ML: Hukou System and Health Outcomes

Marta Bengoa and Thierry Warin [Colin Powell School. City University of New York. University of Johannesburg and CIRANO; SKEMA Business School and CIRANO]

AEA 2019 - Atlanta

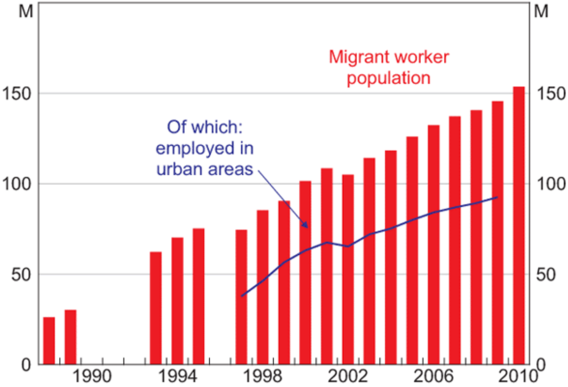
1. Introduction
2. Literature Review
3. Research Questions and Motivation
4. Econometric Model
5. Machine Learning Techniques: Regression trees, random forests, bagging and boosting
6. Conclusions
7. Appendices

Introduction

Introduction

- China's rapid development have spurred large migration from rural areas to urban areas
- Between 1990 and the end of 2015 the proportion of China's population living in urban areas jumped from 26% to 56%
- Currently estimated by census, there are more than 240 million rural-to-urban migrants and more than 160 million working in cities outside of their hukou. That accounts for approx. 30% of total rural labor force (China National Bureau of Statistics).
- The Hukou household registration system imposes restrictions and limits to where to live, which is determined mainly by birth.
- Hukou card is an internal passport that sets access to education and health services. It started in 1956-58, relaxed during the 60s and enforced again since 1978.

Introduction



* The red colums refer to all rural migrants without local 'hukou' where they are living; the blue line by Herd, Koen and Reuterswald (2010) calculates the level of rural–urban migrant employment, RBA estimate for 2009 number following their methodology

Sources: Chan (forthcoming) for 1988, 1989, 1993–1995 data; MOLSS (2002) and RBA for 1997–1999 data; Herd *et al* (2010) for all 2000–2008 data; NBS (2010) and RBA for 2009 data; MOLSS (2011) for 2010 data

Figure 1: China - Number of rural migrant workers

Introduction

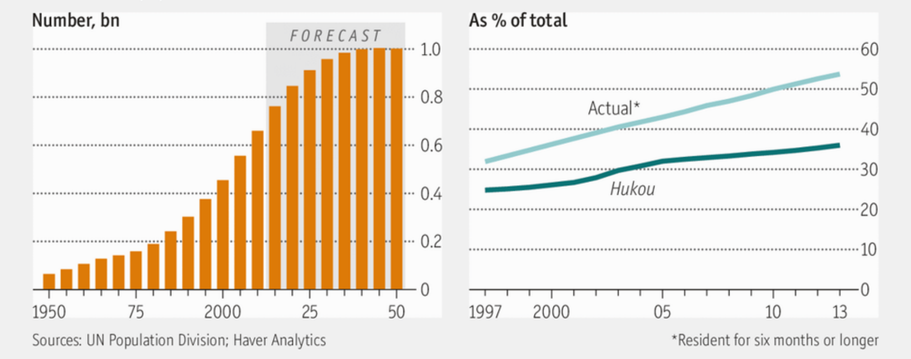


Figure 2: China's population

Introduction

- In 2009, China designed a healthcare reform that was intended to address inequalities in health access. Specific goals include:
  1. Provide wide basic health coverage for legal urban workers (mandatory) and to allow voluntary enrolment for urban residents without jobs (including students, elderly or disabled);
  2. Voluntary cooperative medical system for rural citizens; and
  3. Provide wide healthcare coverage to vulnerable and low-income population groups.
- The reform left the rural-to-urban migrant population outside of the provision of public healthcare.
- Several major cities have tried to introduce policy reforms to address the health challenges confronted by this phenomenon.
- Since 2006, Beijing is offering primary care health services and free healthcare to the children of migrant workers in community centres. There is also a pilot programs

