14.385 Recitation 5: Quantile Regression

Paul Schrimpf

October 10, 2008

1 Inference

This section goes over the asymptotic behavior of quantile regression. It is based on Koenker (2005).

1.1 Setup

Let $\{Y_i\}$ be independent random variables with distributions $\{F_i\}$. Suppose that the τth quantile of Y_i given x_i is linear in x:

$$Q_{Y_i}(\tau|x_i) = x_i'\beta(\tau) \tag{1}$$

By definition, we have

$$F_i^{-1}(\tau|x_i) = Q_{Y_i}(\tau|x_i) \equiv \xi_i(\tau) \tag{2}$$

We will consider the behavior of the quantile regression estimator:

$$\hat{\beta}_n(\tau) = \underset{b\in\Re^p}{\arg\min} \sum \rho_\tau(y_i - x_i'b) \tag{3}$$

where $\rho_\tau(u) = u\left(\tau - \mathbf{1}(u<0)\right)$ is the check function.

1.2 Asymptotic Distribution

Assume:

A1 $\{F_i\}$ are uniformly continuous with $f_i(\xi)$ uniformly bounded away from 0 and ∞ at $\{\xi_i(\tau)\}$.

A2 There exist positive definite D_0 and $D_1(\tau)$ such that

(a) $\lim_{n\to\infty} n^{-1}\sum x_i x_i' = D_0$

(b) $\lim_{n\to\infty} n^{-1}\sum f_i(\xi_i(\tau))x_i x_i' = D_1(\tau)$

(c) $\max ||x_i||/\sqrt{n} \to 0$

Theorem 1 (Asymptotic Normality). *Under these assumptions,*

$$\sqrt{n}\left(\hat{\beta}_n(\tau) - \beta(\tau)\right) \xrightarrow{d} N\left(0, \tau(1-\tau)D_1(\tau)^{-1}D_0D_1(\tau)^{-1}\right)$$

Proof. We show this result by finding the limit distribution of the objective function. Convexity then implies that the limit distribution of the estimator is the distribution of the minimizer limiting objective function.

Rewrite the objetive function as:

$$\begin{aligned} Q_n(\hat{\beta}_n(\tau)) &= \sum \rho_\tau(y_i - x_i'\hat{\beta}_n(\tau)) && (4)\\ &= \sum \rho_\tau(y_i - x_i'\beta(\tau) - x_i'(\sqrt{n}(\hat{\beta}_n(\tau) - \beta(\tau)))/\sqrt{n} && (5)\end{aligned}$$

Let $u_i = y_i - x_i'\beta(\tau)$ and $\delta = \sqrt{n}(\hat{\beta}_n(\tau) - \beta(\tau))$. We can add a constant to the objetive function without changing our estimator. Consider

$$Z_n(\delta) = \sum \rho_\tau(u_i - x_i'\delta/\sqrt{n}) - \rho_\tau(u_i) \tag{6}$$

As in Knight (1998), consider the identity:

$$\rho_\tau(u - v) - \rho_\tau(u) = -v\psi_\tau(u) + \int_0^v (\mathbf{1}(u \le s) - \mathbf{1}(u \le 0))ds \tag{7}$$

where $\psi_\tau(u) = \tau - \mathbf{1}(u < 0)$[1] Using this identity, we can rewrite () as:

$$Z_n(\delta) = -\sum x_i'\delta/\sqrt{n}\psi_\tau(u_i) + \sum \int_0^{x_i'\delta/\sqrt{n}} \mathbf{1}(u_i \le s) - \mathbf{1}(u_i \le 0)ds \tag{8}$$

We can deal with these two terms separately. Let $Z_{1n}(\delta) = -\sum x_i'\delta/\sqrt{n}\psi_\tau(u_i)$ and $Z_{2n}(\delta) = \sum \int_0^{x_i'\delta/\sqrt{n}} \mathbf{1}(u_i \le s) - \mathbf{1}(u_i \le 0)ds$.

