Lecture 1

Quantile Methods

Manuel Arellano

CEMFI

DSE-ES Winter School
Delhi, December 10-11, 2018

Introduction

- In this lecture I provide a review of quantile regression (QR).
- Quantile regression is a useful tool for studying conditional distributions.
- Part 1 provides an informal introduction to conditional quantiles and QR.
- Part 2 contains a more formal development and some large sample results.
- Part 3 is a discussion of instrumental-variable QR and quantile treatment effects.

Part 1

Conditional quantiles and quantile regression: informal introduction

Conditional quantile function

- Econometrics deals with relationships between variables involving unobservables.

- Consider an empirical relationship between two variables Y and X.

- Suppose that X takes on K different values $x_1, x_2, ..., x_K$ and that for each of those values we have M_k observations of Y: $y_{k1}, ..., y_{kM_k}$.

- If the relationship between Y and X is exact, the values of Y for a given value of X will all coincide, so that we could write
$$Y = q(X).$$

- However, in general units having the same value of X will have different values of Y.

- Suppose that $y_{k1} \leq y_{k2} \leq ... \leq y_{kM_k}$, so the fraction of observations that are less than or equal to y_{km} is $u_{km} = m/M_k$.

- It can then be said that a value of Y does not only depend on the value of X but also on the rank u_{km} of the observation in the distribution of Y given $X = x_k$.

- Generalizing the argument:
$$Y = q\left(X, U\right)$$

Conditional quantile function (continued)

- The distribution of the ranks U is always the same regardless of the value of X, so that X and U are statistically independent.

- Also note that $q\left(x, u\right)$ is an increasing function in u for every value of x.

- An example is a growth chart where Y is body weight and X is age (Figure 1).

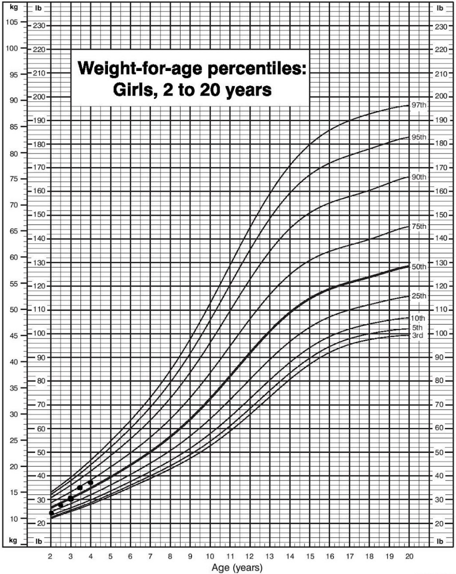
- In this example U is a normalized unobservable scalar variable that captures the determinants of body weight other than age, such as diet or genes.

- The function $q\left(x, u\right)$ is called a conditional quantile function.

- It contains the same information as the conditional cdf (it is its inverse), but is in the form of a statistical equation for outcomes that may be related to economic models.

- $Y = q\left(X, U\right)$ is just a statistical statement: e.g. for $X = 15$ and $U = 0.5$, Y is the weight of the median girl aged 15, but one that can be given substantive content.

CDC Growth Charts: United States



Weight-for-age percentiles: Girls, 2 to 20 years

Published May 30, 2000.

SOURCE: Developed by the National Center for Health Statistics in collaboration with
the National Center for Chronic Disease Prevention and Health Promotion (2000).

CDC

SAFER • HEALTHIER • PEOPLE™

Quantile function of normal linear regression

- If the distribution of Y conditioned on X is the normal linear regression model of elementary econometrics:

$$Y = \alpha + \beta X + V \text{ with } V \mid X \sim \mathcal{N}\left(0, \sigma^2\right),$$

  the variable U is the rank of V and it is easily seen that

$$q\left(x, u\right) = \alpha + \beta x + \sigma \Phi^{-1}\left(u\right)$$

  where $\Phi\left(.\right)$ is the standard normal cdf.

- In this case all quantiles are linear and parallel, a situation that is at odds with the growth chart example.

Linear quantile regression (QR)

- The linear QR model postulates linear dependence on X but allows for a different slope and intercept at each quantile $u \in (0,1)$

$$q(x,u) = \alpha(u) + \beta(u)x \tag{1}$$

- In the normal linear regression $\beta(u) = \beta$ and $\alpha(u) = \alpha + \sigma\Phi^{-1}(u)$.
- In linear regression one estimates α and β by minimizing the sum of squares of the residuals $Y_i - a - bX_i$ $(i = 1, ..., n)$.
- In QR one estimates $\alpha(u)$ and $\beta(u)$ for fixed u by minimizing a sum of absolute residuals where $(+)$ residuals are weighted by u and (-) residuals by $1 - u$.
- Its rationale is that a quantile minimizes expected asymmetric absolute value loss.
- For the median $u = 0.5$, so estimates of $\alpha(0.5)$, $\beta(0.5)$ are least absolute deviations.
- All observations are involved in determining the estimates of $\alpha(u)$, $\beta(u)$ for each u.
- Under random sampling and standard regularity conditions, sample QR coefficients are $\sqrt{n}$-consistent and asymptotically normal.
- Standard errors can be easily obtained via analytic or bootstrap calculations.
- The popularity of linear QR is due to its computational simplicity: computing a QR is a linear programming problem (Koenker 2005).

Linear quantile regression (QR) (continued)

- One use of QR is as a technique for describing a conditional distribution. For example, QR is a popular tool in wage decomposition studies.