Literature Review

Literature Review

- There are not many studies that have addressed this link between migration with limited access to healthcare and health outcomes in developing economies. The most recent for China published by Sun (2015), uses self-reported outcomes (do I feel well or not, have I been sick..) which suffers from measurement errors.
- Other studies suggest that migrants are reasonably healthy at the point of migration but more likely to experience adverse effects than non-migrants. As they get injured and can't have access to health some return home while others remain in urban areas.
- Therefore, increases risk of workplace accidents, other contagious diseases (Chen, 2011; Lu and Quin, 2014; Wallace and Kulu, 2014).

Literature Review

- Studies mainly focus on the US, on the relationship of health of Mexican immigrants and lack of insurance. Goldman et al. (2014, Demography) find significant and strong correlation between non having insurance and decline in health of Mexican immigrants (compared to never migrants), specially immediately after migration.
- Wassink (2008, Demographic Research) analyses health insurance coverage for returned Mexican migrants to determine that lack of access is mainly driven by unemployment, revealing a negative association between lack of insurance and health

Research Questions and Motivation

Research Questions and Motivation

- Motivation: Using Machine Learning techniques (ensemble methods) to study a very complex relationship in Social Sciences: health and migration
- Steps:
  - First, we use traditional econometric techniques for such a topic
  - Sectond, we use ML techniques to move from **correlations** (econometrics) to **predictive modelling** (ML)

Research Questions and Motivation

RQ1: This paper examines whether hukou has a predictive effect on migrants' self-reported health. We assess if there are observable differences in health outcomes migrant workers by hukou status. We use regression analysis, propensity score models and machine learning.

- We first estimate an ordered logit model based on self-reported health status to analyse if there exists a significant differential impact on health outcomes between rural-to-urban migrants.
- Secondly, we compared these results with those based on objectively measured health indicators:
    1. We try to assess part of the measurement error an check if there any differential effect when using objectively-measured health outcomes?
    2. Are health outcomes conditioned on years passed since migration?

Research Questions and Motivation

RQ2: We then use Machine Learning techniques to question whether hukou status plays a role in the health outcome of migrant workers.

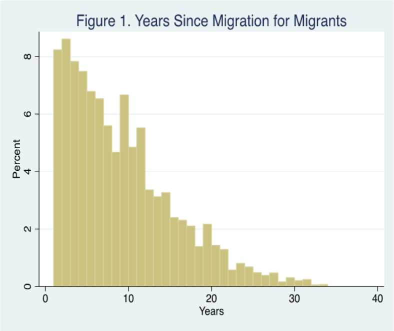
Econometric Model

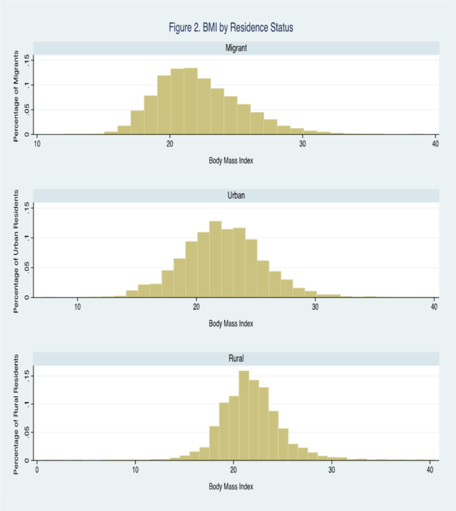
Econometric Model

Dataset 1 for RQ1:

- We use survey data reported in the Longitudinal Survey on Rural Urban Migration in China from the Institute for the Study of Labor (IZA). The survey collects data for 71,074 individuals (29,556 urban persons; 32,171 rural persons; and 9,347 migrants. Aprox 29% of rural persons) in two waves for the years 2008 and 2009.
- The survey contains data on socioeconomic indicators, such as education, income, ethnicity, and hukou registration.
- The RUMiC survey also includes data on many health indicators and outcomes. These include weight (kilograms), height (centimeters), dominant handedness, blood pressure, and grip strength.

