Data Manipulation

January 20, 2020

The goal of this project is to process large and realistic data, and apply basic techniques for data manipulation and descriptive statistics.

There will be three datasets:

- datstu: 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. Students can submit a ranked list of up to six programs.
  - X: unique id for students
  - score: student test score
  - agey: student age
  - male: dummy variable indicating the gender of the student. = 1 for male student
  - schoolcode1: first school
  - schoolcode2: second school
  - choicepgm1: first program
  - schoolpgm2: second program
  - jssdistrict: the district where the student is locate at
  - rankplace: where the student has been admitted to. = 1 means the student has been admitted to its first ranked choice.
- datjss: geographical information indicating the longitude ($point_x$) and latitude ($point_y$) of each district (jssdistrict).
- datsss: a dataset for school name, school code, district, longitude and latitude.

1 Data Overview and Missing data

In order to get an overview of the datasets, the following statistics will be reported:

- Number of students
- Number of schools
- Number of programs
- Number of choices (uniquely identify by school*program)
- If there is missing test score
- How many students applied to the same school (at student-level)

2 Choice-Level Dataset

Create a choice-level dataset, where each row corresponds to a choice(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)

3 Descriptive Characteristics

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

- Cutoff
- Quality
- Distance (between junior high school and senior high school)

  The distance is calculated 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.