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

Yiqing Zheng

June 11, 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 working 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 a supporting HSR line that links the worker's resident province with the province of maximum predicted earning. The kernel density estimation further verifies this negative correlation relationship between HSR and location cost.

keywords: regional inequality, human capital, location cost, high-speed rail

1 Introduction

Inter-provincial migration has been a significant phenomenon in China since 1980s. In some export centers such as Shenzhen and Guangdong, migrant labor accounted for 70%-80% of the labor force in the early years of the 21st century (Chan, 2009). Part of the reason lies in the inequality in payoffs to labor (Park and Wang, 2010). China also witnesses a rapid development of high-speed rail (HSR) since 2008. HSR lines have now formed a network linking important cities and provinces in China. The formation of this network provides workers a great opportunity to improve their earnings by migration. The objective of this paper is to find whether the existence of appropriate HSR lines reduces the inefficient allocation of human capital. To measure the inequality in payoffs to human capital, the paper uses a counterfactual strategy combined with computational methods to calculate workers' potential "location cost" for residing in current provinces. The paper compares several regression methods to accurately predict location cost, including OLS, ridge, lasso regression, and random forest. Random forest turns out to be the best fit model. The paper finds that, according to the kernel density estimation, the distribution of predicted location costs shifts to the right for the cohorts who do not have access to supporting HSR lines. Moreover, the result of a simple fixed effect model demonstrates that a supporting HSR line that directly links resident province with the province of maximum predicted earnings reduces location cost by about 166.30 yuan.

Three aspects of literature are related most closely with the research question of this paper: (i) regional inequality and its relationship with transportation infrastructure in China; (ii) the application of China Household Income Project (CHIP) data set in the topic of human capital; (iii) the adoption of counterfactual strategy.

There exists a large amount of literature discussing regional inequality in China and potential causes of this phenomenon. According to existing literature, China witnessed a decline in regional inequality from approximately 1980 to 1990 (Chen and Fleisher, 1996), but then the regional inequality continues to increase (Zax, 2019; Tian et al., 2016; Yao and Zhang, 2001).

Using residuals from a fixed effect Solow growth model, Fleisher et al. (1997) find that total factor productivities in coastal provinces are twice as high as that in non-

coastal provinces. Fleisher et al. (2010) measure inequality across major regions simply by the coefficient of variation of per-capita real gross domestic product and confirm that inequality in China increases from 1990s. Tian et al. (2016), Yao and Zhang (2001), and Pedroni and Yao (2006) all adopt the nonstationary panel techniques introduced by Evans (1998), using pair-wise convergence between provinces to measure regional inequality and divide provinces into different clubs. Tian et al. (2016) further concludes that the inequality between the two clubs are increasing due to different investment in physical and human capital. Different from all the methods stated above, Zax (2019) uses a counterfactual strategy to calculate people's hypothetical earnings in different provinces and focuses on inequality in human capital payoff by comparing predicted earnings from different regions for a same individual. This paper follows the method developed by Zax and also puts emphasis on inequality in reward to education rather than general economic growth or productivity.

Zax (2019) just demonstrates the regional gaps in payoffs to human capital but does not explain relevant reasons. This paper tries to establish the relationship between regional inequality in returns to human capital and transportation infrastructure, focusing on HSR. Fleisher et al. (1997) discover that the effect of infrastructure on total factor productivity is higher in coastal regions than interior regions in China. Using a panel data covering from 2000 to 2014, Chen and Haynes (2017) find that the development of HSR helps decrease regional disparities. They adopt the rail network density and accessibility to capture quantity and quality change in rail investment. Banerjee et al. (2012) address the endogeneity issue between transportation investment and economic development through utilizing the fact that transportation networks tend to connect historically important cities. They conclude that access to transportation networks is beneficial to regional economic growth.

CHIP data set provides comprehensive details on individual earnings, so it is broadly adopted in literature studying human capital and earnigns in China. Liu (1998) exploits the 1988 CHIP survey to estimate cross-industrial return to human capital. He finds that return to education varies from 3% to 6% and is higher in industrial sectors. The article also identifies two labor reform programs and obtains significant and positive effects of them on earnings. Li (2003), using CHIP survey in 1995, argues that return to education was underestimated in previous literature and

educated people are more rewarded in less-developed, low-income provinces.

Both Whalley and Xing (2014) and Zax (2019) incorporate multiple waves of CHIP surveys in their research. Whalley and Xing (2014) estimate return to education in 1995, 2002, and 2007 and find that only coastal regions witnessed an increase in skill premia during 2002-2007 and urban-rural wage inequalities are also more obvious in such regions. Besides, this kind of variation in earnings are increasing in China. Zax (2019) uses CHIP data from 1988 to 2013 and suggests that cross-provincial inequality in human capital decreases over time.