- However, a linear QR can also be seen as a semiparametric random coefficient model with a single unobserved factor:

$$Y_i = \alpha\left(U_i\right) + \beta\left(U_i\right) X_i$$

  where $U_i \sim \mathcal{U}\left(0,1\right)$ independent of X_i.

- For example, this model determines log earnings Y_i as a function of years of schooling X_i and ability U_i, where $\beta\left(U_i\right)$ represents an ability-specific return to schooling.

- This is a model that can capture interactions between observables and unobservables.

- A special case of model with an interaction between X_i and U_i is the heteroskedastic regression $Y \mid X \sim \mathcal{N}\left[\alpha + \beta X, \left(\sigma + \gamma X\right)^2\right]$.
  - In this case $\alpha\left(u\right) = \alpha + \sigma\Phi^{-1}\left(u\right)$ and $\beta\left(u\right) = \beta + \gamma\Phi^{-1}\left(u\right)$.

Part 2

Quantiles methods: formal development

I. Unconditional quantiles

- Let $F(r) = \Pr(Y \leq r)$. For $\tau \in (0,1)$, the τth population quantile of Y is defined to be
$$Q_\tau(Y) \equiv q_\tau \equiv F^{-1}(\tau) = \inf\{r : F(r) \geq \tau\}.$$
- $F^{-1}(\tau)$ is a generalized inverse function. It is a left-continuous function with range equal to the support of F and hence often unbounded.

Equivariance of quantiles under monotone transformations

- This is an interesting property of quantiles not shared by expectations.
- Let $g(.)$ be a nondecreasing function. Then, for any random variable Y
$$Q_\tau[g(Y)] = g[Q_\tau(Y)].$$
- Thus, the quantiles of $g(Y)$ coincide with the transformed quantiles of Y.
- To see this note that
$$\Pr[Y \leq Q_\tau(Y)] = \tau \Rightarrow \Pr(g(Y) \leq g[Q_\tau(Y)]) = \tau.$$

Asymmetric absolute loss

- Let us define the "check" function (or asymmetric absolute loss function). For $\tau \in (0, 1)$

$$\rho_\tau(u) = [\tau \mathbf{1}(u \geq 0) + (1 - \tau)\mathbf{1}(u < 0)] \times |u| = [\tau - \mathbf{1}(u < 0)] u.$$

- Note that $\rho_\tau(u)$ is a continuous piecewise linear function, but nondifferentiable at $u = 0$. We should think of u as an individual error $u = y - r$ and $\rho_\tau(u)$ as the loss associated with u.
- Using $\rho_\tau(u)$ as a specification of loss, it turns out that q_τ minimizes expected loss:

$$s_0(r) \equiv E[\rho_\tau(Y - r)] = \tau \int_r^\infty (y - r)\, dF(y) - (1 - \tau) \int_{-\infty}^r (y - r)\, dF(y).$$

- Any element of $\{r : F(r) = \tau\}$ minimizes expected loss. If the solution is unique, it coincides with q_τ as defined above. If not, we have an interval of τth quantiles and the smallest element is chosen so that the quantile function is left-continuous.

Sample quantiles

- Given a random sample $\{Y_1, ..., Y_N\}$ we obtain sample quantiles replacing F by the empirical cdf:

$$F_N(r) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}(Y_i \leq r).$$

- That is, we choose $\widehat{q}_\tau = F_N^{-1}(\tau) \equiv \inf\{r : F_N(r) \geq \tau\}$, which minimizes

$$s_N(r) = \int \rho_\tau (y - r)\, dF_N(y) = \frac{1}{N}\sum_{i=1}^{N} \rho_\tau (Y_i - r).$$

- An important advantage of expressing the calculation of sample quantiles as an optimization problem, as opposed to a problem of ordering the observations, is computational (specially in the regression context).
- The optimization perspective is also useful for studying statistical properties.

Linear program representation

- An alternative presentation of the minimization leading to $\widehat{q}_\tau$ is

$$\min_{r, u_i^+, u_i^-} \sum_{i=1}^{N} \left[\tau u_i^+ + (1-\tau) u_i^- \right]$$

  subject to

$$Y_i - r = u_i^+ - u_i^-, \qquad u_i^+ \geq 0, u_i^- \geq 0, \qquad (i = 1, ..., N)$$

  where here $\left\{ u_i^+, u_i^- \right\}_{i=1}^{2N}$ denote $2N$ artificial additional arguments, which allow us to represent the original problem in the form of a linear program.
- We are using the notation $\rho_\tau(u) = \tau u^+ + (1-\tau) u^-$ with $u^+ = \mathbf{1}(u \geq 0)|u|$ and $u^- = \mathbf{1}(u < 0)|u|$.
- Note that

$$u^+ - u^- = \mathbf{1}(u \geq 0)|u| - \mathbf{1}(u < 0)|u| = \mathbf{1}(u \geq 0) u + \mathbf{1}(u < 0) u = u.$$

- A linear program takes the form:

$$\min_x c'x \text{ subject to } Ax \geq b, x \geq 0.$$

- The simplex algorithm for numerical solution of this problem was created by George Dantzig in 1947.

