

Hamzah Ahmed

Econ 613

Assignment 3

General Note: When importing datasets, I got a first column V1 that was just 1-... . I used this as a "ID" quite often.

Problem 1 – Basic Statistics

1. Number of:

a. Students

I assumed each row in datstu was a unique student (no ID variable given so seemed valid - later used V1 as an id (happens when importing .csv in R for me)). Number of Students is: **340823**

b. Schools

I had to clean up the data on datsss for this. Many repeats of school code with different names and jurisdictions (if they had it) in datsss. I first grouped by schoolcode and then chose the row with the longest school name (least abbreviations) and used distinct() for getting rid of duplicates. Then, I noticed that there were still some repeats (68 or so) so I went again and grouped by code and then filtered out the rows where district is blank (for duplicates only) - could have also just checked which district had longest name but this seemed to work. It seemed as though all these duplicates also did not have a school name. I also included schoolcodes with no names. In any case, the number of schools I found was: **898**

c. Programs

I found the number of programs in two ways. First was with unique programs (ie if school A and B offered Program X, that counts as 2 programs (school unique). Second, under the same setup, it is considered 1 program (school non-unique). Both definitions feel valid but have different outcomes.

Number of (school non-unique) Programs is: **33** (I think this is what the question is asking for but I included the other value just in case)

Number of (school unique) Programs is: **3086**

2. Number of Choices (equivalent to number of school unique programs).

Same result as earlier about unique school programs but how I constructed this one may be a bit more useful for future problems (thanks to the hint about converting to long). I included choices with no code (such as NATechnical, NAGeneral Arts, NAHome Economics). Number of Unique Choices is: **3086**

3. Number of Students applying to at least one high school in the same district

I merged the choice data (the long format of datstu) with the new (cleaned up) school information data (originally datsss). I merged by school code and checked if jssdistrict was equal to one of the sssdistrict's for each person (ie they applied to at least one school that was in their home district). I found that the following number of students applied to at least one school in the same home district: **262167**

4. Number of Students each high school admitted

First, I created a subset of the merged data that aimed to show admitted results. This meant I omitted students who had rankplace either equal to 99 or missing (NA or "") as they did not get into one of their choices (or at least we have no information corroborating that). I then created a placement variable that simply stated which choice they got admitted to (choice1 through choice6). I then filtered out the rows where placement wasn't equal to choiceNum and thus got admitted data.

To find the number of students admitted, I grouped by schoolcode and had admittance equal the size of the group.

	schoolcode	admittance
1	10101	398
2	10102	248
3	10103	443
4	10104	220
5	10105	346
6	10106	395
7	10107	306
8	10108	318
9	10109	300
10	10110	535
11	10111	600
12	10112	300
13	10114	350
14	10115	238
15	10116	446
16	10117	471
17	10118	539

First few rows of data on admittance data

*Not part of the question but interestingly, although I saw that 898 were in the sample, only a subset of them admitted students. That is, this many schools admitted students: 517




Hamzah Ahmed

Econ 613

Assignment 3

5. The cutoff of senior high schools

I used my admitted data from earlier and grouped by school code (so looking at each school) and then summarized cutoff as the minimum score

		schoolcode 	cutoff 
	1	10101	284
	2	10102	343
	3	10103	316
	4	10104	245
	5	10105	260
	6	10106	293
	7	10107	281
	8	10108	248
	9	10109	257
	10	10110	343
	11	10111	371
	12	10112	316
	13	10114	319
	14	10115	274
	15	10116	205
	16	10117	330
	17	10118	273

First few rows of cutoff data by school




Hamzah Ahmed

Econ 613

Assignment 3

6. Quality of each high school

I used my admitted data from earlier and grouped by school code (so looking at each school) and then summarized quality as the mean score

	 schoolcode 	quality 
1	10101	320.2312
2	10102	394.1492
3	10103	353.8330
4	10104	296.9182
5	10105	351.2139
6	10106	340.1013
7	10107	311.9542
8	10108	303.9025
9	10109	281.8233
10	10110	408.0785
11	10111	412.5100
12	10112	375.6133
13	10114	346.2229
14	10115	316.3361
15	10116	289.9574
16	10117	369.3163
17	10118	315.1206

First few rows of quality by school (based on average score of admitted student)

Problem 2 – Data

1. Since datsss doesn't tell you what programs are offered at each school, our school + program info comes from the choices each student made (can be found in our mergedData that is long) I broke the choices into schoolcode and program choice so that I could merge with the schoolInfo to get the district, longitude and latitude.
I found the cutoff, size, and quality of each program and merged it with this data. My final output is called program_data

I didn't remove any where either the code or the program was missing so some of that data may be missing. If nobody got into a specific program, its info (cutoff, quality, and size) are also missing (haven't removed them yet).

