Semi and Nonparametric Econometrics

Part I: quantile regression

Xavier D'Haultfœuille

ENSAE - Paris Saclay Master

Outline

Introduction

Model and motivation

Interpreting quantile regressions

Inference in quantile regressions

Computational aspects

Quantile regressions with panel data

Quantile restrictions in nonlinear models

Outline

Introduction

Model and motivation

Interpreting quantile regressions

Inference in quantile regressions

Computational aspects

Quantile regressions with panel data

Quantile restrictions in nonlinear models

Brief history

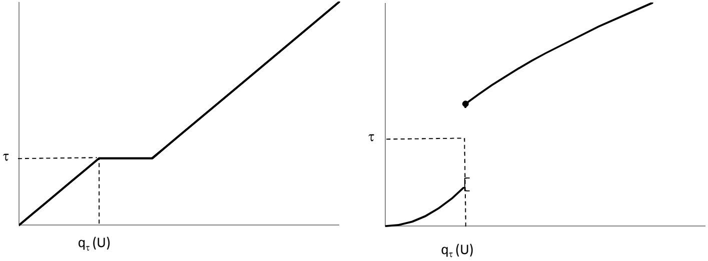
- Median regression is older than linear regression: introduced by Boscovitch in 1760, then Laplace (1789).
- Revisited by Edgeworth by the end of the 19th century. But overall and compared to OLS, totally forgotten for a long time.
- Brought up to date with Koenker's work, starting in the end of the 70's.
- Has gained popularity in applied economics by the end of the 90's, when people realize the importance of heterogeneity.

Basic definitions and properties

- The τ-th quantile ($\tau \in (0,1)$) of a random variable U is defined by

$$q_\tau(U) = \inf\{x / F_U(x) \geq \tau\},$$

where F_U denotes the distribution function of U. Note that when F_U is strictly increasing, $q_\tau(U) = F_U^{-1}(\tau)$. Otherwise, $q_\tau(U)$ satisfies for instance:



Basic definitions and properties

- The quantile function $\tau \mapsto q_\tau(U)$ is an increasing, left continuous function which satisfy, for all $a > 0$ and b:

$$q_\tau(aU + b) = aq_\tau(U) + b. \quad (1)$$

- Caution: $q_\tau(U + V) \neq q_\tau(U) + q_\tau(V)$ in general.
- Conditional quantiles are simply defined as:

$$q_\tau(Y|X) = \inf\{u / F_{Y|X}(u|X) \geq \tau\}.$$

- Similarly to conditional expectations, conditional quantiles are random variables (as they depend on the random variable X).
- Example: Y =monthly wage, $X = \mathbb{1}_{\text{male}}$. Then if median wages are 1,770 for men and 1,420 for women,

$$q_{0.5}(Y|X) = 1,420 + 350X.$$

Outline

Introduction

Model and motivation

Interpreting quantile regressions

Inference in quantile regressions

Computational aspects

Quantile regressions with panel data

Quantile restrictions in nonlinear models

The model

- Let $Y \in \mathbb{R}$ be the dependent variable and $X \in \mathbb{R}^p$ be the explanatory variables, including the intercept. We consider here a model of the form

$$Y = X'\beta_\tau + \varepsilon_\tau, \; q_\tau(\varepsilon_\tau | X) = 0. \tag{2}$$

  Equivalently, we have

$$q_\tau(Y|X) = X'\beta_\tau.$$

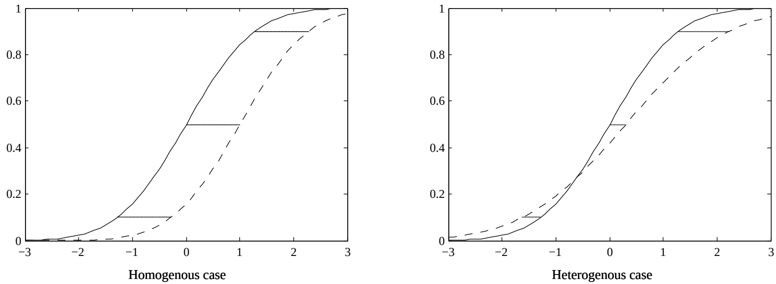
- This model is similar to the standard linear regression, except that we replace the conditional expectation $E(Y|X)$ by a conditional quantile.
- An important point is that β_τ depends on the τ we consider.

First motivation: measuring heterogenous effects

- The effect of a variable may not be the same for all individuals. Ignored in standard linear regressions, which focus on average effects.
- But this heterogeneity may be important for public policy.
- First example: the effect of a class size reduction may have an effect for low achieving students only $\Rightarrow$ may be an effective policy even if does not rise the average level by much.
- Second example: the effect of an increase of the minimum wage (MW) on wages is likely large on low wages and far smaller on other wages (still with some diffusion effects) $\Rightarrow$ effect on inequalities.
- Formally, $\tau \mapsto \beta_{MW,\tau}$ decreases towards 0 as $\tau \uparrow 1$.

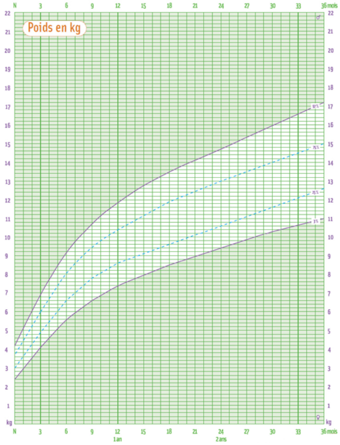
First motivation: measuring heterogenous effects

- Graphical interpretation with Y =wage and $X = (1, \mathbb{1}_{\text{male}})'$.
- In the left plot, the wage gap is similar for each quantile $\Rightarrow \beta_{\text{male},\tau}$ does not depend on τ.
- In the right plot, the wage gap is a function of the quantile we consider. $\beta_{\text{male},\tau}$ is first negative, then positive.



First motivation: measuring heterogenous effects

- Second example: growth curve. 1-year children may gain from $\simeq$ 200 grams per month (bottom curve) to $\simeq$ 400 grams per month (top curve).
- Formally, this means that the age (in month) coefficient satisfies $\beta_{0.03} \simeq 0.2$ and $\beta_{0.97} \simeq 0.4$.
- Note that here, the effect of age is not linear. One would have to add age^2 in the quantile regression.



Interpretation of the heterogeneity

- Consider for instance the "location-scale" model:

$$Y = X'\beta + (X'\gamma)\varepsilon,$$

where ε is independent of X and we suppose $X'\gamma \geq 0$.

- Restriction here: the shape of Y given $X = x$ is the same for all x. Example: wages are (approximately) lognormal for all subpopulations.
- In this case, by (1):

$$q_\tau(Y|X) = X'\left(\beta + \gamma q_\tau(\varepsilon)\right).$$

Hence, (2) holds with $\beta_\tau = \beta + \gamma q_\tau(\varepsilon)$.

- In the location-scale model with $E(\varepsilon) = 0$, $\beta_{OLS} = \beta$. Running OLS, we miss the fact that the effect of X differs according to quantiles of the unobserved variable ε.

Interpretation of the heterogeneity

- Consider the more general random coefficient model:

$$Y = X'\beta_U, \quad U|X \sim \mathcal{U}[0,1], \tag{3}$$

  where for all x, $\tau \mapsto x'\beta_\tau$ is suppose to be strictly increasing.

- We thus consider a random coefficient model with a *unique* underlying random variable, which determines the ranking of each individual in terms of Y, within his "subpopulation" X (e.g., unobserved ability in the class size example).

- Under these assumptions,

$$P(Y \le X'\beta_\tau|X) = P(X'\beta_U \le X'\beta_\tau|X) = P(U \le \tau|X) = \tau.$$

  In other words, (2) holds for all $\tau \in (0,1)$.