Nonsmoothness in sample but smoothness in population

- The sample objective function $s_N(r)$ is continuous but not differentiable for all r.
- Moreover, the gradient or moment condition

$$b_N(r) = \frac{1}{N}\sum_{i=1}^{N}\left[\mathbf{1}(Y_i \le r) - \tau\right]$$

  is not continuous in r.
- Note that if each Y_i is distinct, so that we can reorder the observations to satisfy $Y_1 < Y_2 < ... < Y_N$, for all τ we have

$$\left|b_N(\widehat{q}_\tau)\right| \equiv \left|F_N(\widehat{q}_\tau) - \tau\right| \le \frac{1}{N}.$$

- Despite lack of smoothness in $s_N(r)$ or $b_N(r)$, smoothness of the distribution of the data can smooth their population counterparts.
- Suppose that F is differentiable at q_τ with positive derivative $f(q_\tau)$, then $s_0(r)$ is twice continuously differentiable with derivatives:

$$\frac{d}{dr}E\left[\rho_\tau(Y-r)\right] = -\tau\left[1-F(r)\right] + (1-\tau)F(r) = F(r) - \tau \equiv E\left[\mathbf{1}(Y \le r) - \tau\right]$$

$$\frac{d^2}{dr^2}E\left[\rho_\tau(Y-r)\right] = f(r).$$

Consistency

- Since a sample quantile does not have a closed form expression we need a method for establishing the consistency of an estimator that maximizes an objective function.

- A theorem taken from Newey and McFadden (1994) provides such a method.

- The requirements are boundedness of the parameter space, uniform convergence of the objective function to some nonstochastic continuous limit, and that the limiting objective function is uniquely maximized at the truth (identification).

- The quantile sample objective function $s_N(r)$ is continuous and convex in r.

- Suppose that F is such that $s_0(r)$ is uniquely maximized at q_τ. By the law of large numbers $s_N(r)$ converges pointwise to $s_0(r)$. Then use the fact that pointwise convergence of convex functions implies uniform convergence on compact sets.

Asymptotic normality

- The asymptotic normality of sample quantiles cannot be established in the standard way because of the nondifferentiability of the objective function.
- However, it has long been known that under suitable conditions sample quantiles are asymptotically normal and there are direct approaches to establish the result.
- Here we just re-state the asymptotic normality result for unconditional quantiles following results on nonsmooth GMM around Newey and McFadden's theorems.
- The idea is that as long as the limiting objective function is differentiable the approach for differentiable problems works if a stochastic equicontinuity assumption holds.
- Fix $0 < \tau < 1$. If F is differentiable at q_τ with positive derivative $f\left(q_\tau\right)$, then

$$\sqrt{N}\left(\widehat{q}_\tau - q_\tau\right) = -\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{\mathbf{1}\left(Y_i \leq q_\tau\right) - \tau}{f\left(q_\tau\right)} + o_p\left(1\right).$$

- Consequently,

$$\sqrt{N}\left(\widehat{q}_\tau - q_\tau\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\tau\left(1-\tau\right)}{\left[f\left(q_\tau\right)\right]^2}\right).$$

- The term $\tau\left(1-\tau\right)$ in the numerator of the asymptotic variance tends to make $\widehat{q}_\tau$ more precise in the tails, whereas the density term in the denominator tends to make $\widehat{q}_\tau$ less precise in regions of low density.
- Typically the latter effect will dominate so that quantiles closer to the extremes will be estimated with less precision.

Computing standard errors

- The asymptotic normality result justifies the large N approximation

$$\frac{\widehat{f}\left(\widehat{q}_{\tau}\right)}{\sqrt{\tau\left(1-\tau\right)}}\sqrt{N}\left(\widehat{q}_{\tau}-q_{\tau}\right)\approx\mathcal{N}\left(0,1\right)$$

  where $\widehat{f}\left(\widehat{q}_{\tau}\right)$ is a consistent estimator of $f\left(q_{\tau}\right)$.
- Since

$$f\left(r\right)=\lim_{h\rightarrow 0}\frac{F\left(r+h\right)-F\left(r-h\right)}{2h}\equiv\lim_{h\rightarrow 0}\frac{1}{2h}E\left[\mathbf{1}(\left|Y-r\right|\leq h)\right],$$

  an obvious possibility is to use the histogram estimator

$$\widehat{f}\left(r\right)=\frac{F_{N}\left(r+h_{N}\right)-F_{N}\left(r-h_{N}\right)}{2h_{N}}=\frac{1}{2Nh_{N}}\sum_{i=1}^{N}\left[\mathbf{1}(Y_{i}\leq r+h_{N})-\mathbf{1}\left(Y_{i}\leq r-h_{N}\right)\right]$$

$$=\frac{1}{2Nh_{N}}\sum_{i=1}^{N}\mathbf{1}(\left|Y_{i}-r\right|\leq h_{N})$$

  for some $h_N > 0$ sequence such that $h_N \rightarrow 0$ as $N \rightarrow \infty$. Thus,

$$\widehat{f}\left(\widehat{q}_{\tau}\right)=\frac{1}{2Nh_{N}}\sum_{i=1}^{N}\mathbf{1}(\left|Y_{i}-\widehat{q}_{\tau}\right|\leq h_{N}).$$

- A sufficient condition for consistency is $\sqrt{N}h_{N}\rightarrow\infty$.
- Other alternatives are kernel estimators for $f\left(q_{\tau}\right)$, the bootstrap, or directly obtain an approximate confidence interval using the normal approximation to the binomial.

