Data

January 14, 2019

The goal of this exercise is to familiarize with large and realistic data, and learn basic techniques for data manipulation and descriptive statistics. We will use three datasets:

- 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.
  - score: student test score
  - agey: student age
  - male: student male
  - schoolcode1: first school
  - schoolcode2: second school
  - choicepgm1: first program
  - schoolpgm2: second program
  - jssdistrict:
- datjss: the longitude ($point_x$) and latitude ($point_y$) of each district (jssdistrict).
- datsss: school name, school code, district, longitude and latitude.

Exercise 1 Missing data

Report the following statistics

- Number of students
- Number of schools
- Number of programs
- Number of choices (school,program

- Missing test score
- Apply to the same school (different programs)
- Apply to less than 6 choices

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 of the district
- the 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 $ssslongandsssslat$ are the coordinates of the district of the school (students apply to), while $jsslongandjsslat$ are the coordinates of the junior high school, calculate the distance between junior high school, and senior high school.

Exercise 4 Descriptive Characteristics

Report the average and sd of the following variables for each ranked choice

- Cutoff
- Quality
- Distance

Redo the same table, differentiating by student test score quantiles.

Exercise 5 Diversification

Group schools by decile of selectivity (cutoffs), and compute for each individual the number of groups in the application. Redo this, by student test score (quantile)