Second motivation: robustness to outliers and to heavy tails

- We want to draw inference on a variable Y^* but observe, instead of Y^*, "contaminated" data $Y = CX'\alpha + (1-C)Y^*$, where $C = 1$ if data are contaminated, 0 otherwise (C is unobserved). We suppose that $p = P(C = 1)$ is small but $X'\alpha$ is large.
- Consider first a linear model $E(Y^*|X) = X'\beta$.
  Then, instead of β, OLS estimate $(1-p)\beta + p\alpha$. The bias $p(\alpha - \beta)$ may be large even if p is small.
- Now consider the quantile model $q_\tau(Y^*|X) = X'\beta_\tau$.
  In this case, $q_\tau(Y|X) = X'\beta_{\frac{\tau}{1-p}}$ so instead of β_τ, we estimate $\beta_{\frac{\tau}{1-p}}$.
  It is independent of α and will typically be close to β_τ. If some components of β_τ are independent of τ (homogenous effects), the contamination does not affect their estimation.

Second motivation: robustness to outliers and to heavy tails

- In a similar vein, consider a linear model

$$Y = X'\beta + \varepsilon, \; X \perp\!\!\!\perp \varepsilon.$$

- If ε is symmetric around zero, we can estimate β with OLS or median regression but we may prefer to estimate it with median regression if ε has heavy tails.
- Indeed, if $E(|\varepsilon|) = \infty$ (examples ?), OLS are inconsistent whereas the median is always defined. One can show that the estimator of the median regression is consistent.
- Useful in finance, insurance...

Outline

Introduction

Model and motivation

Interpreting quantile regressions

Inference in quantile regressions

Computational aspects

Quantile regressions with panel data

Quantile restrictions in nonlinear models

Distinguishing several effects

- The interpretation of β in a linear regression $E(Y|X) = X'\beta$ is simple:
$$\beta = \frac{\partial E(Y|X=x)}{\partial x} = E\left[\frac{\partial E(Y|X)}{\partial x}\right].$$
β is thus the average marginal effect of X on Y, either for those s.t. $X = x$ or for the whole population.
- Similarly, β_τ in a quantile regression satisfies
$$\beta_\tau = \frac{\partial q_\tau(Y|X=x)}{\partial x} = E\left[\frac{\partial q_\tau(Y|X)}{\partial x}\right],$$
which is the average marginal effect of X on the conditional quantile of Y.
- It is often tempting to also interpret β_τ as the effect of a small variation in X for individuals at the τ-th quantile of $Y|X = x$.
- But this is possible only under a rank invariance condition.

Individual vs aggregated effects

- To better understand this condition, consider the following potential outcome model:

$$Y(x) = x'\beta_{U_x}, \quad \text{with } U_x \sim \mathcal{U}[0,1] \text{ and } \tau \mapsto x'\beta_\tau \uparrow . \qquad (4)$$

- $Y(x)$ is the outcome an individual would have if his covariate was equal to x. Observed outcome: $Y = Y(X)$.
- Example: $Y(x)$ =wage an individual would get if his education level was equal to $X = x$.
- In this model, for each possible x, an individual "draws" a random term U_x, which then corresponds to his ranking in the distribution of $Y(x)$.
- Note that under the assumptions above, we have $q_\tau(Y|X) = X'\beta_\tau$.

Individual vs aggregated effects

- In this setting, for someone at $U_x = \tau$, we have
$$\frac{dY(x)}{dx} = \beta_\tau + x' \frac{d\beta_\tau}{d\tau} \frac{dU_x}{dx} \neq \beta_\tau \text{ in general.}$$
- But the equality holds if $U_x = U$ for all x, i.e. under a *rank invariance* condition: individuals have the same ranking in the distribution of $Y(x)$, whatever x.
- Sometimes reasonable: e.g. $X =$minimum wage.
- Sometimes harder to swallow: e.g. $X =$education.
- Under the rank invariance condition, β_τ can be interpreted as the effect on Y of an increase of one unit of X among individuals at the rank τ in the distribution of $Y|X = x$.

An illustrative example

- Suppose we are interested in the effect of a new pedagogical method on test score achievement.
- Let $X = \mathbb{1}\{\text{new method}\}$ and $Y(x) = \text{test score when } X = x$.
- We use a randomized experiment to evaluate the effect of this method. We observe X and $Y = Y(X)$.
- Suppose we have 5 equal-sized groups of students who react differently to this method. For simplicity, students are supposed to be identical in terms of $(Y(0), Y(1))$ within each group.

An illustrative example

Using the table below, determine:

- the effect of the new method on the median score.
- the effect of the new method on individuals initially at the median;
- the median effect of the new method.
- what parameter(s) a median regression of Y on X identifies.

Group	$Y\|X = 0$	$Y\|X = 1$
A	1	4
B	2	6
C	4	3
D	7	7
E	9	10

Outline

Introduction

Model and motivation

Interpreting quantile regressions

Inference in quantile regressions

Computational aspects

Quantile regressions with panel data

Quantile restrictions in nonlinear models

The check functions

- It is easy to estimate the τ-th quantile of a random variable Y: we simply consider the order statistic $Y_{(1)} < ... < Y_{(n)}$ and estimate $q_\tau(Y)$ by
$$\widehat{q}_\tau(Y) = Y_{(\lceil n\tau \rceil)},$$
where $\lceil n\tau \rceil \geq n\tau > \lceil n\tau \rceil - 1$.
- It does not seem obvious, however, to generalize this to quantile regression.
- The key observation is the following property:

Proposition

Consider the check function $\rho_\tau(u) = (\tau - \mathbb{1}\{u < 0\})u$. Then:

$$q_\tau(Y) \in \arg\min_a E\left[\rho_\tau(Y - a)\right].$$

The check functions

Proof: suppose for simplicity that Y admits a density f_Y. Then we have

$$E\left[\rho_\tau(Y-a)\right] = \tau(E(Y)-a) - \int_{-\infty}^{a}(y-a)f_Y(y)dy.$$

This function is differentiable, with

$$\frac{\partial E\left[\rho_\tau(Y-a)\right]}{\partial a} = -\tau-(a-a)f_Y(a)+\int_{-\infty}^{a}f_Y(y)dy = F_Y(a)-\tau.$$

This function is increasing, thus $a \mapsto E\left[\rho_\tau(Y-a)\right]$ is convex and reaches its minimum at $q_\tau(Y)$ $\square$

The check functions

- The minimum need not be unique (there may be several solutions to $F_Y(a) = \tau$). When Y is not continuous, there may be no solution to $F_Y(a) = \tau$ but we can still show that $q_\tau(Y)$ is a minimum of $E\left[\rho_\tau(Y - a)\right]$.
- The τ-th quantile minimizes the risk associated with the (asymmetric) loss function $\rho_\tau(.)$. This is similar to the expectation which minimizes the risk corresponding to the L^2-loss :

$$E(Y) = \arg\min_a E\left[(Y - a)^2\right].$$

- Similarly to conditional expectation, we can extend the reasoning to conditional quantiles. We have

$$q_\tau(Y|X = x) \in \arg\min_a E\left[\rho_\tau(Y - a)|X = x\right].$$

Thus,integrating over P^X,

$$(x \mapsto q_\tau(Y|X = x)) \in \arg\min_{h(.)} E\left[\rho_\tau(Y - h(X))\right].$$

Definition of the estimator

- Suppose that $q_\tau(Y|X) = X'\beta_\tau$. We have, by the preceding argument,

$$\beta_\tau \in \arg\min_\beta E\left[\rho_\tau(Y - X'\beta)\right]. \quad (5)$$

- We use this property to define the quantile regression estimators. Suppose that we observe a sample $(Y_i, X_i)_{i=1...n}$ of i.i.d. data, we let

$$\widehat{\beta}_\tau \in \arg\min_\beta \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(Y_i - X_i'\beta). \quad (6)$$

- N.B.: when $\tau = 1/2$ (median), this is equivalent to minimizing

$$\frac{1}{n}\sum_{i=1}^{n} |Y_i - X_i'\beta|.$$

The corresponding solution is called the least absolute deviations (LAD) estimator.

