ECON613: Applied Econometrics in Microeconomics

Problem Set # 3

Josh(Yueru) Li

Due Date: 11 PM, Fri Feb 4

Problem 1. Basic Statistics

1. Number of students, schools, programs.

   Assuming each row in the "datstu_vs.cs" file correspond to 1 unique students. There are, in total, 340823 students.

   For the data from "datsss.csv", I group the rows by school code and keep the longest school name, There are, in total, 898 Schools. However, this data set includes schools with school code but without school names or location data, suggesting that there might be some measurement error or this includes schools that had been closed or no longer used for some reason. If we remove schools without school names, we get 689 schools. If we further delete the schools without location data, we get 650 remaining schools.

   Given that the programs only appear in the "datstu_vs.csv" data set, I can only count the programs that actually have student applicants. There are students with choice program empty. Without interpreting what empty program means, there are 32 unique non-empty programs applied by students. If we count the empty programs in our analysis, there are 33 programs.

2. Number of choices

   I keep the choices where a student apply for a school but does not specify the program (non NA school but NA program choice). I remove the choices with school code being NA or empty. Under these criterion, there are 3080 different choices. If I further remove the program choices that are NA or empty, there would be 2773 different choices.

3. Number of students applying to at least one senior high schools in the same district to home.

   For this problem, I maintain the criterion that a choice associated with NA or empty program choice is valid as long as there is a valid school code.

   I first deal with the problem that the variable "sssdistrict" has many noise in the "datss" data set. For example, "Accra Metropolitan", "Accra Metro" and "ACCRA METRO" seem to be the same school district recorded with different names. I merge duplicated entries in "sssdistrict" to the value in "jssdistrict". In doing this, I try to err on the side of caution and leave the uncertain entries untouched. For example, I do not merge "Ga West (Amasaman)" and "Amasaman" even though I suspect they refer to the same district. Furthermore, in the above example, the location data associated with "Amasaman" is NA, so I cannot verify with location data. Interestingly, this is the case for many of the district names that resemble each other. In total, I merged 31 values. More detail can be found in the R code.

   I then merge the student data set that was converted to long form with the other two data sets. For each student I test if there is a choice of senior high school in the same district as the home district. In total, 261195 student applied to at least one senior high schools in the same district to home. It is worth noting that since I was being conservative in dealing with the noise in "sssdistrict", this could be interpreted as a lower bound for the number of students applying to a senior high in the same distrcit as home.

4. Number of students each senior high school admitted

   I remove those with rankplace NA or 99 to make sure I am capturing the actualy admissions in my data. Using the student dataset that was already converted to long format, I group by schoolcode and calculate the cutoff, quality and number of student admitted.

   Please refer to the R code for implementation. Please refer to the bottom table for the data for six of the schools as examples of the results. The size column records the number of students admitted by the school.

   There are 517 schools for which there is admission data.

School Name	Cutoff	Quality	Size
Krobo Girls Secondary	310	359.9	348
Manya Krobo Secondary	238	286.6	350
Eguafo-Abraem Senior Secondary	240.2	778	175
Akontombra Snior Secondary	218	249.9	27
Bosome Secondary	215	259.1	65
St Jerome Secondary	211	255.6	117

5. The cutoff of senior high schools.

   Please refer to the R code for implementation. Please refer to the same table as in question 4 from the data for six of the schools as examples of the results.

6. The quality of senior high schools.

   Please refer to the R code for implementation. Please refer to the same table as in question 4 from the data for six of the schools as examples of the results.

Problem 2. Data

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

1. The district where the school is located.
2. The latitude and longitude of the district
3. Cutoff
4. Quality
5. Size

For the "datStu" data set, I only keep the rows with "rankplace" equal "Rank", which would be the rows containing students with their choice that they get admitted to. I group this data set by school and program and calculate the three metrics.

Please refer to code for implementation. The data set is constructed from "datsss" and named "datSeniorHighs".

Problem 3. Distance

I dropped choices that involves empty or NA values for "ssslong", "ssslat", "point_x" and "point_y" and performed the calculation as specified by the exercise. 2102 choices were removed in this process. Please refer to the R code for implementation.

As reqruied in the exercise, I calculated a maximum of 6 distances for each students because of they each have a maximum of 6 choices. I did not calculate the distance with each potential school for each student.

