ECON 613 Applied Econometrics Homework1

Promvarat Pradit

January 29, 2019

Exercise 1 Missing data: Report the following statistics

- Number of students

There are a total of **340,823** students in the dataset.
There is no features that can be used to identify a student. Hence, each entry in the dataset $datstu$ is assumed to represent 1 student, and the number of students in this data set is equal to the number of entries in the dataset $datstu$.

- Number of schools

There are a total of **689** schools that have complete data in the dataset $datsss$.
There are many duplicates and missing data in $datsss$, so I kept only the rows that has school name, district, longtitude and latitude. Then the duplicated data was cleaned base on school code.

- Number of Programs

There are a total of **32** programs selected.
All 6 choices of program from $datstu$ were merged, and then the duplicated names were droped.

- Number of Choice

There are a total of **2,006,470** choices made.
Counting all choices that has both school and program designated.

- Missing test score

There are a total of **179,887** students that have no test score reported.
Counting all NA in student's test score column.

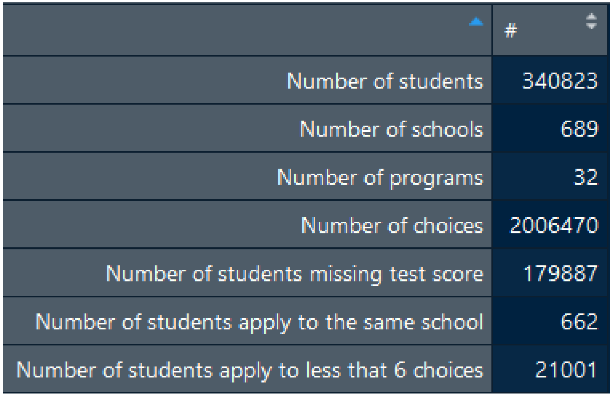
- Apply to the same school

There are a total of **662** students choose only one school.
The criteria used is that the student must has at least 1 school choice, and all other subsequent choice must be either the same school or missing.

- Apply to less than 6 choices

There are a total of **21,001** students who apply to less that 6 choices.
Counting any row that has at least one missing school or program choice.

Figure 1: Table of exercise 1 answers



	#
Number of students	340823
Number of schools	689
Number of programs	32
Number of choices	2006470
Number of students missing test score	179887
Number of students apply to the same school	662
Number of students apply to less that 6 choices	21001

Exercise 2 Data

The code used to create school level dataset is within R code file attached. The school level dataset created is called ”schoolDataset” in the data file.

Exercise 3 Distance

The code used to create school level dataset is within R code file attached. The student dataset created is called ”datstuDist” in the data file.