Brief/Repeated Synopsis:

First, I found all the unique programs that were applied to (not necessarily ones where anybody got accepted/applied to). I then created a data frame where the first three rows were the choice programs, the school, and the program. I then merged with schoolInfo by schoolcode to get the district, longitude and latitude. Then, using my admitted data, I found the cutoff, quality, and size for each program and merged them together. If nobody got in to a program, we have no info on cutoff, quality, and size (so values are NA). An example can be seen in the first line as (at least) one person had specified a choice with the schoolcode but no program but they weren't admitted to that choice."

	choiceProgram	schoolcode	program	sssdistrict	ssslong	ssslat	cutoff	quality	size
1	100101	100101		Wa Municipal	-2.2850304	10.030622	NA	NA	NA
2	100101General Arts	100101	General Arts	Wa Municipal	-2.2850304	10.030622	198	244.3924	79
3	100101Home Economics	100101	Home Economics	Wa Municipal	-2.2850304	10.030622	199	229.4500	40
4	100101Technical	100101	Technical	Wa Municipal	-2.2850304	10.030622	201	235.1020	49
5	100102Agriculture	100102	Agriculture	Wa Municipal	-2.2850304	10.030622	273	292.5556	90
6	100102Business	100102	Business	Wa Municipal	-2.2850304	10.030622	283	303.3444	90
7	100102General Arts	100102	General Arts	Wa Municipal	-2.2850304	10.030622	291	311.1111	90
8	100102General Science	100102	General Science	Wa Municipal	-2.2850304	10.030622	273	298.4333	90
9	100102Home Economics	100102	Home Economics	Wa Municipal	-2.2850304	10.030622	262	278.8667	45
10	100102Visual Arts	100102	Visual Arts	Wa Municipal	-2.2850304	10.030622	250	275.2000	45
11	100103Agriculture	100103	Agriculture	Wa Municipal	-2.2850304	10.030622	NA	NA	NA
12	100103Business	100103	Business	Wa Municipal	-2.2850304	10.030622	NA	NA	NA
13	100103General Arts	100103	General Arts	Wa Municipal	-2.2850304	10.030622	NA	NA	NA
14	100104Business	100104	Business	Wa Municipal	-2.2850304	10.030622	NA	NA	NA
15	100104General Arts	100104	General Arts	Wa Municipal	-2.2850304	10.030622	319	337.4444	45
16	100104General Science	100104	General Science	Wa Municipal	-2.2850304	10.030622	313	334.0000	45
17	100104Home Economics	100104	Home Economics	Wa Municipal	-2.2850304	10.030622	282	309.3556	45

First few rows of program_data – can see where data is missing (such as program or quality info)

Hamzah Ahmed
Econ 613
Assignment 3
Problem 3

1. I looked at my choice data and merged with datjss to get location information about each individual. 6 rows per person (as each person had 6 choices). I then merged with the schoolInfo to get the school location information. I then looked at each row and used the given formula to find distance (after recoding to jsslong and jsslat for for point_x and point_y.