II. Conditional quantiles

- Consider the conditional distribution of Y given X:

$$\Pr\left(Y \leq r \mid X\right) = F\left(r; X\right)$$

  and denote the τth quantile of Y given X as

$$Q_{\tau}\left(Y \mid X\right) \equiv q_{\tau}\left(X\right) \equiv F^{-1}\left(\tau; X\right).$$

- Now quantiles minimize expected asymmetric absolute loss in a conditional sense:

$$q_{\tau}\left(X\right) = \arg\min_{c} E\left[\rho_{\tau}\left(Y - c\right) \mid X\right].$$

- Suppose that $q_{\tau}\left(X\right)$ satisfies a parametric model $q_{\tau}\left(X\right) = g\left(X, \beta_{\tau}\right)$, then

$$\beta_{\tau} = \arg\max_{b} E\left[\rho_{\tau}\left(Y - g\left(X, b\right)\right)\right].$$

- Also, since in general

$$\Pr\left(Y \leq q_{\tau}\left(X\right) \mid X\right) = \tau \quad \text{or} \quad E\left[1\left(Y \leq q_{\tau}\left(X\right)\right) - \tau \mid X\right] = 0,$$

  it turns out that β_{τ} solves moment conditions of the form

$$E\left\{h\left(X\right)\left[1\left(Y \leq g\left(X, \beta_{\tau}\right)\right) - \tau\right]\right\} = 0.$$

Conditional quantiles in a location-scale model

- The standardized variable in a location-scale model of $Y \mid X$ has a distribution that is independent of X.
- Namely, letting $E\left(Y \mid X\right) = \mu\left(X\right)$ and $Var\left(Y \mid X\right) = \sigma^2\left(X\right)$, the variable
$$V = \frac{Y - \mu\left(X\right)}{\sigma\left(X\right)}$$
is distributed independently of X according to some cdf G.
- Thus, in a location scale model all dependence of Y on X occurs through mean translations and variance re-scaling.
- In the location-scale model:
$$\Pr\left(Y \leq r \mid X\right) = \Pr\left(\frac{Y - \mu\left(X\right)}{\sigma\left(X\right)} \leq \frac{r - \mu\left(X\right)}{\sigma\left(X\right)} \mid X\right) = G\left(\frac{r - \mu\left(X\right)}{\sigma\left(X\right)}\right)$$
and
$$G\left(\frac{Q_\tau\left(Y \mid X\right) - \mu\left(X\right)}{\sigma\left(X\right)}\right) = \tau$$
or
$$Q_\tau\left(Y \mid X\right) = \mu\left(X\right) + \sigma\left(X\right) G^{-1}\left(\tau\right)$$
so that
$$\frac{\partial Q_\tau\left(Y \mid X\right)}{\partial X_j} = \frac{\partial \mu\left(X\right)}{\partial X_j} + \frac{\partial \sigma\left(X\right)}{\partial X_j} G^{-1}\left(\tau\right).$$

Conditional quantiles in a location-scale model (continued)

- Under homoskedasticity, $\partial Q_{\tau}\left(Y \mid X\right)/\partial X_j$ is the same at all quantiles since they only differ by a constant term.

- More generally, in a location-scale model the relative change between two quantiles $\partial \ln\left[Q_{\tau_1}\left(Y \mid X\right) - Q_{\tau_2}\left(Y \mid X\right)\right]/\partial X_j$ is the same for any pair $\left(\tau_1, \tau_2\right)$.

Structural representation

- Define U such that
$$F\left(Y;X\right)=U.$$
- It turns out that U is uniformly distributed independently of X between 0 and 1.
- Note that if $\Pr\left(Y\leq r\mid X\right)=F\left(r;X\right)$ then $\Pr\left(F\left(Y;X\right)\leq F\left(r;X\right)\mid X\right)=F\left(r;X\right)$ or $\Pr\left(U\leq s\mid X\right)=s$.
- Also
$$Y=F^{-1}\left(U;X\right)\text{ with }U\mid X\sim\mathcal{U}\left(0,1\right).$$
- This is sometimes called the Skorohod representation.
- For example, the Skorohod representation of the Gaussian linear regression model is $Y=X'\beta+\sigma V$ with $V=\Phi^{-1}\left(U\right)$, so that $V\mid X\sim\mathcal{N}\left(0,1\right)$.

III. Quantile regression

- A linear regression is an optimal linear predictor that minimizes average quadratic loss. Given data $\{Y_i, X_i\}_{i=1}^{N}$ OLS sample coefficients are given by

$$\widehat{\beta}_{OLS} = \arg\min_{b} \sum_{i=1}^{N} \left(Y_i - X_i' b\right)^2 .$$

- If $E\left(Y \mid X\right)$ is linear it coincides with the least squares population predictor, so that $\widehat{\beta}_{OLS}$ consistently estimates $\partial E\left(Y \mid X\right)/\partial X$.
- For robustness in the regression context one may be interested in median regression. That is, an optimal predictor that minimizes average absolute loss:

$$\widehat{\beta}_{LAD} = \arg\min_{b} \sum_{i=1}^{N} \left|Y_i - X_i' b\right| .$$

- If $med\left(Y \mid X\right)$ is linear it coincides with the least absolute deviation (LAD) population predictor, so that $\widehat{\beta}_{LAD}$ consistently estimates $\partial med\left(Y \mid X\right)/\partial X$.
- The idea can be generalized to quantiles other than $\tau = 0.5$ by considering optimal predictors that minimize average asymmetric absolute loss:

$$\widehat{\beta}\left(\tau\right) = \arg\min_{b} \sum_{i=1}^{N} \rho_{\tau}\left(Y_i - X_i' b\right) .$$

- As before if $Q_{\tau}\left(Y \mid X\right)$ is linear, $\widehat{\beta}\left(\tau\right)$ consistently estimates $\partial Q_{\tau}\left(Y \mid X\right)/\partial X$.

