Quantile regression

Framework

Specification

Testing heteroskedasticity

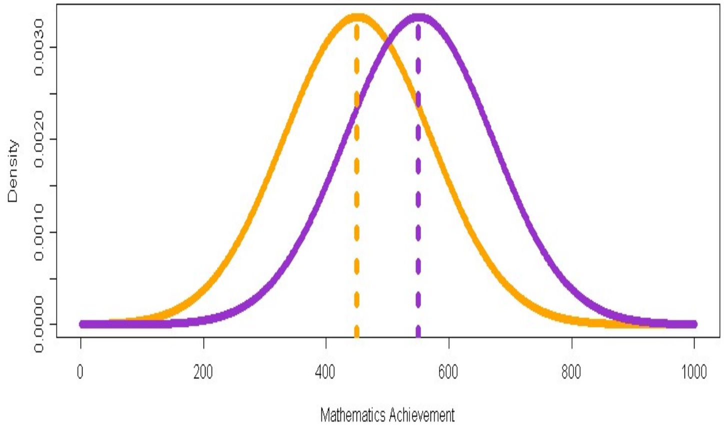
Estimation in STATA

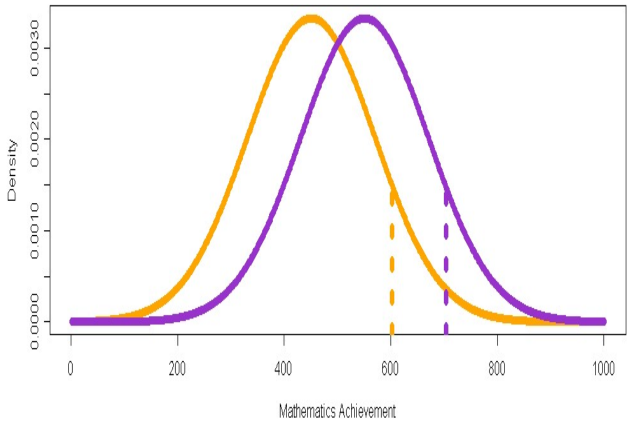
Quantiles for counts

Framework

Aim: modelling the relationship between a response and explanatory variables at different points of the conditional distribution of Y|X

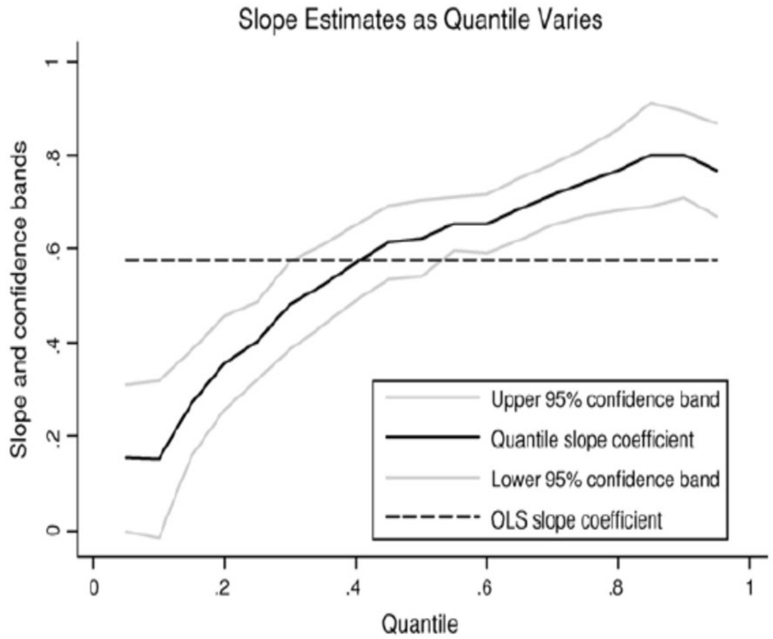
- Conditional mean models focus only on the mean E(Y|X)
- Median or least absolute deviation (LAD) is the most well known form of quantile regression
- Allows focusing in noncentral location