V1	score	agey	male	rankplace	choiceNum	choiceProgram	schoolcode	jssdistrict	jsslong	jsslat	ssslong	ssslat	distance
1	1	NA	16	0	NA	choice2	50107	Bosomtwe/Atwima/Kwanwoma (Kuntanase)	-1.5627517	6.559323	-1.59718716	6.682060	8.813579
2	1	NA	16	0	NA	choice1	50112	Bosomtwe/Atwima/Kwanwoma (Kuntanase)	-1.5627517	6.559323	-1.59718716	6.682060	8.813579
3	1	NA	16	0	NA	choice4	50202	Bosomtwe/Atwima/Kwanwoma (Kuntanase)	-1.5627517	6.559323	-1.80875707	6.681337	18.895053
4	1	NA	16	0	NA	choice3	50202	Bosomtwe/Atwima/Kwanwoma (Kuntanase)	-1.5627517	6.559323	-1.80875707	6.681337	18.895053
5	1	NA	16	0	NA	choice5	50702	Bosomtwe/Atwima/Kwanwoma (Kuntanase)	-1.5627517	6.559323	-1.54142010	6.806778	17.179653
6	1	NA	16	0	NA	choice6	50901	Bosomtwe/Atwima/Kwanwoma (Kuntanase)	-1.5627517	6.559323	-1.36796534	7.462874	63.917746
7	2	NA	17	0	NA	choice1	70102	Ho Municipal	0.5261422	6.717607	0.52614224	6.717607	0.000000
8	2	NA	17	0	NA	choice4	70105	Ho Municipal	0.5261422	6.717607	0.52614224	6.717607	0.000000
9	2	NA	17	0	NA	choice3	70107	Ho Municipal	0.5261422	6.717607	0.52614224	6.717607	0.000000
10	2	NA	17	0	NA	choice2	70602	Ho Municipal	0.5261422	6.717607	0.26738513	6.896852	21.672792
11	2	NA	17	0	NA	choice6	70603	Ho Municipal	0.5261422	6.717607	0.26738513	6.896852	21.672792
12	2	NA	17	0	NA	choice5	70605	Ho Municipal	0.5261422	6.717607	0.26738513	6.896852	21.672792

First few rows of cleanedChoiceData

*** At this point, I also converted the cleanedChoiceData into a wide format (so had schoolcode, choiceprogram, and location stuff for each choice (1-6) for each person. So there, we have 1 row per person and have the distance of each choice included in that row (distance_choice*i*) in the last 6 columns and the values should match up. It was easier to paste the long format as it had fewer columns but the wide format is in my markdown.

Problem 4 – Dimensionality Reduction

1. Recode the schoolcodes

How it made sense for me to do, I didn't make scode_rev and pgm_rev but restarted and changed the initial school codes to the reduced version (in datstu and called it datstu_rev) and all the programs into the new ones (also did it for schoolInfo and called it schoolInfo_rev which was the cleaned datsss).

I have converted the wide data to long data (or plan to) for ease of use. With how I have constructed it, I have done all the recoding but the names are somewhat different. Somewhat ambiguous question so hopefully my methodology is correct. If schoolcode is missing, it uses NA as the recoding.

Creating new dataframes that are revised (_rev). Schoolcode is found in datstu and schoolInfo (cleaned up datsss)

I created datstu_rev and schoolInfo_rev that did the truncations. They both add columns to the pre-existing data (I did not want to delete any information) so please look at the last columns to see the recodings.

2. Recode the program variables

For the recoding of programs, I just checked if each program was in the list for each recoding (ie arts, economics, science, or other. If missing, it should do other as well. New recodes in datstu_rev as well.

3. Create new choice variable

Again, I did it slightly differently and did choiceProgram_rev instead of choice_rev for each person and their (up to) 6 choices (in a long format) and updated datstu_rev with these new recoded choices.

Final stuff with all the dimensionality reductions is in cleanedChoiceData_rev. Again, 6 rows associated with 6 choices for each person (make sure to have V1 in order to see it). Last 2 columns show the choiceNum_rev and choiceProgram_rev (equivalent to choice_rev I think). In cleanedChoiceData_rev, we can see the relevant recodes for schoolcodes and programs (choices 1 through 6).

4. Recalculate cutoff and quality for each recoded choice

Did same process as earlier to get admitted data where I first filtered out people who had no rankplace info (or weren't admitted into one of their choices). Then, I filtered out the choices that didn't match with the placement (associated rankplace) and got my admitted data info (for each person that got into a program they chose (1-6)

I used my admitted data from earlier and grouped by choice (schoolcode + program) and then summarized cutoff as the minimum score.

I used my admitted data from earlier and grouped by choice and then summarized quality as the mean score.

	choiceProgram_rev	cutoff_rev		choiceProgram_rev	quality_rev
1	100arts	194	1	100arts	275.5233
2	100economics	195	2	100economics	264.4993
3	100other	191	3	100other	245.6381
4	100science	228	4	100science	305.1814
5	101arts	243	5	101arts	340.0850
6	101economics	205	6	101economics	326.3979
7	101other	257	7	101other	313.2753
8	101science	203	8	101science	368.7612
9	102arts	216	9	102arts	315.5544
10	102economics	206	10	102economics	308.9986
11	102other	209	11	102other	280.9509
12	102science	242	12	102science	340.3426
13	103arts	260	13	103arts	299.1236
14	103economics	236	14	103economics	288.6645
15	103other	246	15	103other	288.6831
16	103science	292	16	103science	317.7857
17	104arts	209	17	104arts	299.1473

First few rows of recoded cutoffs and quality information on revised programs

Hamzah Ahmed

Econ 613

Assignment 3

By now, I also converted this data to wide again (called cleanedChoiceDataWide_rev) but I'm not sure it was necessary.

