

Perspectives on Computational Research

Project: Methods and Initial Results

Xingyun Wu

5/17/2017

Proposed Research Question

My research topic is the effect of college major on geographical mobility. The specific proposed research questions are: (1) Does college major influence individual's geographical mobility; (2) How (or to what extent) does college major influence individual's geographical mobility?

Hypotheses

Based on the proposed research question, I put forward the College Major Inference Hypotheses:

H₁: College major has impact on whether geographical mobility happens

H₂: College major has impact on distance of geographical mobility

To map the theoretical constructs to variables, I would define original location as the location of high school, and would define current location as working location. If original location is different from current location, then geographical mobility happens on individual. And Hypothesis 2 would be tested by measuring how different these two locations are.

Data

1. Data Source

To conduct analysis on the above hypotheses, this study needs data with information of college degree, occupation, and demographics. The data I would use is from the National Survey of College Graduates, which could be directly downloaded on its official website.

It is a longitudinal biennial survey, particularly focusing on the science and

engineering workforce. The respondents are individuals under the age of 76 by February 1 2015, with at least a bachelor's degree by January 1 2014, and living in the U.S. during the survey reference period. The survey has been conducted since the 1993, replacing the previous Survey of Natural and Social Scientists and Engineers (SSE) which began in 1972. This study would use the most recent collection, the 2015 NSCG.

The initial data was collected with a self-administered web survey. For the occurrence of nonrespondents, a self-administered mail survey would be held. And nonrespondents to the mail survey would receive computer-assisted telephone interviewing (CATI). With these efforts, the 2015 NSCG has a weighted response rate of 70%. According to data documentation on its official site, this dataset contains 135,000 sample cases. However, due to the nonresponses, its public version only contains 91,000 observations.

2. Descriptive Statistics

This study mainly uses 11 variables. 3 of them are used in original setting, 5 of them are adjusted, and the other three of them are constructed based on several other variables in the dataset.

| Table 1: Classification of Variables | | |
|--------------------------------------|-------------------------|-----------------------|
| Original Variables | Reclassified Variables | Constructed Variables |
| Age | Marital Status | Race |
| Gender | College Major | Whether Move |
| Annualized Salary | Location of High School | Distance of Mobility |
| | Locaton of Work | |
| | Location of Respondent | |

Note that the annualized salary uses imputation to deal with missing values. According to data documentation, the general range of nonresponse rate for key items is from 0.0% to 0.6%, but salary has a nonresponse rate around 11%. Since this assignment is to provide initial results and insights, I use unconditional mean imputation to simply adjust this problem. The non-missing values and mean of the annualized salaries are not revised.

The original variable of marital status includes married, living in a marriage-like relationship, widowed, separated, divorced, and never married. To simplify the situation, I classify the married observations and living in a marriage-like relationship

observations into one category with stable relationship. And I classified the others into another category without stable relationship.

Variables of college major and locations are also reclassified. The raw data of college major has more than 100 majors indicated. In this assignment, I categorized them into three categories: (1) engineering, computer science, math, and natural sciences; (2) social sciences; (3) others. And the raw data of locations contains many abroad countries. I converge all the abroad countries into one category: abroad.