As for the method, counterfactual strategy is often applied in regional inequality studies to explore how the differences in some regional structures, such as population, industrial sectors, labor allocation etc., affect cross-regional inequalities.

Zhang et al. (2015) calculate counterfactual per capita output inequality level cross provinces by assuming that all provinces have the same age structure as the national average level. The result suggests that Gini coefficient would decline by 13% under the counterfactual assumption. Démurger et al. (2009) simulate counterfactual job status, earnings, and working hours to decompose urban-rural income inequality into four sources. The study finds that population effect is the most robust and significantly important one in influencing income differences between rural migrants and urban residents. Xing (2014) calculates the counterfactual earnings of rural-urban migrants, assuming that they are paid as rural local workers. The comparison between predicted earnings and actual rural wage densities help examine self-selection in rural-urban migration.

This paper also adopts counterfactual strategy and uses the CHIP2013 data set. The paper mainly contributes to existing literature on two aspects. The first is that the paper combines counterfactual strategy and computational technologies to more precisely measure regional inequality in payoffs to human capital. Most previous studies use ordinary least squares (OLS) to implement counterfactual strategies. The paper compares OLS, lasso regression, ridge regression, and random forest to find the most accurate estimation of regional payoffs to education level according to their mean squared errors. K-fold cross validation is used to find the best parameters for each model.

Besides, the paper also digs more deeply into the topic by building a relationship

between regional inequality in human capital valuation and transportation. Most previous research focuses on the effect of infrastructure on macro indices such as regional economic growth. It is also meaningful to investigate whether transportation would influence personal choice of migration especially when human capital are rewarded higher outside home cities.

2 Empirical model

2.1 Counterfactual strategy

The first part of the empirical model follows Zax (2019) that uses counterfactual strategy to calculate a kind of location cost which reflects inter-provincial inequality in payoffs to human capital. The structure of the model 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 earnings in his province of residence j and any province k assuming that the person lives in province k. Individual i's predicted earning in resident 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 j.

Location cost is the difference between maximum predicted earning and predicted earning in resident 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 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 the resident province j and the province of maximum predicted earning m_i.

2.2 Computational methods

The analysis depends heavily on the computation of location cost, so the accuracy of predicting personal earning is very critical. To promote the accuracy, different from Zax (2019), I compare the mean squared error (MSE) 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 and its corresponding parameters, I then can use it to predict earnings in each province and get location cost for each person in the sample. The relationship between location cost and transportation is then demonstrated by a simple provincial level fixed effect regression and kernel density estimation.

2.2.1 Ridge regression

Ridge regression puts a constraint on the magnitudes of estimated coefficients by adding a penalty term in the cost function. The estimated coefficients satisfy the

following condition:

$$\hat{\beta} = argmin \sum_{1}^{N} (y_i - \alpha - \sum_{j} \beta_j x_{ij})^2 + \lambda \sum_{j} \beta_j^2. \tag{6}$$

2.2.2 Lasso regression

Lasso regression was firstly introduced by Tibshirani (1996). It does variable selection and shrinkage, whereas ridge regression, in contrast, only shrinks (Tibshirani, 2011). Lasso regression minimizes the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant (Tibshirani, 1996). Some coefficients will be exactly zero because of this constraint. The lasso estimate $(\hat{\alpha}, \hat{\beta})$ is defined as

$$\begin{aligned} (\hat{\alpha}, \hat{\beta}) &= argmin \sum_{1}^{N} (y_i - \alpha - \sum_{j} \beta_j x_{ij})^2 \\ \text{subject to } & \sum_{j} |\beta_j| \leq t \end{aligned} \tag{7}$$

where t is a tuning parameter.

This is equivalent to solving the penalized regression problem (Tibshirani, 2011)

$$\hat{\beta} = argmin \sum_{1}^{N} (y_i - \alpha - \sum_{j} \beta_j x_{ij})^2 + \lambda \sum_{j} |\beta_j|. \tag{8}$$

2.2.3 Random forest regression

Random forest was firstly introduced by Breiman (2001). As a ensemble learning method, it adds an additional layer of randomness to bagging (Liaw et al., 2002) and corrects for decision tree's habit of overfitting the training set.

This paper tries traditional OLS regression, ridge regression, lasso regression, and random forest with K-fold cross validation (K=4) and compares their mean square errors. The result shows that random forest regression is the best model, so the maximum predicted earnings and location cost are both computed based on the random forest regression.

3 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 in China. 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.

4 Results

4.1 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.