There seemed to be 4 programs that at least one person chose but nobody got accepted to. They were: NAOther, 907arts, NAArts, and NAEconomics. Makes sense for the NAschoolcode ones but interesting that 907arts was included in that. That means there is no school quality data and it should not be including in any discrete choice models. There seemed to now be 426 unique choices now. I did some setup stuff for problems 5 and 6.

The Top 20,000 scores ends with people with scores of 355. Because it would seem arbitrary to cut off certain people who had 355, I included all individuals who had scores of at least 355. This included slightly more students: 20447"

For condData, I removed the choices of schools with no info and the students who had those schools as their first choice. Two people had their first choice be one of the schools with no admissions so I dropped them too. Basically has column of choices and columns of quality of each choice. Only showing options that were first choices of at least one person (we can't say much about options that were not picked).

There were **246 unique first choices**.

For multData, we just have the test score for each of these students (minimum was 355).

***I practiced writing conditional and multinomial code for margarine data (provided by Hung-Wei) and I aimed to replicate my margarine code here (so there are some things that may seem out of place like productNames (instead of choiceNames) but I didn't want to keep changing things and mess up my code that seems to (slowly) work. Sorry for the inconvenience but I tried to comment my code as much as I could so that my logic flowed.

Hamzah Ahmed
Econ 613
Assignment 3
Problem 5

1. Think it is multinomial logit as we are using variation in individuals (their scores) as opposed to variation in choices (ie the quality). The equation is: $v_{ij} = \alpha_j + \beta_j^1 * Score_i$ where $p_{ij} = \frac{\exp(v_{ij})}{\sum \exp(v_{il})}$. I estimated $(425-1) * 2$ parameters. Reference group was first choice which was 210economics. I ran the code 5 times (took a while) and kept the parameters with the smallest likelihood (as my likelihood function returned the negative as optim aims to minimize).
2. For my parameters, the first 245 are intercepts and the last 245 are the betas (see markdown for values). For parameters that are negative (in the last 245), it means that as scores increase, the probability of picking those options decreases (relative to reference group). Can't say much about magnitude from the parameters (need AME to do that).

I computed the AME using the equation from the notes. Almost all the ME's were negative (but VERY small). The only one that was non-negative was the ME for 501arts.

These values can be interpreted as:

When score increases by 1 unit, the probability of choosing that option changes by that value (generally decreases). A unit increase in score leads to a small decrease in these as the first choice_rev. Everything was very small (but negative) but the change for 501arts was zero. Makes sense that values are small because a 1 unit increase in score is not that much so shouldn't cause that much change so interesting but rationalizable.

	Score
210economics	-6.842041e-202
210arts	-3.373092e-87
210other	-8.088876e-84
501science	-1.123575e-177
502arts	-2.278588e-156
301economics	-1.216921e-60
301arts	-3.085239e-57
211science	-7.446456e-56
301science	-1.383404e-153
213arts	-1.266668e-120
501arts	0.000000e+00
204arts	-1.045585e-78
213economics	-8.340664e-86
201economics	-6.967851e-119
211arts	-1.529517e-177
303science	-5.933717e-121
201arts	-4.620125e-60
203economics	-4.157040e-80
101economics	-3.203333e-143

First couple of rows of AME.