A standard CLT applies to $Z_{1n}(\delta)$. It is a sum of independent terms with expectation 0 (because $E[\psi_\tau(u)|x_i] = 0$) and variance

$$\begin{aligned} E[x_i'\delta\psi_\tau(u_i)^2\delta' x_i] =& E[x_i'\delta E[\psi_\tau(u_i)^2|x_i]\delta' x_i] \\ =& \tau(1-\tau)\delta' E[x_i' x_i]\delta \end{aligned} \tag{9}$$

so,

$$Z_{1n}(\delta) \rightsquigarrow -\delta' W \text{ where } W \sim N(0, \tau(1-\tau)D_0) \tag{10}$$

Now, let $Z_{2ni}(\delta) = \int_0^{x_i'\delta/\sqrt{n}} \mathbf{1}(u_i \le s) - \mathbf{1}(u_i \le 0)ds$. Note that $P(u_i < s) = F_i(\xi_i + s)$ ($u_i < s$ means that the difference between y_i and its τth quantile is less than s, i.e. $y_i - \xi_i < s$ or $y_i < \xi_i + s$). Therefore,

$$\begin{aligned} \sum EZ_{2ni}(\delta) =& \sum \int_0^{x_i'\delta/\sqrt{n}} F_i(\xi_i + s) - F_i(\xi_i)ds \\ =& \frac{1}{n}\sum \int_0^{x_i'\delta} \sqrt{n}\left(F_i(\xi_i + t/\sqrt{n}) - F_i(\xi_i)\right)dt \\ =& \frac{1}{n}\sum \int_0^{x_i'\delta} f_i(\xi_i)tdt + o(1) \\ =& \frac{1}{2n}\sum f_i(\xi_i)\delta' x_i x_i'\delta + o(1) && (11) \\ \to& \frac{1}{2}\delta' D_1(\tau)\delta && (12) \end{aligned}$$

Furthermore, the variance of $(Z_{2n}(\delta))$ is bounded by:

$$V(Z_{2n}(\delta)) \le \frac{1}{\sqrt{n}} \max_i |x_i'\delta| \sum EZ_{2ni}(\delta) \tag{13}$$

[1] This identity is easily verified by plugging in the definitions of the various functions. Begin with the right side:

$$\begin{aligned} -v\psi_\tau(u) + \int_0^v (\mathbf{1}(u \le s) - \mathbf{1}(u \le 0))ds =& -v(\tau - \mathbf{1}(u < 0)) - v\mathbf{1}(u < 0) + \int_0^v \mathbf{1}(u \le s)ds \\ =& -v\tau + \begin{cases} v & u < 0,\, v > u \\ 0 & u < 0,\, v < u \\ (v - u) & u > 0,\, v > u \\ 0 & u > 0,\, v < u \end{cases} \\ =& -v\tau + (v-u)\mathbf{1}(v > u) + u\mathbf{1}(u < 0) \\ =& (u - v)(\tau - \mathbf{1}(u - v < 0)) - u(\tau - \mathbf{1}(u < 0)) \\ =& \rho_\tau(u - v) - \rho_\tau(u) \end{aligned}$$

By assumption 0c, $V(Z_{2n}(\delta)) \to 0$, so $Z_{2n}(\delta) \xrightarrow{p} EZ_{2n}(\delta)$ and we can conlude that

$$Z_n(\delta) \rightsquigarrow Z_0(\delta) = -\delta' W + \frac{1}{2}\delta' D_1 \delta \tag{14}$$

Finally, we obtain a limiting distribution for $\hat{\beta}(\tau)$ by noting that (ignoring some details)

$$\begin{aligned}\sqrt{n}(\hat{\beta}(\tau) - \beta(\tau)) =& \hat{\delta}_n = \arg\min Z_n(\delta) \\ & \rightsquigarrow \arg\min Z_0(\delta) = \hat{\delta}_0 = D_1^{-1} W \sim N\left(0, D_1^{-1} D_0 D_1^{-1} \tau(1-\tau)\right)\end{aligned} \tag{15}$$

□

1.3 Inference in Practice

We can apply the above result to perform Wald or t-tests. The primary difficulty is that we must estimate the inverse of the conditional density (aka the sparsity) at the τth quantile, $f_i(\xi_i(\tau))^{-1} = s(\tau)$. One option would be to use a standard kernel density estimator. Powell suggested using

$$\hat{D}_1(\tau)\frac{1}{nh_n}\sum K(\hat{u}_i(\tau)/h_n) x_i x_i'$$

Another approach recognizes that $\frac{d}{dt}F^{-1}(t|x) = x'\frac{d}{dt}\beta(t) = s(t)$. Then we can use, e.g. (as in Siddiqui (1960))

$$\hat{s}_n(t) = \frac{x'(\hat{\beta}(t+h_n) - \hat{\beta}(t-h_n))}{2h_n}$$

where $h_n \to 0$ as $n \to \infty$. More complicated ways of approximating the derivative are also possible.