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

[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.

4.2 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 every 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 can provide a general concept about regional inequality in payoffs to human capital.

4.3 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);

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

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.

Table 3: Best parameters selected

model	best parameters
lasso regression	$\lambda = 6.1514$
ridge regression	$\lambda = 29.3636$
random forest	max depth=3, max features = 3, min samples leaf = 13, min samples split = 10, n estimators = 200

[1] Each parameter is selected by k-fold cross validation method (k=4).

[2] Once best parameter is selected, MSE is computed by the model with best parameters and k-fold cross validation is not used in this step.

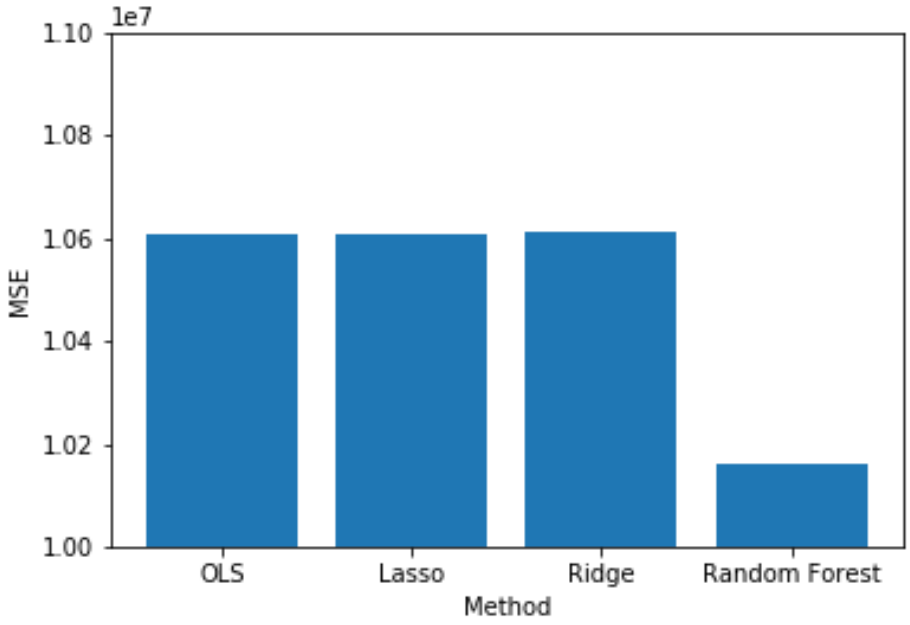

Figure 1: Comparison of MSE

4.4 Location cost and HSR

Figure 2 demonstrates average 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 yield more production in these regions.

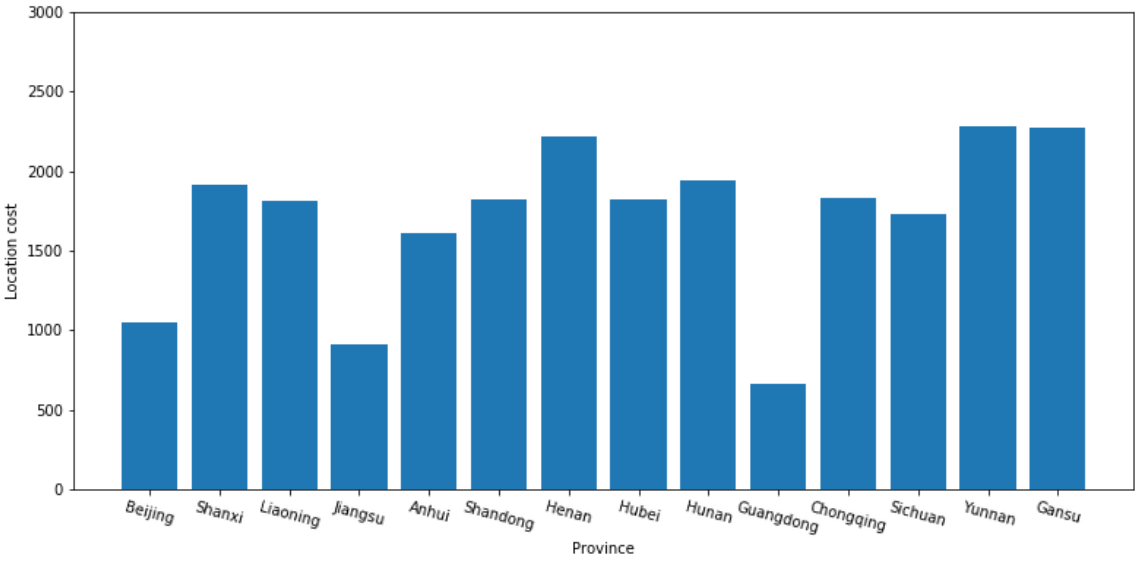

Figure 2: Location cost by province

Table 4 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 the subsample in 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 earnings or they already gain maximum predicted earnings, the location cost decreases by about 617 yuan. For those who have not earned maximum

predicted earning, the existence of a HSR line from resident province to where they earn maximum earnings decreases the location cost by 166.3 yuan.

