The Empirical Study on the Urban-Rural Income Gap in China

Huafu Shen, Pei Mao[1]
Department of Economics
Central University of Finance and Economics
39 South College Road Haidian District
Beijing, China, 100081

Abstract

This paper tries to use the counterfactual decomposition method to study the urban-rural income gap via quantile by adding some factors, such as Chinese characteristic like height and social status. We find education, gender, and age are still the influence factors of urban-rural income, and also are significant influence in each quantile. After introducing the height and social status, we found that the coefficient of height from the urban residents was not significant. However, social status has a positive influence on the low-income and middle-income urban residents, and high-income rural residents, indicating that the urban and rural areas have different views on social status. In the decomposition of income gap, the coefficient effect explains most causes of urban-rural income gap. But the tendency of characteristics effect is consistent with that of the total effect.

***Keywords*：** Urban-rural income gap; Quantile regression; Counterfactual decomposition

1.Introduction

City and the country are the two sides of a coin. But the urban-rural income is different in every country. Less is the developed country, and the large is the developing country. So, the income gap is always the hotspot in the developing country. Especially for the urban-rural dual structure in China，the urban-rural income gap attracts more attention than that in other countries. Since the reform and opening up happened over the past 30 years, China's economy has been developing unprecedentedly, and the residents' average income has also been significantly improved. We can see, However, from the table 1 below, the urban-rural income rose beyond 2.3 times from 1991 to 2015 during the over 20 years, especially beyond 3 times in 2010-2013. Some people get rich rapidly, but someone lives still in poverty. The large income gap shows the goal that when some people and some regions get rich first, others will be brought along, and through this process common prosperity of the entire population will be gradually achieved remains to be implemented. It also has a negative influence on social security system and people's happiness in China, which may lead to the social instability. We can figure further out the ratio of urban and rural income gap is decreasing, which also shows China has taken measure to reduce it recently. In order to reduce the income gap efficiently, we need to figure out which affects the income gap.

Table I. Ratio of urban-rural income gap（1991-2015）

Year	**1991**	**1995**	**2000**	**2005**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**
Ratio	2.4	2.71	2.79	3.22	3.23	3.13	3.10	3.03	2.97	2.32

Data：China Statistical Yearbook 2015-2016

The last decade has witnessed an increasing interest by economists in the analysis of wage inequality. Three thoughts are widely used in wage composition.

First, to decompose the income inequality coefficient like Gini decomposition and Theil decomposition. Some are used from Chinese Data. By decomposing Gini coefficient, the wage contributes most in the urban-rural income gap, and transfer income follows (Qiyun, 2015). And the urban-rural income gap is the most important factor affecting the overall income gap between urban and rural areas by decomposing the Theil coefficient (Hongtao, 2009).

[1] Pei Mao is the correspondent author.

Another two ways are about the wage differentials, with respect to the mean composition and distribution composition. Hence, the second is the mean composition which first proposed Oaxaca(1973).Upon the difference in wage between gender, they take the counterfactual analytical method to construct the counterfactual wage. They assumed that if women had the same characteristics as men, then what would their average salary be. Oaxacade composition contained that the first is that the wage differentials are decomposed into composition effect (the "explained" part) and structure effect (the "unexplained" part) and the second is that the "detailed" decomposition is into the contribution of each individual covariate. The method has been widely used to understand racial and gender wage differentials. From a gender perspective, gender discrimination has always existed in our society; After that, Oaxaca and Ranson (1994) further propose a procedure to estimate the nondiscriminatory wage structure that analyzed union/nonunion wage differentials .

Obviously, the Oaxaca Decomposition describes the differentials in mean. and it does not cope with the differentials in distribution, which is restricted that we intend to know the detailed information from the distribution. Hence the third composition focuses obviously on the distribution composition. In regard to the different regression model, it can be divided into five patterns. The first is from JMP composition based on the classical regression (Juhn，Murphy，and Pierce，1993), but it is questioned by Melly and Yun, Myeong[2];The second is from DFL composition based on the semiparametric estimator (DiNardo, Fortin,and Lemieux, 1996) ;The third is from MM composition based on the unconditional regression(Machado and Mata, 2005);The fourth is from FFL composition based on the unconditional regression(Firpo, Fortin, and Lemieux, 2007); The fifth is from Melly composition based on the semiparametric estimator and unconditional regression(Melly, 2005).The unconditional composition from Melly is that the conditional distribution integrated over the range of covariates. and the decomposition of changes explains wage differentials in terms of differences in individual characteristics differences in the coefficients of wage equations and differences in residuals. More about decomposition might refer to the Fortin, Lemieux, and Firpo (2011).