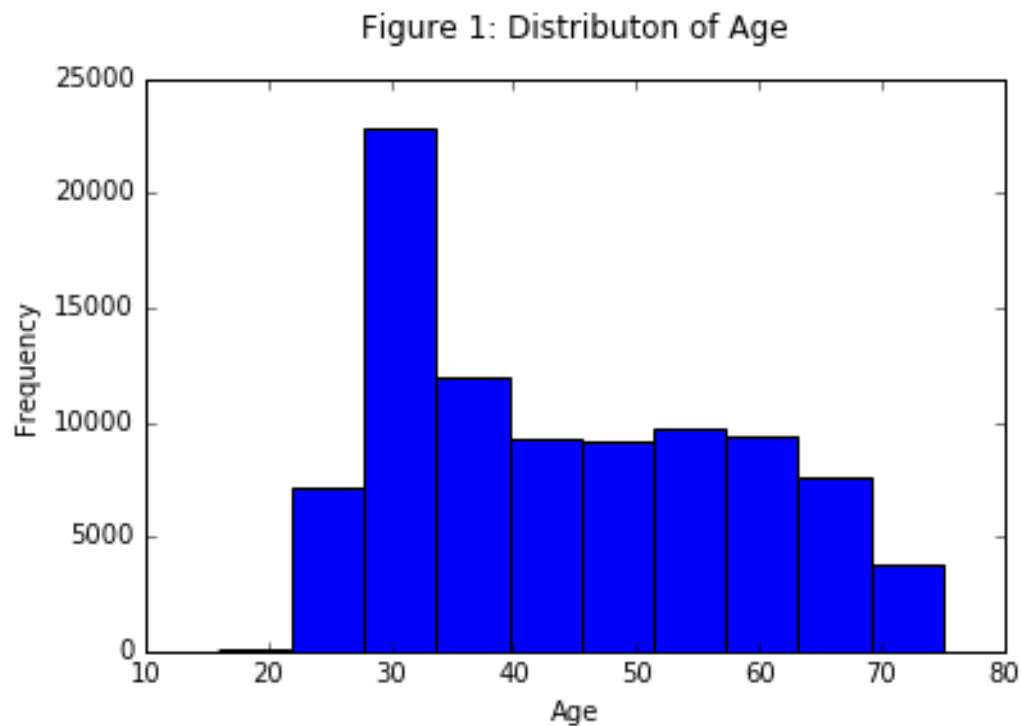
And I constructed three variables based on several variables in the raw data. The variable of race is constructed with variables indicating Asian, black and Hispanic in the raw data. The variable of whether individuals move is constructed on whether the observations report the same location of high school and work.

The variable of mobility distance is constructed on the basis of difference between the location of high school and the location of occupation. The distance is recorded as 1 when individuals move to their neighboring regions, and recorded as 2 when they move to regions next their neighboring regions. The rest is defined as the same rule, with the biggest internal mobility distance recorded as 5. If individuals move abroad, the distance is recorded as 8. Although the values of this variable are discrete, they represent the distance in order. So in data analysis, I would treat this variable as continuous variables.

| Table 2: Descriptive Statistics | | | | | |
|---------------------------------|---|--------------------|-----|---------|--------------|
| Continuous Variables | | | | | |
| Variable | Mean | Standard Deviation | Min | Max | Observations |
| Age | 44.26 | 14.16 | 16 | 75 | 91,000 |
| Annual Salary | 83727.09 | 83349.38 | 0 | 1223166 | 91,000 |
| Mobility Distance | 1.85 | 2.54 | 0 | 8 | 91,000 |
| Categorical Variables | | | | | |
| Variable | Categories | | | Percent | Observations |
| Gender | | | | | 91,000 |
| | Male | | | 53.18 | 48,396 |
| | Female | | | 46.82 | 42,604 |
| Marital Status | | | | | 91,000 |
| | Married or in a married-like relationship | | | 72.58 | 66,052 |
| | Others | | | 27.42 | 24,948 |
| Race | | | | | 91,000 |

