High-speed Rail and Inter-provincial Inequality in payoffs to human capital

Yiqing Zheng

May 27, 2020

Abstract. This paper investigates the relationship between inter-provincial inequality in payoffs to human capital and high-speed rail (HSR). The paper uses counterfactual strategy to calculate the opportunity cost for people to work in their resident province. The paper compares the prediction accuracy of several regression models and adopts the random forest estimation. The location cost decreases about 166.30 yuan with the existence of supporting HSR lines that links the worker's resident province and the province of maximum predicted earning. The kernel density estimation further verifies this negative correlation relationship between HSR and location cost.

1 Empirical model

1.1 Counterfactual strategy

The first part follows Zax (2019) and uses counterfactual strategy to calculate a kind of location cost which reflects inter-provincial inequality in payoffs to human capital. The strategy is shown as following.

Firstly, I run the regression below for each province separately

$$y_{ji} = \beta_{j0} + \beta_{j1} gender_{ji} + \beta_{j2} age_{ji} + \beta_{j3} age_{ji}^2 + \beta_{j4} school_{ji} + \epsilon_{ji} \tag{1}$$

where j indexes provinces and i indexes individuals within province j. The dependent variable is monthly earning, and the independent variables are gender, age, quadratic term of age, and education level. Since there are 14 provinces in the data set, this step yields 14 estimation results.

According to those estimation coefficients, I predict individual monthly earning in his province of residence j and any province k assuming that the person lives in province k. Individual i's predicted earning in resided province is

$$\hat{y}_{jji} = \hat{\beta}_{j0} + \hat{\beta}_{j1} gender_{ji} + \hat{\beta}_{j2} age_{ji} + \hat{\beta}_{j3} age_{ji}^2 + \hat{\beta}_{j4} school_{ji}. \tag{2}$$

The first subscript denotes the province of residence and the second subscript denotes the province within which the earning is predicted. Similarly, for individual i now living in province j, his predicted earning in province k is

$$\hat{y}_{jki} = \hat{\beta}_{k0} + \hat{\beta}_{k1} gender_{ji} + \hat{\beta}_{k2} age_{ji} + \hat{\beta}_{k3} age_{ji}^2 + \hat{\beta}_{k4} school_{ji}. \tag{3}$$

For individual i living in province j, his maximum predicted earning is

$$\hat{y}_{jm_i i} = \max_{k \in K} \hat{y}_{jki} \ , \ K = \textit{the set of all provinces.} \tag{4}$$

So m_i denotes the province of maximum predicted earning for worker i living in province j. The province of maximum predicted earning may or may not be the province of residence.

Location cost is difference between maximum predicted earning and predicted earning in resided province, so it measures the opportunity cost for individual i to work in province j. If location cost is greater than zero, it means that the same person can

gain higher earnings in another province, so location cost also reflects inter-provincial inequality in payoffs to human capital.

$$Location\ cost_i = \hat{y}_{jm_ii} - \hat{y}_{jji} \tag{5}$$

This paper investigates the relationship between location cost and high speed rail (HSR). After calculating location cost for every observation, we also know m_i, the province of maximum predicted earning. Then a dummy variable $transportation_i$ can be created to represent whether there is an HSR line linking province j and m_i.

1.2 Computational methods

The analysis depends heavily on the computation of location cost, so the accuracy of prediction is very critical. To promote the accuracy, different from Zax (2019), I compare the mean squared error of several regression methods, including ordinary least squares (OLS), lasso regression, ridge regression, and random forest regression. Best-fit parameters for these regression methods are found through k-fold cross validation.

After selecting best-fit regression model, I then can use it to predict earnings in each province and get location cost. The relationship between location cost and transportation is then demonstrated by a simple regression and kernel density estimation.

2 Data

I use the Chinese Household Income Project (CHIP) for the income part.CHIP has conducted five waves of surveys in 1989, 1996, 2003, 2008, and 2013. The survey provides detailed information on household and individual income and expenditure as well as basic demographic features.

This paper uses the CHIP2013 data because this is the only wave of survey conducted after 2008 when the development of HSR in China was formally initiated. CHIP2013 reports total individual income from primary job and other jobs. The wave covers 14 provinces from eastern, western, and center regions in China. Following Zax (2019), I only focus on individuals who hold urban hukou and exclude those whose working time is less than six months a year, 15 days a month, or six hours a day. The sample also drops those whose annual total earning from primary job is less than

4,000 yuan. The annual earning level is rescaled to monthly earning by their reported working months a year. There are 8759 observations in total.

HSR data are from the Chinese Research Data Services Platform. The data include the name, opening time, and passing stations of each HSR line. Since CHIP2013 was conducted in 2014, I only include HSR lines opened before 2014. I created a dummy variable called transportation for eveary observation from the CHIP2013 database. The dummy equals to one if the person can earn predicted maximum earning in the current resident province or there exists an HSR line directly sending the person to the province where he earns his predicted maximum earning; otherwise zero.

