisited,” *Journal of Econometrics*, 97, 189-202.

Parente, P.M.D.C. and Santos Silva, J.M.C. (2013). “Quantile Regression with Clustered Data”, Department of Economics, University of Essex, Discussion Paper No. 728.

Powell, J.L. (1984), “Least Absolute Deviation Estimation for the Censored Regression Model,” *Journal of Econometrics*, 25, 303-325.

Rogers, W.H. (1993), “sg11.2: Calculation of Quantile Regression Standard Errors,” *Stata Technical Bulletin* 13, 18-19.

Wooldridge, J.M. (2009), *Introductory Econometrics*, 4th edition, Mason (OH): South Western.