The literatures about decomposing the income gap give us a good reference on seeking for the factors affecting the urban-rural income gap in China. As we know, the large income gap has taken place in China. And in order to figure out which factor is the most prominent and which effect contributes most to the total effects in every quantile, we intend to make some changes based on the Melly(2005). For the factors, the variables such as education, gender and age, as we know, appeared in many articles. However, when controlling for the age, they most neglect the fact that with the age increasing, and the influence of age on income may decrease. Hence, we intend to add the age square into the control variable. Second, as the object of study is from China, some Chinese characteristics like social statue are valued most. Chinese think highly of reputation which can bring them wealth, so the social statues as a social capital will be used into the variable. Third, another factor introduced is height. The city's health care will be better than the countryside, and that some research has proved the height is related with the income（Judge, 2004）.So a healthy height will bring them benefits. Based on the three above and with the help of quantile regression and counterfactual decomposition, we use the data from CFPS(China Family Panel Studies, CFPS) to decompose the urban-rural income gap again and figure out which factor is the most prominent and which effect contributes to the total effects.

2. Counterfactual Decomposition

It is well known to us all that the counterfactual decomposition is first proposed by Nobel Laureate Fogel who studied the relationship between the United States railway and the economic growth in the 19th century. Generally speaking, if there is no railway in the United States, the economy did not grow so fast, so the economic growth is closely related to the railway. Under the hypothesis, he proposed that if the railway did not exist at that time, what would happen. He started to focus on the American clinometric and came to a conclusion that other things are equal, the American GNP would decrease 3% than that in fact in 1890. From this, we can know that this counterfactual method means that in order to compare the contribution of each influencing factor, among the overall factors, we often suppose some factor is not contained and observe the change of the corresponding value and real value of the dependent variable. This thought is widely used in the income distribution accordingly.

[1] Melly thought it neglected the heteroscedasticity. And Yun, Myeong（2007）thought JMP have to rely on a few strong assumptions. First, OLS estimates of one group are not biased; Second, discrimination is stable over time.

The counterfactual thought is used in this paper, which is built in the quantile regression. The quantile hypothesis comes from Koenker(1978) below.

$F_{y|x}^{-1}(\tau|x_i) = x_i\beta(\tau), \forall\tau \in (0,1)$

where the $F_{y|x}^{-1}(\tau|x_i)$ is the conditional distribution y on x under the τ quantile. Meanwhile, Koenker(1978) proposed the $\beta(\tau)$ estimator below.

$$\hat{\beta}(\tau) = \arg min\frac{1}{N}\sum_{i=1}^{N}(y_i\text{-}x_ib)\big(\tau\text{-}1(y_i\text{-}x_ib)\big)$$

As under the conditional quantile $\tau_i \leq \tau_k$, we can't follow that $x_i\hat{\beta}(\tau_i) \leq x_k\hat{\beta}(\tau_k)$ which can't satisfy the monotony. Melly(2005) integrated the conditional distribution over the whole range of the distribution of the regressors as follow.

$$\mathrm{q}_0 = \mathrm{F}^{\text{-}1}(\theta) \Leftrightarrow \int 1(\mathrm{y_i}\text{-}\mathrm{q}_0)\, \mathrm{dF_y}(\mathrm{y}) = \theta \Leftrightarrow \int\int 1(\mathrm{y_i}\text{-}\mathrm{q}_0)\mathrm{f_{Y|X}}(\mathrm{y|x})\mathrm{d\,F}_X(\mathrm{x}) = \theta$$

$$\Leftrightarrow \int(\int_0^1 1\big(\mathrm{F_{Y|X}^{-1}}(\tau|\mathrm{x_i}) \leq \mathrm{q}_0\big)\mathrm{d}\tau)\mathrm{dF}_X(\mathrm{x}) = \theta$$

where $q_0$is the population's θ quantile of y. Finally, after taking the infimum of the set, the sample analog of q_0 is given by

$$\hat{q}(\hat{\beta},x) = \inf\Big\{q{:}\tfrac{1}{N}\textstyle\sum_{i=1}^{N}\sum_{j=1}^{J}(\tau_j\text{-}\tau_{j\text{-}1})1\big(x_i\hat{\beta}(\tau) \leq \mathrm{q}\big) \geq \theta\Big\} \qquad (1)$$

