Quantile Regression

Nikolaj Nasenko

2021/09/10

1. Introduction to quantile regression

The issue of income inequality, among many others, has become a major concern in 21st century politics. For policy makers, a crucial question is how certain measures, such as greater investment in a country's educational system, can improve the financial situation of its citizens.

Suppose now that we want to describe this relationship with a simple linear regression model.

$$y = X\beta + \epsilon$$

Let y be an $n \times 1$ vector of income. Let X be an $n \times 2$ matrix including constant term and years of schooling and let ϵ be an $n \times 1$ vector of errors for the population of n people. To estimate 2×1 coefficient vector β, from which we get the effect of education on income, we would usually choose **least square estimator** (**LSE**) to do the job. The LSE minimizes least square loss function (L2 loss).

$$\text{given that} \quad \hat{\epsilon}_i = y_i - X_i\hat{\beta} \qquad \arg\min_{\hat{\beta}} \sum_{i=1}^{n} (y_i - X_i\hat{\beta})^2$$

From that we end up with a **conditional mean**.

$$E[y_i|X_i] = X_i\hat{\beta} \quad \text{or} \quad \mu + \hat{\beta}_1 x_i \quad \text{(scalar form)}$$

Thus, the LSE would only provide us with a description of the effect of education on a person with an **average income**. But what about the effect on people in the lower decile of the income distribution? Will the poor without adequate education become less poor with education? What about the upper class? What would be the implications for them? In this case, from a statistical point of view, the interest of the policy makers lies in describing changes in the distribution of a nation's income as a function of education. This is where quantile regression comes in handy.

The (linear) **quantile regression** model for the continuously-distributed random variable y can be defined as

$$Q_q[y_i|X_i] = X_i\hat{\beta}_q \quad \text{such that} \quad Prob[y_i \leq X_i\hat{\beta}_q|X_i] = q, \quad 0 < q < 1$$

The quantiles of the $y|x$ distribution would be defined by **variation of the constant term**, e.g.

$$\text{for } q = 0.9\text{:} \quad Q_{0.9}[y_i|X_i] = \mu_{0.9} + \hat{\beta}_1 x_i \quad \text{with} \quad Prob[y \leq \mu_{0.9} + \hat{\beta}_1 x_i] = 0.9$$
$$\text{for } q = 0.5\text{:} \quad Q_{0.5}[y_i|X_i] = \mu_{0.5} + \hat{\beta}_1 x_i \quad \text{with} \quad Prob[y \leq \mu_{0.5} + \hat{\beta}_1 x_i] = 0.5$$

The quantile regression is achieved by minimizing least absolute deviations function (L1 loss), so called **LAD estimator**.

$$\text{given that} \quad \hat{\epsilon}_i = y_i - X_i\hat{\beta} \qquad \arg\min_{\hat{\beta}} \sum_{i=1}^{n} |y_i - X_i\hat{\beta}|$$

In this form, the LAD estimator gives us the **median regression**. In order to derive additional conditional quantiles, we adjust the LAD function by assigning asymmetrical weights to positive and negative losses based on the desired quantile q.

$$\arg\min_{\hat{\beta}_q}\Big(\sum_{i:y_i\geq X_i\hat{\beta}_q}^{n} q|y_i - X_i\hat{\beta}_q| \quad + \sum_{i:y_i<X_i\hat{\beta}_q}^{n} (1-q)|y_i - X_i\hat{\beta}_q|\Big)$$

Now that we know how to estimate our quantile regression coefficients, the interpretation is a little bit tricky. Quantile coefficients tell us about effects on distributions and not on individuals. If we discover that education raises the lower decile of the income distribution, this does not necessarily mean that someone who would have been poor (i.e. at the lower decile without education) is now less poor. It only means that those who are poor with education are less poor than the poor would be without education. Thus, we call quantile regression to be so called **rank-preserving**.

2. Some properties of the quantile regression

We already know form the introduction that quantile regression allows us to understand the relationship of covariates on independent variable across the entire conditional distribution, not just on conditional mean. In addition, there are several other advantages worth mentioning.

(1) Due to its weighting scheme, the LAD estimator is **robust to outliers**, unlike the LSE, which puts larger weights to larger losses when the sample is small, and is thus more sensitive. In that case, the median regression offers a better choice.

$$\text{L2 loss: } \sum_{i=1}^{n} \hat{\epsilon_i}^2 = \sum_{i=1}^{n} \hat{\epsilon_i} w_i \quad \text{where weight } w_i = \hat{\epsilon_i}$$

$$\text{L1 loss: } \sum_{i=1}^{n} |\hat{\epsilon_i}| = \sum_{i=1}^{n} |\hat{\epsilon_i}| w_i \quad \text{where weight } w_i = 1$$

(2) There is no need to make any assumptions regarding conditional $y|x$ distribution or conditional variance.

(3) Quantile regression can be used when part of the conditional distribution of the dependent variable is hidden. For example for confidentiality reasons. (censored quantile regression).

3. Empirical analysis: QR application on wage distribution

We started with an example of politicians asking whether it is possible to reduce income inequality by supporting a country's educational system. Now let us empirically investigate whether we can find a relationship between years of education and a person's income (taking wage as equivalent for income). For simplicity, we assume a naive model that defines wages solely based on education. We take our data from the wooldridge package of R and apply quantile regression to it. The results are shown in Figure 1 and Figure 2.