3 Results

3.1 Summary statistics

The result of descriptive statistics is reported in Table 1. The mean of monthly earning is 3554.43 yuan and the maximum is 150,000 yuan. 43% of the observations are female. The range of age is from 15 to 78 and the mean is about 40 years old. The average years of formal education are about 12 years, corresponding to the high-school education level in China. About one quarter of the observations can utilize HSR to directly travel to the province where they earn highest predicted earning or can earn highest predicted earning in their resident province.

3.2 Regression result

Table 2 gives the coefficients of OLS regression for each province. As it is shown in the result, all the variables are highly significant in almost every province and education is significant in each regression. Payoff to one year of formal education is higher in developed regions such as Beijing and Guangdong and becomes lower in less developed region such as Gansu province. Although I do not use OLS regression to calculate predicted earning and location cost, the coefficients from OLS provide a general concept about regional inequality in payoffs to human capital.

Table 1: Descriptive statistics

variable	mean	sd	min	p25	p50	p75	max
earning	3554.43	3522.67	333.33	2000.00	3000.00	4166.67	150000
gender	0.43	0.49	0.00	0.00	0.00	1.00	1.00
age	40.43	9.96	15.00	33.00	41.00	48.00	78.00
school	11.87	3.28	0.00	9.00	12.00	15.00	21.00
transportation	0.25	0.43	0.00	0.00	0.00	1.00	1.00

[1] The number of observations for all variables is 8759.

[2] The variable gender equals to one if the individual is female; otherwise zero.

[3] The method used here to find maximum predicted earning and the corresponding province is random forest regression. It is shown below that random forest regression is the best-fit method. The dummy transportation can only be evaluated after finding maximum predicted earning and the province where it happens.

Table 2: OLS regression coefficients for fourteen provinces

province	gender	age	age^2	school	intercept	$\#obs$
Beijing	-808.01***	296.80***	-295.85***	546.40***	-9,111.83***	873
Shanxi	-1009.67***	193.98***	-230.53***	111.55***	-1,809.82*	719
Liaoning	-531.82	406.12*	-465.32*	268.72**	-8,027.06	504
Jiangsu	-1557.10***	373.96***	-412.66***	268.59***	-6,236.14***	964
Anhui	-1087.25***	149.72*	-141.33	180.32***	-1,944.17	495
Shandong	-485.74***	169.14***	-185.56**	174.81***	-2,169.05*	614
Henan	-556.16***	148.34***	-162.79***	188.29***	-2,247.48**	707
Hubei	-950.13***	95.78	-87.85	268.17***	-1,818.30	638
Hunan	-818.24***	264.69***	-316.75***	116.79**	-2,889.62	478
Guangdong	-1,437.56***	321.64***	-323.14**	449.17***	-7,555.37***	796
Chongqing	-724.80***	126.32**	-141.97*	164.25***	-873.09	613
Sichuan	-711.62***	130.04**	-130.47*	185.08***	-1,586.66	454
Yunnan	-304.05**	200.70***	-214.64***	189.25***	-3,598.67***	475
Gansu	-513.36***	36.79	-28.35	161.06***	81.56	429

[1] Dependent variable: monthly earning.

[2] *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

3.3 Comparison of methods

Figure 1 compares MSE of different regression methods. The best-fit parameters of lasso regression, ridge ression, and random forest regression are chosen using k-fold cross validation, setting k=4. As it is shown in the figure, MSE of random forest regression is much smaller than MSE of the other three methods. Actually, according to the computation result, the order of MSE is as following: random forest (10162166.03) < OLS (10607887.00) < lasso regression (10608163.70) < ridge regression (10613054.00); thus, I choose random forest regression with its best-fit parameters to calculate counterfactual earning level and the location cost. All the analysis below is based on the random forest estimation.

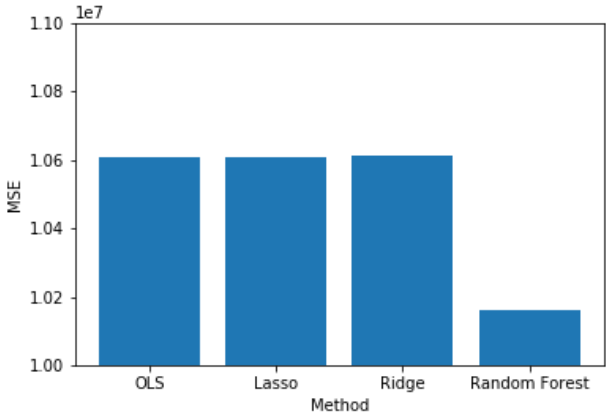

Figure 1: Comparison of MSE

3.4 Location cost and HSR

Figure 2 demonstrates location cost in different provinces. Location cost in Guangdong province is the lowest among the fourteen provinces, meaning that the opportunity cost for working in Guangdong province is the lowest. The large migration wave of workers to Guangdong province in the past few years in China verifies this conclusion. The second and third lowest location costs happen in Jiangsu and Beijing both of which are more developed provinces compared to the rest. Location costs in less developed regions such as Gansu and Henan are very high. The phenomenon indicates that regional inequality in payoffs to human capital is related to regional economic development. The reason may be that more developed regions enjoy richer resources,

infrastructure, and other factors, so people with same education years and ages can produce more in these regions.

