An Introduction to Quantile Regression

Philipp Burckhardt

December 31, 2012

To date, the most comprehensive treatment of quantile regression is given by Koenker (2005). He starts his book with a quote from Galton, who was bewildered about the obsession many statisticians have with averages, noting that their "souls seem as dull to the charm of variety as that of the native of one of our flat English counties, whose retrospect of Switzerland was that, if its mountains could be thrown into its lakes, two nuisances would be got rid of at once" (Galton, 1889, p.62).

Yet, statisticians are not the only culprits here. Economists and political scientists are predominantly concerned with averages, too. However, Angrist and Pischke (2009) stress that an analysis of the whole distribution is often warranted. As an example, they point to the wage distribution, which in the last decades has seen a growing gap of the incomes between the rich and poor, a fact which cannot be brought to light by looking solely at averages.

Quantile regression provides a means to go beyond analysis of the conditional mean in regression. With ordinary least squares (OLS) as one of its main building blocks, it is still the most widely used approach in empirical social science research to model the conditional expectation as a linear function of different explanatory variables.

Consequently, we begin our discussion of quantile regression with a short reminder of how the linear models based on OLS had been motivated. We then point out the areas in which this method which so pervades econometrics is lacking and demonstrate to which extent quantile regression is superior.

Imagine being confronted with the task to predict the value of a variable y for a given entity (say the body weight of a person). It seems to be a reasonable approach to use the mean of all ys you have observed as a predictor. Carl Friedrich Gauss provided a rationale for this assertion in his influential astronomical treatise "Theoria motus corporum coelestium in sectionibus conicis solem ambientium" from 1809. In this book, he proves that if the errors, that is the deviations of the observed values from the true one, follow a normal distribution, the arithmetic mean will be the most probable value.

The proof makes use of the fact that the sample mean is the argument which solves

$$\min_{\hat{y}} \sum_{i=1}^{n} \rho\left(e_i\right) \quad with \quad \rho\left(e_i\right) = e_i^2 = (y_i - \hat{y}\ )^2. \tag{1}$$

So minimizing the sum of all squared residuals e of a sample of the size n, whereat residual is defined as the difference between the observed value y and the predictor $\hat{y}$, yields as a solution for $\hat{y}$ the sample mean.

Obviously, when predicting the value y for an entity, one can make use of other information one possesses about said entity. In our example, knowing whether the person is a man or a woman surely helps in guessing the weight. This is the main idea of every econometrician's favorite tool, regression analysis, in which the relationship between y and one or more independent variables xs (called regressors) is modeled.

So instead of using the sample mean as an estimate, we now derive the conditional mean function $\hat{y}(x) = E[y|x]$ which in the classical linear model is assumed to be linear in parameters and thus takes the following form:

$$E[y|x] = x^{\intercal}\beta.$$

In here, $x^{\intercal}$ is a $1 \times k$ vector containing the values of the respective attributes, and β denotes a $k \times 1$ vector of so called regression coefficients which measure the impact of an unit increase in the particular x on y. As in the case of the unconditional mean, differentiating

$$\sum_{i=1}^{n} \rho(e_i) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}(x_i, \beta))^2 = \sum_{i=1}^{n} (y_i - x_i^{\intercal}\beta)^2$$

with respect to β now yields the conditional mean as an optimal predictor. In this context, we implicitly face the standard assumption that the ys arise from the following data generating process:

$$y_i = x_i^{\intercal}\beta + \varepsilon_i \quad i = 1...n, \tag{2}$$

where the shape of the distribution of y does not change with varying xs except for a shift in the mean under the standard assumptions of the linear model. Of these, the most important ones are (in the notation of Hayashi (2000)):

1. strict exogeneity:
$$E[\varepsilon_i |x_1, \ldots, x_n] = 0 \quad (i = 1, 2, \ldots, n)$$

2. spherical error variance, i.e. homoscedasticity and no autocorrelation:
$$\begin{aligned} &E\left[\varepsilon_i^2 |x_1, \ldots, x_n\right] = \sigma^2 > 0 \quad (i = 1, 2, \ldots n) \\ &E[\varepsilon_i \varepsilon_j |x_1, \ldots, x_n] = 0 \quad (i, j = 1, 2, \ldots n; i \neq j). \end{aligned}$$

Empirical data often do not exhibit these characteristics, and the choice of the mean as a predictor rests on the equally dubious assumption that the error terms follow a normal distribution. For Koenker and Bassett (1978), the "Gaussian Law of Errors" is an "ex post rationalization for the use of the sample mean" which originates from a "kind of wishful thinking" (pp.34f.).

As a matter of fact, focusing exclusively on the mean of a given distribution often leads to false conclusions. Imagine Bill Gates entering a shabby bar in the middle of nowhere. The bartender, as a trained statistician, will notice immediately that the mean net worth of the guests has risen to unimagined heights after Bill Gates stepped across the doorsill. In such cases, other more robust location parameters like the median deliver a better representation of the distribution.

Recall that the median is the 50%-quantile. For any continuous random variable y, the qth quantile is defined as the value $Q_q(y)$ such that y is smaller or equal to $Q_q(y)$ with probability q. If $F(\cdot)$ denotes the cumulative distribution function (cdf) of y, then this is equivalent to specifying that

$$P(y \leqslant Q_q(y)) = F(Q_q(y)) = q.$$

Hence, the qth quantile can be written as

$$Q_q\left(y\right) = F^{-1}\left(q\right) = \inf\left\{y : F\left(y\right) \geqslant q\right\}.$$

Note that if one randomly draws a substantial amount of samples from the distribution, half of them will be smaller and the remaining half will be greater than the median. Therefore, to compute the sample median, one normally arranges all observations according to their size and selects the middle one in case of an odd number of observations. When the sample size is even, the median is calculated as the average of the two middle values.