| | | | |
|-------------------------|---|-------|--------|
| | None of the below | 64.87 | 59,028 |
| | Asian | 16.62 | 15,122 |
| | Black | 8.35 | 7,594 |
| | Hispanic | 10.17 | 9,256 |
| College Major | | | 91,000 |
| | Engineering, CS, Math, and Natural Sciences | 48.42 | 44,063 |
| | Social Sciences | 21.07 | 19,171 |
| | Others | 30.51 | 27,766 |
| If Moved | | | 91,000 |
| | No | 56.32 | 51,251 |
| | Yes | 43.68 | 39,749 |
| Location of High School | | | 91,000 |
| | New England | 4.96 | 4,516 |
| | Middle Atlantic | 14.91 | 13,571 |
| | East North Central | 15.74 | 14,324 |
| | West North Central | 7.39 | 6,722 |
| | South Atlantic | 11.76 | 10,706 |
| | East South Central | 3.41 | 3,106 |
| | West South Central | 6.72 | 6,116 |
| | Mountain | 4.48 | 4,075 |
| | Pacific & US Territories | 13.84 | 12,596 |
| | Abroad | 16.78 | 15,268 |
| Location of Employer | | | 76,814 |
| | New England | 6.1 | 4,684 |
| | Middle Atlantic | 0.09 | 71 |
| | East North Central | 14.51 | 11,144 |
| | West North Central | 14.99 | 11,518 |
| | South Atlantic | 7.27 | 5,588 |
| | East South Central | 18.24 | 14,014 |
| | West South Central | 3.54 | 2,717 |
| | Mountain | 8.95 | 6,873 |
| | Pacific & US Territories | 6.38 | 4,899 |
| | Abroad | 19.93 | 15,306 |
| Location of Respondent | | | 91,000 |
| | New England | 5.98 | 5,446 |
| | Middle Atlantic | 14.45 | 13,146 |
| | East North Central | 14.89 | 13,550 |
| | West North Central | 7.18 | 6,530 |
| | South Atlantic | 18.26 | 16,614 |
| | East South Central | 3.61 | 3,287 |
| | West South Central | 8.92 | 8,121 |
| | Mountain | 6.54 | 5,955 |
| | Pacific & US Territories | 20.08 | 18,275 |
| | Abroad | 0.08 | 76 |

Table 2 shows descriptive statistics of the variables. And the plots below show the distribution of some key variables and some visualized relationships.



According to Figure 1, the distribution of age in the sample is not normal. This is because the 2015 NSCG has an oversample of young graduates. The group around the age of 30 has higher observations than the other groups. Simply adjust the age distribution may cause more serious problems. By now, I have not found a suitable way to adjust this problem. So I would not apply variable transformation on age in data analysis.

