Quantile regression: Basics and recent advances

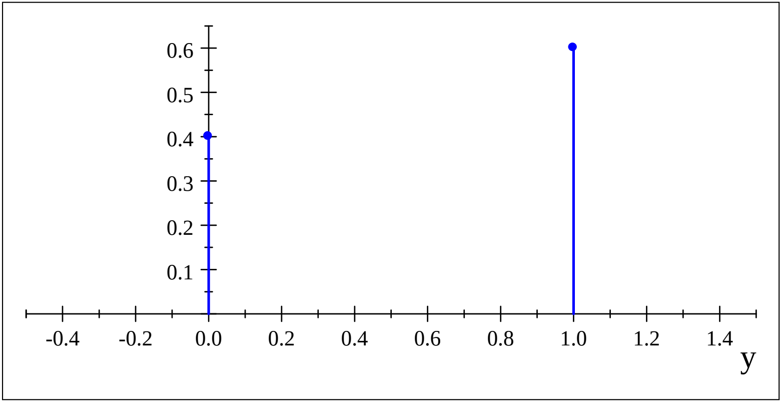
J. M.C. Santos Silva
University of Surrey

2019 UK Stata Conference
06/09/19

- Quantile regression (Koenker and Bassett, 1978) is increasingly used by practitioners but it is still not part of the standard econometric/statistics courses.
- Road map:
  - general introduction to quantile regression
  - two topics from recent research:
    - models with time-invariant individual ("fixed effects") effects
    - structural quantile function.
- I will present the approach to these problems proposed by Machado and Santos Silva (2019), and illustrate the use of the corresponding Stata commands `xtqreg` and `ivqreg2`.

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

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



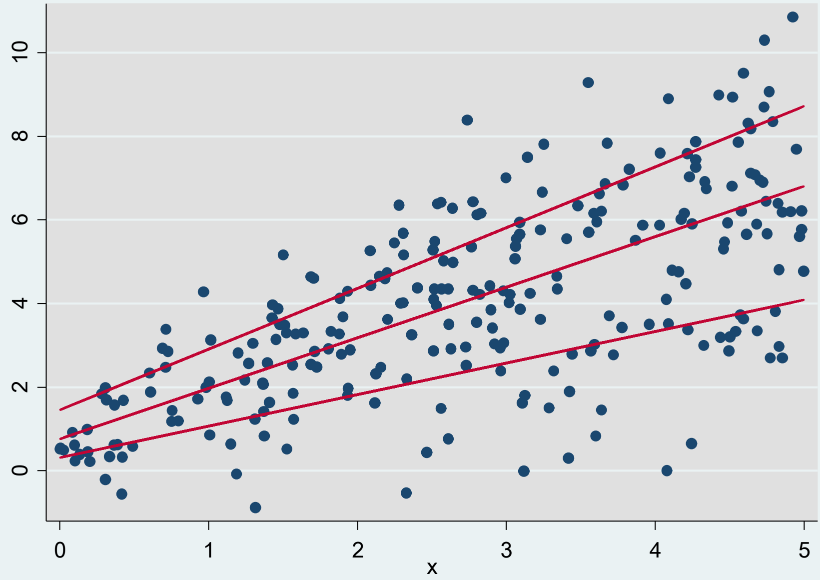
Bernoulli probability mass function with $\Pr(y = 1) = 0.6$

3. Basics of quantile regression

- Quantile regression estimates $\mathrm{Q}_y(\tau|x)$.
- Throughout we assume linearity: $\mathrm{Q}_y(\tau|x) = x'\beta(\tau)$.
- With linear quantiles, we can write

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

  - Note that the **errors** and the **parameters** depend on τ.
  - For $\tau = 0.5$ we have the median regression.
  - We need to restrict the **support** of x to ensure that quantiles do not cross.



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

- The estimator of $\beta(\tau)$ is defined by

$$\hat{\beta}(\tau) = \arg\min_b \frac{1}{n} \left\{ \sum_{y_i \geq x_i' b} \tau \left| y_i - x_i' b \right| + \sum_{y_i < x_i' b} (1 - \tau) \left| y_i - x_i' b \right| \right\}.$$

- The **F.O.C**. can be written as

$$\frac{1}{n} \sum_{i=1}^{n} \left(\left(\tau - \mathbf{1} \left(\left(y_i - x_i' \hat{\beta}(\tau) \right) < 0 \right) \right) \right) x_i = 0.$$

- $\hat{\beta}(\tau)$ is **invariant** to perturbations of y_i that do not change the sign of $\left(y_i - x_i' \hat{\beta}(\tau) \right)$.
- $\hat{\beta}(\tau)$ can be estimated by **linear programming** (see qreg).