Econometric Model





Migrants BMI distribution is skewed to the left, 31% of migrants have a BMI below 18.0, which categorizes them as being underweight and subject to higher health risks

Figure 3

Econometric Model

The baseline specification is as follows with coefficients estimated by maximum likelihood (Williams, 2006 and Kleinbaum and Klein, 2010):

$$\begin{aligned} Pr(Y_j = i) = & \\ & Pr(\tau_{i-1} < \beta_1 MHukou_j \\ & + \beta_2 MHukou * YearsinceM_j \\ & + X_j \phi + Z_j \phi \\ & + \epsilon_j \leqslant \tau_i) \end{aligned} \tag{1}$$

- Where Y_j states the self-reported health status for all j individuals living in urban areas, migrants and non-migrants.

Econometric Model

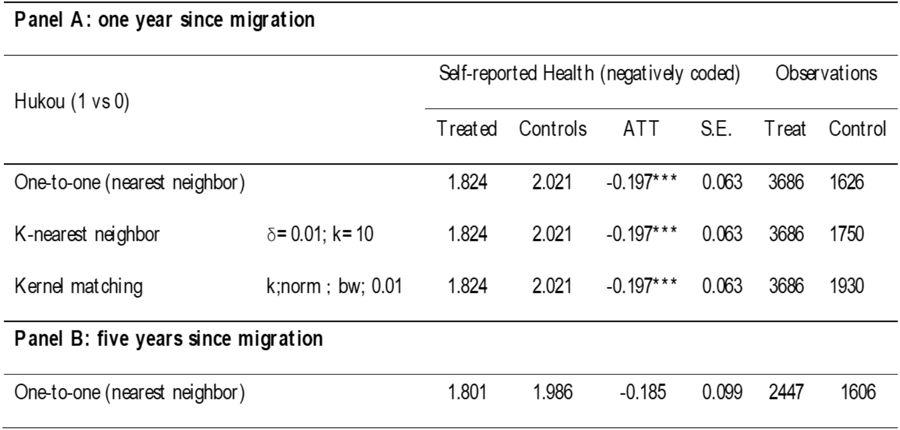
$$\begin{aligned} BloodPressure = \quad & \\ & \beta_1 MHukou_i \\ & + \beta_2 Mhukou * YrsSinceM_i \\ & + K_i \psi + Z_i \phi \\ & + \epsilon_i \end{aligned} \tag{2}$$

- Equation (2)s include a vector K of health indicators such as BMI, or being a smoker.

Econometric Model

- Health indicators could also be impacted by the length of time since a person migrated. of time since a person migrated. The direction of the relationship is unknown, but the longer the individual remains in the city, they can create more networks and gather more information about primary care clinics to access healthcare.
- We add health risk factor variables as well as socioeconomic variables to control for migration bias.