Figure 2: Distributon of Annualized Salary

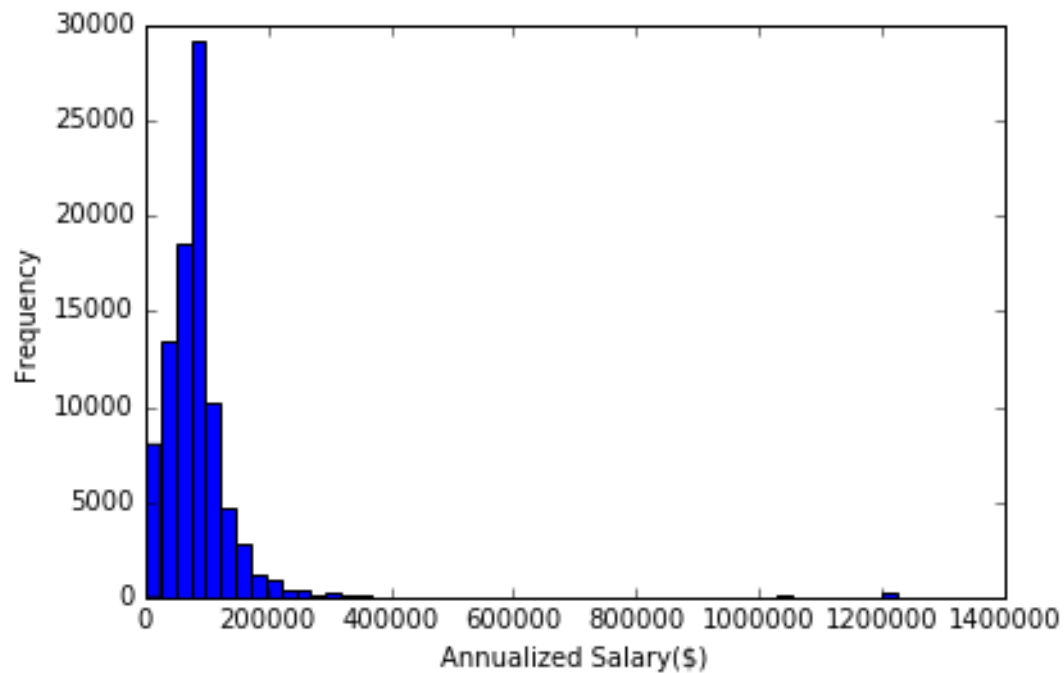


Figure 3: Distributon of Logged Annualized Salary

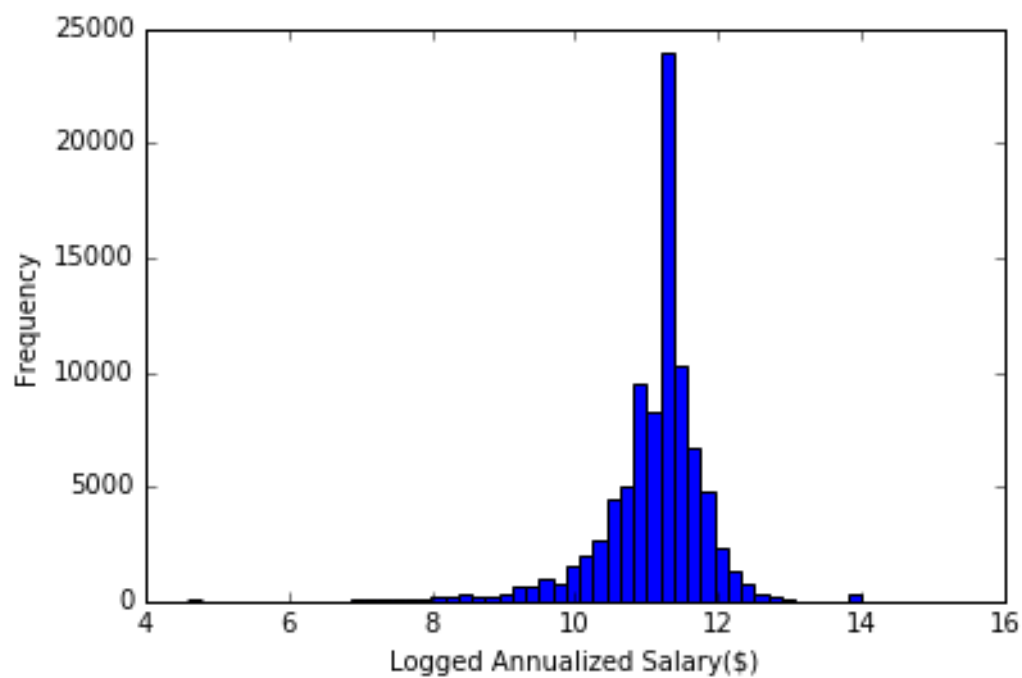


Figure 2 shows that annualized salary does not have a normal distribution. I take the logarithm of annualized salary to adjust this problem. And Figure 3 shows an approximately normal distribution of logged annualized salary. So I would use the logged annualized salary in data analysis.

