Homework 3: Data and multinomial choices

Due on March 23th 2022 at 11 PM

Your final output should consist of raw code and a pdf file with answers. Output should be uploaded on Github before the due date/hours. For this assignment,

Definitions

We will use three datasets in this exercise:

- datstu: is an administrative data on students from junior high school applying for admission to senior high school through a centralized application system. Students apply to specific academic programs within a school and can submit a ranked list of up to six programs. The dataset consists of the following variables: **score** (student test score), **agey** (student age), **male** (male indicator), **schoolcode1** (first school) **schoolcode2** (second school) **choicepgm1** (first program) **schoolpgm2** (second program) **rankplace** (admission outcome, when rankplace=1, the student got admitted to its first choice, when rankplace=99, the individual could not get assigned to any of her choices), and finally **jssdistrict** (the home district location of the student.)
- datjss is a dataset that provides the **point_x** (longitude) and **point_y** (latitude) of each **jssdistrict**(district).
- datsss is school data that consists of **schoolname**(school name), **schoolcode**(school code), **sssdistrict** (district of the school), **ssslong**(longitude of the school) **ssslat**(latitude of the school).

Exercise 1 Basic Statistics

Calculate the following statistics.

- Number of students, schools, programs
- Number of choices (school, program)
  (Hint: Create a matrix of school, programs. Convert data from Wide to Long)
- Number of students applying to at least one senior high schools in the same district to home
  (Suppose students live in the same district to their junior high schools)
- Number of students each senior high school admitted
- The cutoff of senior high schools (the lowest score to be admitted)
- The quality of senior high schools (the average score of students admitted)

Exercise 2 Data

Create a school level dataset, where each row corresponds to a (school,program) with the following variables:

- the district where the school is located
- the latitude and longitude of the district
- cutoff (the lowest score to be admitted)
- quality (the average score of the students admitted)
- size (number of students admitted)

Exercise 3 Distance

- Using the formula

$$dist(sss, jss) = \sqrt{(69.172 * (ssslong - jsslong) * cos(jsslat/57.3))^2 + (69.172 * (ssslat - jsslat))^2)}$$

  where ssslong and ssslat are the coordinates of the district of the school (students apply to), while jsslong and jsslat are the coordinates of the junior high school, calculate the distance between junior high school, and senior high school. You should generate a value of distance for each of students' choices.

In the rest of the assignment, we want to understand the determinants of school choice. The first exercise, we reduce the number of choices, then the following exercises apply techniques to understand the effect of individual test score and school quality.

Exercise 4 Dimensionality Reduction

- Recode the schoolcode into its first three digits (*substr*). Call this new variable **scode_rev**.
- Recode the program variable into 4 categories: arts (general arts and visual arts), economics (business and home economics), science (general science) and others. Call this new variable **pgm_rev**.
- Create a new choice variable **choice_rev**.
- Recalculate the cutoff and the quality for each recoded choice.
- Consider the 20,000 highest score students.
- The rest of the assignment uses the recoded choices and the 20,000 highest score students.

Exercise 5 First Model

Using the new data with recoded choices, we want to understand the effect of the student test score on his first choice.

- Propose a model specification. Write the Likelihood function.
- Estimate parameters and compute the marginal effect of the proposed model.

Exercise 6 Second Model

Using the new data with recoded choices, we want to understand the effect of the school quality on the first choice.

- Propose a model specification. Write the Likelihood function.
- Estimate parameters and compute marginal effect of the proposed model.

Exercise 7 Counterfactual simulations

In this exercise, we are interested in the effect of excluding choices where the program is "Others".

- Explain and justify, which model (first or second model) you think is appropriate to conduct this exercise.

- Calculate choice probabilities under the appropriate model.
- Simulate how these choice probabilities change when these choices are excluded.