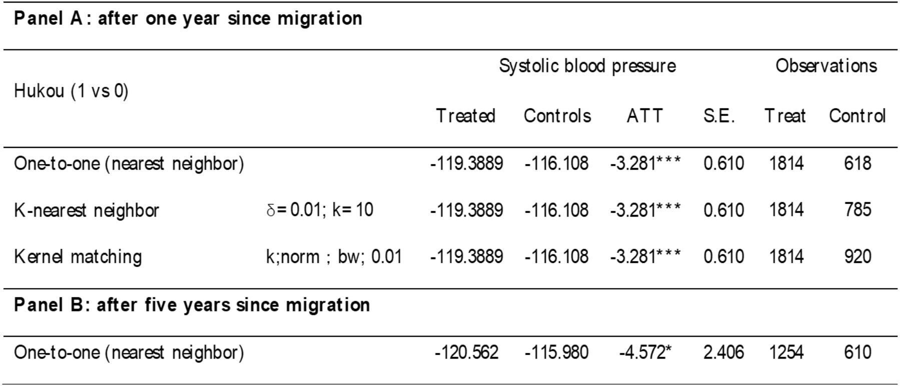
Econometric Model



Panel A: one year since migration							
Hukou (1 vs 0)		Self-reported Health (negatively coded)				Observations	
		Treated	Controls	ATT	S.E.	Treat	Control
One-to-one (nearest neighbor)		1.824	2.021	-0.197***	0.063	3686	1626
K-nearest neighbor	δ= 0.01; k= 10	1.824	2.021	-0.197***	0.063	3686	1750
Kernel matching	k;norm ; bw; 0.01	1.824	2.021	-0.197***	0.063	3686	1930
Panel B: five years since migration							
One-to-one (nearest neighbor)		1.801	1.986	-0.185	0.099	2447	1606

Figure 4: Propensity score model: estimated effects of hukou on self-reported health for rural-to-uban migrants

Econometric Model



Panel A: after one year since migration							
Hukou (1 vs 0)		Systolic blood pressure				Observations	
		Treated	Controls	ATT	S.E.	Treat	Control
One-to-one (nearest neighbor)		-119.3889	-116.108	-3.281***	0.610	1814	618
K-nearest neighbor	δ= 0.01; k= 10	-119.3889	-116.108	-3.281***	0.610	1814	785
Kernel matching	k;norm ; bw; 0.01	-119.3889	-116.108	-3.281***	0.610	1814	920
Panel B: after five years since migration							
One-to-one (nearest neighbor)		-120.562	-115.980	-4.572*	2.406	1254	610

Figure 5: Propensity score model: estimated effects of hukou on health outcomes -systolic blood pressure- for rural-to-urban migrants

Preliminary Results

- Our preliminary results state that holding and urban hukou increases the chance for self-evaluated good health compared to dwellers with rural hukou in urban areas, although is only significant at the margin. Problems: mismeasurement, self-selection of migrants, do not observe the health of migrants before migrating.
- When using objective health measures the effect increases in magnitude and significance, but tends to disappear as migrants remain in the urban cities, suggesting a network effect of informal access to health providers, increase in incomes and access to private health.
- Migration will require adjustments in health provisions to accommodate the changing spatial demographics. Restricting migrants access to healthcare will clearly have an effect in the long run, including on migrant's health, productivity, and potential economic growth.

Machine Learning Techniques: Regression trees, random forests, bagging and boosting

Data

Dataset 2 for RQ2:

- 27 variables
- 10,478 observations

Data

Dependent Variable	Independent Variables
daysmissedforsick	migrants, city, age, hukou, hukoudate, reasonforhukouchange, healthrating, commercialinsurance, publicinsurance, ruralinsurance, comprehensiveinsurance, womenandchildinsurance, immunizationinsurance, otherinsurance, yearsofeduc, maledummy, systolicavg, diastolicavg, rightgrip, leftgrip, yrssincemigration, monthlyincome, smokerdummy, marrydummy, BMI, workinsurance, insuranceforfamilies

Fitting Regression Trees

- Split data 50-50 into training, test sets.
- First, we fit the regression tree model on the training data only and plot the tree.