We can represent 1 and 0 respectively represents different dimensions, such as the different year, gender and so on. If we code the urban and rural income as 1 and 0respectively. On the basis of this unconditional quantile, the urban income is distributed as below.

$$\hat{q}(\hat{\beta}^1,x^1) = \inf\{q{:}\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{J}(\tau_j\text{-}\tau_{j\text{-}1})1(x_i^1\hat{\beta}^1(\tau_j) \leq \mathrm{q}) \geq \theta\}$$

We can use the idea of counterfactual analysis furtherly. Therefore, if the rural residents lived in the city, they would have the property as the urban ones, the counterfactual distributed income constructed for the rural residents is as below.

$$\hat{q}(\hat{\beta}^1,x^0) = \inf\{q{:}\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{J}(\tau_j\text{-}\tau_{j\text{-}1})1(x_i^o\hat{\beta}^1(\tau_j) \leq \mathrm{q}) \geq \theta\}$$

If we consider the urban-rural income gap as D, by using the equation (1), D can be expressed in the τ quantile as below

$D = \hat{q}(\ddot{\beta}^1,x^1)\text{-}\hat{q}(\ddot{\beta}^0,x^0)$

Then, D can be decomposed as

$$\begin{aligned}D &= \hat{q}(\ddot{\beta}^1,x^1)\text{-}\hat{q}(\ddot{\beta}^0,x^0)\\ &= \Big(\hat{q}(\hat{\beta}^1,x^1)\text{-}\hat{q}(\hat{\beta}^{m1,r0},x^1)\Big) + \Big(\hat{q}(\hat{\beta}^{m1,r0},x^1)\text{-}\hat{q}(\hat{\beta}^0,x^1)\Big) + \Big(\hat{q}(\hat{\beta}^0,x^1)\text{-}\hat{q}(\hat{\beta}^0,x^0)\Big)\\ &= D_r + D_\beta + D_x \qquad (2)\end{aligned}$$

where Melly (2005) constructed the $\hat{\beta}^{\mathrm{m1,r0}} = (\hat{\beta}^1(0.5) + \hat{\beta}^0(\tau_j)\text{-}\hat{\beta}^0(0.5))$.

From equation (2), we can know D_r results in the changes in residuals, so it is called as Residuals Effect; The second $D_\beta$the effects of changes in (median) coefficients, it is called as Coefficients Effect; The third is the effects of changes in the distribution of the covariates, so it is called as Characteristics Effect.

3. The quantile regression on urban-rural income from China

3.1 Introduction

With the help of CFPS (China Family Panel Studies) data[3] which aims at studying the changes in Chinese society, economy, population, education and health by collecting the personal, family and community data, we choose five factors the "Age", "gender", "education", "height", "social status" (So-statue) as independent variables, and the

[3] The data is open. http://opendata.pku.edu.cn/dataset.xhtml?persistentId=doi:10.18170/DVN/45LCSO.

independent variable for personal income, and take its logarithm (linc).The variable “age” is counted in fact, and the male is label as “1” for the “gender”, and “education” is divided into eight grades from the uneducated to Doctor, and the centimeter is unit for "height", and “social status ”is a subjective variable. It asks you "how do you think of your social status locally". one is very low, and five is very high.

The latest data is from 2014 survey. In our sample processing, as the income from student come from his/her parents generally, so we first removed the identity of the student data; Second, the urban residents begin to retire for the men and the female at the age of 60 and 55 respectively. And for the people in the countryside who most will work until their body can't support, they can get the regular income beyond the age of 55. So, we removed the data that the rural residents get the income beyond the age of 55. Finally, the sample we can use is 6736, among which the sample size of the urban and rural areas was 3202 and 3534 respectively.