Identification

- Before proving consistency of the estimator, we have to prove identification of β_τ by (5).
- In other words, is β_τ the *unique* minimizer of

$$\beta \mapsto E\left[\rho_\tau(Y - X'\beta)\right]?$$

- Sufficient condition: the residuals are continuously distributed conditional on X and the matrix $E\left[f_{\varepsilon_\tau|X}(0)XX'\right]$ is positive definite.
- Very similar to the rank condition in linear regression ($=E\left[XX'\right]$ positive definite).
- N.B.: this fails to hold when $f_{\varepsilon_\tau|X}(0) = 0$. In the case without covariate, this is close to being necessary because the minimizer of (5) is not unique when the d.f. of ε_τ is flat at τ.

Consistency

- Achieving consistency of $\widehat{\beta}_\tau$ is not as easy as with OLS because we have no explicit form of the estimator.
- We may use the special feature of ρ_τ, or use general consistency theorems on M-estimators defined as

$$\widehat{\theta} = \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \psi(U_i, \theta). \tag{7}$$

Theorem

(van der Vaart, 1998, Theorem 5.7) Let Θ denote the set of parameters θ and suppose that for all $\delta > 0$:

$$\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} \psi(U_i, \theta) - E(\psi(U_1, \theta)) \right| \xrightarrow{\mathbb{P}} 0, \tag{8}$$

$$\inf_{\theta / d(\theta, \theta_0) \geq \delta} E(\psi(U_1, \theta)) > E(\psi(U_1, \theta_0)). \tag{9}$$

Then any sequence of estimators $\widehat{\theta}_n$ defined by (7) converges in probability to θ_0.

Consistency

- Here $U_i = (Y_i, X_i)$ and $\psi(U, \theta) = \rho_\tau(Y - X'\theta)$.
- Condition (9) is a "well-separated" minimum condition, which is typically satisfied in our case under the identification condition above and if we restrict Θ to be compact.
- The first condition is the most challenging. By the law of large numbers, we have pointwise convergence but not, *a priori*, uniform convergence. To achieve this, we may use *Glivenko-Cantelli* theorems.
- The idea behind is that if the set of functions $(\psi(., \theta))_{\theta \in \Theta}$ is not "too large", one can approximate the supremum by a maximum over a finite subset of Θ and applies the law of large numbers to each of the elements of this subset.

Consistency

Example: the standard Glivenko-Cantelli theorem. Let us consider the functions $\psi(x,t) = \mathbb{1}\{x \leq t\}$. Then:

$$\sup_{t\in\mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^{n} \psi(Y_i, t) - E(\psi(Y_1, t)) \right| \xrightarrow{\mathbb{P}} 0.$$

N.B.: letting F_n denote the empirical d.f. of Y, this can be written in a more usual way as

$$\sup_{t\in\mathbb{R}} |F_n(t) - F(t)| \xrightarrow{\mathbb{P}} 0.$$

Proof (here for continuous Y): fix $\delta > 0$ and consider $t_0 = -\infty < ... < t_K = \infty$ such that $F(t_k) - F(t_{k-1}) < \delta$. Then for all $t \in [t_{k-1}, t_k]$,

$$F_n(t) - F(t) \leq F_n(t_k) - F(t_{k-1}) \leq F_n(t_k) - F(t_k) + \delta$$

Similarly, $F_n(t) - F(t) \geq F_n(t_{k-1}) - F(t_{k-1}) - \delta$. Thus,

$$|F_n(t) - F(t)| \leq \max\{|F_n(t_k) - F(t_k)|, |F_n(t_{k-1}) - F(t_{k-1})|\} + \delta.$$

Consistency

As a result,

$$\sup_{t\in\mathbb{R}} |F_n(t) - F(t)| \leq \max_{i\in\{0,\dots,K\}} |F_n(t_i) - F(t_i)| + \delta.$$

By the weak law of large numbers, the maximum tends to zero. The result follows $\square$

This proof can be generalized to classes of functions different from $(\mathbb{1}\{. \leq t\})_{t\in\mathbb{R}}$. A δ-bracket in L_r is a set of functions f with $l \leq f \leq u$, where l and u are two functions satisfying $\left(\int |u-l|^r dF\right)^{1/r} < \delta$. For a given class of functions $\mathcal{F}$, define the *bracketing number* $N_{[\,]}(\delta, \mathcal{F}, L_r)$ as the minimum number of δ-brackets needed to cover $\mathcal{F}$.

Proposition

(van der Vaart, 1998, Theorem 19.4) Suppose that for all $\delta > 0$, $N_{[\,]}(\delta, \mathcal{F}, L_1) < \infty$. Then

$$\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^n f(X_i) - E(f(X_1)) \right| \xrightarrow{\mathbb{P}} 0.$$

Consistency

The proposition applies to many cases, see van der Vaart (1998), chapter 19, for examples. In particular, it holds with parametric families satisfying

$$|\psi(U_i, \theta_1) - \psi(U_i, \theta_2)| \leq m(U_i)||\theta_1 - \theta_2||, \ E(m(U_1)) < \infty. \tag{10}$$

In quantile regression,

$$\begin{aligned} |\rho_\tau(Y - X'\beta_1) - \rho_\tau(Y - X'\beta_2)| &\leq \max(\tau, 1-\tau)|X'(\beta_1 - \beta_2)| \\ &\leq ||X|| \times ||\beta_1 - \beta_2||. \end{aligned}$$

Thus (10) holds provided that $E(||X||) < \infty$. This establishes consistency of $\widehat{\beta}_\tau$ since we can then apply the theorem above.

Asymptotic normality

- We now investigate the asymptotic distribution of $\widehat{\beta}_\tau$.
- The usual method for smooth M-estimator is to use a Taylor expansion. The first order condition writes as

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\psi}{\partial\theta}(U_i,\widehat{\theta}) = 0. \tag{11}$$

Then expanding around $\widehat{\theta}$, we get

$$0 = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial\psi}{\partial\theta}(U_i,\theta_0) + \left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2\psi}{\partial\theta\partial\theta'}(U_i,\theta_0)\right](\widehat{\theta}-\theta_0) + o_P(||\widehat{\theta}-\theta_0||).$$

Hence, provided that one can show that $||\widehat{\theta}-\theta_0|| = O_P(1/\sqrt{n})$, we have

$$\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2\psi}{\partial\theta\partial\theta'}(U_i,\theta_0)\right]\sqrt{n}(\widehat{\theta}-\theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial\psi}{\partial\theta}(U_i,\theta_0) + o_P(1).$$

Asymptotic normality

- By the weak law of large numbers, the central limit theorem and Slutski's lemma, we get:

$$\sqrt{n}\left(\widehat{\theta}-\theta_0\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, J^{-1}HJ^{-1}),$$

  where $J = E\left[\frac{\partial^2\psi}{\partial\theta\partial\theta'}(U_i, \theta_0)\right]$ and $H = V(\frac{\partial\psi}{\partial\theta}(U_i, \theta_0))$. This kind of variance is often called a "sandwich formula".
- N.B.: in the maximum likelihood case, $-J = H = I_0$, the Fisher information matrix, and the formula simplifies.
- In quantile regression, we cannot use such a Taylor expansion directly since the derivative of ρ_τ (for $u \neq 0$) is the step function $\rho'_\tau(u) = \tau - \mathbb{1}\{u < 0\}$, which is not differentiable.
- The first order condition (11) may not hold exactly either. However, 0 can be replaced by $o_P\left(\frac{1}{\sqrt{n}}\right)$, which will be sufficient subsequently.