Fitting Regression Trees

Regression Tree:	daysmissedforsick
Selected variables:	healthrating (**qualitative**), diastolicavg, BMI, age, maledummy
Number of terminal nodes:	6
Residual mean deviance:	16.79

Fitting Regression Trees

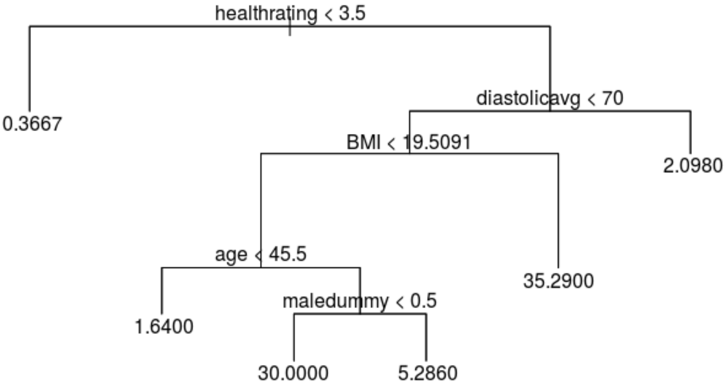


Figure 6

Fitting Regression Trees

Regression Tree:	daysmissedforsick
Selected variables:	age, monthlyincome
Number of terminal nodes:	3
Residual mean deviance:	19.79

Fitting Regression Trees

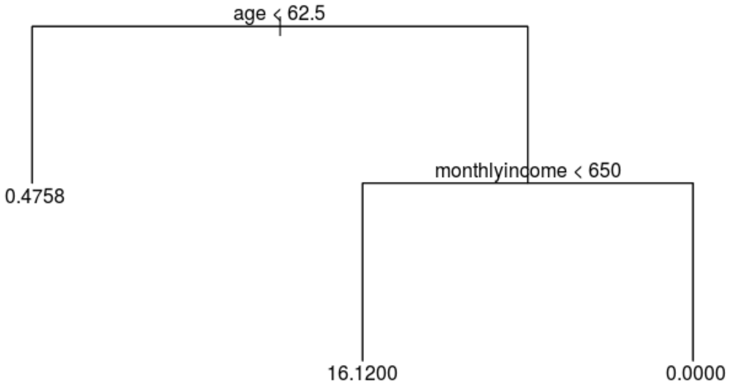


Figure 7

Random Forests and Bagging

- Recalling that bagging is a special case of a random forest.
- Bagging first:

Regression Tree:	daysmissedforsick
Type of random forest:	regression
Number of trees:	500
No. of var tried at each split:	3

Random Forests and Bagging

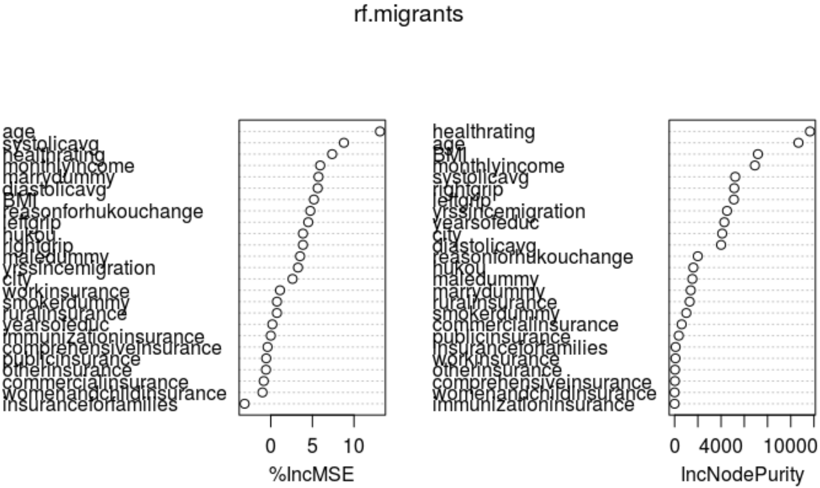


Figure 8

Random Forests and Bagging

- The results indicate that across all trees considered in the random forest, **age**, **healthrating** and **monthly income** are by far the three most important variables.
- Now we try yet another method: Boosting