From Figure 1, we see that the slope of the median, upper and lower-decile regression is positive and not much different.[1] This tells us that people with education are better off compared to people without education regardless of whether you are in the 10% of the population with the highest income or the 10% with the lowest income.

Figure 2 shows the estimated regression coefficients (y-axis) for multiple quantiles (x-axis). The grey area represents the confidence interval calculated by bootstrapping and the red lines represent the results for the

[1] $\hat{\beta}_{0.1} = 0.044$, $\hat{\beta}_{0.5} = 0.06$, $\hat{\beta}_{0.9} = 0.071$

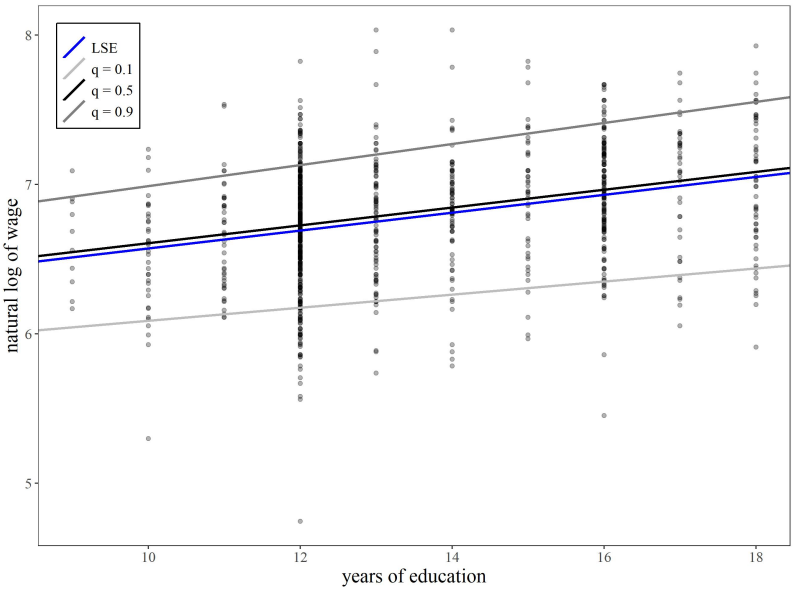

Figure 1: Quantile regression: Wage vs Education

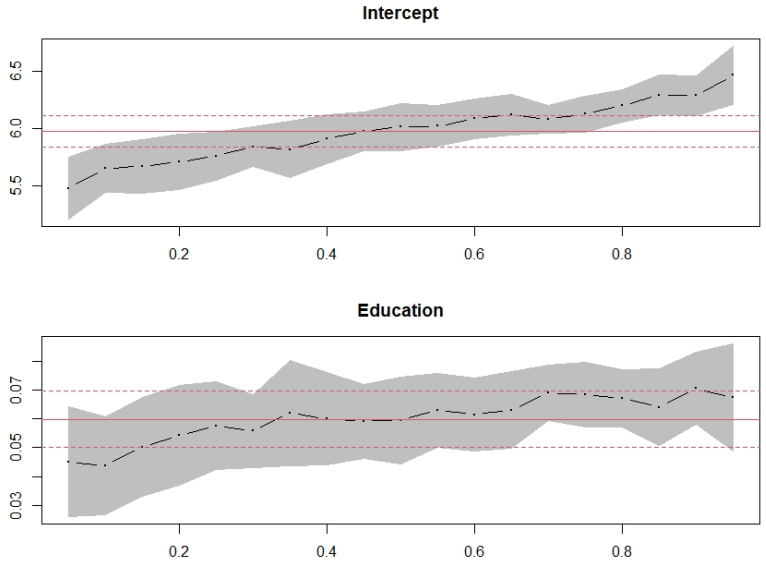

Figure 2: Estimated coefficients for quantile q

LSE. The findings seems counter-intuitive, as an additional year of education has a larger impact as the quantile increases, rather than the other way around, as would be probably more expected. The results suggest that the rich with education are paid more than the rich without education, and the same is true for the poor, but the effect on the rich is slightly larger than on the poor.

Finally, note that the LSE coefficient is almost similar to the median regression coefficient. This indicates the case where the distribution of $y|x$ is normal.

4. Summary

This paper briefly introduces the theoretical background of quantile regression and provides its empirical application to the relationship between wages and education. Quantile regression is a useful tool for gaining insights into the effect of independent variables on different segments of the dependent-variable distribution, not just the effect on its conditional mean. Apart from that, quantile regression is a robust alternative to traditional linear regression as it is not as affected by extreme values as traditional regression estimated using LSE.

5. References

GREEN, W. H. (2018), *Econometric Analysis*, 8th edition

ANGRIST, J. D. and PISCHKE, J. S. (2008), *Mostly Harmless Econometrics: An Empiricist's Companion*

KOENKER, R. (2021), *quantreg: Quantile Regression*, R package version 5.86, https://CRAN.R-project.org/package=quantreg

SHEA, J. M. (2021), *wooldridge: 115 Data Sets from "Introductory Econometrics: A Modern Approach, 7e" by Jeffrey M. Wooldridge*, R package version 1.4-1, https://CRAN.R-project.org/package=wooldridge