Asymptotic normality

Two key ideas for these kinds of situations:

- Even if $\theta \mapsto \frac{\partial \psi}{\partial \theta}(U_i, \theta)$ is not differentiable at θ_0, $\theta \mapsto Q(\theta) = E\left[\frac{\partial \psi}{\partial \theta}(U_i, \theta)\right]$ is usually (continuously) differentiable.
- Starting from (11), we then write:

$$
\begin{aligned}
0 &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\frac{\partial \psi}{\partial \theta}(U_i, \widehat{\theta}) - Q(\widehat{\theta})\right] + \sqrt{n}\left(Q(\widehat{\theta}) - Q(\theta_0)\right) \\
&= G_n(\widehat{\theta}) + Q'(\widetilde{\theta})\sqrt{n}(\widehat{\theta} - \theta_0). \qquad (12)
\end{aligned}
$$

  where $\widetilde{\theta} \in (\theta_0, \widehat{\theta})$ and $G_n(\theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\frac{\partial \psi}{\partial \theta}(U_i, \theta) - Q(\theta)\right]$. G_n is a stochastic process (i.e., a random function) which is called the *empirical process*.
- To show asymptotic normality of $\sqrt{n}(\widehat{\theta} - \theta_0)$, it suffices to show that $G_n(\widehat{\theta})$ converges to a normal distribution.

Asymptotic normality

- By the central limit theorem, for any fixed θ, $G_n(\theta)$ converges to a normal distribution. Here however, $\widehat{\theta}$ is random.
- The idea is to extend "simple" central limit theorem to convergence of the whole process G_n to a continuous gaussian process G. This is achieved through *Donsker theorems*.
- Such theorems may be seen as uniform CLT, just as Glivenko-Cantelli were uniform LLN. Under such conditions, we can prove that $G_n(\widehat{\theta}) \xrightarrow{\mathcal{L}} G(\theta_0)$, a normal variable.
- As previously, Donsker theorems can be obtained when the class of functions $\mathcal{F}$ is not too large. For instance:

Proposition

(van der Vaart, Theorem 19.5) G_n, as a process indexed by $f \in \mathcal{F}$, converges to a continuous gaussian process if

$$\int_0^1 \sqrt{\ln N_{[\,]}(\delta, \mathcal{F}, L_2)}\, d\delta < \infty.$$

Asymptotic normality

- Like previously, many classes of functions satisfy the *bracketing integral* condition. In parametric classes where (10) holds, for instance, one can show that for δ small enough,

$$N_{[\,]}(\delta, \mathcal{F}, L_2) \leq \frac{K}{\delta^d}.$$

  Thus the bracketing integral is finite and one can apply the previous theorem.
- Coming back to (12), we have, under the bracketing integral condition,

$$\sqrt{n}\left(\widehat{\theta} - \theta_0\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, Q'(\theta_0)^{-1} V\left(\frac{\partial \psi}{\partial \theta}(U_i, \theta_0)\right) Q'(\theta_0)^{-1}\right)$$

Asymptotic normality

- Application to the quantile regression: the bracketing integral condition is satisfied, thus it suffices to check the differentiability of $Q(\beta)$ at β_τ. Here, $\partial\psi/\partial\theta(U_i,\theta) = -(\tau - \mathbb{1}\{Y - X'\theta < 0\})X$. Thus,

$$\begin{aligned} -Q(\beta) &= \tau E(X) - E\left[\mathbb{1}\{\varepsilon_\tau < X'(\beta - \beta_\tau)\}X\right] \\ &= \tau E(X) - E\left[F_{\varepsilon_\tau|X}(X'(\beta - \beta_\tau)|X)X\right] \end{aligned}$$

- Thus, provided that ε_τ admits a density conditional on X at 0, $Q(.)$ is differentiable and

$$Q'(\beta_\tau) = E\left[f_{\varepsilon_\tau|X}(0|X)XX'\right].$$

- Besides,

$$\begin{aligned} V\left(\frac{\partial\psi}{\partial\theta}(U_i,\theta_0)\right) &= E\left\{V\left[(\tau - \mathbb{1}\{Y - X'\beta_\tau < 0\})X|X\right]\right\} \\ &= \tau(1-\tau)E\left[XX'\right]. \end{aligned}$$

Asymptotic normality

- Finally, we get:

$$\sqrt{n}\left(\widehat{\beta}_\tau - \beta_\tau\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \tau(1-\tau)E\left[f_{\varepsilon_\tau|X}(0|X)XX'\right]^{-1} E\left[XX'\right] E\left[f_{\varepsilon_\tau|X}(0|X)XX'\right]^{-1}\right).$$

- Remark 1: if $Y = X'\beta + \varepsilon$ where ε is independent of X (location model), $\varepsilon_\tau = \varepsilon - q_\tau(\varepsilon)$ and the asymptotic variance V_{as} reduces to

$$V_{\text{as}} = \frac{\tau(1-\tau)}{f_\varepsilon(q_\tau(\varepsilon))^2} E\left[XX'\right]^{-1}.$$

This formula is similar to the one for the OLS estimator, except that σ^2 is replaced by $\tau(1-\tau)/f_\varepsilon(q_\tau(\varepsilon))^2$. In general, as we let $\tau \to 1$ or 0, $f_\varepsilon(q_\tau)^2$ becomes very small and thus $\widehat{\beta}_\tau$ becomes imprecise. This is logical since data are often more dispersed at the tails.

Asymptotic normality

- Remark 2: this result applies in particular to simple quantiles $\widehat{q}_\tau$, in which case we have:

$$\sqrt{n}\left(\widehat{q}_\tau - q_\tau\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{\tau(1-\tau)}{f_Y^2(q_\tau)}\right).$$

- Remark 3: we can also generalize it to parameters $(\beta_{\tau_1}, ..., \beta_{\tau_m})$ corresponding to different quantiles:

$$\sqrt{n}\left(\widehat{\beta}_{\tau_k} - \beta_{\tau_k}\right)_{k=1}^{m} \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, V\right), \tag{13}$$

where V is a $m \times m$ block-matrix, whose (k, l) block $V_{k,l}$ satisfies

$$V_{k,l} = \left[\tau_k \wedge \tau_l - \tau_k \tau_l\right] H(\tau_k)^{-1} E\left[XX'\right] H(\tau_l)^{-1}$$

and as before, $H(\tau) = E\left[f_{\varepsilon_\tau|X}(0)XX'\right]$.

Confidence intervals and testing

- This result is useful to build confidence intervals or test assumptions on β_τ.
- However, to obtain estimators of the asymptotic variance, one has to estimate $f_{\varepsilon_\tau|X}(0|X)$, which is a difficult task.
- Alternative solutions have thus been proposed for inference:
  - using rank tests (not presented here);
  - using bootstrap or, more generally, resampling methods;
  - making finite sample inference.

Asymptotic variance estimation

- In the location model, $V_{as} = \tau(1-\tau)E(XX')^{-1}/f_\varepsilon(q_\tau(\varepsilon))$, and the only problem is the denominator. Note that

$$\begin{aligned}\frac{1}{f_\varepsilon(q_\tau(\varepsilon))} &= \frac{1}{f_\varepsilon(F_\varepsilon^{-1}(\tau))} = \frac{\partial F_\varepsilon^{-1}}{\partial \tau}(\tau) \\ &= \lim_{h\to 0} \frac{F_\varepsilon^{-1}(\tau+h) - F_\varepsilon^{-1}(\tau-h)}{2h}.\end{aligned}$$

- Thus we can estimate this term by, e.g., $(\widehat{F}_\varepsilon^{-1}(\tau+h_n) - \widehat{F}_\varepsilon^{-1}(\tau-h_n))/2h_n$.
- Like often, h_n must be chosen so as to balance bias and variance. Several choices have been proposed. Minimally, we must have, $h_n \to 0$ and $nh_n \to \infty$.
- This is (roughly) the estimator provided by default in Stata. However, the corresponding variance estimator is inconsistent in general when ε is *not* independent of X.