Boosting

Predictor	Rel. Inf.
BMI	15.393776962
age	13.242068594
healthrating	13.053958866
monthlyincome	12.864551956
city	7.901481885
systolicavg	6.724572933
yearsofeduc	6.492462610
yrssincemigration	6.027275970
diastolicavg	4.280571577
rightgrip	3.902531257
reasonforhukouchange	2.786321958

Boosting

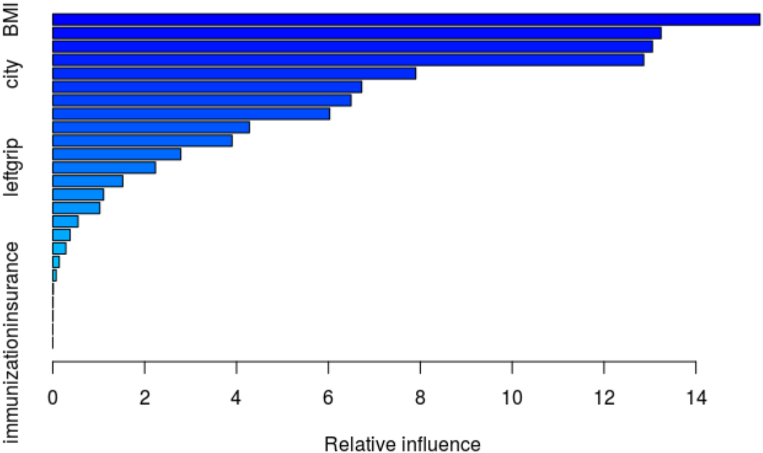


Figure 9

Boosting

- Predictive modelling: 41.50

Conclusions

Conclusions

- China's Hukou reform is a move in the right direction. Pilot programs in 29 provinces are helping to raise awareness about the necessity to eliminate barriers to health access which are now linked to geography. In 2015, Shanghai had approximately 9.8 million migrant workers holding a rural hukou.
  1. Allow insurance transferability by designing a method to break down some of the institutional barriers in obtaining care;
  2. Design a system of government transfers across provinces. Therefore, those provinces that support more health related expenditures receiving appropriate funding;
  3. Migrants' contributions to social security should be linked to place of residence and be transferrable;
  4. Incentivize general practitioners to provide treatment in local communities where migrants live and build a strong network of primary health providers and local clinics. This last policy has the potential to enhance trust by engaging with the local communities, raising awareness on health risk, and fostering preventive care.

Conclusions

- ML techniques are SO! interesting techniques for **predictive modelling**

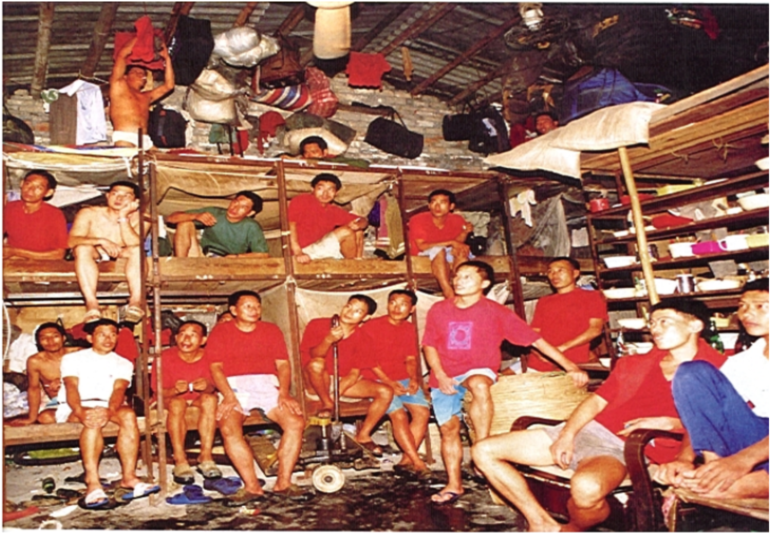
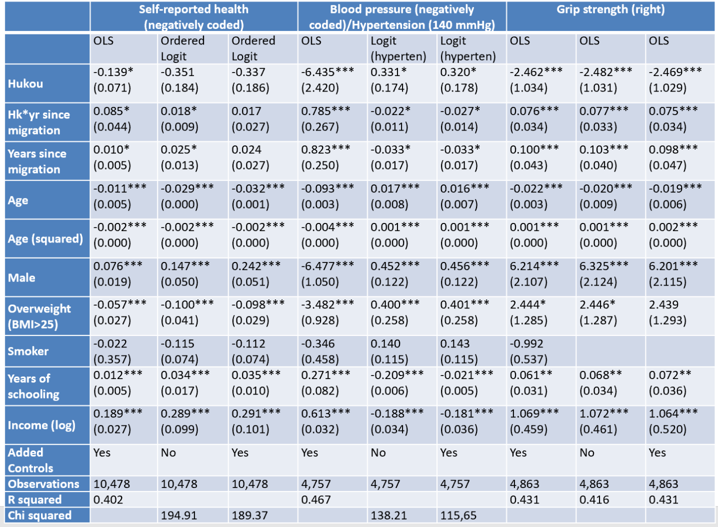


