Quantile Regression and Empirical Analysis

–An Example on Impact of Internet Use on Income

Zhuoma Zhang, Lina Xu

May 8, 2022

Contents

1 Introduction to quantile regression

1.1 A brief introduction to quantile regression

The previous regression models actually studied conditional expectations of explained variables. But sometimes people also care about the relationship between the explanatory variables and the quantiles of the distribution of explained variables, for example, the people in lower quantile of wage would sacrifice more when increasing the minimum wage level compared with people in higher quantile. So as an extension of OLS regression, Koenker and Bassett firstly proposed the quantile regression in 1978. The quantile regression is a modeling method to estimate the linear relationship between a set of regression variables X and the quantile of the explained variable Y.

Compared with OLS regression, quantile regression can more comprehensively describe the **full picture** of the conditional distribution of the explained variable. It could analyze how the explained variable affects any quantiles of the explained variable, rather than just analyzing the conditional expectation value. Because the estimators of regression coefficients under different quantiles are often different, the quantile regression model could explain the different influence on explanatory variables at different levels.

OLS regression estimators are calculated based on minimizing the squared residuals. And quantile regression estimators are also calculated based on minimization of absolute residuals in an asymmetric form. Compared with the least square method, the estimation method of median regression is more **robust for outliers**.

1.2 The theory explanation of quantile regression model

Before introducing quantile regression, let's talk about how we figure out the OLS regression. Basically, we assume a function, and then let the function fit the training data y as much as possible, and determine the unknown parameters of the function. Fitting training data as closely as possible is generally done by minimizing MSE:

$$\min \frac{1}{n} * \sum_{i=1}^{n} (y_i - \hat{y_i})^2 = E(y - \hat{y_i})^2 \tag{1}$$

$\hat{y}$ combines a set of unknown parameters$\hat{\beta}$ which we try to predict. And we can also

minimize the absolute value of y and the predicted y, which is called **L1 loss function**:

$$\min \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{2}$$

So as the same logic, we could define the loss function of quantile regression. Since the symmetry of absolute values yields medians, so that minimizing the sum of asymmetrically weighted absolute residuals by simply giving different weights to positive and negative residuals will yield quantiles. Solving

$$\min_{\beta} \sum \rho_\tau(y_i - \xi(x_i, \beta)) \tag{3}$$

Where the function $\rho_\tau(\cdot)$ is denoted as the Figure 1 below.

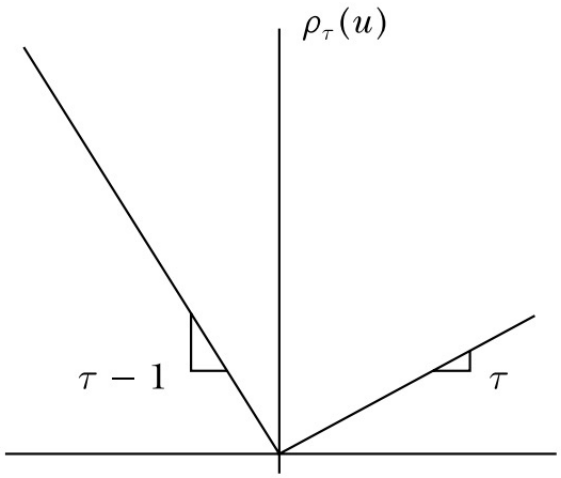

Figure 1: Quantile regression $\rho function$

And this graph could also write in formula as $\rho_\tau(u) = u(\tau - I(u \leq 0))$, where u denotes the residual.

Essentially, this is a weighted least square method that gives different weights to different values of y. For example, now we have a data set that includes 10 integers ranging from 1 to 10, and we want to find the 0.8 quantile. Let's assume the 0.8 quantile is q, then all numbers greater than q are assigned a weight of 0.8, and those less than q are assigned a weight of 0.2. We minimize the function above to find the quantile, and then verify that q equals 8, which is also the predicted value of y.

2 Empirical implementation

2.1 Introduction to research question

The impact of Internet use income distribution is uncertain. On one hand, it may reduce the cost of trade in poor and remote areas by providing information and communication media, which can drive the economic development in these areas and then increase people's income here. On the other hand, the Internet can be seen as a technology, which is a complement of skilled workers but a substitution for unskilled workers. So the Internet use may widen the income inequality between skilled workers and unskilled workers.

In other words, we do know that the Internet use has a positive effect on people's income in a general way, however, the more important question is, for people from different income groups, do they obtain the same benefits from Internet use? Or does Internet use benefit the rich more or the poor more? To answer this question, our old friend, simple OLS regression can not help, because OLS only estimates the average treatment effect on all population. We need quantile regression to estimate the different effects on different income groups.