Asymptotic variance estimation

- In this general case, main difficulty: estimate $J = E(f_{\varepsilon_\tau|X}(0|X)XX']$. A simple solution (Powell, 1991) relies on the following idea:

$$J = \lim_{h\to 0} E\left[\frac{\mathbb{1}\{|\varepsilon_\tau| \le h\}}{2h}XX'\right].$$

- Letting $\widehat{\varepsilon}_{i\tau} = Y_i - X_i'\widehat{\beta}_\tau$, we thus may estimate J by (with also h_n "small but too small"):

$$\widehat{J} = \frac{1}{2nh_n}\sum_{i=1}^{n} \mathbb{1}\{|\widehat{\varepsilon}_{i\tau}| \le h_n\}X_iX_i'. \tag{14}$$

- Other solution (cf. Koenker and Machado, 1999): if $q_{\tau'}(Y|X) = X'\beta_{\tau'}$ for τ' close to τ,

$$f_{\varepsilon_\tau|X}(0|X) = \frac{1}{\partial q_\tau(Y|X)/\partial\tau} = \lim_{h\to 0}\frac{2h}{X'\beta_{\tau+h} - X'\beta_{\tau-h}}.$$

Asymptotic variance estimation

- With a consistent estimator of V_{as} in hand, we can easily make inference on β_τ.
- Confidence interval on β_τ:

$$IC_\alpha = \left[\widehat{\beta}_\tau - z_{1-\alpha/2}\sqrt{\widehat{V}_{\text{as}}}, \widehat{\beta}_\tau + z_{1-\alpha/2}\sqrt{\widehat{V}_{\text{as}}}\right],$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$-th quantile of the $\mathcal{N}(0,1)$ distribution.

- The Wald statistic test of $g(\beta_\tau) = 0$ writes

$$T = ng(\widehat{\beta}_\tau)' \left[\frac{\partial g}{\partial \beta'}(\beta_\tau)\widehat{V}_{\text{as}}\frac{\partial g}{\partial \beta}(\beta_\tau)\right]^{-1} g(\widehat{\beta}_\tau),$$

and it tends to a $\chi^2_{\text{dim}(g)}$ under the null hypothesis.

Bootstrap

- The previous approach requires to choose a smoothing parameter h_n, and results may be sensitive to this choice.
- Alternatively, we can use bootstrap by implementing the algorithm:
  For $b = 1$ to B:
  - Draw with replacement a sample of size n from the initial sample $(Y_i, X_i)_{i=1...n}$. Let $(k^*_{b1}, ..., k^*_{bn})$ denote the corresponding indices of the observations;
  - Compute $\widehat{\beta}^*_{\tau b} = \arg\min_\beta \sum_{j=1}^n \rho_\tau(Y_{k^*_{bj}} - X'_{k^*_{bj}}\beta)$.

Bootstrap

- Then we can estimate the asymptotic variance by

$$V_{\text{as}}^* = \frac{1}{B}\sum_{b=1}^{B}(\widehat{\beta}_{\tau b}^* - \widehat{\beta})^2.$$

- Confidence intervals or hypothesis testing may be conducted as before, using the normal approximation.
- Alternatively (*percentile bootstrap*), you can compute the empirical quantiles q_u^* of $(\widehat{\beta}_{\tau 1}^*, \ldots, \widehat{\beta}_{\tau B}^*)$ and then define a confidence interval as

$$IC_{1-\alpha} = [q_{\alpha/2}^*, q_{1-\alpha/2}^*].$$

- N.B.: there are other (quicker) resampling methods specialized for the quantile regression, see Koenker (1994), Parzen et al. (1994) and He and Hu (2002).

Finite sample inference

- Simple yet very recently developed idea (Chernozhukov et al., 2009, Coudin and Dufour, 2009): if $\beta_\tau = \beta_0$, then $B_i(\beta_0) = \mathbb{1}\{Y_i - X_i'\beta_0 \leq 0\}$ is such that

$$B_i(\beta_0)|X_i \sim \text{Be}(\tau).$$

- As a result, for all $g(.)$ and positive definite W_n, under the hypothesis $\beta_\tau = \beta_0$, the distribution of

$$T_n(\beta_0) = \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\tau - B_i(\beta_0))g(X_i)\right)' W_n \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\tau - B_i(\beta_0))g(X_i)\right)$$

is known (theoretically at least). Letting $z_{1-\alpha}$ denote its $(1-\alpha)$-th quantile, we reject the null hypothesis if $T_n(\beta_0) > z_{1-\alpha}$.

- In practice, the distribution of $T_n(\beta_0)$ under the null can be approximated by simulations.

Finite sample inference

- We can then define a confidence region by *inverting* the test: $CR_{1-\alpha} = \{\beta / T_n(\beta) \leq z_{1-\alpha}\}$. Indeed, letting β_τ denote the true parameter,

$$\begin{aligned} \Pr(CR_{1-\alpha} \ni \beta_\tau) &= \Pr\left(T_n(\beta_\tau) \leq z_{1-\alpha}\right) \\ &\geq 1-\alpha. \end{aligned}$$

- This is a general procedure to build confidence regions from a test.
- To obtain confidence interval on a real-valued parameter $\psi(\beta_\tau)$, we let
$$IC_{1-\alpha} = \{\psi(\beta), \beta \in CR_{1-\alpha}\}.$$
This is known as the *projection method* (see, e.g., Dufour and Taamouti). Corresponding confidence intervals are conservative.
- The computation of such confidence regions / intervals may be demanding. See Chernozhukov et al. (2009) for MCMC methods that partially alleviate this issue.

Testing homogeneity of effects

- As mentioned before, an interesting property of quantile regression is that it allows for heterogeneity of effects of X across the distribution of Y. A byproduct is that they also provide tests for the homogeneity hypothesis.
- Let $X = (1, X_{-1})$ and $\beta_\tau = (\beta_{1\tau}, \beta_{-1\tau})$ and $\mathcal{T}$ denote a set included in $[0, 1]$, the test formally writes as

$$\beta_{-1\tau} = \beta \quad \forall t \in \mathcal{T}.$$

  This may be seen as testing for the location model $Y = X'\beta + \varepsilon$, with $\varepsilon \perp\!\!\!\perp X$.
- If the set $\mathcal{T}$ is finite, we can use (13) to implement such a test. If the set is infinite, this is far more complex and can be achieved using the convergence of $\tau \mapsto \widehat{\beta}_\tau$ as a process (see Koenker and Xiao, 2002).

Outline

Introduction

Model and motivation

Interpreting quantile regressions

Inference in quantile regressions

Computational aspects

Quantile regressions with panel data

Quantile restrictions in nonlinear models

Computation of $\widehat{\beta}_\tau$.

- There is no explicit solution to (6) so one has to solve the program numerically.
- An issue is the non differentiability of the objective function. Standard algorithms such as the Newton-Raphson cannot be used here.
- The key idea is to reformulate (6) as a linear programming problem:

$$\min_{(\beta,u,v)\in\mathbb{R}^p\times\mathbb{R}_+^{2n}} \tau\mathbf{1}'u + (1-\tau)\mathbf{1}'v \quad s.t.\ \mathbf{X}\beta + u - v - \mathbf{Y} = 0,$$

  where $\mathbf{X} = (X_1, ..., X_n)'$, $\mathbf{Y} = (Y_1, ..., Y_n)'$ and $\mathbf{1}$ is a n-vector of 1.
- Such linear programming problems can be efficiently solved by simplex methods (for small n) or interior point methods (large n).

Computation of $\widehat{\beta}_\tau$.

- Simplex method: consider a linear programming problem of the form

$$\min_{x\in\mathbb{R}^n} c'x \quad \text{s.t. } x \in S = \{u/Au \geq b, Bu = d\}, \tag{15}$$

where $c \in \mathbb{R}^n$, A and B are two matrices and "$\geq$" is considered elementwise.

- Then one can show that (i) S is a convex polyhedron and (ii) if solutions exist, then they are vertices of S.
- Basically, the simplex method consists of going from one vertex to another, choosing each time the steepest descent.
- Interior point methods: consider (15) with $A = I_n$ and $b = 0$, the idea is to replace (15) by

$$\min_{x\in\mathbb{R}^n} c'x - \mu \sum_{k=1}^{n} \ln x_k \quad \text{s.t. } Bx = d. \tag{16}$$

(16) can be solved easily with a Newton method. Then let $\mu \to 0$.

