High-speed Rail and Inter-provincial Inequality in Payoffs to Human Capital

Yiqing Zheng

Master in Computational Social Science, Social Science Division, the University of Chicago

Research Question

- How does high-speed rail development influence regional inequality in payoffs to human capital?
- How to accurately measure regional inequality in payoffs to human capital?

Introduction

- A large wave of migration in China

↓ indicates

- People earn different levels of earning in different provinces

↓ i.e.

- Regional inequality in payoffs to human capital.

↑ How to affect?

- More convenient to migrate

↑ causes

- Fast development of high-speed rail (HSR) in China

Data

- Income data: the China Household Income Project (CHIP) [2013]
- HSR data: Chinese Research Data Services Platform [2002 – 2013]

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.

Empirical Models

- **First Part:** Counterfactual strategy
- Step 1: run the following regression within each province. Finally get 14 regression results. i indexes individual and j indexes the province of residence for i.

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

- Step 2: predict individual i's earning inside his province of residence j and any other province k as if i lives in k. Get 14 predicted earnings for each i.

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

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

- Step 3: find the province of maximum predicted earning m_i and get location cost for i.

$$\hat{y}_{jm_i i} = \max_{k \in K} \hat{y}_{jki} \ , \ K = the\ set\ of\ all\ provinces. \quad Location\ cost_i = \hat{y}_{jm_i i} - \hat{y}_{jji}$$

- **Computation methods:** compare MSE of different methods in the first step with k-fold cross validation, including OLS, Lasso regression, Ridge regression, and random forest.
- **Second Part:** generate a dummy "transportation" to indicate whether there is a direct HSR line linking province j with m_i. Find the relationship between location cost and transportation.

Results

First part result

- Random forest is the best model.