2.2 Data definition and model specification

In this paper, we use CFPS (China Family Panel Study) data in survey year 2010, 2014, 2016, 2018 and 2020. We don't use data in 2012 because there is no survey question about Internet use. We append data in these 5 years together to get a panel data. The main variables used in this paper are shown in the following Table 1.

Type of variable	Variable	Name of variable	Value
Explained variable	ln(income)	log_income	continuous
Explaining variable	Treat	treat	=1 if in treated group, =0 o.w.
	post	Post	=1 if after access to the Internet, =0 o.w.
	treat*post	treat_Post	=1 if treat*Post=1, =0 o.w.
Control variable	Male	Male	=1 if male, =0 if female
	Age	age	continuous
	Education year	EducYear	continuous
	Year of working expirience	WorkExperience	continuous
	Urban	Urban	=1 if living in urban areas, =0 o.w.

Table 1: Data definition

This paper basically uses DID method to estimate the effect of Internet use on residents' income. The treated group (treat = 1) includes people who have used Internet in any of the survey years. We assume once a person has access to Internet, he/she can use Internet after that year as he/she likes, and hence the variable Post = 1 after that year. In other words, this person will always has access to Internet once he/she uses it.

In this paper, the model specification is as below:

$$ln(income_{ict}) = \beta_0 + \beta_1 * treat_i * post_i + X_{it} + \eta_c + \eta_t + \epsilon_{it} \tag{4}$$

Where β_1 is the coefficient of interest and $treat_i * post_i$ is the interaction term of $treat_i$ and $Post_i$. X_{it} includes all the observed characteristics of individuals, which are the control variables shown in above table. η_c is the city fixed effect, which is of province-level. η_t refers to the time fixed effect, which is of year level. Moreover, β_0 is the intercept and ϵ_{it} is the error term. In this paper, we take quantile 0.1, 0.5 and 0.9.

2.3 Data description and empirical results

In this paper, we only use samples with age in the interval [16, 65], and people out of this range are seen as out of the labor force. The following Table 2 is descriptive statistics of our main variables.

VARIABLES	obs	mean	s.d.	min	max
age	142,849	41.16	13.84	16	65
log_income	58,642	9.624	1.431	0	16.15
Male	142,849	0.497	0.500	0	1
EducYear	133,353	8.301	4.629	0	24
WorkExperience	31,199	9.211	12.73	0	60
Urban	142,849	0.464	0.499	0	1
treat	142,849	0.630	0.483	0	1
Post	142,849	0.419	0.493	0	1
treat_Post	142,849	0.419	0.493	0	1

Table 2: Discriptive statistics

The following Table 3 shows the results of quantile regression, with quantile 0.1 in

the column (1), 0.5 in the coliumn (2) and 0.9 in the column (3).

Variable: ln(income)	(1) q10	(2) q50	(3) q90
treat_Post	0.373***	0.276***	0.258***
	(0.0436)	(0.0213)	(0.0246)
Male	0.708***	0.478***	0.442***
	(0.0362)	(0.0115)	(0.0161)
age	0.175***	0.100***	0.0778***
	(0.00516)	(0.00486)	(0.00408)
age2	-0.00207***	-0.00122***	-0.000920***
	(6.14e-05)	(6.47e-05)	(5.04e-05)
EducYear	0.0761***	0.0619***	0.0516***
	(0.00573)	(0.00216)	(0.00202)
WorkExperience	-0.0230***	-0.0109***	-0.00838***
	(0.00202)	(0.000851)	(0.00114)
Urban	0.458***	0.239***	0.152***
	(0.0400)	(0.0144)	(0.0157)
Province FE	YES	YES	YES
Year FE	YES	YES	YES
Constant	3.922***	6.804***	8.370***
	(0.268)	(0.147)	(0.135)
Observations	23,450	23,450	23,450
Standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1			

Table 3: Quantile regression results

The result in column (1) can be intepreted in this way: the use of Internet can increase a person's income by about 37.3%, holding other factors constant, for people with income at 10% quantile. The interpretations for other two columns are similar. From the regression table we can also see that from column (1) to (3), the coefficient of $treat * post$ decreases, which means that the effect of Internet use on income is larger for people with relatively lower income. It turns out that, the popularity of Internet use can help reduce income inequality. Moreover this effect is positive and significant at 1% level at all quantiles.

2.4 Parallel trend test

The basic strategy of identification is DID (difference in difference), so we need to do parallel trend test to make sure the validity of our model. The condition must be held that in the absence of the policy shocks the treatment group and the control group follow the same trend. In this paper, the parallel trend test are comparing the mean of the dependent variable of treatment group and control group over time, which is shown in the following Figure 2.

