Robust covariance estimation for quantile regression

J. M.C. Santos Silva
School of Economics, University of Surrey

UK STATA USERS' GROUP, 21st MEETING,
10 Septeber 2015

1. Summary

- Quantile regression (Koenker and Bassett, 1978) is increasingly used by practitioners, but there are still some **misconceptions** about how difficult it is to obtain valid standard errors in this context.
- In this presentation I discuss the estimation of the covariance matrix of the quantile regression estimator, focusing special attention on the case where the regression errors may be **heteroskedastic and/or "clustered"**.
- **Specification tests** to detect heteroskedasticity and intra-cluster correlation are also discussed.
- The presentation concludes with a brief description of `qreg2`, which is a **wrapper** for `qreg` that implements all the methods discussed in the presentation.

- For $0 < \alpha < 1$, the α-th quantile of y given x is defined by

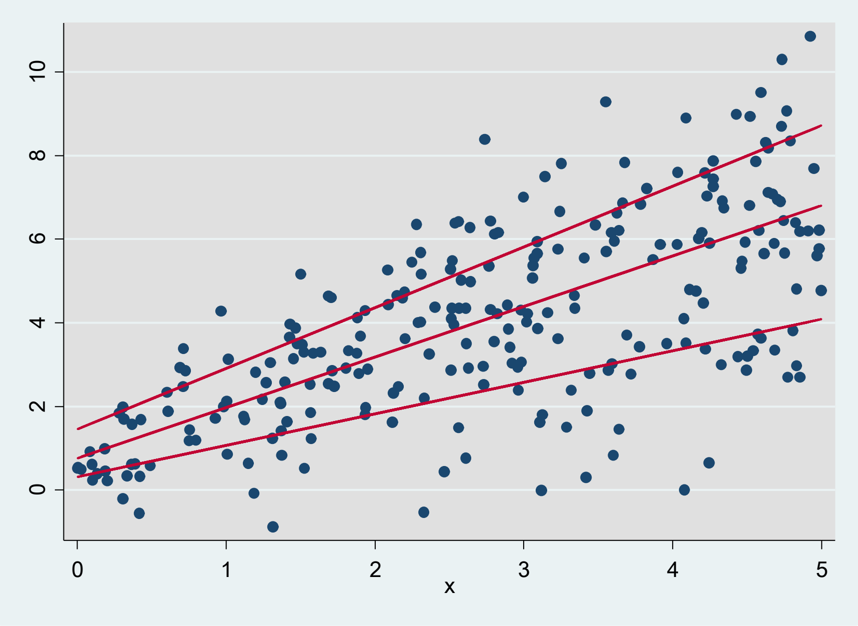
$$\mathrm{Q}_y(\alpha|x) = \min\{\eta | P(y \leq \eta|x) \geq \alpha\}.$$

- Assume that $\mathrm{Q}_y(\alpha|x)$ is linear, so that

$$\mathrm{Q}_y(\alpha|x) = x'\beta(\alpha),$$

which is equivalent to

$$y = x'\beta(\alpha) + u(\alpha); \qquad \mathrm{Q}_{u(\alpha)}(\alpha|x) = 0.$$



10
8
6
4
2
0
0
1
2
3
4
5
x

- Let the data be $\{(y_{gi}, x_{gi}),\ \ g = 1, \ldots, G,\ i = 1, \ldots, n_g\}$, where g indexes a set of G **clusters**, each with n_g elements (for simplicity, we set $n_g = n$).
- It is assumed that the disturbances are **conditionally independent** across clusters (but can be correlated within clusters).
- Note that for $n_g \equiv 1$ we have the usual (heteroskedastic) case.
- So the model to be estimated is:

$$y_{gi} = x_{gi}' \beta(\alpha) + u(\alpha)_{gi}\,.$$

- **Examples include**: Cross-sectional regression with clustered data (by regions, industry, etc.), Pooled quantile regression, Quantiles with correlated random effects.

- $\beta(\alpha)$ can be estimated as

$$\hat{\beta}(\alpha) = \arg\min_b \frac{1}{G} \sum_{g=1}^{G} \left\{ \sum_{y_{gi} \geq x_{gi}' b} \alpha \left| y_{gi} - x_{gi}' b \right| + \sum_{y_{gi} < x_{gi}' b} (1-\alpha) \left| y_{gi} - x_{gi}' b \right| \right\},$$

- $\hat{\beta}(\alpha)$ is usually estimated by **linear programming** methods.
- Asymptotic theory is **non-standard** because the objective function is not differentiable.

- It is possible to show that (Parente and Santos Silva, 2016):

$$\sqrt{G}\left(\hat{\beta}(\alpha) - \beta(\alpha)\right) \xrightarrow{D} \mathcal{N}\left(0, \mathrm{B}^{-1}\mathrm{A}\mathrm{B}^{-1}\right).$$

where

$$\mathrm{A} = \mathrm{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} x_{gi}x_{gj}'\left(\alpha - \mathrm{I}\left[u_{gi} < 0\right]\right)\left(\alpha - \mathrm{I}\left[u_{gj} < 0\right]\right)\right],$$

$$\mathrm{B} = \sum_{i=1}^{n} \mathrm{E}[x_{gi}x_{gi}' f(0|x_{gi})]$$

- Notice that the **asymptotic** results depend on $G \to \infty$.