- Asymptotic theory is **non-standard** because the objective function is not differentiable.
- However, under certain regularity conditions, $\hat{\beta}(\tau)$ has standard properties:

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

$$D = \mathrm{E}\left[f_{u(\tau)}(0|x_i)\, x_i x_i'\right], \quad A = \mathrm{E}\left[(\tau - \mathbf{1}(u(\tau)_i \leq 0))^2 x_i x_i'\right].$$

- It is possible to estimate A and D under different assumptions (see `qreg` and `qreg2`).

- The main advantage of quantile regression is the **informational gains** they provide.
- Quantiles are "**robust**" measures of location and are estimated using a "**robust**" estimator.
- Quantiles and means have very **different** properties.
  - Quantiles are not **additive**; the quantile of the sum is not the sum of the quantiles.
  - Quantiles are **equivariant** to non-decreasing transformations; for example, if y_i is non-negative with
  $$\mathrm{Q}_{y_i}(\tau|x_i) = \exp\left(x_i'\beta\left(\tau\right)\right),$$
  then,
  $$\mathrm{Q}_{\ln(y_i)}(\tau|x_i) = x_i'\beta\left(\tau\right).$$

- The plain-vanilla quantile regression estimator has been extended to different settings:
    - Censored regression; Powell (1984)
    - Binary data; Manski (1975, 1985), Horowitz (1992)
    - Ordered data; M.-j. Lee (1992)
    - Count data; Machado and Santos Silva (2005)
    - Corner-solutions data; Machado, Santos Silva, and Wei (2016)
    - Clustering; Parente and Santos Silva (2016)
- Two areas of active research are:
    - quantile regressions with time-invariant individual ("fixed") effects, and
    - structural quantile function.

- Consider a location-scale model

$$y_i = x_i'\beta + \left(x_i'\gamma\right) u_i,$$

where x_i and u_i are independent and $\Pr\left(x_i'\gamma > 0\right) = 1$.

- In this case the mean and all conditional quantiles are linear

$$\begin{aligned} \mathrm{Q}_y(\tau|x) &= x_i'\beta + \left(x_i'\gamma\right)\mathrm{Q}_u(\tau|x_i) \\ &= x_i'\beta\left(\tau\right) \end{aligned}$$

$$\beta\left(\tau\right) = \beta + \gamma \mathrm{Q}_u(\tau).$$

- In this model, the information provided by β, γ, and $\mathrm{Q}_u(\tau)$ is equivalent to the information provided by regression quantiles.

- Machado and Santos Silva (2019) noted that, assuming $E(U) = 0$ and using the normalization $E(|U|) = 1$, β and γ are identified by conditional expectations:

$$E\left[y_i|x_i\right] = \beta_0 + \beta_1 x_i$$

$$E\left[\left|y_i - \beta_0 - \beta_1 x_i\right| |x_i\right] = \gamma_0 + \gamma_1 x_i$$

- $\mathrm{Q}_u(\tau|x_i)$ can be estimated from the scaled errors

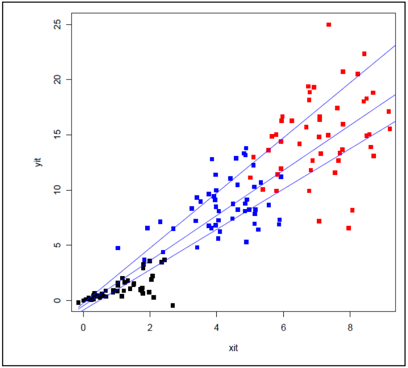
$$\frac{y_i - \beta_0 - \beta_1 x_i}{\gamma_0 + \gamma_1 x_i}$$

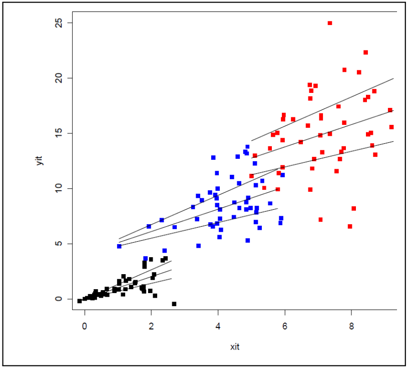
- This provides a way to estimate quantile regression using two OLS regressions and the computation of a univariate quantile.