Figure 10

Appendices

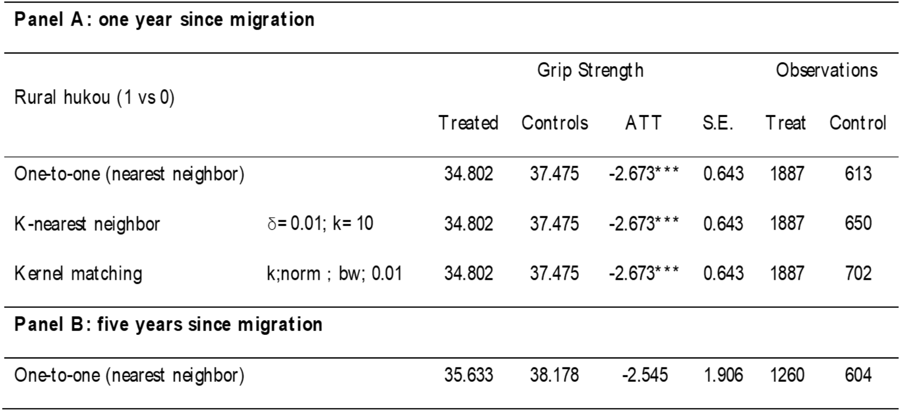
Econometric Model



	Self-reported health (negatively coded)			Blood pressure (negatively coded)/Hypertension (140 mmHg)			Grip strength (right)		
	OLS	Ordered Logit	Ordered Logit	OLS	Logit (hyperten)	Logit (hyperten)	OLS	OLS	OLS
Hukou	-0.139* (0.071)	-0.351 (0.184)	-0.337 (0.186)	-6.435*** (2.420)	0.331* (0.174)	0.320* (0.178)	-2.462*** (1.034)	-2.482*** (1.031)	-2.469*** (1.029)
Hk*yr since migration	0.085* (0.044)	0.018* (0.009)	0.017 (0.027)	0.785*** (0.267)	-0.022* (0.011)	-0.027* (0.014)	0.076*** (0.034)	0.077*** (0.033)	0.075*** (0.034)
Years since migration	0.010* (0.005)	0.025* (0.013)	0.024 (0.027)	0.823*** (0.250)	-0.033* (0.017)	-0.033* (0.017)	0.100*** (0.043)	0.103*** (0.040)	0.098*** (0.047)
Age	-0.011*** (0.005)	-0.029*** (0.000)	-0.032*** (0.001)	-0.093*** (0.003)	0.017*** (0.008)	0.016*** (0.007)	-0.022*** (0.003)	-0.020*** (0.009)	-0.019*** (0.006)
Age (squared)	-0.002*** (0.000)	-0.002*** (0.000)	-0.002*** (0.000)	-0.004*** (0.000)	0.001*** (0.000)	0.001*** (0.000)	0.001*** (0.000)	0.001*** (0.000)	0.002*** (0.000)
Male	0.076*** (0.019)	0.147*** (0.050)	0.242*** (0.051)	-6.477*** (1.050)	0.452*** (0.122)	0.456*** (0.122)	6.214*** (2.107)	6.325*** (2.124)	6.201*** (2.115)
Overweight (BMI>25)	-0.057*** (0.027)	-0.100*** (0.041)	-0.098*** (0.029)	-3.482*** (0.928)	0.400*** (0.258)	0.401*** (0.258)	2.444* (1.285)	2.446* (1.287)	2.439 (1.293)
Smoker	-0.022 (0.357)	-0.115 (0.074)	-0.112 (0.074)	-0.346 (0.458)	0.140 (0.115)	0.143 (0.115)	-0.992 (0.537)		
Years of schooling	0.012*** (0.005)	0.034*** (0.017)	0.035*** (0.010)	0.271*** (0.082)	-0.209*** (0.006)	-0.021*** (0.005)	0.061** (0.031)	0.068** (0.034)	0.072** (0.036)