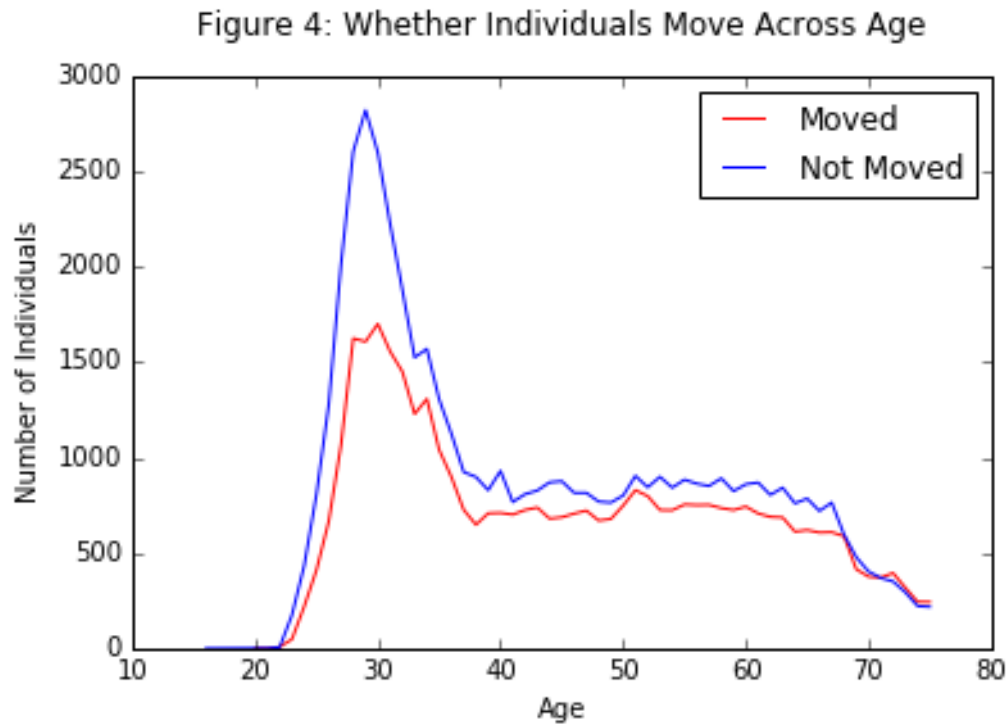


Figure 4 visualized the relationship between age and whether individuals move. Age distributions are generally similar for individual moved and not moved. Individuals with age around 30 have much higher frequency of mobility, which may due to the oversampling of the young graduates. Approximately after the age of 67, the frequencies of both moved observations and the not-moved observations drop, which may due to the age distribution of the sample. And almost across all ages, the not-moved observations are more than the moved observations.