Asymptotic inference for quantile regression

- The first and second derivatives of the limiting objective function are:

$$\begin{aligned}\frac{\partial}{\partial b} E\left[\rho_\tau\left(Y - X'b\right)\right] &= E\left\{X\left[1\left(Y \leq X'b\right) - \tau\right]\right\} \\ \frac{\partial^2}{\partial b \partial b'} E\left[\rho_\tau\left(Y - X'b\right)\right] &= E\left[f\left(X'b \mid X\right) XX'\right] = H\left(b\right)\end{aligned}$$

- Moreover, under some regularity conditions we can use Newey and McFadden's asymptotic normality theorem, leading to

$$\sqrt{N}\left[\widehat{\beta}\left(\tau\right) - \beta\left(\tau\right)\right] = -H_0{}^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N} X_i\left\{1\left[Y_i \leq X_i'\beta\left(\tau\right)\right] - \tau\right\} + o_p\left(1\right).$$

where $H_0 = H\left(\beta\left(\tau\right)\right)$ is the Hessian of the limit objective function at the truth, and

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N} X_i\left\{1\left[Y_i \leq X_i'\beta\left(\tau\right)\right] - \tau\right\} \xrightarrow{d} \mathcal{N}\left(0, V_0\right)$$

where

$$V_0 = E\left(\left\{1\left[Y_i \leq X_i'\beta\left(\tau\right)\right] - \tau\right\}^2 X_i X_i'\right) = \tau\left(1 - \tau\right) E\left(X_i X_i'\right).$$

- The last equality follows under the assumption of linearity of conditional quantiles. Thus,

$$\sqrt{N}\left[\widehat{\beta}\left(\tau\right) - \beta\left(\tau\right)\right] \xrightarrow{d} \mathcal{N}\left(0, W_0\right) \qquad \text{with } W_0 = H_0^{-1} V_0 H_0^{-1}.$$

Getting consistent standard errors

- To get a consistent estimate of W_0 we need consistent estimates of H_0 and V_0.
- A simple estimator of H_0 suggested in Powell (1984, 1986), which mimics the histogram estimator discussed above, is as follows:

$$\widehat{H} = \frac{1}{2Nh_N} \sum_{i=1}^{N} \mathbf{1}\left(\left|Y_i - X_i'\widehat{\beta}(\tau)\right| \leq h_N\right) X_i X_i'.$$

- This estimator is motivated by the following iterated expectations argument:

$$\begin{aligned} H_0 &= E\left[f\left(X'\beta(\tau) \mid X\right) XX'\right] \equiv \lim_{h\to 0} \frac{1}{2h} E\left\{E\left[\mathbf{1}(\left|Y - X'\beta(\tau)\right| \leq h) \mid X\right] XX'\right\} \\ &= \lim_{h\to 0} \frac{1}{2h} E\left[\mathbf{1}(\left|Y - X'\beta(\tau)\right| \leq h) XX'\right]. \end{aligned}$$

- If the quantile function is correctly specified a consistent estimate of V_0 is

$$\widehat{V} = \tau(1-\tau)\frac{1}{N}\sum_{i=1}^{N} X_i X_i'.$$

- Otherwise, a fully robust estimator can be obtained using

$$\widetilde{V} = \frac{1}{N}\sum_{i=1}^{N}\left\{1\left[Y_i \leq X_i'\widehat{\beta}(\tau)\right] - \tau\right\}^2 X_i X_i'$$

Getting consistent standard errors (continued)

- Finally, if $U_\tau = Y - X'\beta(\tau)$ is independent of X (as in the location model) it turns out that

  $$H_0 = f_{U_\tau}(0)\, E\left(X_i X_i'\right)$$

  so that

  $$W_0 = \frac{\tau(1-\tau)}{\left[f_{U_\tau}(0)\right]^2}\left[E\left(X_i X_i'\right)\right]^{-1},$$

  which can be consistently estimated as

  $$\widehat{W}_{NR} = \frac{\tau(1-\tau)}{\left[\widehat{f}_{U_\tau}(0)\right]^2}\left(\frac{1}{N}\sum_{i=1}^{N} X_i X_i'\right)^{-1}$$

  where

  $$\widehat{f}_{U_\tau}(0) = \frac{1}{2Nh_N}\sum_{i=1}^{N}\mathbf{1}\left(\left|Y_i - X_i'\widehat{\beta}(\tau)\right| \leq h_N\right).$$

- In summary, we have considered three different alternative estimators for standard errors:
  - A non-robust variance matrix estimator under independence $\widehat{W}_{NR}$,
  - a robust estimator under correct specification: $\widehat{W}_R = \widehat{H}^{-1}\widehat{V}\widehat{H}^{-1}$,
  - and a fully robust estimator under misspecification: $\widehat{W}_{FR} = \widehat{H}^{-1}\widetilde{V}\widehat{H}^{-1}$.

Part 3

Flexible QR, instrumental variables and quantile treatment effects

Introduction

- As a model for causal analysis, linear QR faces similar challenges as ordinary linear regression. Namely, linearity, exogeneity and rank invariance.