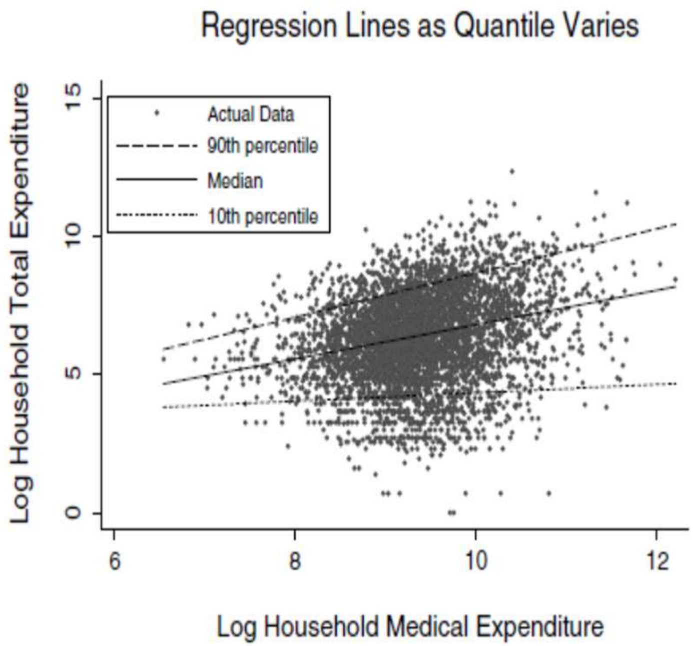




Quantile regression Framework

Figure CT, pp. 89, 90





- See also: http://www.econ.uiuc.edu/~roger/research/intro/jep.pdf and the corresponding survey paper: Koenker & Hallok (2001)

Advantages of QR:

- Median is more robust to outliers that the mean

- No distribution is assumed for the error

- Analysis can be made at different locations, providing a richer characterization of the data

- Retransformation for recovering original scale is straightforward:
  - While for the mean $E[h(y)] \neq h[E(y)]$
  - $Q[h(y)] = h[Q(y)]$

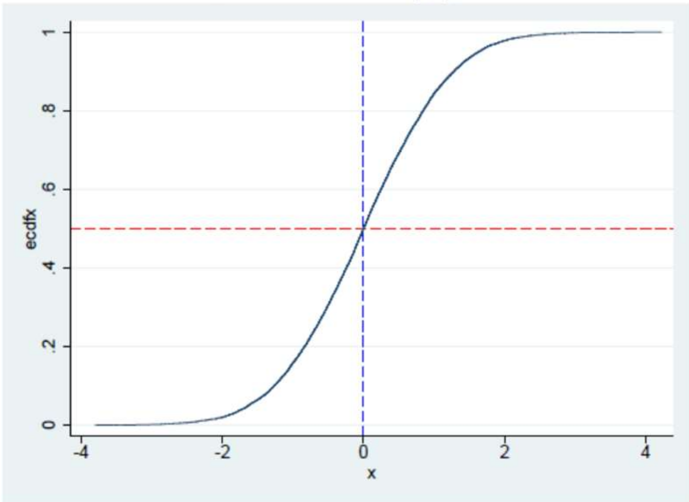
Quantile regression
Specification

Let $q, q \in (0,1)$ be a quantile. q splits the sample into a proportion q bellow and a proportion $(1-q)$ above

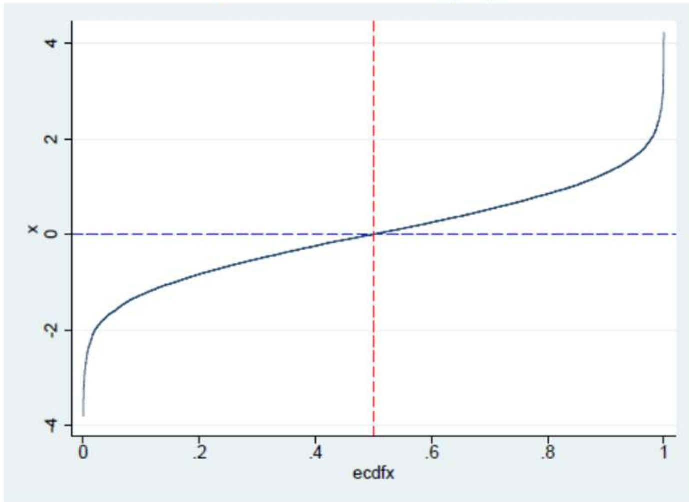
With a cdf $\mathrm{F}(y) = P(Y \leq y)$