Income (log)	0.189*** (0.027)	0.289*** (0.099)	0.291*** (0.101)	0.613*** (0.032)	-0.188*** (0.034)	-0.181*** (0.036)	1.069*** (0.459)	1.072*** (0.461)	1.064*** (0.520)
Added Controls	Yes	No	Yes	Yes	No	Yes	Yes	No	Yes
Observations	10,478	10,478	10,478	4,757	4,757	4,757	4,863	4,863	4,863
R squared	0.402			0.467			0.431	0.416	0.431
Chi squared		194.91	189.37		138.21	115,65			

Figure 11: Hukou status on health measures for rural-to-urban migrants

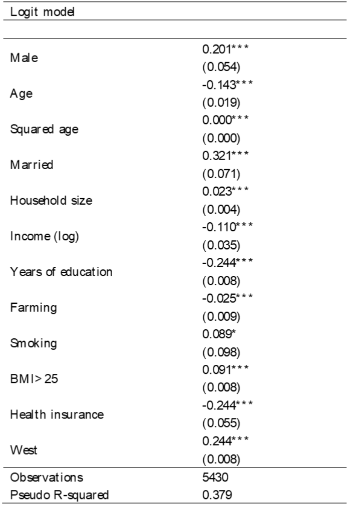
Econometric Model



Panel A: one year since migration							
Rural hukou (1 vs 0)		Grip Strength				Observations	
		Treated	Controls	ATT	S.E.	Treat	Control
One-to-one (nearest neighbor)		34.802	37.475	-2.673***	0.643	1887	613
K-nearest neighbor	δ= 0.01; k= 10	34.802	37.475	-2.673***	0.643	1887	650
Kernel matching	k;norm ; bw; 0.01	34.802	37.475	-2.673***	0.643	1887	702
Panel B: five years since migration							
One-to-one (nearest neighbor)		35.633	38.178	-2.545	1.906	1260	604

Figure 12: Propensity score model: estimated effects of hukou on health outcomes -grip strength- for rural-to-urban migrants

Econometric Model



Logit model	
Male	0.201***
	(0.054)
Age	-0.143***
	(0.019)
Squared age	0.000***
	(0.000)
Married	0.321***
	(0.071)
Household size	0.023***
	(0.004)
Income (log)	-0.110***
	(0.035)
Years of education	-0.244***
	(0.008)
Farming	-0.025***
	(0.009)
Smoking	0.089*
	(0.098)
BMI> 25	0.091***
	(0.008)
Health insurance	-0.244***
	(0.055)
West	0.244***
	(0.008)
Observations	5430
Pseudo R-squared	0.379

Figure 13: Determinants of being a rural-to-urban migrant with rural hukou in