From the Figure 1 below, we can know the kernel density curve of urban person logarithmic income stay in the right of that of urban in the upper income. Furtherly we also can know, from the Table I, the difference between urban income and rural average income was 6619 yuan significantly, and the logarithmic income was also significantly different. In a word, the income wage obviously exists in the urban-rural residents

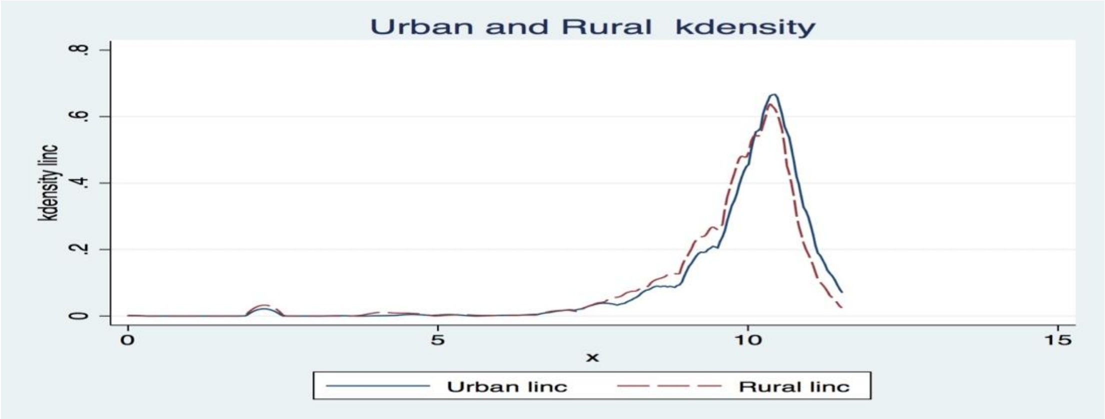


Figure 1. log(income) densities for urban and rural residents

It also shows from the Table II. that the amount of male is more than that of female both in the urban and rural region. And the amount of male in rural is more than that in urban. Maybe the rural residents prefer to giving birth to a boy, which suggests men can give more labor to their families. In term of education, the city provides more education resource. They also have an advantage in height. But in regard to social statue, the rural people value more.

Table II. Covariates: descriptive statistics

	Income			Log(income)			Gender		Education		Height		So-statue	
	Urban	Rural	Difference	Urban	Rural	Difference	Urban	Rural	Urban	Rural	Urban	Rural	Urban	Rural
Mean	18469	11850	6619***	9.95	9.72	0.23***	0.51	0.53	3.7	3.1	166.3	165.36	2.94	3.1
Sd	27129	18231	(16)	1.2	1.33	(7.3)	0.5	0.5	1.2	1.3	8.3	7.8	0.85	0.9
Max	408400	220000		11.5	11.5		1	1	7	7	216	192	5	5
Min	0	0		0	0		0	0	0	0	100	112	1	1
P75-25	30000	20000		0.981	1.08		1	1	2	2	12	10	0	1
N	5775	7599		3202	3498		5782	7614	4577	4946	5198	5913	4577	4946

t value in parentheses, *** p<0.01, ** p<0.05, * p<0.1

3.2 The quantile regression on urban-rural income from China

We regress the logarithmic income on the variables gender, education, age, age square, height and social statue via quantile in the urban and rural respectively. As the coefficients of the three factors gender, education, age are significantly positive in all quantiles both in the urban and the rural, so we display them in figure as bellows in Figure 2.

First, as for the urban and rural income, the age has positive significantly influence in every quantile. But it decreases when the income increases, which indicates the age is more important for the lower-income people than the high-income ones. Age can accumulate experience, so age maybe is the only source to gain skill for the low-income group. The big gap happens in the 10th quantile, which indicates the most different group between them for age is the low-income ones. For most quantiles, the coefficient of the age from the urban residents is upper than that from the rural. Maybe the people in the urban get skill faster than one in the rural with age. As the coefficient of age square is significantly negative[4],it says the influence of age on income will decrease after some age in every quantile. Second, in term of gender, it has the positive significantly influence on income from the residents both in the urban and in the rural. Male has more economic status than female.

However, in every quantile, the coefficient from the urban is large than that from the rural. It indicates that the male in the urban has more importance on income. Third, in term of education, the two coefficients curve are above line “0”, it indicates education can add the income for people both in the urban and the rural. The two curves do not intersect, and the urban curve is above the rural curve, which illustrates the effect of education from the urban is more outstanding. we also can know from the figure that the coefficients gap of education is smaller than that of gender.