Figure 5: Distance of Move Across Age

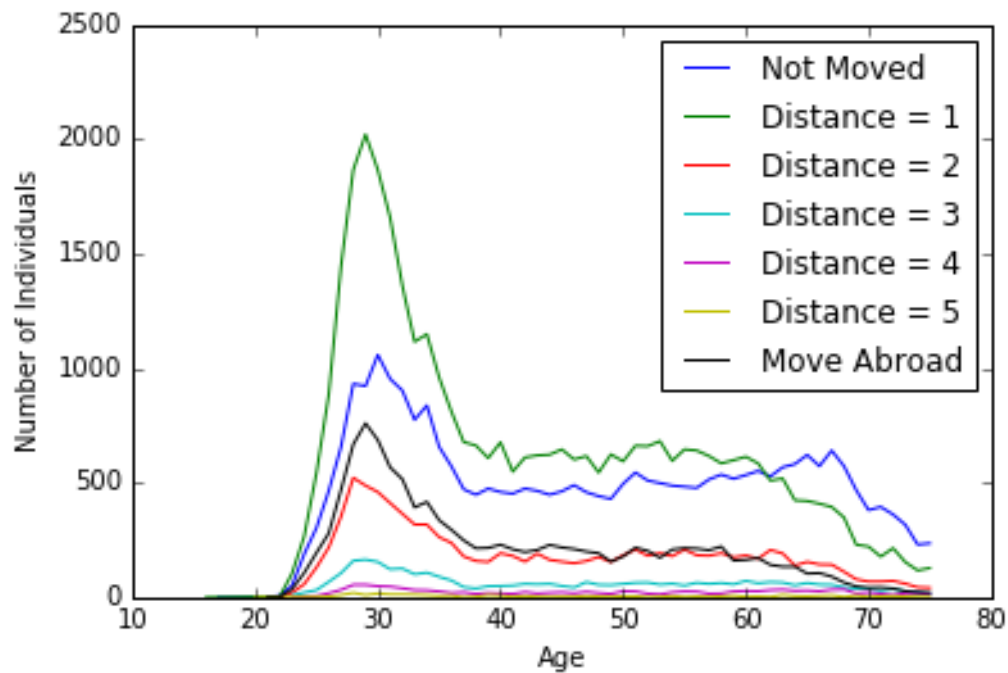


Figure 5 shows the distance of move across age. Similar to the patterns of Figure 4, there are more observations around the age of 30 than any other cohorts. For the mobility distance range from 2 to 5, the more the distance, the less the frequency. Many observations move to their neighboring regions.

Figure 6: Mobility by Major

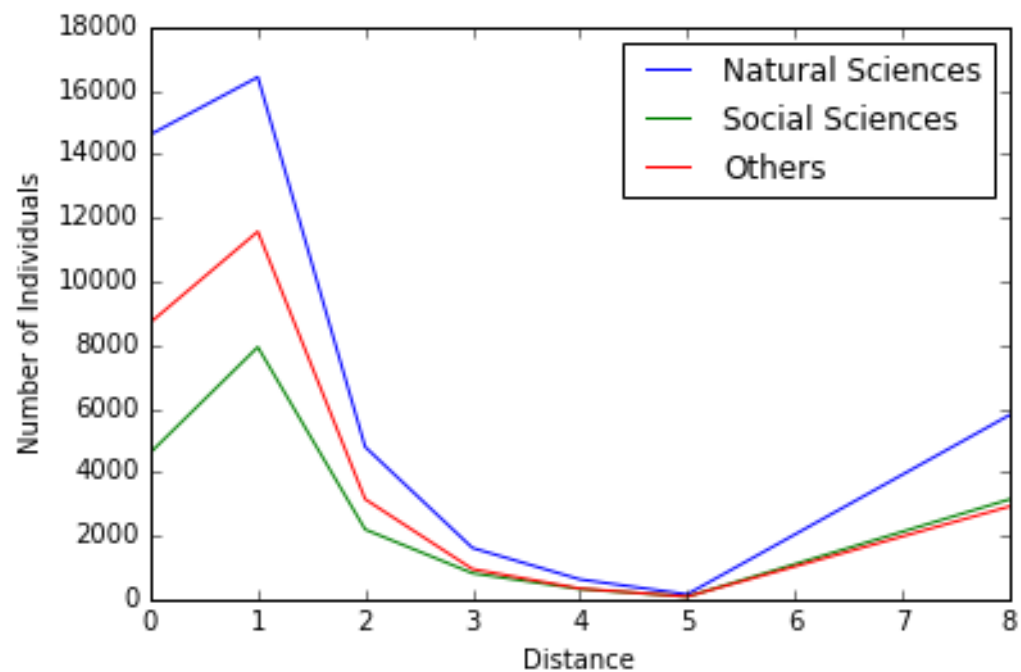


Figure 6 indicates that there is far more individuals majored in natural sciences

than in social sciences and other majors. And there are more individuals graduated with other majors than with the social science. However, for internal mobility within US, with the increase of mobility distance, the amounts of people move of each category are getting closer. And they converge when the distance is 5.

Method

For tentative data analysis, I would apply logistic regression to examine Hypothesis 1, and apply OLS regression to examine Hypothesis 2. These two techniques are very traditional and have long been used, so I would not discuss the usage of them.