Software programs

- SAS: `proc quantreg`.

```
proc quantreg data=(dataset) algorithm=(choice of algo.)  ci=
   (method for performing confidence intervals);
    class (qualitative variables);
    model (y) = (x) /quantile = (list of quantiles or ALL);
run;
```

- By default, the simplex method is used. One should switch to an interior point method (by letting `algorithm=interior`) for $n \geq 1000$.
- By default, the confidence intervals are computed by inverting rank-score tests when $n \leq 5000$ and $p \leq 20$, and resampling method otherwise (N.B.: the latter provide more robust standard error estimates).

Software programs

- Stata: command `sqreg`:

```
sqreg depvar indepvars , quantiles(choice of quantiles)
```

- Standard errors are obtained by bootstrap $\Rightarrow$ can be long.
- N.B: the command `qreg` computes only one quantile regression, with standard errors valid for the location model only. The command `bsqreg` computes only one quantile regression, with bootstrap standard errors.

Software programs

- A very complete R package has been developed by R. Koenker: `quantreg`.

```
library(quantreg)
rq(y ~ x1 + x2, tau = (single quantile or vector of
  quantiles), data=(dataset), method=("br" or "fn"))
```

- To obtain inference on all quantiles put tau = -1 (or any number outside $[0, 1]$).
- `method ="br"` corresponds to the Simplex (default), while "fn" is an interior point method.
- a tutorial is available at Roger Koenker's webpage.

An example

- I look at the impact of various factors on birth weight, following Abreveya (2001). Indeed, a low birth weight is often associated with subsequent health problems, and is also related to educational attainment and labor market outcomes.
- Quantile regression provides a more complete story than just running a probit on the dummy variable (*birth weight* $<$ *arbitrary threshold*).
- The analysis is based on exhaustive 2001 US data on birth certificates. I restrict the sample to singleton births with mothers black or white, between the ages of 18 and 45, resident in the US (roughly 2.9 million observations).
- Apart from the gender, information on the mother is available: marital status, age, being black or white, education, date of the first prenatal visit, being a smoker or not, number of cigarettes smoked per day...

An example

- SAS code:

```
ods graphics on;
proc quantreg data=birth_weights ci=sparsity/iid alg=interior(tolerance=1e-4);
  model birth_weight = boy married black age age2 high_school some_college
    college prenatal_second prenatal_third no_prenatal smoker
    nb_cigarettes /quantile= 0.05 to 0.95 by 0.05 plot quantplot;
run;
ods graphics off;
```

- Stata code:

```
sqreg birth_weigh boy married black age age2 high_school some_college prenatal_second
   prenatal_third no_prenatal smoker nb_cigarettes, quantiles(0.05 0.1 0.2 0.3 0.4
   0.5 0.6 0.7 0.8 0.9 0.95)
```

- Stata is quite long here (1 hour for a single quantile with 20 bootstrap replications). To run SAS on large databases like this one, you may have to increase the available memory.

An example

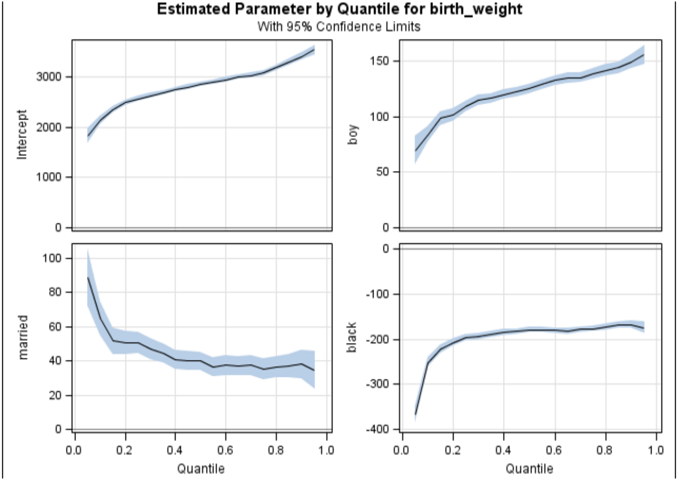
Quantile and Objective Function

Quantile	0.1
Objective Function	31108564.261
Predicted Value at Mean	2727.4037

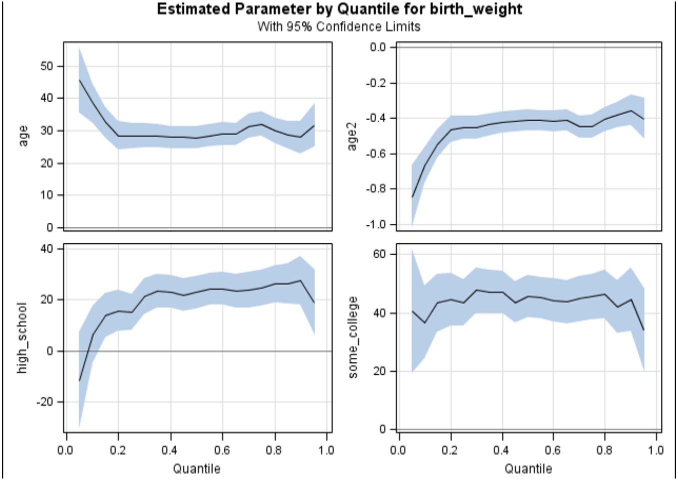
Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		t Value	Pr > |t|
Intercept	1	2150.419	41.9615	2068.176	2232.662	51.25	<.0001
boy	1	83.8925	3.8034	76.4380	91.3471	22.06	<.0001
married	1	64.9045	4.9650	55.1734	74.6357	13.07	<.0001
black	1	-251.465	5.4947	-262.234	-240.696	-45.77	<.0001
age	1	38.3584	3.0443	32.3916	44.3251	12.60	<.0001
age2	1	-0.6657	0.0523	-0.7682	-0.5631	-12.73	<.0001
high_school	1	6.5725	5.7090	-4.6170	17.7620	1.15	0.2496
some_college	1	36.6800	6.4022	24.1319	49.2281	5.73	<.0001
college	1	76.1075	6.7700	62.8384	89.3765	11.24	<.0001
prenatal_second	1	-4.1840	5.9940	-15.9321	7.5641	-0.70	0.4852
prenatal_third	1	22.2022	12.2669	-1.8405	46.2449	1.81	0.0703
no_prenatal	1	-472.532	19.1648	-510.095	-434.970	-24.66	<.0001
smoker	1	-156.928	10.6564	-177.815	-136.042	-14.73	<.0001
nb_cigarettes	1	-5.8266	0.8140	-7.4221	-4.2311	-7.16	<.0001

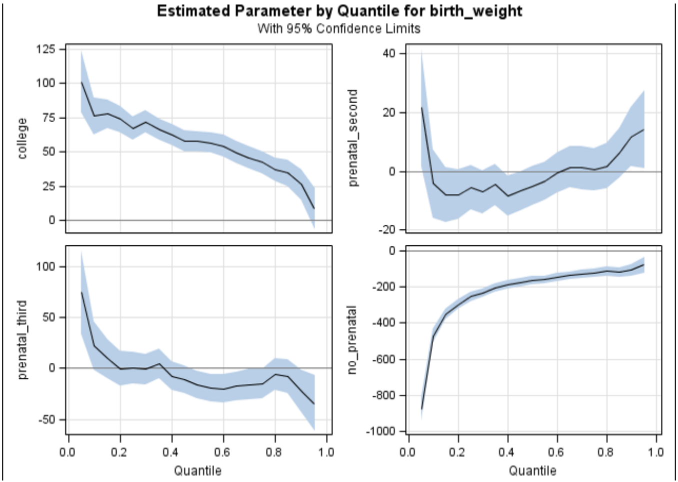
An example


Estimated Parameter by Quantile for birth_weight
With 95% Confidence Limits
Intercept
3000
2000
1000
0
boy
150
100
50
0
married
100
80
60
40
20
0
black
0
-100
-200
-300
-400
0.0
0.2
0.4
0.6
0.8
1.0
Quantile