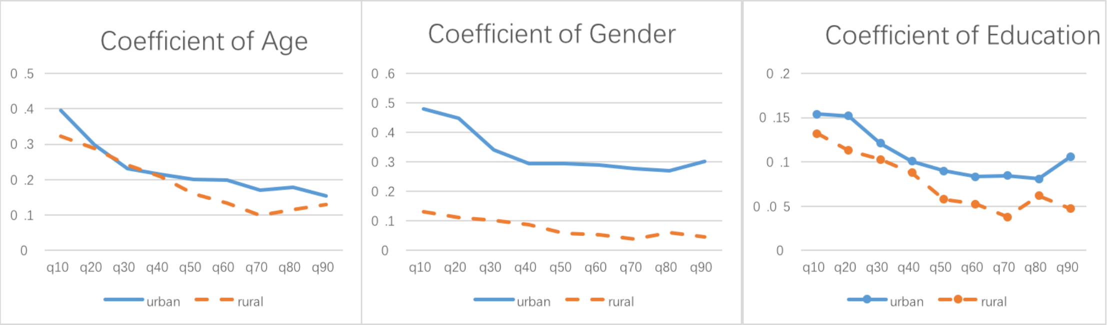


Figure2. Coefficients of Age, Education and Gender from urban-rural income

Except the three factors above, next we consider the height and social statue from Table III as below. It states that the impact on income is not significant in every quantile. First, the height is not significant for the urban resident. It illustrates that short men are also able to earn high income as the same as tall. From the other side, it illustrates that the urban residents regard other capitals as more important than height. But it has only significant positive influence on the rural residents in the 60th and 70th quantile. This may be a reflection of the "appearance" capital of the middle-income and high-income group in rural areas.

As for the social statue, it has the significant positive influence on the low-and middle-income residents, but not on the high-income residents. It states that social status is of no worth for the high-income group in the urban. Maybe age and education are more important for them. However, on the contrary, the social statue only has the significant positive influence on the high-income residents in the rural. It states that the social statue network is more valuable for the specific group, which is more likely to generate income for them.

Table III. The coefficient of height and social statue

	Variables	q10	q20	q30	q40	q50	q60	q70	q80	q90
Urban	Height	0.00965	0.00423	0.00468	0.00514	0.00395	0.0029	0.00297	0.00251	0.000987
		(0.0076)	(0.00534)	(0.00465)	(0.00344)	(0.00289)	(0.00221)	(0.00215)	(0.00218)	(0.00291)
	So-statue	0.154***	0.101***	0.0635***	0.0383**	0.0136	0.0173	0.0248	0.0205	0.0379
		(0.0495)	(0.032)	(0.0208)	(0.0163)	(0.0144)	(0.0136)	(0.0189)	(0.0192)	(0.0245)
Rural	Height	0.000625	0.00026	0.00244	0.00146	0.00476	0.00564**	0.00591**	0.00212	-0.000254
		(0.01)	(0.00841)	(0.00633)	(0.00646)	(0.00329)	(0.00287)	(0.0024)	(0.00287)	(0.0037)
	So-statue	-0.122	-0.0247	0.0244	0.0256	0.0232	0.0184	0.0362*	0.0194	0.0332*
		(0.0979)	(0.049)	(0.0302)	(0.0376)	(0.0253)	(0.0213)	(0.0196)	(0.0171)	(0.0178)

[4] The coefficient is not displayed in Figure.

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

4. The counterfactual decomposition in the urban-rural income

With the help with the equation (2) and combined with the quantile regression above, we can decompose the total effect for urban-rural income gap into three effects in Table IV as follow.

Table IV. The decomposition effects

	Total difference	Residuals	Coefficients	Characteristics
p10	0.374***	0.128***	0.101***	0.144***
p20	0.312***	0.085***	0.101***	0.126***
p30	0.246***	0.044	0.105***	0.098***
p40	0.196***	0.012	0.105***	0.079***
p50	0.167***	-0.004	0.108***	0.064***
p60	0.155***	-0.006	0.109***	0.052***
p70	0.155***	0.0005	0.107***	0.048***
p80	0.161***	0.006	0.107***	0.048***
p90	0.181***	0.022	0.111***	0.047***

We also provide the same results in the attached figure. According to the Table IV and the attached figure, we can figure it out below. First, for the residuals effect, it is significant positive only in the 10th and 20th quantile. As the residuals effect accounts for the unobserved part for the total effects except for the coefficients effect and characteristics effect. Hence the five factors can't explain the total effect completely, and there is something unobserved such as institution, hukou and so on. As for the other quantile, the residuals effect is not significant again. So, five factors can explain the most total effect. The total effect can be decomposed into the covariates.