4. Robust covariance matrix estimation

- One way to perform robust inference is to use **bootstrap**.
- This, however, can be quite **expensive** especially for large models for large datasets.
- Parente and Santos Silva (2016) show that it is possible to obtain consistent estimators of A and B :

$$\widehat{\mathrm{A}} = \frac{1}{G}\sum_{g=1}^{G}\sum_{i=1}^{n}\sum_{j=1}^{n} x_{gi}x'_{gj}\psi_{\alpha}(i)\psi_{\alpha}(j),$$

$$\widehat{\mathrm{B}} = \frac{1}{2\delta_G G}\sum_{g=1}^{G}\sum_{i=1}^{n}\mathbf{1}\left(-\delta_G \leq \left(y_{gi} - x'_{gi}\hat{\beta}(\alpha)\right) \leq \delta_G\right) x_{gi}x'_{gi},$$

$$\psi_{\alpha}(i) = \alpha - \mathrm{I}\left[\left(y_{gi} - x'_{gi}\hat{\beta}(\alpha)\right) < 0\right],$$

where δ_G is a **bandwidth** parameter.

- As in Koenker (2005, p. 81), we can define

$$\delta_G = \kappa \left[\Phi^{-1} \left(\alpha + h_G \right) - \Phi^{-1} \left(\alpha - h_G \right) \right],$$

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

$$h_G = \left(nG \right)^{-1/3} \left(\Phi^{-1} \left(1 - \frac{0.05}{2} \right) \right)^{2/3} \left(\frac{1.5 \left(\phi \left(\Phi^{-1} \left(\alpha \right) \right) \right)^2}{2 \left(\Phi^{-1} \left(\alpha \right) \right)^2 + 1} \right)^{1/3},$$

and κ is a **robust estimate of scale**.

- For example, κ can be defined as the MAD (median absolute deviation) of the α-th quantile regression residuals.

- When there is **no intra-cluster** correlation the proposed covariance estimator is **equivalent** to a standard "heteroskedasticity robust" estimator (see Powell, 1984, Chamberlain, 1994, and Kim and White, 2003).
- This is also the case when $n_g \equiv 1$.
- When the **errors are i.i.d.**, the estimator is **equivalent** to the one originally proposed by Koenker and Bassett, 1978).
- Specification tests can be used to detect intra-cluster correlation and heteroskedasticity.

- Parente and Santos Silva (2016) proposed a test to check for **intra-cluster correlation**.
  - Jeff Wooldridge proposed a similar test based on the OLS residuals.
  - These are robust versions of Breusch and Pagan's (1980) error components test
- Machado and Santos Silva (2000) proposed a test to check for **heteroskedasticity** in quantile regression.
  - For $\alpha = 0.5$, this is the well-known Glejser (1969) test for heteroskedasticity.
- Simulation results suggest the tests have good performance both under the null and under the alternative.

6. The `qreg2` command

- `qreg2` is a wrapper for `qreg` which estimates quantile regression and reports robust standard errors and t-statistics.
- By default the **standard errors** are asymptotically valid under **heteroskedasticity** and misspecification.
    - Standard errors that are also robust to intra-cluster correlation can be obtained with the option `cluster`.
- By default, the Machado-Santos Silva (2000) **test for heteroskedasticity** is reported.
    - When the option cluster is used the Parente-Santos Silva (2016) test for intra-cluster correlation is reported.

`qreg2 depvar [indepvars] [if] [in] [weight] [, options]`

`quantile(#)`: estimates # quantile; default is quantile(.5)

`cluster(clustvar)`: standard errors are computed allowing for intra-cluster correlation

`mss(varlist)`: use varlist in the MSS heteroskedasticity test

`silverman`: uses Silverman's rule-of-thumb as a scaling factor for the bandwidth

`epsilon(#)`: controls the number of residuals set to zero; default is epsilon(1e-7)

- `qreg` has an option to compute robust standard errors and t-statistics (but not clustered-robust).
  - However, it is not clear to me how this is implemented.
  - Simulations suggest that our estimator performs much better.
- The discussion of quantile (median) regression in the Stata manual could be much improved.
  - The comparison with `regress` and `rreg` is very misleading.

Further reading

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

- 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. (2005). *Quantile Regression*, Cambridge: Cambridge University Press.

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

- Machado, J.A.F., Parente, P.M.D.C., and Santos Silva, J.M.C. (2013). `qreg2`: Stata module to perform quantile regression with robust and clustered standard errors, Statistical Software Components S457369, Boston College Department of Economics.

- 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. (2016). "Quantile Regression with Clustered Data," *Journal of Econometric Methods*, forthcoming.

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