But just like the mean, the median can be estimated via an optimization problem. To do this, one does not minimize the sum of squared residuals as in (1), but the sum of absolute residuals. So one uses $\rho(e) = |e|$ to weigh each residual. Actually, not only the median, but *any* quantile can be estimated via optimizing. This result, which is attributed to Fox and Rubin, is rather impressive as the task of finding the median and the other sample quantiles "might seem inherently tied to the notion of an ordering of the sample obervations" (Koenker, 2005, p.6).

For this purpose, instead of using $\rho(e) = e^2$ or $\rho(e) = |e|$ to weigh each residual in Formula (1), one chooses the following loss function:

$$\rho(e) = \begin{cases} q \cdot |e| & \text{for } e \geqslant 0 \\ (1-q) \cdot |e| & \text{for } e \leqslant 0 \end{cases} \qquad q \in [0,1]\,. \tag{3}$$

Although at first glance this weighting of the residuals might seem arbitrary, it can be shown that using this function in Equation (1) leads to an estimate of the qth quantile. For a mathematical proof, see Koenker (2005). Note that in the special case of $q = 0.5$, in which overprediction and underprediction are assessed equally, we are led back to estimation of the median. To estimate the other quantiles, one puts a different price on over- and underprediction. To gain some intuition, Koenker (2005) points to the fact that for high q it is much more costly if y is greater than the predictor $\hat{y}$. Hence $\hat{y}$ will take a greater value in order to compensate. Specifically, for $q = 0.75$, it is marginally three times more costly if $y > \hat{y}$ compared to the case when $y < \hat{y}$, and hence $\hat{y}$ will be chosen such that $P(y > \hat{y})$ is three times less likely than $P(y < \hat{y})$. This results in $P(y < \hat{y}) = 0.75$. So $\hat{y}$ is chosen to be the 75% sample quantile, just as predicted.

We have noted that in the context of ordinary least squares regression, one utilizes information about explanatory variables and estimates the conditional mean instead of using the unconditional mean as a predictor. The same approach can be applied to the quantiles. This is the main idea elaborated on in the seminal paper by Koenker and Bassett (1978).

Adapting the notation of Koenker (2005), we define the conditional quantile function as follows:

$$Q_q\left(y\,|x\right) = x^{\intercal}\beta\left(q\right).$$

For this equation to hold, the error terms of the underlying dgp (2) must meet the condition that $Q_q\left(\varepsilon\,|x\right) = 0$. Besides, no distributional assumptions have to be placed on the error terms.

Minimizing Formula (4) then yields the q-th conditional quantile as the optimal linear predictor:

$$\min_{\beta} \sum_{i=1}^{n} \rho\left(e_i\right) = \min_{\beta} \sum_{i=1}^{n} \rho\left(y_i - x_i^{\intercal}\beta\right) = \min_{\beta} \sum_{y_i \geqslant x_i^{\intercal}\beta} q\cdot\left(y_i - x_i^{\intercal}\beta\right) + \sum_{y_i \leqslant x_i^{\intercal}\beta} (q-1)\cdot\left(y_i - x_i^{\intercal}\beta\right). \tag{4}$$

Equation (4) can be reformulated as a linear program and be solved with the simplex algorithm. This leads to an estimator for $\beta\left(q\right)$ in the respective conditional quantile function. The bracketed q indicates that the values of β may vary depending on which quantile function is estimated. Consequently, quantile regression can be used as a method to test for heteroskedasticity. Only if all slope coefficients take roughly the same value for all quantiles $q \in [0, 1]$, the for OLS necessary assumption of homoskedasticity will be satisfied.

If such heteroskedasticity was present, the use of quantile regression would give rise to valuable insights and provide a more complete picture of the distribution of the data. Both prediction of y based on the values of the regressors and "causal" inference of the regression coefficients can be done. In this context, the interpretation of the regression coefficients is the same as in OLS regression, namely that a marginal increase of regressor x_i leads to a $\beta_{i,q}$ increase of the response variable y, as can be seen by taking the partial derivative of the estimated conditional quantile function with respect to x_i:

$$\frac{\partial Q_q\left(y\left|x\right.\right)}{\partial x_i} = \hat{\beta}_{i,q}.$$

The only, but major difference is that one is now confronted with different $\hat{\beta}_{i,q}$ depending on the particular q. This enables the researcher to detect how the magnitude of the relationship between the response and the respective regressor varies across the different quantiles. Returning to the example of body weight, one would assume that for the most corpulent persons the influence of the gender is rather significant, whereas for the lower quantiles the difference is smaller. Unlike in the OLS setting, the $\beta_{i,q}$ are not assumed to be constant over the quantiles. Because of that, it should be emphasized that they represent a *marginal* effect whose value changes when the object of interest advances towards a higher quantile.

References

ANGRIST, J. AND J. PISCHKE (2009): *Mostly harmless econometrics*, Princeton, NJ: Princeton Univ. Press.

GALTON, F. (1889): *Natural inheritance*, London, Macmillan and co.

HAYASHI, F. (2000): *Econometrics*, Princeton, New Jersey: Princeton University Press, 1 ed.

KOENKER, R. (2005): *Quantile Regression*, Cambridge: Cambridge University Press.

KOENKER, R. AND G. BASSETT (1978): “Regression Quantiles,” *Econometrica*, 46, 33–50.