- Suppose now that we are interested in estimating

$$Q_{y_{it}}\left(\tau | x_{it}, \eta_i\right) = x_{it}'\beta\left(\tau\right) + \eta\left(\tau\right)_i \text{, with } i = 1, \ldots, n;\ t = 1, \ldots, T.$$

- As in mean regression, "**fixed effects"** can be important.





- Estimation of quantile regression with fixed effects is difficult because there is **no transformation** that can be used to eliminate the incidental parameters.
- Therefore, due to the **incidental parameter problem**, consistency requires that both $n \rightarrow \infty$ and $T \rightarrow \infty$.
- For fixed T, the only realistic option is the "**correlated random effects**" (Mundlak) estimator; see Abrevaya and Dahl (2008).
- Roger Koenker (2004) and Canay (2011) proposed estimators based on the assumption that $\eta(\tau)_i = \eta_i$ but this goes against the spirit of quantile regression.

- Kato, Galvão, and Montes-Rojas (2012) studied the properties of quantile regression in a model where the fixed effects are explicitly included as **dummies**.
- The estimator is consistent and asymptotically normal when both $n \rightarrow \infty$ and $T \rightarrow \infty$ with $n^2 \left[\ln(n)\right]^3 / T \rightarrow 0$.
- This is an issue because in many applications n is much larger than T (e.g. for $T = 40$, $n = 100$, $n^2 \left[\ln(n)\right]^3 / T = 24,416$).
- An alternative is to use the quantiles-via-moments estimator.

- Consider the location-scale model for panel data

$$y_{it} = \alpha_i + x_{it}'\beta + (\delta_i + x_{it}'\gamma) u_{it}$$

$$\eta(\tau)_i = \alpha_i + \delta_i \mathrm{Q}_u(\tau), \qquad \beta(\tau) = \beta + \gamma \mathrm{Q}_u(\tau),$$

  where x_i and u_i are independent and $\Pr((\delta_i + x_{it}'\gamma) > 0) = 1$.

- Estimation is performed using two fixed effects regressions (`xtreg`) and computing a univariate quantile.
- Consistency requires $(n, T) \rightarrow \infty$ with $n = o(T)$.
- For fixed T the estimator will have a bias but:
  - simulations suggest that the bias is negligible for $n/T \leq 10$;
  - the bias can be removed using **jackknife**.
- The estimator is implemented in the `xtqreg` command (available from SSC)

```
xtqreg depvar [indepvars] [if] [in] [, options]
```

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

`id`: specifies the variable defining the panel

`ls`: displays the estimates of the location and scale parameters

- Suppose that we have a structural relationship defined by

$$\begin{aligned} y &= d\alpha + x'\beta + u, \\ d &= \delta(x, z, v) \end{aligned}$$

where v may not be independent of u

- We are interested in

$$S_y(\tau|d, x) = d\alpha(\tau) + x'\beta(\tau),$$

the structural quantile function such that:

  - $\Pr[y < S_y(\tau|d, x)\,|z, x] = \tau$,
  - $S_y(\tau|d, x) = Q_y(\tau|z, x) \neq Q_y(\tau|d, x)$.

- **Chernozhukov and Hansen (2008)** propose an estimator of $S_Y(\tau|d,x)$ based on the observation that

$$Q_{y-d\alpha(\tau)}(\tau|z,x) = x'\beta(\tau) + z\gamma(\tau)$$

with $\gamma(\tau) = 0$.
- We can implement the estimator by:
  - estimating $\beta(\tau)$ and $\gamma(\tau)$ for a range of values of $\alpha(\tau)$
  - and choosing as estimates the ones corresponding to the value of $\alpha(\tau)$ for which $\gamma(\tau)$ is in some sense closer to zero.
- Chernozhukov and Hansen (2008) prove the consistency and asymptotic normality of the estimator.
- The estimator is difficult to implement when there are multiple endogenous variables, but there have been a number of recent **developments** on this.