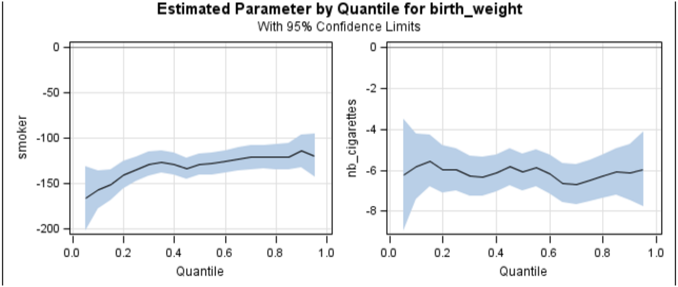
An example


Estimated Parameter by Quantile for birth_weight
With 95% Confidence Limits
age
age2
high_school
some_college
Quantile
Quantile
0.0
0.2
0.4
0.6
0.8
1.0
50
40
30
20
10
0
-0.2
-0.4
-0.6
-0.8
-1.0
-20
60

An example


Estimated Parameter by Quantile for birth_weight
With 95% Confidence Limits
college
prenatal_second
prenatal_third
no_prenatal
Quantile
Quantile
125
100
75
50
25
0
40
20
0
-20
100
50
0
-50
0
-200
-400
-600
-800
-1000
0.0
0.2
0.4
0.6
0.8
1.0

An example


Estimated Parameter by Quantile for birth_weight
With 95% Confidence Limits
smoker
0
-50
-100
-150
-200
0.0
0.2
0.4
0.6
0.8
1.0
Quantile
nb_cigarettes
0
-2
-4
-6
-8
0.0
0.2
0.4
0.6
0.8
1.0
Quantile

Outline

Introduction

Model and motivation

Interpreting quantile regressions

Inference in quantile regressions

Computational aspects

Quantile regressions with panel data

Quantile restrictions in nonlinear models

The model and problems

- A way to address endogeneity is to follow units through time, using panel data.
- Idea in the linear model with mean restrictions: introducing a fixed effect that captures this endogeneity and getting rid of it through differencing:

$$\begin{aligned} Y_{it} &= X_{it}\beta + \alpha_i + \varepsilon_{it},\ E(\varepsilon_{it}|X_{i1},...,X_{iT}) = 0 \\ \Rightarrow WY_{it} &= WX_{it}\beta + W\varepsilon_{it},\ E(W\varepsilon_{it}|WX_{it}) = 0. \end{aligned} \quad (17)$$

where W is the within operator, $WU_{it} = U_{it} - \overline{U}_i$.
$E(W\varepsilon_{it}|WX_{it}) = 0$ implies that the OLS estimator (=within estimator) of (17) is consistent.

The model and problems

- $E(W\varepsilon_{it}|WX_{it}) = 0$ holds by linearity of the expectation. This is not true for quantiles, however. Thus, if

$$Y_{it} = X_{it}\beta_\tau + \alpha_{i\tau} + \varepsilon_{it\tau}, \; q_\tau(\varepsilon_{it}|X_{i1}, ..., X_{iT}) = 0, \tag{18}$$

  a quantile regression on the within equations does not provide a consistent estimator of β_τ in general.

- Moreover, making the "large" quantile regression of Y_{it} on $(X_{it}, (\mathbb{1}_j)_{j=1...n})$ does not work because of the *incidental parameters problem*: the number of parameters to estimate $(\beta_\tau, \alpha_{1\tau}, ..., \alpha_{n\tau})$ tends to infinity as $n \to \infty$.
- This problem makes the asymptotic properties of estimators nonstandard. In general the estimators are inconsistent.
- Another issue is the computational burden, because one has to optimize over a very large space.

A solution: Canay (2011)

- A solution has been proposed by Canay (2011). Suppose that

$$Y_{it} = X_{it}\beta_{U_{it}} + \alpha_i, \tag{19}$$

  where α_i and U_{it} are unobserved, $U_{it}|X_{it}, \alpha_i \sim U[0,1]$. Then, Eq. (18) holds with $\varepsilon_{it} = X_{it}(\beta_{U_{it}} - \beta_\tau)$.
- The main restriction is that individual heterogeneity correlated with X_{it} should have a pure location effect. No scale effect for instance (as in a model $Y_{it} = X_{it}(\beta_{U_{it}} + \gamma_i) + \alpha_i$).
- Canay (2011) proposes the following simple two-step estimator:
  1. Within estimation of the linear regression

     $$Y_{it} = X_{it}\beta_\mu + \alpha_i + u_{it}, \text{ with } E(u_{it}|X_{it}, \alpha_i) = 0.$$

     From this estimation of $\beta_\mu = E[\beta_U]$, we can estimate individual fixed effects: $\hat{\alpha}_i = \frac{1}{T}\sum_{t=1}^{T}(Y_{it} - X_{it}\hat{\beta}_\mu)$.
  2. Standard quantile regression of $\widetilde{Y}_{it} = Y_{it} - \hat{\alpha}_i$ on X_{it}.

A solution: Canay (2011)

- Canay shows that the corresponding estimator is consistent and asymptotically normal estimator, but only as $T \to \infty$. Very strong condition (very often $T \leq 10$...).
- Koenker (2004) proposes an estimator based on the "large" quantile regression, with an L^1 penalization of the fixed effects. But it is more cumbersome and suffers from the same limitations (location effect, consistency only as $T \to \infty$).
- Both have been implemented on `R` (see Ivan Canay's website and the package `rqpd` for Koenker's solution).
- For the moment no consistent estimator has been proposed for fixed T.

Outline

Introduction

Model and motivation

Interpreting quantile regressions

Inference in quantile regressions

Computational aspects

Quantile regressions with panel data

Quantile restrictions in nonlinear models

Introduction

- We consider here extensions of the quantile linear regression to nonlinear models of the form

$$Y = g(X'\beta_0 + \varepsilon), \tag{20}$$

where g is a nonlinear function.
- It is difficult to use restrictions of the kind $E(\varepsilon|X) = 0$ in (20) because in general, $E(Y|X) \neq g(X'\beta_0)$.
- On the other hand, by an equivariance property, quantile restrictions are easy to use in such models.

The basic idea

The equivariance property can be stated as follows:

Proposition

Let g be an increasing, left continuous function, then

$$g(q_\tau(Y)) = q_\tau(g(Y)).$$

Proof: recall that $q_\tau(g(Y)) = \inf\{x \in \mathbb{R} / F_{g(Y)}(x) \geq \tau\}$. we have

$$\tau \leq P(Y \leq q_\tau(Y)) \leq P(g(Y) \leq g(q_\tau(Y))).$$

Thus, $g(q_\tau(Y)) \geq q_\tau(g(Y))$. Conversely, let $u = q_\tau(g(Y))$ and $g^-(v) = \sup\{x/g(x) \leq v\}$. Then

$$\tau \leq P(g(Y) \leq u) \leq P(Y \leq g^-(u)).$$

As a result, $g^-(u) \geq q_\tau(Y)$. Because g is left continuous, $g(g^-(u)) \leq u$. Thus, $q_\tau(g(Y)) = u \geq g(q_\tau(Y))$, which ends the proof.

The basic idea

- Now consider Model (20) with $q_\tau(\varepsilon|X) = 0$. If g is increasing and left continuous, we have

$$q_\tau(Y|X) = g(q_\tau(X'\beta_0 + \varepsilon|X)) = g(X'\beta_0).$$

- By the same argument as previously, it follows that

$$\beta_0 \in \arg\min_\beta E\left[\rho_\tau\left(Y - g(X'\beta)\right)\right].$$

- Thus, compared to a linear quantile regression, we simply add g in the program.
- This comes however at the cost of some identification, estimation and implementation issues, as we shall see below.

The basic idea

- Although this idea is general, we study in details two examples: binary and tobit models. In the first, $g(x) = \mathbb{1}\{x > 0\}$ and in the second, $g(x) = \max(x, 0)$.
- Note that an alternative nonlinear model would be

$$Y = \mu(X, \beta_0) + \varepsilon, \quad q_\tau(\varepsilon|X) = 0.$$

Such an extension leads to a similar optimization program as above and is thus not considered afterwards.