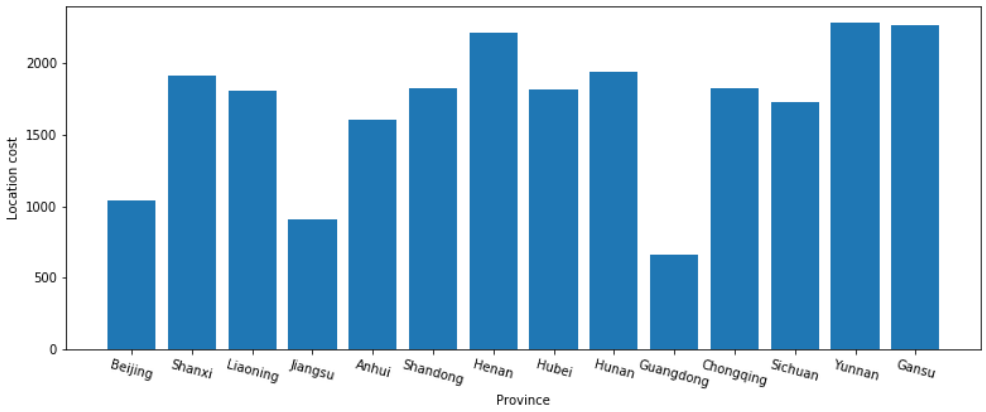

Figure 2: Location cost by province

Table 3 presents the coefficients of regression where location cost is the dependent variable and the dummy transportation serves as one of the independent variables. Column (1) is the regression result for the total sample. Column (2) shows the coefficients for the subsample whose location cost is not zero (i.e. they do not earn maximum predicted earning in their current resident province), so for regression (2) the value of the dummy variable transportation becomes one only if there exists an HSR line directly sending the person from the current province to the province where he earn maximum predicted earning.

It is shown in (1) that if people have direct access to the province where they earn maximum predicted earning or they already gain maximum predicted earning, the location cost decreases by about 617 yuan. For those who have not earned maximum predicted earning, the existence of HSR line from resided province to where they earn maximum earning, the location cost decreases by 166.3 yuan.

Figure 3 is the kernel density estimation of location cost. The the subsample with transportation = 1, location cost clusters in low values, while for transportation = 0, the mean of location cost shifts to the right. The kernel density estimation further suggests that location cost is lower with access to supporting HSR lines. Note that the regression result and kernel density estimation both verify the correlation relationship rather than the causal effects between HSR and location cost, for many potential

Table 3: Regression coefficients for location cost

	(1)	(2)
variable	total sample	subsample (location cost $\neq 0$)
gender	-338.60***	-407.20***
age	165.26***	177.28***
age^2	-179.62***	-193.80***
school	175.53***	196.30***
transportation	-616.74***	-166.30***
#obs	8,759	8059

[1] Dependent variable: location cost.

[2] Controlling for province fixed effect.

[3] *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

factors may influence location cost and the construction of HSR simultaneously. There are mainly two possible explanations for the correlation relationship:

(1) The decision to open HSR in a province may depend on its regional economic development. Provinces passed by more HSR lines are more economically developed, so location costs in such regions are also lower.

(2) This is the result of "self selection". People who suffer from high location cost choose to move to regions that reward them higher payoff if the movement is convenient with appropriate HSR lines. For those who can directly travel to the province of maximum predicted earning by HSR but still choose to stay, the reason is that the location cost of staying is not high enough to drive them to migrate. Because people do not have the right to choose to move or stay if there is no supporting transportation, the consequence of this "self selection" procedure is that access to appropriate HSR lines always co-occurs with lower location cost.

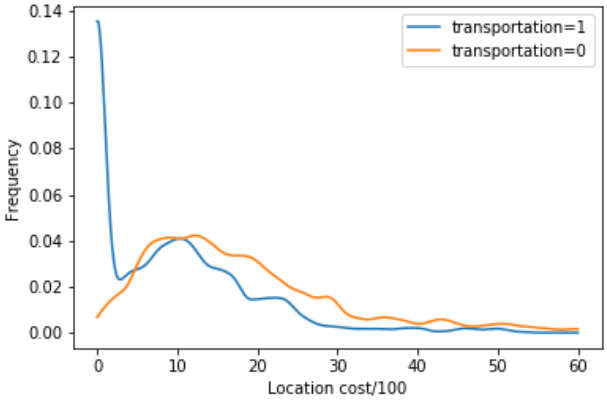

Figure 3: Kernel density estimation of location cost by transportation

References

Zax, J. S. (2019), 'Provincial valuations of human capital in urban china, inter-provincial inequality and the implicit value of a guangdong hukou', *working paper (March 29, 2019)* .