- Next we discuss each of these aspects in turn.

- Other topics not covered:
  - Censored regression quantiles
  - Crossings and rearrangements.
  - Decompositions: Machado and Mata (2005) counterfactuals.
  - Quantile regression under misspecification (Angrist et al, 2006).
  - Functional inference.

Flexible QR

- Linearity is restrictive. It may also be at odds with the monotonicity requirement of $q(x, u)$ in u for every value of x.

- Linear QR may be interpreted as an approximation to the true quantile function (Angrist, Chernozhukov, and Fernández-Val 2006).

- An approach to nonparametric QR is to use series methods:

$$q(x, u) = \theta_0(u) + \theta_1(u) g_1(x) + ... + \theta_P(u) g_P(x).$$

- The g's are anonymous functions without an economic interpretation. Objects of interest are derivative effects and summary measures of them.

- In practice one may use orthogonal polynomials, wavelets or splines (Chen 2007).

- This type of specification may be seen as an approximating model that becomes more accurate as P increases, or simply as a parametric flexible model of the quantile function.

- From the point of view of computation the model is still a linear QR, but the regressors are now functions of X instead of the Xs themselves.

Exogeneity and rank invariance

- To discuss causality it is convenient to use a single $0-1$ binary treatment X_i and a potential outcome notation Y_{0i} and Y_{1i}.
- Let U_{0i}, U_{1i} be ranks of potential outcomes and $q_0(u)$, $q_1(u)$ the quantile functions.
- Note that unit i may be ranked differently in the distributions of the two potential outcomes, so that $U_{0i} \neq U_{1i}$. The causal effect for unit i is given by

$$Y_{1i} - Y_{0i} = q_1(U_{1i}) - q_0(U_{0i}).$$

- Under exogeneity X_i is independent of (Y_{0i}, Y_{1i}).
- The implication is that the quantile function of $Y_i \mid X_i = 0$ coincides with $q_0(u)$ and the quantile function of $Y_i \mid X_i = 1$ coincides with $q_1(u)$, so that

$$\beta(u) = q_1(u) - q_0(u).$$

- This quantity is often called a quantile treatment effect (QTE). In general it is just the difference between the quantiles of two different distributions.
- It will only represent the gain or loss from treatment of a particular unit under a rank invariance condition. i.e. that the ranks of potential outcomes are equal to each other.
- Under rank invariance treatment gains may still be heterogeneous but a single unobservable variable determines the variation in the two potential outcomes.
- Next we introduce IV endogeneity in a quantile model with rank invariance.

Instrumental variable QR

- The linear instrumental variable (IV) model of elementary econometrics assumes

$$Y_i = \alpha + \beta X_i + V_i$$

  where X_i and V_i are correlated, but there is an instrumental variable Z_i that is independent of V_i and a predictor of X_i.

- Potential outcomes are of the form $Y_{x,i} = \alpha + \beta x + V_i$ so that rank invariance holds.

- If x is a $0 - 1$ binary variable, $Y_{0,i} = \alpha + V_i$ and $Y_{1,i} = \alpha + \beta + V_i$.

- A QR generalization subject to rank invariance is to consider

$$Y_{x,i} = q\left(x, U_i\right).$$

- A linear version of which is

$$Y_{x,i} = \alpha\left(U_i\right) + \beta\left(U_i\right)x.$$

Instrumental variable QR (continued)

- Chernozhukov and Hansen (2006) propose to estimate $\alpha(u)$ and $\beta(u)$ for given u by directly exploiting the IV exclusion restriction.
- Specifically, if we write the model as

$$Y_i = \alpha(U_i) + \beta(U_i) X_i + \gamma(U_i) Z_i,$$

  the IV assumption asserts that Z_i only affects Y_i via X_i so that $\gamma(u) = 0$ for each u.
- Now let $\widehat{\gamma}_u(b)$ be the estimated slope coefficient in a u-quantile regression of $(Y_i - bX_i)$ on Z_i and a constant term.
- The idea, which mimics the operation of 2SLS, is to choose as estimate of $\beta(u)$ the value of b that minimizes $|\widehat{\gamma}_u(b)|$, hence enforcing the exclusion restriction.
- In the absence of rank invariance the treatment effects literature (e.g. Abadie 2003) has focused on QTEs for compliers in the context of a binary treatment that satisfies a monotonicity assumption.

Matching and quantile treatment effects

- Experiments guarantee the independence condition
$$(Y_1, Y_0) \perp D$$
but with observational data this is not very plausible.
- A less demanding condition for nonexperimental data is:
$$(Y_1, Y_0) \perp D \mid X.$$
- Conditional independence implies
$$\begin{aligned} E(Y_1 \mid X) &= E(Y_1 \mid D=1, X) = E(Y \mid D=1, X) \\ E(Y_0 \mid X) &= E(Y_0 \mid D=0, X) = E(Y \mid D=0, X). \end{aligned}$$
Therefore, for α_{ATE} we can calculate (and similarly for α_{TT}):
$$\begin{aligned} \alpha_{ATE} &= E(Y_1 - Y_0) = \int E(Y_1 - Y_0 \mid X)\, dF(X) \\ &= \int [E(Y \mid D=1, X) - E(Y \mid D=0, X)]\, dF(X). \end{aligned}$$
- The following is a matching expression for $\alpha_{TT} = E(Y_1 - Y_0 \mid D=1)$:
$$E[Y - E(Y_0 \mid D=1, X) \mid D=1] = E[Y - \mu_0(X) \mid D=1]$$
where $\mu_0(X) = E(Y \mid D=0, X)$ is used as an imputation for Y_0.