Second, for the coefficients effect, it changes within a small rang. It states that the coefficients effect is not an important factor influencing the change of total effect. When we consider the coefficients effect, it means that characteristics and residuals are kept at the same level. The coefficients effect assumes that they would have had the same gender, the same height, the same age, the same education and the same social statue. They are all the same in the over quantile. Therefore, we think the tiny differentials in every quantile may be is from the identity between the urban and the rural residents. Third, for the characteristics effect, it decreases when the income gap increases. The tendency of characteristics effect is consistent with that of the total effect. The characteristics effect comes from the variables, which is related to the gender differentials, the average education level and the age distribution between the two group. In the lowest quantile, it is more obvious for the differentials. Since it is hard to control the age distribution and gender, so the education is more important to reduce the characteristics effect to ultimately reduce the total income gap effect by providing the low-income group, especially people in the remote region with more education service. Finally, for the total effect, the large effect appears in the lowest quantile. And the characteristics effect explains most causes of urban-rural income gap. The next one is the residuals effect. The coefficients effect contributes least. In a word, the amount changes of variables concerning the characteristics effect is the key to lessening the total effect in the low quantile.

5. Conclusion

The income gap in China is a hotspot, especially for the Chinese researcher. This paper tries to use the counterfactual decomposition method to study the urban-rural income gap for the adult residents in the urban and rural areas in the quantile by adding some factor variables, such as China's characteristic like height and status in society, finding education, gender, and age is still the influence factors of urban and rural income, and also are significant influence in each quantile. After introducing the height variable, we found that the height is regarded as an important capital to gain income for the rural people in the 60th and 70th quantile, but not an important factor affecting the income for the urban people. Maybe the city provides the same medical service. However, social status has a positive influence on the urban residents of low-income and middle-income and high-income rural residents, indicating that the urban and rural areas have different views on social status. In the decomposition of income gap, the coefficient effect explains most causes of urban-rural income gap. But the tendency of characteristics effect is consistent with that of the total effect. The residuals effect follows. The coefficients effect contributes the least. The amount change of variables is the key to lessening the total effect in the low quantile.

Therefore, in order reduce the income gap, we propose the government should provide the low-income group, especially people in the remote region, with more education expenditure and high-quality education service.

References

Firpo,Sergio,Fortin,N . and Lemieux, T . (2007).“Decomposing Wage Distributions Using Recentered Influence Function Ｒegressions” , Mimeo , Department of Economics , University of PUC - ＲIO.

Dinardo J, Fortin N M, Lemieux T. (1996).Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach. Econometrica, 64(5):1001-1044.

Hongmei Li(2012).The counterfactual decomposition of the income gap between the urban and the rural In China[J]. The practice and understanding of mathematics.(10)

Hongtao Wang(2009).The study on the urban-rural income gap—from the Theil decomposition [J], Economic Forum, (12):4-8.

Judge T A, Cable D M.(2014). The effect of physical height on workplace success and income: preliminary test of a theoretical model.. Journal of Applied Psychology, 89(3):428-41.

Juhn C, Murphy K M, Pierce B.(1993). Wage Inequality and the Rise in Returns to Skill.[J]. Journal of Political Economy, 101(3):410-442.

Machado J A F, Mata J. (2005).Counterfactual Decomposition of Changes in Wage Distributions Using Quantile Regression. Journal of Applied Econometrics, 20(4):445-465.

Melly B.(2005). Decomposition of differences in distribution using quantile regression. Labour Economics, 12(4):577-590.

Nicole Fortin,Thomas Lemieux,Sergio Firpo(2011).Decomposition methods in economics[M],Handbook of Labor Economics.Volume 4A.

Oaxaca R L, Ransom M R.(2004). On discrimination and the decomposition of wage differentials ☆[J]. Journal of Econometrics, 61(1):5-21.

Qiyun li, Cheng Chi(2015). Research on the urban-rural income gap under the perspective of income sources[J]The Reform of Economic System, (6):47-54.

Roger Koenker (2005).Quantile regression[M],Cambridge University Press.

Ronald Oaxaca(1973). Male-Female Wage Differentials in Urban Labor Markets. International Economic Review, 14(3):693-709.