Problem 4. Dimension Reduction

Please refer to the R-code for the implementation. There are a number of students who tie at the $20,000$ place. I keep all of them so the resulting data set contains a little over $20,000$ students, in particular, there are $20,443$ students.

Problem 5. First Model

We are interested in understanding the demographics of people choosing schools, so I propose to fit a multinomial logit model. Because of the size of the data set and the limited computational capabilities of my laptop, I will fit the following simplest specification for the limited computational capability of my laptop. (Rewriting the code in higher performance languages, for example Julia, would be faster. But for the purpose of this assignment, I am limited to programming in R.)

$$p_{ij} = \frac{\exp(v_{ij})}{\sum_k \exp(v_{ik})}$$
$$v_{ij} = \alpha_j + \beta_j * \text{ score}_i$$

The likelihood function and the optimization procedure can be found in the R file. For the $20,443$ students remaining in our data set, they make 246 distinct first choices. Holding the first choice "100 arts" as the baseline choice. I estimated 245 intercepts and 245 slopes. In total, there are 490 parameters estimated. The result is too big to be shown here, please refer to R code.

Because of computational complexity of involving more than 200 choices, it takes 24 hours for one round of optimization. Therefore I was unable to run multiple optimizations to avoid local minimum. The following marginal effects is calculated with the result from one round of optimization which is likely to be a local optimum.

Let's first derive the formula for marginal effect.

$$\begin{aligned}\frac{\partial p_{ij}}{\partial score_i} &= \frac{\left(\sum_k \exp\left(v_{ik}\right)\right)\beta_j \exp\left(v_{ij}\right) - \exp\left(v_{ij}\right)\left(\sum_k \beta_k \exp\left(v_{ik}\right)\right)}{\left(\sum_k \exp\left(v_{ik}\right)\right)^2} \\ &= \frac{\exp\left(v_{ij}\right)}{\sum_k \exp\left(v_{ik}\right)} \times \frac{\beta_j \sum_k \exp\left(v_{ik}\right) - \sum_k \beta_k \exp\left(v_{ik}\right)}{\sum_k \exp\left(v_{ik}\right)} \\ &= p_{ij} \times \left(\beta_j - \sum_k \beta_k p_{ik}\right)\end{aligned}$$

Let's take this formula to the data. Please find the calculation for all marginal effects in the R code. For example, I calculated $\frac{\partial p_{12}}{\partial score_1} = -0.000141312$. As a sanity check, notice that student 1 actually choose choice 1, which is "100arts", with a quality score of 378.25. Choice two, however, is "100economics", with a quality score of 367. Notice that choice 2 is associated with lower quality, so it makes sense that this marginal effect is negative.

Problem 6. Second Model

Now we are interested in how a change in the characteristic of product, which in our case would be schools, affect its demand. I propose to fit a conditional logit model to answer this problem. Because of the size of the data set and the limited computational capabilities of my laptop, I will fit the following simplest specification allowing for no heterogeneity across students(Hence too simplistic but more computationally managable on my laptop. Rewriting the code in higher performance languages, for example Julia, would be faster. But for the purpose of this assignment, I am limited to programming in R.).

$$\begin{aligned}p_{ij} &= \frac{\exp\left(v_{ij}\right)}{\sum_k \exp\left(v_{ik}\right)} \\ v_{ij} &= \alpha_j + \beta * \text{ quality}_j\end{aligned}$$

The likelihood function and the optimization procedure can be found in the R file. For the $20,443$ students remaining in our data set, they make 246 distinct first choices. Holding the first choice "100 arts" as the baseline choice. I estimated 245 intercepts and 1 slope for the quality of school. In total, there are 246 parameters estimated. The result is too big to be shown here, please refer to R code attachment.

Because of computational complexity of involving more than 200 choices, it takes more than 4 hours for one round of optimization. Therefore I was unable to run multiple optimizations to avoid local minimum. The following marginal effects is calculated with the result from one round of optimization which is likely to be a local optimum.