Bootstrap To avoid having to estimate the sparsity function, one can also use the bootstrap for inference. Note that if we do not estimate the sparsity function, we will be bootstrapping a statistic that is not asymptotically pivotal. This has led to a number of papers about variants of the bootstrap and their rates. The residual bootstrap converges slower than standard asymptotic distribution. Smoothed variants of the bootstrap do as well as the standard asymptotic distribution. Another approach based on resampling of the subgradient condition has attractive computational properties, especially for non-convex problems such as censored quantile regression. See Koenker (2005) for more information.

Rank Based FIXME: rank based inference

Inference on $\beta(\cdot)$ Some interesting hypotheses depend on the entire function $\beta(\cdot)$ instead of just $\beta(\tau)$ at some fixed τ. For example, we might want to test whether x has an effect at any quantile, $H_0 : \beta(\tau) = 0 \forall \tau$, or whether x has a constant effect, $H_0 : \beta(\tau) = \beta(0.5)\forall\tau$. To test hypotheses of this form, we need to derive the limit distribution of $\sqrt{n}(hat\beta(\tau) - \beta(\tau))$ veiwed as a function of τ. From the result above we know that for a finite set of points, $\sqrt{n}D_0^{-1/2}D_1 \begin{pmatrix}\hat{\beta}(\tau_1) - \beta(\tau_1) \\ \vdots \\ \hat{\beta}(\tau_k) - \beta(\tau_k)\end{pmatrix}$, is asymptotically normal with variances $\tau_j(1-\tau_j)$ and covariances given by:

$$\begin{aligned}E\psi_{\tau_j}(u_i)\psi_{\tau_k}(u_i) =& E\left[(\tau_j - \mathbf{1}(y_i - x\beta(\tau_j) \le 0))(\tau_k - \mathbf{1}(y_i - x\beta(\tau_k) \le 0))\right] \\ =& (\tau_j \wedge \tau_k) - \tau_j \tau_k\end{aligned}$$

In fact, this convergence is true for all τ. $\sqrt{n}D_0^{-1/2}D_1(hat\beta(\tau) - \beta(\tau))$ converges to a random function, $\nu(\tau)$, which is normally distributed at any finite set of points with the variance above. This sort of random function is called a Brownian bridge. The sense in which this convergence occurs is called weak convergence, and is often denoted by $\rightsquigarrow$ or $\Rightarrow$. One definition of weak convergence is that for all bound continuous functions, $f : T \to \mathbb{R}$ (where T is the space in which $\nu(\cdot)$ lies. In this case, $T = \{g : [\epsilon, 1-\epsilon] \to \Re^k,\ g \text{ continuous}\}$ with the ℓ^∞ metric), $Ef(\sqrt{n}D_0^{-1/2}D_1(hat\beta(\cdot) - \beta(\cdot))) \to Ef(\nu(\cdot))$.

In fact, we already relied on this sort of convergence when we said the limit distribution of $\hat{\beta}(\tau)$ is the distribution of the minimum of the limiting objective function. For the problem at hand, weak convergence implies that

$$\sup_{\tau} n(\hat{\beta}(\tau) - \beta(\tau))' D_1' D_0^{-1} D_1 (\hat{\beta}(\tau) - \beta(\tau)) \xrightarrow{d} \sup_{\tau} \nu(\tau)'\nu(\tau)$$

The distribution of the later can be simulated to obtain critical values. This result can be used to test the hypothesis of no effect, $H_0 : \beta(\tau) = 0 \forall \tau$. Testing the hypothesis of constant effect is slightly more complicated since it involves the value of the constant effect as a nuisance parameter.

2 Applications and Extensions

2.1 Decomposition of Distribution Changes

One popular application of quantile regression has been the analysis of the change in inequality. During the 80s and 90s, income inequality increased in the US and much of the rest of the world. Labor economists have been interested in the mechanism through which this happened. One way of thinking about the increase in inequality is to try to break it into (1) changes in the observed distribution of characteristics, (2) changes in the prices of worker characteristics, and (3) residual changes. If I recall correctly, Katz and Murphy (1992?) were the first to consider this sort of decomposition, but they had a rather crude and ad-hoc method. DiNardo, Fortin, and, Lemieux (1996?) proposed a more complicated method based on kernel reweighting. Autor, Katz, and Kearney (2008) use a similar method. Machado and Mata (2005) use quantile regression to perform the decomposition. In a quantile regression,

