(The challenge question)

A comparison of quantile regression and OLS regression in empirical study

–An example on the relationship between Chinese Female Education Level and Family Income

Jiawen Ke 15220182202469

04/12/2021

[Abstract]This paper focuses on the difference of quantile regression and OLS regression on empirical study. The relationship of female educational level and family's income is taken as an example. The results show that the families with better-educated female is going to have higher income and the quantile regression performs better than the OLS model in estimation.

[Keyword] Female, family income, quantile regression, OLS regression

Contents

1 The overview of quantile regression

1.1 The Introduction of Quantile Regression

The idea of quantile regression generated by a esuit Catholic priest from Dubrovnik[1]. He was interested in the ellipticity of the earth, which was built on Isaac Newton's suggestion that its rotation could cause it to bulge at the equator with a corresponding flattening at the poles.He finally produced the first geometric procedure for determining the equator of a rotating planet from three observations of a surface feature.

Therefore, The idea of estimating a median regression slope, a major theorem about minimizing sum of the absolute deviances and a geometrical algorithm for constructing median regression was proposed. More importantly for quantile regression, he was able to develop the first evidence of the least absolute criterion and preceded the least squares introduced by Legendre in 1805 by fifty years.

Consider about a class of 50 children finished their final exam. Suppose Tom, one of the students in the class, got 90 points in the exam which was the 10th highest score among the 50 students[2]. Hence, we say Tom performs better than 80% students and worse than 20% students in the class. And Tom scored at then 20th quantile of the exam in the class.

1.2 The Comparison of quantile regression and OLS regression

In empirical study, the least square method (OLS) is the most commonly used regression method. The standard least-squares linear regression focuses on the effect of the explanatory variable x on the *conditional mean* $E(Y|X)$ of the explained variable y. However, empirically, researchers may be interested in other important quantiles of the $y|x$ distribution. When labor economists study gender pay inequality, for example, they need to look more closely at the 90 per cent of women's income if they want to explore whether high-earning women are less likely to get promoted.

What the linear regression model obtains is a conditional mean, which does not take into account the overall distribution characteristics of the dependent variable. When the information of the position (quantile) of the dependent variable is needed, the linear regression shows its insufficiency.

One of the most basic assumptions of linear (mean) regression model(OLS) is that under the conditions of normal distribution, random error and independence, the parameter estimated by the least square method is the least variance unbiased estimate. However, most of the data in real life do not meet the normal distribution. At this time, if the linear regression model is still used for analysis, since the calculation of the value in the hypothesis test depends on the hypothesis of normality, it may cause the bias of the value, which leads to the invalid hypothesis test. If there is heteroscedasticity in the sample data or the distribution of the data is sharp peak and thick tail, the least square estimator does not have the good properties with the least variance and unbias.

For example, when there are outliers in the sample data, using the linear regression model to calculate the parameters of the estimates may have larger deviation. As a result, the regression fitting is usually right after remove outliers establish linear regression model, but that would make the outliers lose their value to social scientific research.

Unlike OLS, quantile regression estimates a linear relationship between the quantile of the explaining variable x and the quantile of the explained variable y. In terms of the loss functions, OLS regression aims at minimizing the square residual $\sum_i e^2$, and median regression aims at minimizing the absolute deviation $\sum_i |e|$. In the case of quantile regression, the goal is to minimize the asymmetric absolute residual value.

The quantile level is denoted by q, and the quantile regression estimator β_q minimizes the loss function:

$$Q(\beta_q) = \sum_{i:y_i > {x_i}^{'}\beta_q}^{N} q|y_i - {x_i}^{'}\beta_q| + \sum_{i:y_i < {x_i}^{'}\beta_q}^{N} q|y_i - {x_i}^{'}\beta_q| \tag{1}$$

Compared with OLS, quantile regression has some advantages. The scientists can have a more comprehensive description of the relationship between the explaining variable x and the explained variable y . In empirical study, since the explaining variable x may have different effects on the explained variable y at different quantiles, the quantile regression coefficient may be different from the OLS regression coefficient. Hence, the quantile regression model has broader conditions and can describe the global characteristics of the dependent variable, rather than just the mean value. On the other hand, the quantile regression model is robust and the estimated value of the model is usually not affected by outliers. From this point of view, the quantile regression has strong robustness.

2 Introduction in Female Educational Level and Family Income in China

This article we will use the quantile regression to compare with the OLS model. We are going to estimate the relation between the educational level of the female and the income of their family. The data used here is from the CFPS(China Family Panel Studies) in 2018[3]. The information of male, unmarried female and some defualt data in the dataset were eliminated and finally a 50,002 sample size is available.

To make the data easier to be used, the logarithm of income is taken to the regression. The logarithm of family income is the explained variable. And the education level[1] of female in each family is the main explaining variables. The age, marriage, family size and the type of hukou. The following table are some details.