Now let's derive the expression for marginal effect.

$$\begin{aligned}
\frac{\partial p_{ij}}{\partial quality_j} &= \frac{\left(\sum_k \exp\left(v_{ik}\right)\right)\beta\exp\left(v_{ij}\right) - \exp\left(v_{ij}\right)\beta\exp\left(v_{ij}\right)}{\left(\sum_k \exp\left(v_{ik}\right)\right)^2} \\
&= p_{ij} \times \frac{\sum_k \exp\left(v_{ik}\right) - \exp\left(v_{ij}\right)}{\sum_k \exp\left(v_{ik}\right)} \times \beta \\
&= p_{ij}(1 - p_{ij})\beta \\
\frac{\partial p_{ij}}{\partial quality_{k\neq j}} &= \frac{-\exp\left(v_{ij}\right)\beta\exp\left(v_{ik}\right)}{\left(\sum_k \exp\left(v_{ik}\right)\right)^2} \\
&= -p_{ij}p_{ik}\beta
\end{aligned}$$

For the general calculation, please refer to R code. In this PDF, Let's look at an example for both cases, we calculate $\frac{\partial p_{12}}{\partial quality_2}$ and $\frac{\partial p_{12}}{\partial quality_4}$. The second choice is "100economics", the forth choice is "100science". So we are calculating the change in probability for student 1 to choose "100economics" in response to the quality of "100economics" and "100science" respectively.

From the estimation conducted in R, $\beta = 0.1683429$, $\alpha_2 = 1.459307$, $\alpha_4 = 1.514822$,

$quality_2 = 367$ and $quality_4 = 373.7674$.

$$\begin{aligned}
p_{12} &= \frac{\exp(v_{12})}{\sum_k \exp(v_{ik})} \\
&= \frac{\exp(\alpha_2 + \beta \times quality_2)}{\sum_k \exp(v_{ik})} \\
&= \frac{2.919354e+27}{4.577798e+30} \\
&= 0.0006377218 \\
p_{14} &= \frac{\exp(v_{14})}{\sum_k \exp(v_{ik})} \\
&= \frac{\exp(\alpha_4 + \beta \times quality_4)}{\sum_k \exp(v_{ik})} \\
&= \frac{9.642017e+27}{4.577798e+30} \\
&= 0.002106257
\end{aligned}$$

Thus we can calculate

$$\begin{aligned}
\frac{\partial p_{12}}{\partial quality_4} &= p_{12}(1 - p_1 2)\beta \\
&= 0.0006377218 * (1 - 0.0006377218) * 0.1683429 \\
&= 0.000107287 \\
\frac{\partial p_{12}}{\partial quality_2} &= -p_{i2}p_{i4}\beta \\
&= -0.002106257 * 0.0006377218 * 0.1683429 \\
&= -2.26119e-7
\end{aligned}$$

Problem 7. Counterfactual Simulations

Here we simulate the situation where all the offerings "others" are removed.

1. Explain and justify which model is appropriate for this task.

   Multinomial logit models allow us to look at how different student respond to school characteristics, in other words, it allows us to study demand side properties. Conditional logit models allow us to look at how changes in school characteristics affect the market share on average. When the choice of "others" is removed and we are interested

in how market share of schools as first choices change, we should use the conditional logit model estimated in Exercise 6.

2. Calculate choice probabilities under the appropriate model

   Let S be the set of choices with "others". The choice probabilities we should calcualte are

$$p_{i,j\notin S} = \frac{\exp(v_{ij})}{\sum_k \exp(v_{ik})}$$
$$v_{i,j\notin S} = \alpha_j + \beta * \text{ quality}_j$$

   The calculation of these probabilities are implemented in the attached R code. The solution is too long to be shown here.

   Notice that for all $i = 1, 2, \ldots, 20,443$, p_{ij} is the same. So we can suppress the notation and write p_j instead. We have a 246×1 vector.

3. Simulate how these choice probabilities change when these choices are excluded.

   The probabilities calculated in this section is just a subset of all the probabilities from the conditional logit model.

   The new probabilities are characterized by the following expressions.

$$p_{j\notin S} = \frac{\exp(v_j)}{\sum_{k\notin S} \exp(v_k)}$$
$$v_{j\notin S} = \alpha_j + \beta * \text{ quality}_j$$

   Notice the difference is in the sum in the numerator of p_j. The exact calculation can be found in the R code. It appears that all probabilities are 1.035513 times their original value. The market share of the removed school choices are distributed evenly across the remaining school choices, in line with the IIA property of logit models.