Table 4: 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] Controlling for province-level fixed effect.

[2] Dependent variable: location cost.

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

Figure 3 is the kernel density estimation of location cost for all the sample. For those with transportation = 1, location cost clusters in low values, while for the cohort with transportation = 0, the mean of location cost shifts to the right. Figure 4 shows the kernel density estimation for the subsample whose location cost $\neq 0$ (i.e those who need to migrate to realize maximum predicted earnings). The distribution of location cost for the cohort without supporting HSR lines (i.e. transportation = 0) shifts to the right compared with the cohort with HSR lines. The kernel density estimation further suggests that location cost is lower with access to supporting HSR lines.

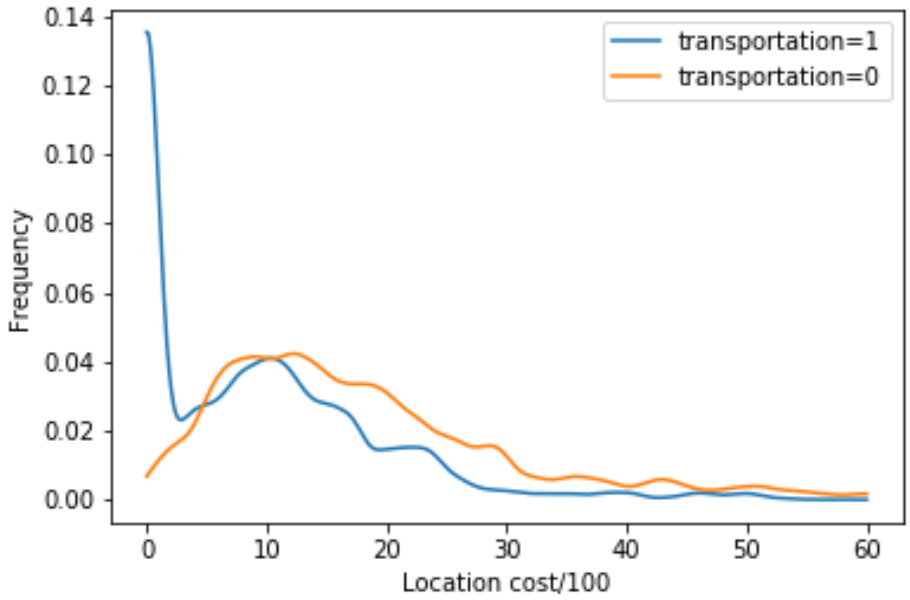

Figure 3: Kernel density estimation of location cost by transportation (all samples)

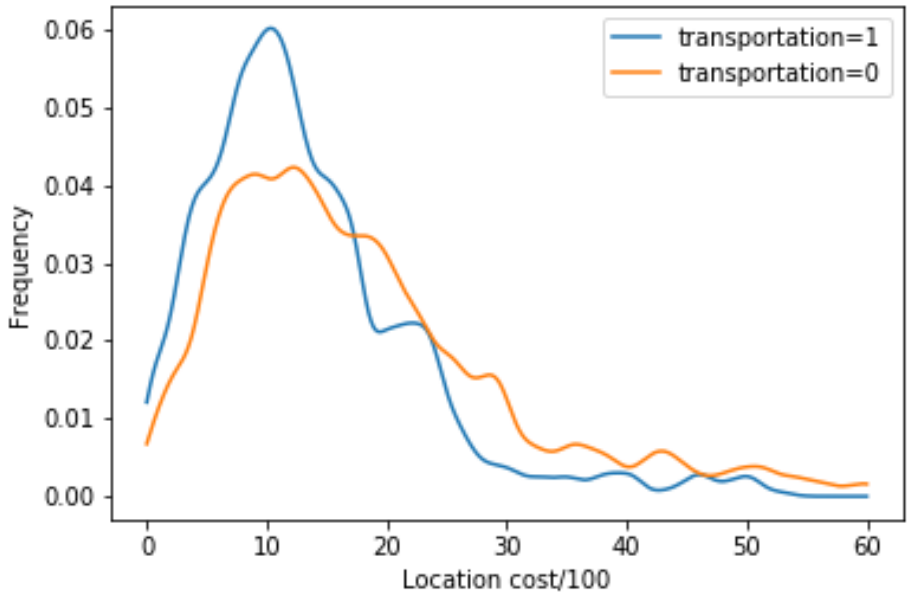

Figure 4: Kernel density estimation of location cost by transportation (subsample: location cost $\neq 0$)