Type of variable	Variable	Name of variable	Value
Explained variable	ln(income)	lnincome	continuous
Main explaining variable	educational level	edu	discrete (see $footnote_1$)
Control variable	Number of child	num_child	continuous
	Age	age	continuous
	Marriage	bi_marriage	0= divorced or widowed,1= married
	Family size	familysize	continuous
	Type of Hukou	bi_hukou	0= rural, 1= town

Table 1: Data definition

In this paper, the empirical strategy used is as below:

$$lnincome = \alpha_0 + \alpha_1 \cdot edu + \alpha_2 \cdot age + \alpha_3 \cdot marriage + \alpha_4 \cdot family + \alpha_5 \cdot hukou + \alpha_6 \cdot num\_child + \epsilon_1 \tag{2}$$

Where the ϵ_1 is the error term, and the α_0 is the intercept. The quantiles at 0.25, 0.5,0.75 are taken.

[1] 0= illiteracy/semi-illiteracy, 1= primary school, 2= junior high school, 3= high school, 4= junior college,5= university undergraduate course, 6= master and above education

3 Empirical results

3.1 The description of data

The standard deviation of the income of Chinese female in 2018 is high, which means that the discrepancy of Chinese income is quite large. For the control variables, most of the subject female have 2.4 childen. Among the subjects, most of them are married since the mean of the variable *bi_marriage* is almost 1. The average age of them is 49, and the mean familysize is 4.5. Also, a higher proportion of the subject female come from rural area.

Variable	Obs	Mean	Std. Dev.	Min	Max
income	50002	96619.196	196697.47	15	9158800
lnincome	50002	10.978	1.02	2.708	16.03
edu	50002	2.378	1.683	-9	7
num_child	50002	1.384	1.231	0	9
age	50002	48.54	16.45	18	106
income_p	50002	24254.427	56615.207	5	4100000
lnincome_p	50002	9.57	1.006	1.609	15.226
bi_marriag	50002	0.975	0.155	0	1
familysize	50002	4.562	2.138	1	21
bi_hukou	50002	0.249	0.432	0	1

Table 2: Data description

To see the availability of the quantile regression here, first we plot the box graph of lnincome in figure 1. We can see that the medium of lnincome is high with either large and small outliers.

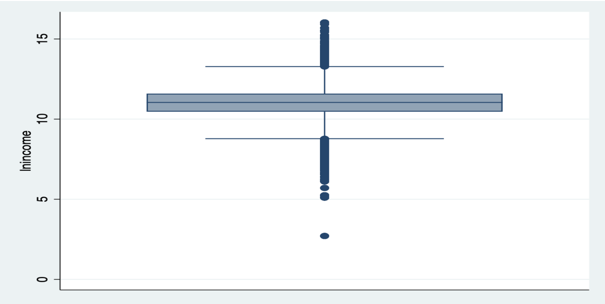

Figure 1: Box graph of lnincome

To see whether the distribution of family income is normal, figure 2 plot the Q-Q graph of lnincome.

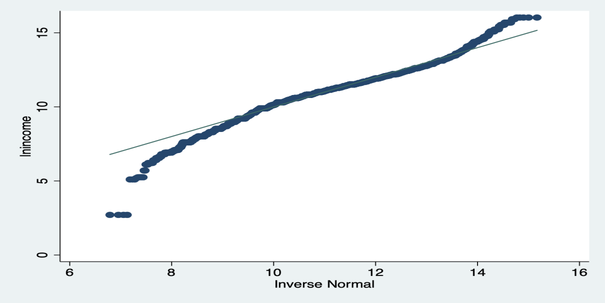

Figure 2: Q-Q graph of lnincome

From the Q-Q graph, the distribution of lnincome is not exactly the same as normal distribution which is an asymmetric line, hence there is the possibility to use the quantile regression to optimize the finding.

3.2 The regression results

Table3 reports our benchmark results from specification. The results illustrated in the three models are at the quantiles of 0.25, 0.5 and 0.75, respectively.

	(1)	(2)	(3)
	lnincome	lnincome	lnincome
edu	0.172***	0.143***	0.0997***
	(46.63)	(48.94)	(31.57)
num_child	-0.0513***	-0.0447***	-0.0445***
	(-11.43)	(-12.62)	(-11.61)
age	0.000511	0.000744*	-0.000224
	(1.38)	(2.55)	(-0.71)
familysize	0.144***	0.119***	0.100***
	(56.22)	(58.77)	(45.76)
bi_marriage	-0.0708*	-0.0697*	-0.0554
	(-2.01)	(-2.50)	(-1.84)
bi_hukou	0.606***	0.539***	0.555***
	(45.58)	(51.38)	(48.90)
_cons	9.385***	10.11***	10.83***
	(214.21)	(292.64)	(289.57)
N	50002	50002	50002
t statistics in parentheses			
* p<0.05, ** p<0.01, *** p<0.001			

Table 3: The quantile regression results

Table 3 shows that the education level of female in Chinese families is statistically significant, playing an important role. A unit increase of female education level, the logarithm of the income of a family will rise 0.172 unit in the families with income at 25% quantile. And the number of children does affect the income of a family. Families with less children seem to have higher income. That is reasonable and consistent with intuition that the resources in a family with only one kid can be more concentrated than others. The regression on control variables shows that with the increase with the age and familysize, a successful marriage and a town living, the family income does increase too.