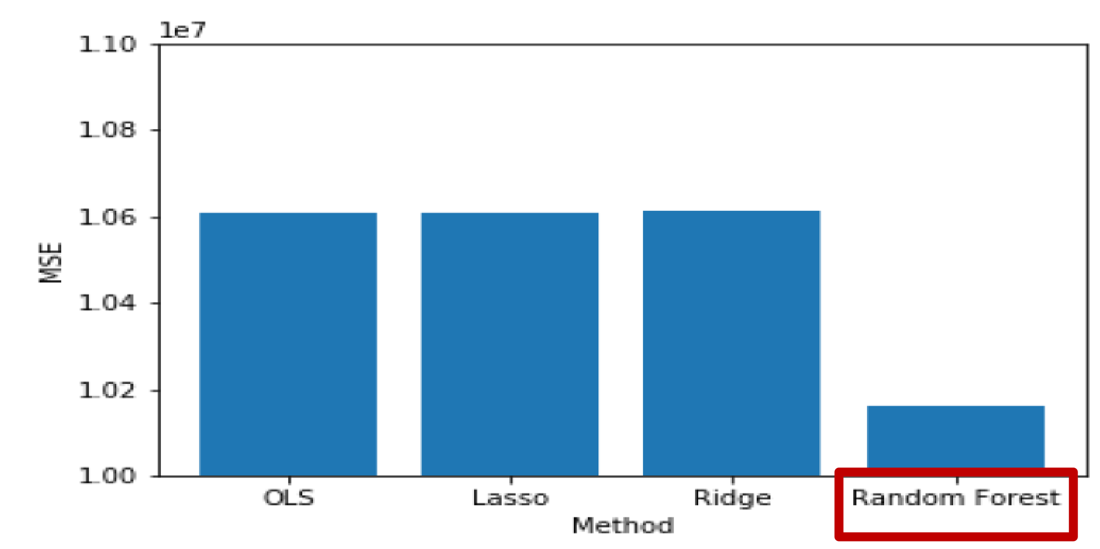

Figure 1: Comparison of MSE

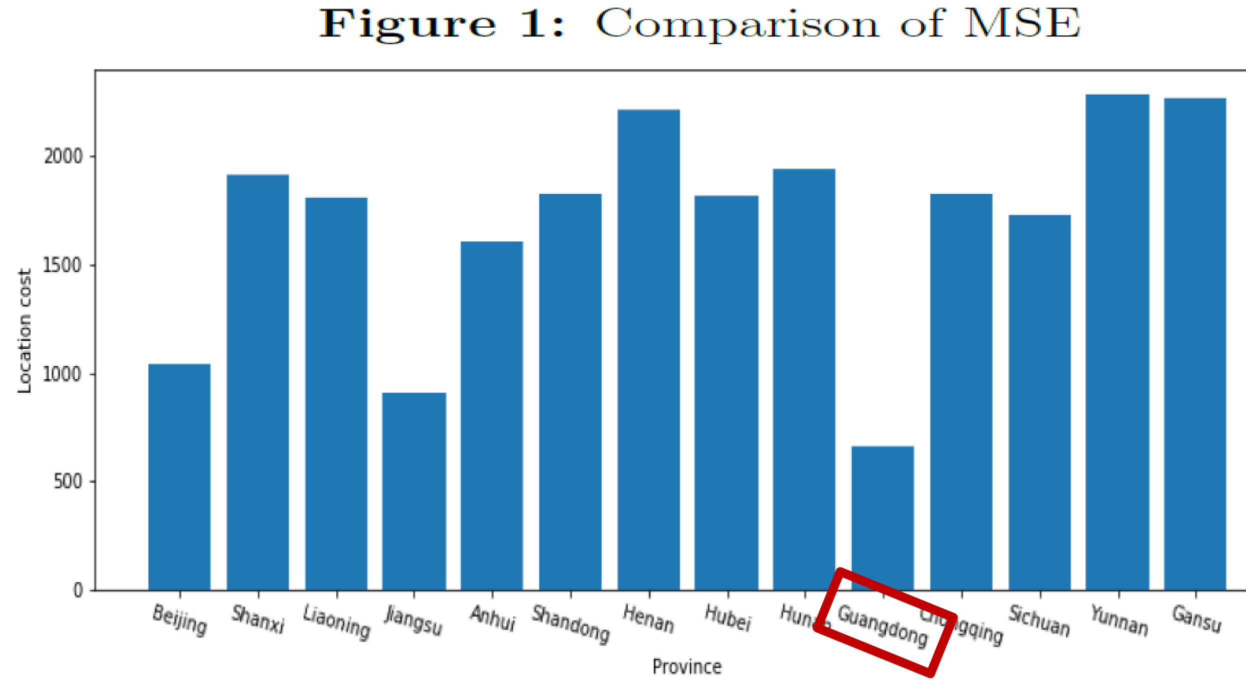

Figure 2: Location cost by province

- Payoff to human capital is highest in Guangdong, so its location cost is the lowest.

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

Second part result

- Transportation reduces location cost.

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

- The distribution of location cost shifts to the right without direct HSR line.

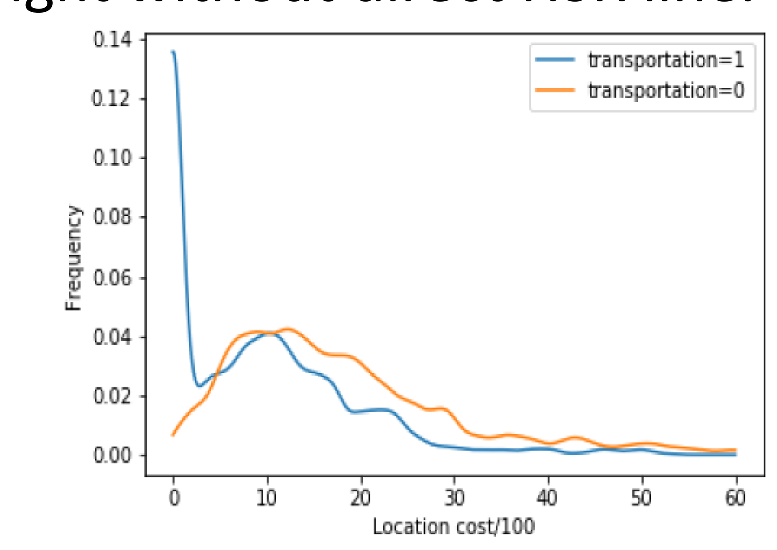

Figure 3: Kernel density estimation of location cost by transportation

Conclusions

- Supporting HSR line reduces location cost.
- It might be the result of "self selection": with supporting HSR, people still choose to stay because the location cost is not high enough to drive them to move.

Limitation

- Need to deal with endogeneity problem if evaluating causal effects.

Main Reference

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.

Acknowledgements

Thank Dr. Evans for his patient guidance and helpful suggestions. Thank all my classmates in MACS 30250.

3rd, June, 2020

Contact: yiqingzheng@uchicago.edu