- $\mathrm{F}(y_q) = q \rightarrow y_q = \mathrm{F}^{-1}(q)$
- For the median: $\mathrm{F}(y_{0.5}) = 0.5 \rightarrow y_{0.5} = \mathrm{F}^{-1}(0.5)$
- Conversely, for a given $y_{0.9} = 10 \rightarrow P(Y \leq 10) = 0.9$

E.c.d.f. $\mathrm{F}(y)$



Quantile $F^{-1}(\tau)$



Objective function

- While for
  - OLS: $\sum(y_i - x_i\beta)^2$ with $\widehat{E[Y|X]} = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots \hat{\beta}_k\, x_{ik}$
  - LAD: $\sum|y_i - x_i\beta|$ with $\widehat{Q_{0.5}[Y|X]} = \hat{\beta}_{0,0.5} + \hat{\beta}_{1,0.5} x_{i1} + \cdots \hat{\beta}_{k,0.5}\, x_{ik}$

- For QR, for a given β_q:

$$\sum_{i|y_i \geq x_i\beta_q} q\left|y_i - x_i\beta_q\right| + \sum_{i|y_i < x_i\beta_q} (1-q)\left|y_i - x_i\beta_q\right|$$

  - Uses asymmetric penalties, except for LAD ($q = 0.5$)
  - Yields an extremum estimator, but due to the nondifferenciable nature of the objective function uses linear programming (simplex) as optimization method

Asymptotic distribution

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N[0, A^{-1}BA'^{-1}]$$

where $A = plim \frac{1}{N}\sum_{i=1}^{N} f_{u_q}(0|x_i)x_i x_i'$, $\mathrm{B} = plim \frac{1}{N}\sum_{i=1}^{N} q(1-q)x_i x_i'$, and $f_{u_q}(0|x_i)$ is the conditional pdf of the error u_q evaluated at 0.

- To avoid obtaining an estimator for f_{u_q} the paired boostrap is usually employed

Conditional quantile function & partial effects

$$Q_q[y_i|x_i] = \beta_0^q + \beta_1^q X_{i1} + \cdots + \beta_k^q X_{ik} + F_{u_i}^{-1}(q)$$

- With hereroscedasticity, all partial effects will depend on X, which is contained in u_i

- With homoscedasticity $F_{u_i}^{-1}(q) = F_u^{-1}(q)$:

$$Q_q(y_i|x_i) = \left[\beta_0^q + F_u^{-1}(q)\right] + \beta_1^q X_{i1} + \cdots + \beta_k^q X_{ik}$$

partial effects will not depend on X, $F_u^{-1}(q)$ is included in the intercept

- Partial effects of dummies: assume that the individual remains in the same quantile

Transformation & retransformation

Consider the transformation $ln(y)$, which yields the conditional quantile function

$$Q_q[ln(y_i)|x_i] = \beta_{q0} + \beta_{q1}X_{i1} + \cdots + \beta_{qk}X_{ik} + F_{u_i}^{-1}(q)$$

- The corresponding quantile in the original scale is

$$Q_q(y_i|x_i) = exp\{Q_q[ln(y_i)|x_i]\}$$
$$= \exp\left(\beta_{q0} + \beta_{q1}X_{i1} + \cdots + \beta_{qk}X_{ik} + F_{u_i}^{-1}(q)\right)$$

