Chapter 7

Quantile Regression

> Here's a prayer for you. Got a pencil? . . . 'Protect me from knowing what I don't need to know. Protect me from even knowing that there are things to know that I don't know. Protect me from knowing that I decided not to know about the things I decided not to know about. Amen.' There's another prayer that goes with it. 'Lord, lord, lord. Protect me from the consequences of the above prayer.'
>
> Douglas Adams, *Mostly Harmless* (1995)

Rightly or wrongly, 95 percent of applied econometrics is concerned with averages. If, for example, a training program raises average earnings enough to offset the costs, we are happy. The focus on averages is partly because obtaining a good estimate of the average causal effect is hard enough. And if the dependent variable is a dummy for something like employment, the mean describes the entire distribution. But many variables, like earnings and test scores, have continuous distributions. These distributions can change in ways not revealed by an examination of averages, for example, they can spread out or become more compressed. Applied economists increasingly want to know what's happening to an entire distribution, to the relative winners and losers, as well as to averages.

Policy-makers and labor economists have been especially concerned with changes in the wage distribution. We know, for example, that flat average real wages are only a small part of what's been going on in the labor market for the past 25 years. Upper earnings quantiles have been increasing, while lower quantiles have been falling. In other words, the rich are getting richer and the poor are getting poorer. But that's not all - recently, inequality has grown asymmetrically; for example, among college graduates, it's mostly the rich getting richer, with wages at the lower decile unchanging. The complete story of the changing wage distribution is fairly complicated and would seem to be hard to summarize.

Quantile regression is a powerful tool that makes the task of modeling distributions easy, even when the underlying story is complex and multi-dimensional. We can use this tool to see whether participation in a training program or membership in a labor union affects earnings inequality as well as average earnings. We can also check for interactions, like whether and how the relation between schooling and inequality has been changing over time. Quantile regression works very much like conventional regression: confounding factors can be held fixed by including covariates; interaction terms work the same as with regular regression, too. And sometimes we can even use instrumental variables methods to estimate causal effects on quantiles when a selection-on-observables story seems implausible.

7.1 The Quantile Regression Model

The starting point for quantile regression is the conditional quantile function (CQF). Suppose we are interested in the distribution of a continuously-distributed random variable, Y_i, with a well-behaved density (no gaps or spikes). Then the CQF at quantile τ given a vector of regressors, x_i, can be defined as:

$$Q_\tau(\mathrm{Y}_i|\mathrm{X}_i) = F_Y^{-1}(\tau|\mathrm{X}_i)$$

where $F_Y(y|\mathrm{X}_i)$ is the distribution function for Y_i conditional on X_i. When $\tau = .10$, for example, $Q_\tau(\mathrm{Y}_i|\mathrm{X}_i)$ describes the lower decile of Y_i given X_i, while $\tau = .5$ gives us the conditional median.[1] By looking at

[1] More generally, we can define the CQF for discrete random variables and random variables with less-than-well-behaved densities as

$$Q_\tau(\mathrm{Y}_i|\mathrm{X}_i) = \inf\{y : F_Y(y|\mathrm{X}_i) \geq \tau\}.$$

changes in the CQF of earnings as a function of education, we can tell whether the dispersion in earnings goes up or down with schooling. By looking at changes in the CQF of earnings as a function of education and time, we can tell whether the relationship between schooling and inequality is changing over time.

The CQF is a quantile-analog for the CEF. Recall that the CEF can be derived as the solution to a mean-squared error prediction problem,

$$E\left[\mathrm{Y}_i|\mathrm{X}_i\right] = \underset{m(\mathrm{X}_i)}{\arg\min} E\left[\left(\mathrm{Y}_i - m\left(\mathrm{X}_i\right)\right)^2\right].$$

In the same spirit, the CQF solves the following minimization problem,

$$Q_\tau(\mathrm{Y}_i|\mathrm{X}_i) = \arg\min_{q(X)} \; E\left[\rho_\tau(\mathrm{Y}_i - q(\mathrm{X}_i))\right], \tag{7.1.1}$$

where $\rho_\tau(u) = (\tau - 1(u \leq 0))u$, called the "check function". If $\tau = .5$, this becomes least absolute deviations because $\rho_{.5}(u) = \frac{1}{2}(\operatorname{sign} u)u = \frac{1}{2}|u|$. In this case, $Q_\tau(\mathrm{Y}_i|\mathrm{X}_i)$ is the conditional median since the conditional median minimizes absolute deviations. Otherwise, the check function weights positive and negative terms asymmetrically:

$$\rho_\tau(u) = 1(u > 0) \cdot \tau u + 1(u \leq 0) \cdot (1 - \tau)u.$$

This asymmetric weighting generates a minimand that picks out conditional quantiles away from the median.

As a practical tool, the CQF shares the disadvantages of the CEF with continuous or high-dimensional X_i: it may be hard to estimate and summarize. We'd therefore like to boil this function down to a small set of numbers, one for each element of X_i. Quantile regression accomplishes this by substituting a linear model for $q(\mathrm{X}_i)$ in (7.1.1), producing

$$\beta_\tau \equiv \arg\min_{b \in \mathbb{R}^d} \; E\left[\rho_\tau(\mathrm{Y}_i - \mathrm{X}_i'b)\right]. \tag{7.1.2}$$

The quantile regression *estimator*, $\hat{\beta}_\tau$, is the sample analog of (7.1.2). It turns out this is a linear programming problem that is fairly easy (for computers) to solve.

Just as OLS fits a linear model to Y_i by minimizing expected squared error, quantile regression fits a linear model to Y_i using the asymmetric loss function, $\rho_\tau(\cdot)$. If $Q_\tau(\mathrm{Y}_i|\mathrm{X}_i)$ is in fact linear, the quantile regression minimand will find it (just as if the CEF is linear, OLS will find it). The original quantile regression model, introduced by Koenker and Bassett (1978), was motivated by the assumption that the CQF is linear. As it turns out, however, the assumption of a linear CQF is unnecessary - quantile regression is useful whether or not we believe this.

Before turning to a more general theoretical discussion of quantile regression, we illustrate the use of this tool to study the wage distribution**.** The motivation for the use of quantile regression to look at the wage distribution comes from labor economists' interest in the question of how inequality varies conditional on covariates like education and experience (see, e.g., Buchinsky, 1994). The overall gap in earnings by schooling group (e.g., the college/high-school differential) grew considerably in the 1980s and 1990s. Less clear, however, is how the wage distribution has been changing *within* education and experience groups. Many labor economists believe that increases in so-called "within-group inequality" provide especially strong evidence of fundamental changes in the labor market, not easily accounted for by changes in institutional features like the percent of workers who belong to labor unions.