Exercise 4 Descriptive Characteristics

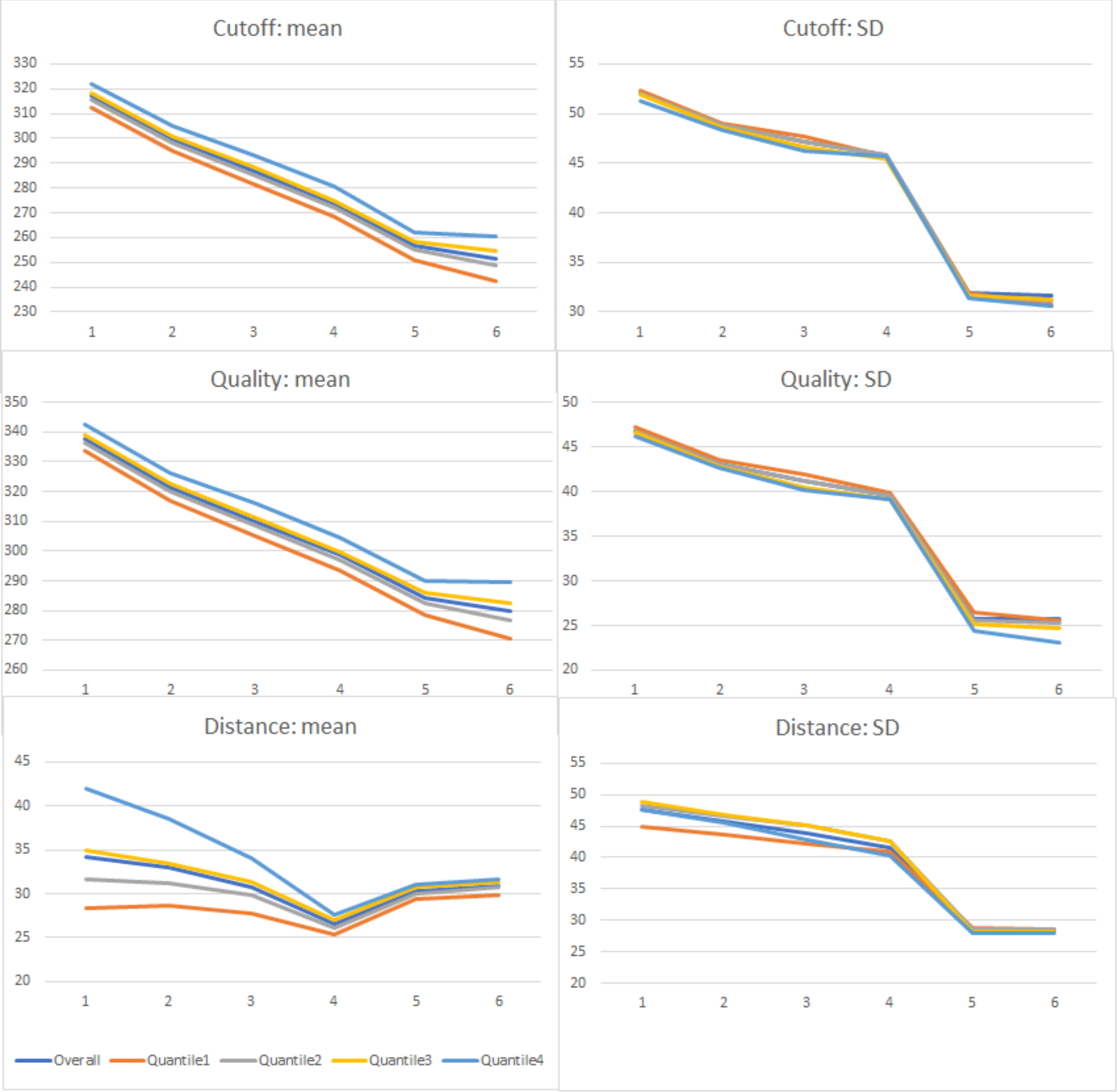
Table which contains descriptive statistics of cutoff, quality and distance can be found in the data file with the name ”reportTable”. Figure 2 shows line graphs of those numbers; x-axis represents each ranked choice, y-axis represents value of each statistic, and line color represents sample group which includes overall and each of the 4 quantiles.

Cutoff and quality decrease as ranked choice is higher. This is natural because students will put harder schools and programs up front, and put in easier choices in later ranks to hedge their risks. In addition, students in higher quantile, apply for higher cutoff and quality school. Again, This is intuitive as students with higher test score have higher probability to get into schools that have high cutoff which implies higher quantity.

For Distance, I observe decreasing trend from ranked 1 through 4 before the slope turns upward in ranked 5 and 6. My conjecture is that students first put in their most favorite schools, and then gradually price-in the distance for choice 1-4. For choice 5 and 6, they sacrifice some distance for certainty. Students choose schools that has very low cutoff though they might need to travel further to prevent from missing all the choices.

SD for all 3 statistics have decreasing trends as ranked choice increase. However, there is a kink at choice 5. This is due to the fact that many students apply to less than 6 choices and they did not put in their choice for rank 5 and 6.

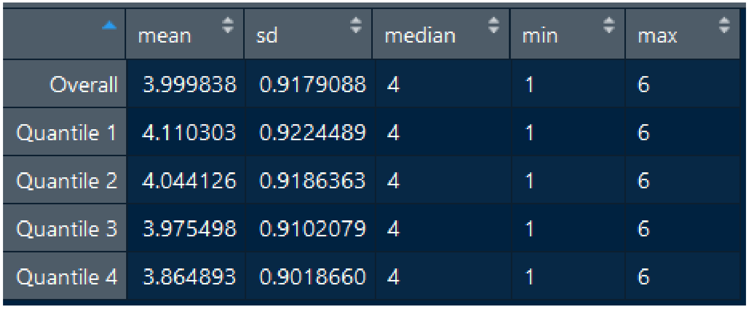
Figure 2: Descriptive statistic(mean, SD) of cutoff, quality and distance for each rank choice)



Exercise 5 Diversification

Figure 3 demonstrates that students, on average, choose around 4 different school groups to diversify their risk. In addition, students in higher quantile, who have higher test score, choose slightly less group than those in lower quantile. This is probably because they have higher chance to get into a school so that they need less diversification.

Figure 3: Descriptive statistic of the number of school group student chosen



	mean	sd	median	min	max
Overall	3.999838	0.9179088	4	1	6
Quantile 1	4.110303	0.9224489	4	1	6
Quantile 2	4.044126	0.9186363	4	1	6
Quantile 3	3.975498	0.9102079	4	1	6
Quantile 4	3.864893	0.9018660	4	1	6