Distributional effects and quantile treatment effects

- Most of the literature focused on average effects, but the matching assumption also works for distributional comparisons.
- Under conditional independence the full marginal distributions of Y_1 and Y_0 can be identified.
- To see this, first note that we can identify not just α_{ATE} but also $E(Y_1)$ and $E(Y_0)$:

  $$E(Y_1) = \int E(Y_1 \mid X)\, dF(X) = \int E(Y \mid D = 1, X)\, dF(X)$$

  and similarly for $E(Y_0)$.
- Next, we can equally identify the expected value of any function of the outcomes $E[h(Y_1)]$ and $E[h(Y_0)]$:

  $$E[h(Y_1)] = \int E[h(Y_1) \mid X]\, dF(X) = \int E[h(Y) \mid D = 1, X]\, dF(X)$$
- Thus, setting $h(Y_1) = \mathbf{1}(Y_1 \leq r)$ we get

  $$E[\mathbf{1}(Y_1 \leq r)] = \Pr(Y_1 \leq r) = \int \Pr(Y \leq r \mid D = 1, X)\, dF(X)$$

  and similarly for $\Pr(Y_0 \leq r)$.
- Given identification of the *cdf*s we can also identify quantiles of Y_1 and Y_0.
- Quantile treatment effects are differences in the marginal quantiles of Y_1 and Y_0.
- More substantive objects are the joint distribution of (Y_1, Y_0) or the distribution of gains $Y_1 - Y_0$, but their identification requires stronger assumptions.

The common support condition

- Suppose for the sake of the argument that X is a single covariate whose support lies in the range $\{X_{MIN}, X_{MAX}\}$.
- The support for the subpopulation of the treated $(D = 1)$ is $\{X_{MIN}, X_I\}$ whereas the support for the controls $(D = 0)$ is $\{X_0, X_{MAX}\}$ and $X_0 < X_I$, so that

$$\Pr(D = 1 \mid X \in \{X_{MIN}, X_0\}) = 1$$

$$0 < \Pr(D = 1 \mid X \in \{X_0, X_I\}) < 1$$

$$\Pr(D = 1 \mid X \in \{X_I, X_{MAX}\}) = 0$$

- The implication is that $E(Y \mid D = 1, X)$ is only identified for values of X in the range $\{X_{MIN}, X_I\}$ and $E(Y \mid D = 0, X)$ is only identified for values of X in the range $\{X_0, X_{MAX}\}$.
- Thus, we can only calculate the difference $[E(Y \mid D = 1, X) - (Y \mid D = 0, X)]$ for values of X in the intersection range $\{X_0, X_I\}$, which implies that α_{ATE} is not identified. Only the average treatment effect of units with $X \in \{X_0, X_I\}$ is identified.
- If we want to ensure identification, in addition to conditional independence we need the overlap assumption:

$$0 < \Pr(D = 1 \mid X) < 1 \qquad \text{for all } X \text{ in its support}$$

Imputing missing outcomes (discrete X)

- Suppose X is discrete, takes on J values $\left\{\xi_j\right\}_{j=1}^{J}$ and we have a sample $\{X_i\}_{i=1}^{N}$. Let

$$\begin{aligned} N^j &= \text{number of observations in cell } j. \\ N^j_\ell &= \text{number of observations in cell } j \text{ with } D = \ell. \\ \overline{Y}^j_\ell &= \text{mean outcome in cell } j \text{ for } D = \ell. \end{aligned}$$

- Thus, $\left(\overline{Y}^j_1 - \overline{Y}^j_0\right)$ is the sample counterpart of

$$E\left(Y \mid D = 1, X = \xi_j\right) - E\left(Y \mid D = 0, X = \xi_j\right),$$

  which can be used to get the estimates

$$\widehat{\alpha}_{ATE} = \sum_{j=1}^{J}\left(\overline{Y}^j_1 - \overline{Y}^j_0\right)\frac{N^j}{N}, \qquad \widehat{\alpha}_{TT} = \sum_{j=1}^{J}\left(\overline{Y}^j_1 - \overline{Y}^j_0\right)\frac{N^j_1}{N_1}$$

- The formula for $\widehat{\alpha}_{TT}$ can also be written in the form

$$\widehat{\alpha}_{TT} = \frac{1}{N_1}\sum_{D_i=1}\left(Y_i - \overline{Y}^{j(i)}_0\right)$$

  where $j\left(i\right)$ is the cell of X_i. Thus, $\widehat{\alpha}_{TT}$ matches the outcome of each treated unit with the mean of the nontreated units in the same cell.
- To see this note that $E\left[E\left(Y \mid D = 1, X\right) - E\left(Y \mid D = 0, X\right) \mid D = 1\right] = E\left[Y - E\left(Y \mid D = 0, X\right) \mid D = 1\right]$.

