Econ 613: Applied Econometrics

Assignment 1: Data Manipulation - OLS/Discrete Choice

Due on March 1st.

Part 1

This exercise aims to familiarize you with a 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 (jhs) applying for admission to senior high school (shs) 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 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.

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.

Part 2

Exercise 5 Data creation

After setting a seed, construct the following objects

- X_1: vector of 10,000 draws from a uniform distribution with range 1:3.
- X_2: vector of 10,000 draws from a gamma distribution with shape 3 and scale 2
- X_3: vector of 10,000 draws from a binomial distribution (one trial) with probability 0.3
- ϵ: vector of 10,000 draws from a normal distribution with mean 2 and sd 1.

Create the variables

- $Y = 0.5 + 1.2X_1 - 0.9X_2 + 0.1X_3 + \epsilon$
- $ydum = \begin{cases} 1, & \text{if } Y > \bar{Y} \\ 0, & \text{otherwise} \end{cases}$

Exercise 6 OLS

You are not allowed to use the pre-programmed OLS function in this assignment.

- Calculate the correlation between Y and X_1. How different is it from 1.2?
- We are interested in the outcome of the regression of Y on X where $X = (1, X_1, X_2, X_3)$.
- Calculate the coefficients on this regression.
- Calculate the standard errors using the standard formulas of the OLS.

Exercise 7 Discrete choice

We consider the determinants of ydum.

- Write and optimize the probit, logit, and the linear probability model. You can use pre-programmed optimization packages.
- Interpret and compare the estimated coefficients. How significant are they?

Exercise 8 Marginal Effects

We consider the determinants of ydum.

- Compute the marginal effect of X on Y according to the probit and logit models.
- Compute the standard error of the marginal effects.