I would use demographic variables (age, gender, and race and marital status) nested models to examine the effect of majors. I would first apply:

$$IS_MOVED_i =$$

$$\beta_0 + \beta_1 * AGE_i + \beta_2 * GENDER + \beta_3 * RACE_i + \beta_4 * MARRITAL_STATUS_i + \beta_5 * SALARY + \varepsilon_i \quad (1), \text{ and}$$

$$DISTANCE_i = \beta_0 + \beta_1 * AGE_i + \beta_2 * GENDER + \beta_3 * RACE_i + \beta_4 * MARRITAL_STATUS_i + \beta_5 * SALARY + \varepsilon_i \quad (2).$$

Then add the variable of college majors to examine its effect:

$$IS_MOVED_i =$$

$$\beta_0 + \beta_1 * AGE_i + \beta_2 * GENDER + \beta_3 * RACE_i + \beta_4 * MARRITAL_STATUS_i + \beta_5 * SALARY + \beta_6 * MAJOR_i + \varepsilon_i \quad (3), \text{ and}$$

$$DISTANCE_i = \beta_0 + \beta_1 * AGE_i + \beta_2 * GENDER + \beta_3 * RACE_i + \beta_4 * MARRITAL_STATUS_i + \beta_5 * SALARY + \beta_6 * MAJOR_i + \varepsilon_i \quad (4).$$

These approaches would be too simple for a real research. But they are useful for this tentative analysis. Further steps would be taken to explore in detail, if the hypothesized relationships occur in these regression models.

Tentative results

Table 3 shows tentative results of the logistic regression and the OLS. Model (1) and Model (3) are nested models for Hypothesis 1, while Model (2) and Model (4) are nested models for Hypothesis 2.

| Table 3: Regression | | | | |
|---|-------------------------|-------------------------|--------------------------|--------------------------|
| VARIABLES | (1) | (3) | (2) | (4) |
| | Whether Moved | | Mobility Distance | |
| Age | 0.0120*** (0.000507) | 0.0129*** (0.000513) | -0.0152*** (0.000610) | -0.0146*** (0.000615) |
| Female | -0.143*** (0.0142) | -0.0832*** (0.0147) | -0.0763*** (0.0171) | -0.0913*** (0.0177) |
| Asian | 1.234*** (0.0198) | 1.207*** (0.0199) | -0.0920*** (0.0232) | -0.0873*** (0.0233) |
| Black | 0.105*** (0.0258) | 0.117*** (0.0259) | -0.502*** (0.0310) | -0.530*** (0.0310) |
| Hispanic | 0.170*** (0.0233) | 0.173*** (0.0234) | 0.782*** (0.0283) | 0.755*** (0.0283) |
| Married | 0.280*** (0.0161) | 0.272*** (0.0162) | -0.235*** (0.0192) | -0.221*** (0.0192) |
| Logged incom | 0.119*** (0.00907) | 0.105*** (0.00911) | -0.0367*** (0.0107) | -0.0305*** (0.0107) |
| Social sciences (vs. natural sciences) | | -0.273*** (0.0187) | | 0.344*** (0.0224) |
| Others (vs. natural sciences) | | -0.213*** (0.0167) | | -0.0907*** (0.0202) |
| Constant | -2.485*** (0.103) | -2.256*** (0.104) | 3.111*** (0.121) | 2.975*** (0.122) |
| Observations | 90,660 | 90,660 | 90,660 | 90,660 |
| R-squared | | | 0.024 | 0.028 |
| Standard errors in parentheses | | | | |
| *** p<0.01, ** p<0.05, * p<0.1 | | | | |

The results of Model (1) and Model (3) supports Hypothesis 1. College major has impact on whether geographical mobility happens. Compared to majors in natural sciences, both social sciences and other majors have negative impact on the probability to move. These relationships are statistically significant. And the results of Model (2) and Model (4) supports Hypothesis 2. College major influences the distance of geographical mobility. Compared to majors in natural sciences, individuals with social sciences majors would on average have greater motility distance, while individuals with other majors would on average have less mobility distance.

I would do more detailed data analysis to explore the mechanisms in the relationship between college major and geographical mobility.