Note that the fixed effect regression result and kernel density estimation both verify the correlation relationship rather than the causal effects between HSR and location

cost, for many potential factors may simultaneously influence location cost and the construction of HSR. 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) The correlation relationship is the result of "self selection". People who suffer from high location cost choose to move to regions that can reward them with higher payoff if the movement is convenient with appropriate HSR lines. For those who can directly travel to the province of maximum predicted earnings 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.

5 Conclusion

The paper develops a more accurate prediction of location cost than Zax (2019) by using computational methods. With predicted location cost and HSR data, the paper finds that if person cannot earn his maximum predicted earnign in his home province and there is a direct HSR line which links his home province and the province of maximum predicted earning, then the location cost is about 166.30 lower in average. The distribution of location cost is also shifted to the left for this cohort compared with the cohort without supporting HSR lines.

This result may be a consequence of self-selection. People have right to choose whether to move to reduce their location cost if they have access to a supporting HSR line. People do not have this right to choose if there is no such a HSR line. The consequence is that the location cost is generally lower for those who can choose. They reject to move to eliminate location cost because location cost is not high enough. Further investigation can be conducted to find how HSR affects people's choice of migration and the caused change in location cost.

References

Banerjee, A., Duflo, E. and Qian, N. (2012), On the road: Access to transportation infrastructure and economic growth in china, Technical report, National Bureau of Economic Research.

Breiman, L. (2001), 'Random forests', *Machine learning* **45**(1), 5–32.

Chan, K. W. (2009), 'The chinese hukou system at 50', *Eurasian geography and economics* **50**(2), 197–221.

Chen, J. and Fleisher, B. M. (1996), 'Regional income inequality and economic growth in china', *Journal of comparative economics* **22**(2), 141–164.

Chen, Z. and Haynes, K. E. (2017), 'Impact of high-speed rail on regional economic disparity in china', *Journal of Transport Geography* **65**, 80–91.

Démurger, S., Gurgand, M., Li, S. and Yue, X. (2009), 'Migrants as second-class workers in urban china? a decomposition analysis', *Journal of Comparative Economics* **37**(4), 610–628.

Evans, P. (1998), 'Using panel data to evaluate growth theories', *International Economic Review* pp. 295–306.

Fleisher, B., Li, H. and Zhao, M. Q. (2010), 'Human capital, economic growth, and regional inequality in china', *Journal of development economics* **92**(2), 215–231.

Fleisher, B. M., Chen, J. et al. (1997), 'The coast-noncoast income gap, productivity, and regional economic policy in china', *Journal of comparative economics* **25**(2), 220–236.

Li, H. (2003), 'Economic transition and returns to education in china', *Economics of education review* **22**(3), 317–328.

Liaw, A., Wiener, M. et al. (2002), 'Classification and regression by randomforest', *R news* **2**(3), 18–22.

Liu, Z. (1998), 'Earnings, education, and economic reforms in urban china', *Economic development and cultural change* **46**(4), 697–725.

Park, A. and Wang, D. (2010), 'Migration and urban poverty and inequality in china', *China Economic Journal* **3**(1), 49–67.

Pedroni, P. and Yao, J. Y. (2006), 'Regional income divergence in china', *Journal of Asian Economics* **17**(2), 294–315.

Tian, X., Zhang, X., Zhou, Y. and Yu, X. (2016), 'Regional income inequality in china revisited: A perspective from club convergence', *Economic Modelling* **56**, 50–58.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society: Series B (Methodological)* **58**(1), 267–288.

Tibshirani, R. (2011), 'Regression shrinkage and selection via the lasso: a retrospective', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**(3), 273–282.

Whalley, J. and Xing, C. (2014), 'The regional distribution of skill premia in urban china: Implications for growth and inequality', *International Labour Review* **153**(3), 395–419.

Xing, C. (2014), 'Migration, self-selection and income distributions: Evidence from rural and urban china', *Economics of Transition* **22**(3), 539–576.

Yao, S. and Zhang, Z. (2001), 'On regional inequality and diverging clubs: a case study of contemporary china', *Journal of Comparative Economics* **29**(3), 466–484.

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

Zhang, H., Zhang, H. and Zhang, J. (2015), 'Demographic age structure and economic development: Evidence from chinese provinces', *Journal of Comparative Economics* **43**(1), 170–185.