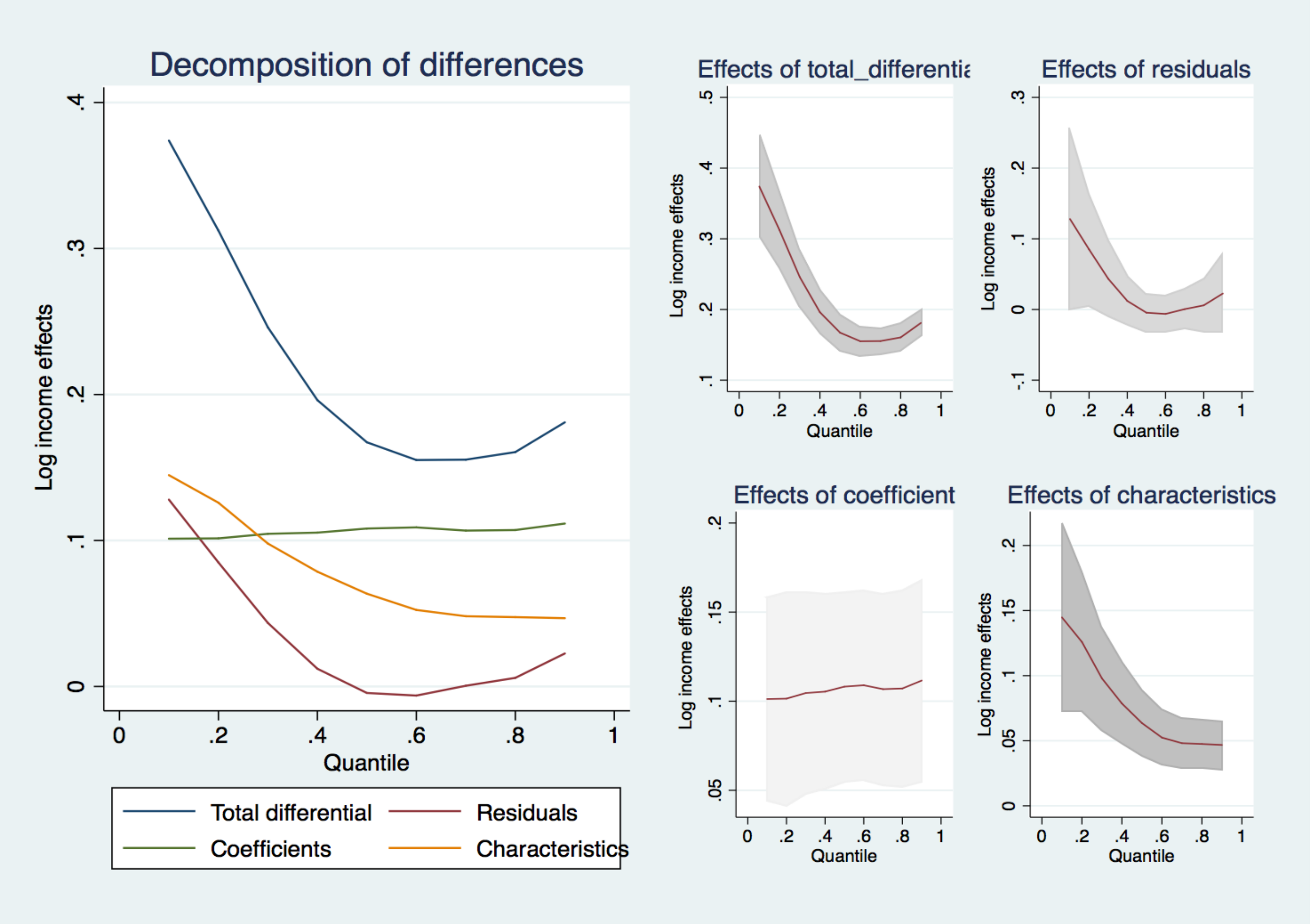
Sergio Firpo, Nicole M. Fortin, Thomas Lemieux(2009). Unconditional Quantile Regression. Econometrica, 77(3):953-973.

Shulan Fei, Jiqiang Guo(2014). Distributional Decomposition on the Income Gap between Urban and Rural Workers[J], Zhejiang Social Sciences, (11):13-24.

Yun,Myeong-su.Wage differentials discrimination and inequality: a cautionary note on the Juhn,Murphy and Pierce decomposition method[J],https://trove.nla.gov.au/work/47413060?q&versionId=60323314

Appendix

Figure. The decomposition effects



Note: The shadow area is the confidence interval