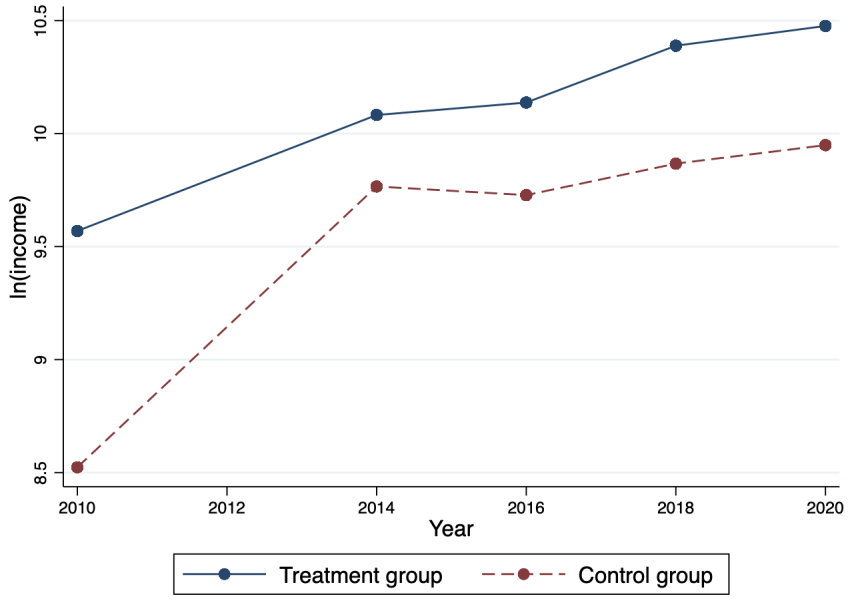

Figure 2: Parallel trend test

This two curves in the graph directly reflect the consistent trend between the treatment group and the control group before the use of Internet, indicating that the two groups are comparable before impact, so they pass the parallel trend test.

2.5 Robustness checks

For robustness check, we replace the dependent variable by log of monthly income, change the quantiles to 0.25, 0.5 and 0.75 and add the square of education year as another regressor. The results are as Table 4.

As shown in this table, the effect of Internet use on monthly income is also positive and significant and decreases with income, which is consistent with the findings in

Variable: ln(monthly income)	(1) q25	(2) q50	(3) q75
treat_Post	0.316***	0.259***	0.212***
	(0.0215)	(0.0183)	(0.0174)
Male	0.587***	0.497***	0.469***
	(0.0164)	(0.0116)	(0.0160)
age	0.136***	0.103***	0.0866***
	(0.00564)	(0.00365)	(0.00356)
age2	-0.00164***	-0.00126***	-0.00107***
	(7.11e-05)	(4.34e-05)	(4.41e-05)
EducYear	0.0294***	0.0241***	0.0170***
	(0.00580)	(0.00541)	(0.00618)
EducYear2	0.00208***	0.00224***	0.00243***
	(0.000335)	(0.000293)	(0.000285)
WorkExperience	-0.0158***	-0.0123***	-0.00889***
	(0.00114)	(0.000788)	(0.000789)
Urban	0.348***	0.227***	0.168***
	(0.0245)	(0.0198)	(0.0131)
Province FE	YES	YES	YES
Year FE	YES	YES	YES
Constant	3.118***	4.392***	5.254***
	(0.140)	(0.0799)	(0.0924)
Observations	23,450	23,450	23,450

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 4: Robustness check

section 2.3. This means our model specification is robust.

2.6 Conclusion

After empirical analysis above, we can come to the conclusion that, the use if Internet has positive impact on the income of all population, and the effect is larger for people with lower income, which can help to reduce the income inequality.

3 Summary

According to our empirical study, the use of quantile regression could show the different influences of the use of Internet on the different level or quantile of income, and the empirical results tell us that people with higher level of income have a lower increasing rate of income when they starting to use the Internet, and vice versa. Then it could helps us analyze the effect of using the Internet on the reduction of income inequality, for the people with lower income have a much larger increment of income than the people with higher income. And this could not be realized with the simple OLS. Therefore, quantile regression is not a specific regression model, but a class of regression model, or an improved idea, which we can apply to linear regression, polynomial regression, kernel regression and so on to help us improve our models.

References

[1] Kevin F. Hallock Roger Koenker. Quantile regression. JOURNAL OF ECONOMIC PERSPECTIVES, 15(4):143â156, December 2001.

[2] Sun Yiping, Xu Yingbo. Research on the Impact of Internet popularization on Income distribution of Chinese residents – An empirical analysis based on CFPS Data [J]. Macroeconomic Research,2021,(07):161-175.]