- Again, the quantile-via-moments estimator can be useful.
- Consider a location-scale structural relationship

$$y = d\alpha + x'\beta + \left(d\delta + x'\gamma\right) u, \quad d = \delta\left(x, z, v\right),$$

where v may not be independent of u but u is independent of x and z.

- Because $S_y(\tau|d,x)$ is such that $\Pr\left[y < S_y(\tau|d,x)|z,x\right] = \tau$,

$$\begin{aligned} S_y(\tau|d,x) &= d\alpha + x'\beta + \left(d\delta + x'\gamma\right) \mathrm{Q}_u(\tau) \\ &= d\left(\alpha + \delta \mathrm{Q}_u(\tau)\right) + x\left(\beta + \gamma \mathrm{Q}_u(\tau)\right). \end{aligned}$$

- GMM can be used to estimate the structural parameters:

$$E\left[\left(\frac{y_i - d\alpha - x'\beta}{d\delta + x'\gamma}\right)\middle| z_i\right] = 0,$$
$$E\left[\left(\frac{|y_i - d\alpha - x'\beta|}{d\delta + x'\gamma} - 1\right)\middle| z_i\right] = 0.$$

- $\mathrm{Q}_u(\tau)$ can be estimated from the standardized errors

$$\left(y_i - d\hat{\alpha} - x'\hat{\beta}\right) / \left(d\hat{\delta} + x'\hat{\gamma}\right).$$

- The estimator has the usual properties.

- The estimator is implemented in the `ivqreg2` command (available from SSC) `ivqreg2 depvar [indepvars] [if] [in] [, options]`

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

`instruments(varlist)`: list of instruments, including control variables; by default no instruments are used and restricted quantile regression is performed

`ls`: displays the estimates of the location and scale parameters

- Quantile regression can be very useful and it is now easy to implement in a variety of cases.
- In some contexts, however, quantile regression can be challenging.
- The Method of Moments-Quantile Regression estimator can be useful in some of these cases.
- `xtqreg` and `ivqreg2` make it easy to estimate quantile regressions with "fixed effects" or endogenous variables.

References

- Abrevaya, J. and Dahl, C.M. (2008). "The Effects of Birth Inputs on Birthweight," *Journal of Business & Economic Statistics*, 26, 379-397.
- Canay, I.A. (2011). "A Simple Approach to Quantile Regression for Panel Data," *Econometrics Journal*, 14, 368-386.
- Chernozhukov, V. and Hansen, C. (2008). "Instrumental Variable Quantile Regression: A Robust Inference Approach," *Journal of Econometrics*, 142, 379–398.
- Horowitz, J.L. (1992). "A Smooth Maximum Score Estimator for the Binary Response Model", *Econometrica*, 60, 505-531.
- Kato, K., Galvão, A.F. and Montes-Rojas, G. (2012). "Asymptotics for Panel Quantile Regression Models with Individual Effects," *Journal of Econometrics*, 170, 76–91.

- Koenker, R. (2004). "Quantile Regression for Longitudinal Data," *Journal of Multivariate Analysis* 91, 74–89.
- Koenker, R. and Bassett Jr., G.S. (1978). "Regression Quantiles," *Econometrica*, 46, 33-50.
- Lee, M.-j. (1992). "Median Regression for Ordered Discrete Response," *Journal of Econometrics*, 51, 59-77.
- Machado, J.A.F. and Santos Silva, J.M.C. (2005), "Quantiles for Counts", *Journal of the American Statistical Association*, 100, 1226-1237.
- Machado, J.A.F., Santos Silva, J.M.C., and Wei, K. (2016), "Quantiles, Corners, and the Extensive Margin of Trade," *European Economic Review*, 89, 73–84.

- Machado, J.A.F. and Santos Silva, J.M.C. (2019), "Quantiles via Moments," *Journal of Econometrics*, forthcoming.
- Manski, C.F. (1975). "Maximum Score Estimation of the Stochastic Utility Model of Choice", *Journal of Econometrics*, 3, 205-228.
- Manski, C.F. (1985). "Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator", *Journal of Econometrics*, 27, 313-333.
- Parente, P.M.D.C. and Santos Silva, J.M.C. (2016). "Quantile Regression with Clustered Data," *Journal of Econometric Methods*, 5, 1-15.
- Powell, J.L. (1984). "Least Absolute Deviation Estimation for the Censored Regression Model," *Journal of Econometrics*, 25, 303-325.