The regression results imply that with the tendency that women have higher and higher educational level, less of them are willing to have as much children as the past time because of the high opportuniry cost. Instead, they will work much harder to increase the family income as men always do.

3.3 The comparison between quantile regression and OLS regression

Based on what have discussed in the parts before, it is highly possible that the quantile regression tend to perform better then OLS regression in this empirical example. Figure 3 shows the advantages of quantile regression in this research.

In order to detect the different estimations made by OLS regression and quantile regression, here we illustrated some figures to make it more clear. The following figure depicts the confidence intervals of the regression model coefficients at different quantile levels, with the quantiles varying isometric from 0.01 to 0.99. The dark curve in the figure represents the estimated values of the coefficients corresponding to each variable at different quantile levels; the gray area represents the 95% confidence intervals of the coefficients; the dark dotted line represents the estimated values of the coefficients in the OLS regression model; and the light dotted line on both sides represents the 95% confidence intervals of the coefficients in the OLS regression model.

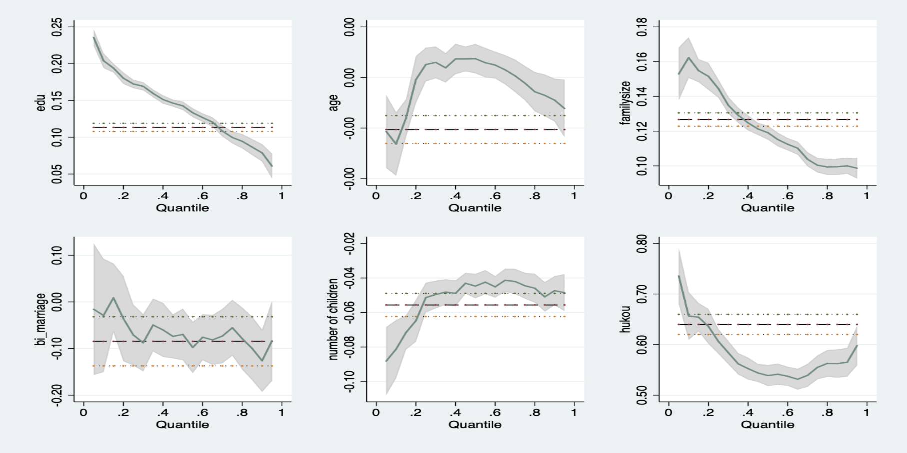

Figure 3: Confidence intervals at different quantile levels

In figure 3, except that the coefficient estimates corresponding to marriage concentration are basically within the coefficient confidence interval of the OLS regression model, the other coefficient estimates are basically not within the coefficient confidence interval of the OLS regression model, especially the large gap in low and high quantiles. That is to say, the quantile regression is more reasonable for this dataset and it can explain the relationship between variables better.

3.4 Robustness checks

Table 4 shows robustness to different explained variables. In the robustness check, the explained variable is changed into the logarithm of family income per person, which is denoted as *lnincome_p*. The outcome is consistent with the former regression. Hence, the main regression results are basically robust.

	(1)	(2)	(3)
	lnincome_p	lnincome_p	lnincome_p
edu	0.177	0.141	0.108
	(0.000)***	(0.000)***	(0.000)***
num_child	-0.0460	-0.0432	-0.0479
	(0.000)***	(0.000)***	(0.000)***
age	0.00153	0.00168	0.000530
	(0.000)***	(0.000)***	(0.090)
familysize	-0.0614	-0.0840	-0.0970
	(0.000)***	(0.000)***	(0.000)***
bi_marriage	-0.133	-0.101	-0.133
	(0.000)***	(0.000)***	(0.000)***
bi_hukou	0.629	0.584	0.590
	(0.000)***	(0.000)***	(0.000)***
_cons	8.896	9.586	10.32
	(0.000)***	(0.000)***	(0.000)***
N	50002	50002	50002
t statistics in parentheses			
* p<0.05, ** p<0.01, *** p<0.001			

Table 4: The robustness check

4 Summary

In empirical study, the OLS regression may not be the best solution to some estimation problems. While the quantile regression can fit more different distributions of data and will be a better way of more accurate estimation instead of the OLS.

As for the regression results, it is possible that the cancellation of the basic state policy on family planning in China can not change the fertility status quo in the short term without any encouragement or policy incentives.

References

[1] Kevin F. Hallock Roger Koenker. Quantile regression. *JOURNAL OF ECONOMIC PERSPECTIVES*, 15(4):143–156, December 2001.

[2] Roger Koenker. Quantile regression: 40 years on. *Annual Review of Economics*, 9(4):155–176, August 2017.

[3] Huang Rongjie Xia Guoxiang. Research on the relationship between female education level, number of children and family income. *Journal of Hainan Normal University (Social Science Edition)M*, 34(02)(201):104–111, 05 2020.