Imputing missing outcomes (continuous X)

- A matching estimator can be regarded as a way of constructing imputations for missing potential outcomes so that gains $Y_{1i} - Y_{0i}$ can be estimated for each unit.
- In the discrete case

$$\widehat{Y}_{0i} = \overline{Y}_0^{j(i)} \equiv \sum_{k \in (D=0)} \frac{\mathbf{1}(X_k = X_i)}{\sum_{\ell \in (D=0)} \mathbf{1}(X_\ell = X_i)} Y_k$$

- In general

$$\widehat{Y}_{0i} = \sum_{k \in (D=0)} w(i,k)\, Y_k$$

- Different matching estimators use different weighting schemes.
- Nearest neighbor matching:

$$w(i,k) = \begin{cases} 1 \text{ if } X_k = \min_i \|X_k - X_i\| \\ 0 \text{ otherwise} \end{cases}$$

  with perhaps matching restricted to cases where $\|X_i - X_k\| < \varepsilon$ for some ε. Usually applied in situations where the interest is in α_{TT} but also applicable to α_{ATE}.
- Kernel matching:

$$w(i,k) = \frac{1}{\sum_{\ell \in (D=0)} K\left(\frac{X_\ell - X_i}{\gamma_{N_0}}\right)} K\left(\frac{X_k - X_i}{\gamma_{N_0}}\right)$$

  where $K(.)$ is a kernel that downweights distant observations and γ_{N_0} is a bandwidth parameter. Local linear approaches provide a generalization.

Methods based on the propensity score

- Rosenbaum and Rubin called "propensity score" to

$$\pi(X) = \Pr(D = 1 \mid X)$$

and proved that if $(Y_1, Y_0) \perp D \mid X$ then

$$(Y_1, Y_0) \perp D \mid \pi(X)$$

provided $0 < \pi(X) < 1$ for all X.

- We want to prove that provided $(Y_1, Y_0) \perp D \mid X$ then $\Pr(D = 1 \mid Y_1, Y_0, \pi(X)) = \Pr(D = 1 \mid \pi(X)) \equiv \pi(X)$. Using the law of iterated expectations:

$$\begin{aligned} E(D \mid Y_1, Y_0, \pi(X)) &= E[E(D \mid Y_1, Y_0, X) \mid Y_1, Y_0, \pi(X)] \\ &= E[E(D \mid X) \mid Y_1, Y_0, \pi(X)] = \pi(X) \end{aligned}$$

- The result tells us that we can match units with very different values of X as long as they have similar values of $\pi(X)$.
- These results suggest two-step procedures in which we begin by estimating the propensity score.

Weighting on the propensity score

- Under unconditional independence

$$\alpha_{ATE} = E\left(Y \mid D=1\right) - E\left(Y \mid D=0\right) = \frac{E\left(DY\right)}{\Pr\left(D=1\right)} - \frac{E\left[\left(1-D\right)Y\right]}{\Pr\left(D=0\right)}$$

- Similarly, under conditional independence

$$\begin{aligned} E\left(Y_1 - Y_0 \mid X\right) &= E\left(Y \mid D=1, X\right) - E\left(Y \mid D=0, X\right) \\ &= \frac{E\left(DY \mid X\right)}{\Pr\left(D=1 \mid X\right)} - \frac{E\left[\left(1-D\right)Y \mid X\right]}{\Pr\left(D=0 \mid X\right)} \\ &= E\left(\frac{DY}{\pi\left(X\right)} - \frac{\left(1-D\right)Y}{1-\pi\left(X\right)} \mid X\right) \end{aligned}$$

so that

$$\alpha_{ATE} = E\left(\frac{DY}{\pi\left(X\right)} - \frac{\left(1-D\right)Y}{1-\pi\left(X\right)}\right) = E\left(Y\frac{\left[D-\pi\left(X\right)\right]}{\pi\left(X\right)\left[1-\pi\left(X\right)\right]}\right)$$

- A simple estimator is

$$\widehat{\alpha}_{ATE} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{D_i Y_i}{\widehat{\pi}\left(X_i\right)} - \frac{\left(1-D_i\right)Y_i}{1-\widehat{\pi}\left(X_i\right)}\right)$$

where $\widehat{\pi}\left(X_i\right)$ is a nonparametric series estimator of the propensity score (Hirano, Imbens, and Ridder, 2003).

Quantile treatment effects (Firpo 2007)

- Let (Y_1, Y_0) be potential outcomes with marginal cdfs $F_1(r), F_0(r)$ and quantile functions $Q_{1\tau} = F_1^{-1}(\tau), Q_{0\tau} = F_0^{-1}(\tau)$. The QTE is defined to be

$$\theta_0 = Q_{1\tau} - Q_{0\tau}$$

- Under conditional exogeneity $F_j(r) = \int \Pr(Y \leq r \mid D = j, X)\, dG(X)$, $(j = 0, 1)$. Moreover, $Q_{1\tau}, Q_{0\tau}$ satisfy the moment conditions:

$$\begin{aligned} E\left[\frac{D}{\pi(X)} 1(Y \leq Q_{1\tau}) - \tau\right] &= 0 \\ E\left[\frac{1-D}{1-\pi(X)} 1(Y \leq Q_{0\tau}) - \tau\right] &= 0 \end{aligned}$$

  and

$$Q_{1\tau} = \arg\min_q E\left[\frac{D}{\pi(X)} \rho_\tau(Y - q)\right], Q_{0\tau} = \arg\min_q E\left[\frac{1-D}{1-\pi(X)} \rho_\tau(Y - q)\right].$$

  where $\rho_\tau(u) = [\tau - 1(u < 0)] \times u$ is the "check" function.

- Firpo's method is a two-step weighting procedure in which the propensity score $\pi(X)$ is estimated first.