- Under homoscedasticity, partial effects will depend on X but not on $F_{u_i}^{-1}(q)$

$$\nabla_{\beta_{qj}} Q_q(y_i|x_i) = \exp\left(\left[\beta_0^q + F_u^{-1}(q)\right] + \beta_{q1}X_{i1} + \cdots + \beta_{qk}X_{ik}\right)\beta_{qj}$$

Quantile regression
Testing heteroskedasticity

- Test BP in the framework of a linear model

  Consider

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + u_i$$

  - Estimate by OLS to obtain $\hat{u}^2$
  - Estimate the auxiliary regression $\hat{u}^2 = \gamma_0 + \gamma_1 X_1 + \cdots + \gamma_k X_k + e$
  - Test the joint significance of regressors (F or LM)

- Heteroskedasticity may often be removed by using the transformation ln(y): once heteroskesdasticity is eliminated the retransformation, under homoskedasticity, is simple

Quantile regression
Estimation in STATA

- Estimating QR with analytic standard errors (default quantile: 0.5). No vce() option available

Stata

qreg $Y\ X_1 \dots X_k$

qreg $Y\ X_1 \dots X_k$, quantile(.25)

- Estimating QR with bootstrap standard errors, retaining the assumption of independent errors but relaxing the assumption of identically distributed errors. Analogous to robust standard errors in linear regression

Stata

set seed(123)

bsqreg $Y\ X_1 \dots X_k$, reps(400) quantile(.25)

- Estimating QR for several values of q simultaneously, allowing for differences between coefficients for different quantiles to be tested. Bootstrap standard errors are produced

Stata

set seed(123)

sqreg $Y\ X_1 \dots X_k$, reps(400) quantile(.25,.5,.75)

Quantile regression
Estimation in STATA

- Testing the equality of a regression coefficient at different quantiles (Wald test)

Stata

set seed(123)
sqreg Y X_1 … X_k, reps(400) quantile(.25,.5,.75)
test [q25=q50=q75]:X_1

- Illustrating the coefficient estimates over different quantiles. Install the command `grqreg` (`ssc install grqreg`)

Stata

qreg Y X_1 … X_k, quantile(.50) nolog
grqreg, cons ci ols olsci reps(40)

Seminal paper: Machado & Santos Silva (2005)

Let $y \in \{0,1,2,\dots\}$. The idea is transforming Y into a continuous variable, estimate as usual in QR, and then retransformate to the original scale

- Transformation ("jittering the count"):

$$z = y + u$$

where $u \sim U(0,1)$. The resulting conditional quantile is

$$Q_q(z_i|x_i) = q + exp\left(\beta_{q0} + \beta_{q1}X_{i1} + \cdots + \beta_{qk}X_{ik}\right)$$

which includes q as the lower limit of $Q_q(z_i|x_i)$, required due to the jittering

- The quantile $Q_q(z_i|x_i)$ is linearized: the dependent variable in QR is

$$\begin{cases} \ln(z-q) \; for \; z-q>0 \\ \ln(\varepsilon) \; for \; z-q \leq 0 \end{cases},$$

where ε is a small constant

- Independent replications are taken from u and the resulting coefficient estimates averaged, giving rise to $\bar{\hat{\beta}}_q$

Stata

ssc install qcount

qcount Y X_1 ... X_k, quantile(.50) reps(500)

Quantile regression
Quantiles for counts

- Retransformation to the original scale uses a ceilling function

$$Q_q(y_i|x_i) = \lceil Q_q(y_i|x_i) - 1 \rceil$$

where $\lceil . \rceil$ denotes the smallest integer $\geq \left(Q_q(y_i|x_i) - 1\right)$

Taking into account the averaging of the coefficients, we obtain

$$\hat{Q}_q(y_i|x_i) = \lceil \hat{Q}_q(y_i|x_i) - 1 \rceil$$
$$= \left\lceil q + exp\left(\bar{\hat{\beta}}_{q0} + \bar{\hat{\beta}}_{q1} X_{i1} + \cdots + \bar{\hat{\beta}}_{qk} X_{ik}\right) - 1 \right\rceil$$