Chapter 1

Quantile Regression

Quantile regression is an important topic in theoretical econometrics. It is particularly useful if we are interested in the tail of the distribution, instead of the center. For example, in financial risk management, we are concerned about some rare events, rather than everyday routines. This line of research started from Roger Koenker.

1.1 Univariate quantile estimation

Given a sample $(y_1, y_2, ..., y_n)$, we are interested in estimating its τ-th quantile, where $\tau \in (0,1)$. To find the quantile from the sample, we can look for q such that

$$\frac{1}{n}\sum \mathbb{I}\{y_i \leq q\} \approx \tau. \tag{1.1}$$

If we ignore discreteness on the left-hand side, we solve the equation $\frac{1}{n}\sum \mathbb{I}\{y_i \leq q\} = \tau$. In this chapter, we always work with continuously distributed y. In the population model, q_τ^0 that solves $E[y \leq q] = \tau$ is the population parameter.

1.1.1 Asymptotic Result

(1.1) characterize the estimation by a method of moment. Now we cast the problem into an m-estimation. Introduce the check function

$$\rho(z) = z(\tau - \mathbb{I}(z \leq 0)).$$

Define

$$\psi(z) = \tau - \mathbb{I}(z \leq 0),$$

which be considered as a **subgradient** of $\rho(z)$. Notice $\rho(z)$ is continuous but $\psi(z)$ is discontinuous.

Let

$$S_n(q) = \frac{1}{n}\sum \rho(y_i - q) = \frac{1}{n}\sum (y_i - q)(\tau - \mathbb{I}\{y_i - q \leq 0\}).$$

The first-order condition is

$$\frac{\partial}{\partial q} S_n(q) = \frac{1}{n}\sum (\tau - \mathbb{I}\{y_i - q \leq 0\}) \times (-1) = \frac{1}{n}\sum \mathbb{I}\{y_i \leq q\} - \tau \xrightarrow{p} F(y \leq q) - \tau.$$

The "second derivative" is

$$\begin{aligned}\frac{\partial^2}{\partial q^2} S_n(q) &\xrightarrow{p} \frac{\partial}{\partial q}\{F(y \leq q) - \tau\} \\ &= \frac{\partial}{\partial q}\{F(y - q_\tau \leq q - q_\tau) - \tau\} \\ &= \frac{\partial}{\partial q}\{F(e \leq q - q_\tau) - \tau\} = f_e(q - q_\tau)\end{aligned}$$

where we define $e = y - q_\tau$, and the above heuristic calculation implicitly assumes the exchangeability between $\lim_{\delta \to 0}$ and $\xrightarrow{p}$. (See Chapter 21 of var der Vaart (1998) for a rigorous treatment.)

If true coefficient q_τ^0 is identified, then by ULLN we have

$$\hat{q} \xrightarrow{p} q_\tau^0.$$

Identification is equivalent to $f_y(q_\tau) = f_e(0) > 0$.

Evaluated at the true value $q = q_\tau$, the binary random variable $\mathbb{I}\{y_i \leq q_\tau\} - \tau$ has variance $\tau(1-\tau)$, and $f_e(q - q_\tau)|_{q=q_\tau}$. As a result,

$$\sqrt{n}\left(\hat{q} - q_\tau^0\right) \xrightarrow{d} N\left(0, \frac{\tau(1-\tau)}{f_e^2(0)}\right).$$

In the expression of the asymptotic variance, τ is known but the density $f_e^2(0)$ must be estimated based on observed "quantile residual" $\widehat{e}_i = y_i - \hat{q}$. The problem of density estimation is fundamentally a nonparametric estimation.

1.2 Quantile Regression

The above univariate quantile estimation is similar to a regression with intercept only. When other regressors X_i are present, we use $X_i'\beta$ to mimic θ in the quantile estimation:

$$S_n(\beta) = \frac{1}{n}\sum \rho_\tau\left(y_i - X_i'\beta\right)$$

The first order condition

$$\begin{aligned}\frac{\partial}{\partial \beta} S_n(\beta) &= -\frac{1}{n}\sum X_i \psi\left(y_i - X_i'\beta\right) \\ &= -\frac{1}{n}\sum X_i \psi\left(y_i - X_i'\beta_\tau + X_i'\beta_\tau - X_i'\beta\right) \\ &= -\frac{1}{n}\sum X_i \psi\left(e_i + X_i'\beta_\tau - X_i'\beta\right) \\ &\xrightarrow{p} -E\left[X\psi\left(e + X'\beta_\tau - X'\beta\right)\right] \\ &= -E\left[XE\left[\psi\left(e_i + X'\left(\beta_\tau - \beta\right)\right)|X\right]\right] \\ &= -E\left[XE\left[\tau - \mathbb{I}\left\{e \leq X'\left(\beta - \beta_\tau\right)\right\}|X\right]\right] \\ &= E\left[X\left(F_{e|X}\left(X'\left(\beta - \beta_\tau\right)\right) - \tau\right)\right]\end{aligned}$$

where the fourth equality follows by the law of iterated expectations.

SOC with respect to β in the population version is $E\left[XX'f_{e|X}\left(X'\left(\beta - \beta_\tau\right)\right)\right]$. Evaluate it at $\beta = \beta_\tau$, the Hessian is $E\left[XX'f_{e|X}(0)\right]$.

Similarly, by ULLN and ID we have consistency

$$\hat{\beta} \xrightarrow{p} \beta_\tau$$

The identification condition is that $Q_\tau = E\left[XX' f_{e|X}(0)\right]$ must positive definite.

Again evaluated at $\beta = \beta_\tau$, the variance of the score function is $\Omega_\tau = E\left[XX'\psi^2(e)\right]$. We have asymptotic normality

$$\sqrt{n}\left(\hat{\beta} - \beta_\tau\right) \xrightarrow{d} N\left(0, Q_\tau^{-1}\Omega_\tau Q_\tau^{-1}\right)$$

1.2.1 Linear Conditional Quantile

Let $Q_{y|X}(\tau)$ be the τ-th conditional quantile. If the linear function is correct specified for the τ-th conditional quantile, then

$$\tau = F_{y|X}\left(X'\beta_\tau\right) = E\left[\mathbb{I}\left\{y \le X'\beta_\tau\right\} \mid X\right] = E\left[\mathbb{I}\left\{e \le 0\right\} \mid X\right] = F_{e|X}(0).$$

This condition simplifies the expression of the variance of the score function as

$$\Omega_\tau = E\left[XX'E\left[\left(\mathbb{I}\left\{y \le X'\beta_\tau\right\} - \tau\right)^2 \mid X\right]\right] = \tau(1-\tau)E\left[XX'\right].$$

As a result, the asymptotic variance.

$$\sqrt{n}\left(\hat{\beta} - \beta_\tau\right) \xrightarrow{d} N\left(0, \tau(1-\tau)Q_\tau^{-1}E\left[XX'\right]Q_\tau^{-1}\right)$$

If we further assume e is statistically independent of X, then the Hessian is simplified as $Q_\tau = E\left[XX'\right] f_e(0)$, and we end up with

$$\sqrt{n}\left(\hat{\beta} - \beta_\tau\right) \xrightarrow{d} N\left(0, \frac{\tau(1-\tau)}{f_e^2(0)}\left(E\left[XX'\right]\right)^{-1}\right).$$

1.3 Summary

The derivations in this chapter are heuristic, but they deliver the essence.

It is helpful to compare quantile regression with our familiar linear regression. The univariate mean model is $y = \mu + \varepsilon$, where $E[y] = u$, or equivalently $E[\varepsilon] = 0$. The univariate quantile model is $y = q_\tau + e$, where $Q_y(\tau) = q_\tau$, or equivalently $Q_e(\tau) = 0$.

In regression model, the conditional mean $E[y|X]$ is in general a nonlinear function of X, and we approximate it by the linear function $X'\beta$. Identification is determined by the minimum eigenvalue of $E[XX']$. The conditional quantile $Q_{y|X}(\tau)$ is in general a nonlinear function of X too, while we approximate it with a linear function $X'\beta$ for simplicity. Identification is determined by the minimum eigenvalue of $E\left[XX' f_{e|X}(0)\right]$.

In regression models, correct specification $E[y|X] = X'\beta$ or equivalently $E[e|X] = 0$ gives unbiasedness to the OLS estimator, and homoskedasticity simplifies the variance. In quantile regression, correct specification $Q_{y|X}(\tau) = X'\beta$ provides an explicit form of the variance of the score function, and independence between e and X simplifies the sandwich-form variance into one piece.

Zhentao Shi. Feb 21, 2023. Transcribed by Shu Shen.