$$y_{it} = x_{it}\beta_t(\tau)$$

x_{it} are the observed characteristics, β_t are the prices, and τ captures residual changes. For each year, t, we can estimate $\beta_t(\tau)$. We can use these estimates to simulate what y_{it} would have been had x's been distributed as in year s,

$$\hat{y}_{it|s} = x_{is}\hat{\beta}_t(\tau)$$

We can seperate out residual inequality by looking at the distribution of

$$\hat{u}_{it|s} = \hat{y}_{it|s} - E[\hat{y}_{it|s}|x_{is}] = x_{is}(\hat{\beta}_t(\tau) - \hat{\beta}_{1/2}(\tau))$$

Angrist, Chernozhukov, and Fernandez-Val (2006) is, in part, about how to interpret this sort of quantile regression if the true conditional quantile is not linear. A main result is that the $x\beta(\tau)$ minimizes a weighted squared difference from the true conditional quantile function. Victor also has a recent paper with Fernandez-Val and Melly about how to do inference on counterfactual distributions estimated in this way.

2.2 Selection

Consider the selection discussed in recitation 3, you have an outcome, y, that is a function of some regressors, x,

$$y = \mu(x) + \epsilon \tag{16}$$

but you do not observe y for the entire population, instead you only observed y if

$$g(z) - \nu > 0 \tag{17}$$

where ν and ϵ are potentially correlated. Assume that ν and ϵ are independent of x and z. We stated that it is possible to identify this model without making any parametric assumptions about the distribution of ν and ϵ. The key to showing identification is to assume that there is a set of values of z, Z^∞ that occur with positive probability such that $P(g(z) > \nu | z \in Z^\infty) = 1$. In this set, there is no selection problem since the fact that $z \in Z^\infty$ and y is observed tells us nothing about the value of ν. This means that $\mu(x)$ could be consistently estimated by standard methods using just the observations with $z \in Z^{\infty}$. Given $\mu(x)$, the rest of the parameters are easy to identify.

This sort of identification argument is often referred to identification at infinity because it relies on pushing z off to an extreme value. It is a fairly common method of proving identification. Unfortunately, at least for selection models, it is quite fragile. If there is no z with $P(g(z) > \nu) = 1$, then identification completely breaks down and there is no finite bound on $\mu(x)$. In practice this means that estimating the model nonparametrically can be very sensitive to the exact choice of method.

Manski (1989) pointed out that though we cannot bound the conditional mean of y, we can always bound the conditional distribution. We can write the conditional distribution of y given x as:

$$F(y|x) = F(y|x, z = 1)P(z = 1|x) + F(y|x, z = 0)P(z = 0|x)$$

The only unobserved part of the right side of the equation is $F(y|x, z = 0)$. Without some assumptions, all we know about $F(y|x, z = 0)$ is that it is between 0 and 1. This suggests the following bound on the conditional distribution of y given x:

$$F(y|x, z = 1)P(z = 1|x) \leq F(y|x) \leq F(y|x, z = 1)P(z = 1|x) + P(z = 0|x) \tag{18}$$

We can invert these bounds on the distribution function to obtain a bound on the conditional quantile function of y. I first heard about this idea in some lecture notes by Koenker that Victor showed. I don't think anyone has actually implemented it. The lower bound, $Q_0(\tau|x)$ solves:

$$\begin{aligned} \tau =& F(Q_0|x, z = 1)P(z = 1|x) + P(z = 0|x) \\ \frac{\tau - P(z = 0|x)}{P(z = 1|x)} =& F(Q_0|x, z = 1) \\ Q_0(\tau|x) =& \begin{cases} Q_Y\left(\frac{\tau - P(z=0|x)}{P(z=1|x)}|x, z = 1\right) & \text{if } \tau \geq P(z = 0|x) \\ \underline{y} & \text{otherwise} \end{cases} \end{aligned} \tag{19}$$

where $\underline{y}$ is the smallest possible value of y (possibly $-\infty$). Similarly, the upper bound is

$$Q_1(\tau|x) = \begin{cases} Q_Y\left(\frac{\tau}{P(z=1|x)}|x, z = 1\right) & \text{if } \tau \leq P(z = 1|x) \\ \overline{y} & \text{otherwise} \end{cases} \tag{20}$$