First example: binary models

- Consider the following model:

$$Y = \mathbb{1}\{X'\beta_0 + \varepsilon > 0\}.$$

- We would like to identify and estimate β without imposing arbitrary assumptions such as $\varepsilon | X \sim \mathcal{N}(0,1)$ (Probit models).
- In particular, we would like to allow for heteroskedasticity and leave the distribution of ε unspecified.
- Note that a scale normalization is necessary. We suppose for instance that the first component of β_0 is equal to 1 or -1.

First example: binary models

- First attempt: $E(\varepsilon|X)=0$.
- We have

$$P(Y=1|X=x)=\int_{-x'\beta_0}^{\infty} dF_{\varepsilon|X=x}(u),$$

and the model imposes that $\int_{-\infty}^{\infty} u dF_{\varepsilon|X=x}(u)=0$.

- Consider $\beta\neq\beta_0$. For all x, it is possible (exercise...) to build a distribution function $G_x\neq F_{\varepsilon|X=x}$ such that:

$$\begin{aligned}\int_{-x'\beta}^{\infty} dG_x(u) &= P(Y=1|X=x)\\ \int_{-\infty}^{\infty} u dG_x(u) &= 0.\end{aligned}$$

- This implies that β_0 is not identified here.

First example: binary models

- Second attempt: $q_\tau(\varepsilon|X) = 0$. In this case, by the equivariance property:

$$q_\tau(Y|X) = \mathbb{1}\{X'\beta_0 > 0\}.$$

- To achieve identification, we must therefore have:

$$\mathbb{1}\{X'\beta > 0\} = \mathbb{1}\{X'\beta_0 > 0\} \text{ a.s.} \Rightarrow \beta = \beta_0.$$

- The following conditions are sufficient for that purpose (Manski, 1988):
  - A1 there exists one variable (say X_1) which is continuous and whose density (conditional on X_{-1}) is almost everywhere positive.
  - A2 The $(X_k)_{1 \leq k \leq K}$ are linearly independent.

First example: binary models

- We use the standard characterization and consider:

$$\widehat{\beta} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(Y_i - \mathbb{1}\{X_i'\beta > 0\}).$$

- When $\tau = 1/2$, the estimator is called the *maximum score estimator*, because one can show that:

$$\widehat{\beta} = \arg\max_{\beta} \frac{1}{n} \sum_{i=1}^{n} Y_i \mathbb{1}\{X_i'\beta > 0\} + (1 - Y_i)\mathbb{1}\{X_i'\beta \leq 0\}.$$

- Note that this program is neither differentiable in β, nor even continuous. This raises trouble in both the asymptotic behavior of $\widehat{\beta}$ and its computation.

First example: binary models

- Kim and Pollard (1990) show that

$$n^{1/3}(\widehat{\beta} - \beta_0) \xrightarrow{\mathcal{L}} Z = \arg \max_{\theta \in \text{Vect}(\beta_0)^{\perp}} W(\theta),$$

  where W is a multidimensional gaussian process (see Kim and Pollard for its exact distribution).

- The reason why we get a nonstandard convergence rate is that contrary to previously, $\widehat{\beta}$ does not solve a (even approximate) first order condition. For general discussion on rates of convergence of M-estimator, see e.g. Van der Vaart (1998), Section 5.8.

- Inference is difficult because the distribution of Z has no exact form and depends on nuisance parameters. Moreover, bootstrap fails in this context (see Abrevaya and Huang, 2005). Instead, one may use subsampling (see Delgado, Rodriguez-Poo and Wolf, 2001).

First example: binary models

- There are also some computational issues, because
  (i) the objective function is a step function and
  (ii) we cannot rewrite the program as a linear programming problem.
- A first algorithm is provided by Manski and Thompson (1986), but it may reach a local solution only. A recent solution based on mixed integer programming has been proposed by Florios and Skouras (2008).
- To my knowledge, it has not been implemented yet in standard softwares.

First example: binary models

- To circumvent the trouble caused by the nonregularity of the objective function, Horowitz (1992) has proposed to replace $\mathbb{1}\{X'\beta > 0\}$ by $K(X'\beta/h_n)$, where K is a smooth distribution function and $h_n \to 0$, in the objective function.
- He shows under mild regularity conditions that his estimator has a faster rate of convergence (still lower than $\sqrt{n}$ yet) and is asymptotically normal. He also shows the validity of the bootstrap.
- Implementation is also easier as the objective function is smooth.

Second example: Tobit models

- Consider the simple tobit model:

$$Y = \max(0, X'\beta_0 + \varepsilon).$$

- Such a model is useful for consumption or top-coding (in which case max and 0 are replaced by min and $\overline{y}$), among others.
- The standard Tobit estimator is the ML estimator of a model where $\varepsilon | X \sim \mathcal{N}(0, \sigma^2)$.
- Powell (1984) considers instead the quantile restriction: $q_\tau(\varepsilon | X) = 0$.

Second example: Tobit models

- In this case, as mentioned before:

$$q_\tau(Y|X) = \max(0, X'\beta_0).$$

- Thus, identification of β_0 is ensured as soon as:

$$\max(0, X'\beta) = \max(0, X'\beta_0) \Rightarrow \beta = \beta_0.$$

- This is true for instance if $E(XX'\mathbb{1}\{X'\beta_0 \geq \delta\})$ (for some $\delta > 0$) is full rank and the distribution of ε conditional on X admits a density at 0.

Second example: Tobit models

- The estimator satisfies

$$\widehat{\beta} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \rho_{\tau} \left(Y_i - \max(0, X_i'\beta) \right).$$

- Contrary to the previous binary model, the program is continuous (and differentiable except on some points). A consequence is that the behavior of $\widehat{\beta}$ is more standard.
- Powell shows indeed that

$$\sqrt{n}\left(\widehat{\beta} - \beta_0\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, J^{-1}HJ^{-1}\right)$$

where

$$\begin{aligned} J &= E\left[f_{\varepsilon_\tau|X}(0|X)\mathbb{1}\{X'\beta_0 \geq 0\}XX'\right], \\ H &= E\left[\mathbb{1}\{X'\beta_0 \geq 0\}XX'\right]. \end{aligned}$$

Second example: Tobit models

- Buchinsky (1991, 1994) proposes an iterative linear programming algorithm based on the decomposition:

$$\widehat{\beta} = \arg\min_{\beta} \frac{1}{n} \left[\sum_{i/X_i'\beta \geq 0} \rho_\tau(Y_i - X_i'\beta) + \sum_{i/X_i'\beta < 0} \rho_\tau(Y_i) \right].$$

1. Set $D_0 = \{1, ..., n\}$, $\widehat{\beta}_0 = 0$ (for instance) and $m = 1$.
2. Repeat until $\widehat{\beta}_m = \widehat{\beta}_{m-1}$:
   Estime a quantile regression on D_{m-1}. Let $\widehat{\beta}_m$ be the corresponding estimator and $D_m = \{i/X_i'\widehat{\beta}_m \geq 0\}$. Set $m = m + 1$.

- Buchinsky (1994) shows that if this algorithm converges, then it converges to a local minimum of the objective function.
- This algorithm is implemented in Stata for $\tau = 1/2$ (clad).

Second example: Tobit models

- Inference can be based on the estimation of the asymptotic variance, as in quantile regression.
- Alternatively, one may use a modified bootstrap proposed by Bilias, Chen and Ying (2000):
  For $b = 1$ to B:
  - Draw with replacement a sample of size n from the initial sample $(Y_i, X_i)_{i=1...n}$. Let $(k^*_{b1}, ..., k^*_{bn})$ denote the corresponding indices of the observations;
  - Compute $\widehat{\beta}^*_b = \arg\min_\beta \sum_{j=1}^n \rho_\tau(Y_{k^*_{bj}} - X'_{k^*_{bj}}\beta)\mathbb{1}\{X'_{k^*_{bj}}\widehat{\beta} > 0\}$.
- Note that each bootstrap estimator $\widehat{\beta}^*_b$ can be obtained easily by a standard quantile regression since the indicator term does not depend on β.