1. "Think it is conditional logit as we are using variation in choices (their quality) as opposed to variation in individuals (ie their scores)." The equation is: $v_{ij} = \alpha_j + \beta * Quality_j$ where $p_{ij} = \frac{\exp(v_{ij})}{\sum \exp(v_{il})}$. I estimated (246-1) + 1 parameters where the last parameter is the elasticity. I ran the code 5 times (took a while) at kept the parameters with the smallest likelihood (as my likelihood function returned the negative as optim aims to minimize).
2. Here, β was roughly .0327 which meant that, for each choice, if quality increased, the probability of it being picked increases (more demand for higher quality programs which intuitively makes sense). I computed the AME using the equation from the notes. It seemed as though all the diagonals were positive and the non-diagonals are negative which makes sense. If the quality of a specific choice increases, the probability of picking that specific choice should increase while the others decrease (if their quality stays the same) as its relative quality increases. The symmetric off-diagonals also seem to be equivalent which is reassuring. Each square tells you the relationship between probability of choices. For example, if 210arts's quality increases by 1 unit, the probability of choosing 210economics decreases by -2.027e-06 percentage points. Again, a small effect since a 1 unit increase in quality is not that much. For $A[i,j]$ where i is the row choice and j is the column choice, if the quality of j increases by 1 unit, the probability of choosing i increases by $A[i,j]$ and the same holds for if the quality of i increases (by symmetry as $A[i,j] = A[j,i]$).

	210economics	210arts	210other	501science	502arts	301economics	301arts	211science	301science
210economics	2.366390e-04	-2.027117e-06	-4.995291e-07	-1.920202e-05	-3.673535e-06	-1.075223e-05	-1.734730e-05	-2.029346e-06	-1.870956e-05
210arts	-2.027117e-06	2.758438e-04	-5.830029e-07	-2.241077e-05	-4.287400e-06	-1.254898e-05	-2.024612e-05	-2.368460e-06	-2.183602e-05
210other	-4.995291e-07	-5.830029e-07	6.841371e-05	-5.522538e-06	-1.056516e-06	-3.092364e-06	-4.989118e-06	-5.836440e-07	-5.380906e-06
501science	-1.920202e-05	-2.241077e-05	-5.522538e-06	2.423074e-03	-4.061272e-05	-1.188712e-04	-1.917829e-04	-2.243541e-05	-2.068433e-04
502arts	-3.673535e-06	-4.287400e-06	-1.056516e-06	-4.061272e-05	4.964010e-04	-2.274123e-05	-3.668995e-05	-4.292115e-06	-3.957116e-05
301economics	-1.075223e-05	-1.254898e-05	-3.092364e-06	-1.188712e-04	-2.274123e-05	1.409117e-03	-1.073894e-04	-1.256278e-05	-1.158226e-04
301arts	-1.734730e-05	-2.024612e-05	-4.989118e-06	-1.917829e-04	-3.668995e-05	-1.073894e-04	2.207554e-03	-2.026838e-05	-1.868644e-04
211science	-2.029346e-06	-2.368460e-06	-5.836440e-07	-2.243541e-05	-4.292115e-06	-1.256278e-05	-2.026838e-05	2.761445e-04	-2.186003e-05
301science	-1.870956e-05	-2.183602e-05	-5.380906e-06	-2.068433e-04	-3.957116e-05	-1.158226e-04	-1.868644e-04	-2.186003e-05	2.366236e-03

First couple of rows of AME

Problem 7

1. I believe we use conditional logit here because the choices available (and the information about quality) changes as the options are reduced (dropping others program). I believe the second model (conditional logit) should be used to conduct this exercise because we are interested in the effect of excluding choices where the program is other which means we are changing the choices and their respective variation. The information set on individual scores doesn't change so I don't think the first model would be helpful.

There were 50 choices that included “other” which mean I now am comparing $246 - 50 = 196$ choices and their respective probabilities under the full and restricted models.

- For the Restricted Data, I used the same optimized parameters and quality data but I only kept them for the choices that did not have others (ie they weren’t allowed to be picked). I computed the choice probabilities as per usual after dropping “other” program choices. For the Full version, I just selected the choice probabilities without dropping “other” program choices.

Since none of the individuals were unique, the prob vector of each row was the same so I just looked at the first person.

- I then put the choice probabilities in a table by the choice and calculated the ratio of restricted / full and got 1.036 for (it seems like) all of the choices. That is a good sign as it indicates that the probabilities all increased equally (in proportion) when a certain set of options was removed which lends credence to IIA assumption (which is usually violated).

	choice	full	restricted	ratio
1	210economics	7.286293e-03	7.552528e-03	1.036539
2	210arts	8.503868e-03	8.814592e-03	1.036539
3	501science	8.055352e-02	8.349688e-02	1.036539
4	502arts	1.541068e-02	1.597378e-02	1.036539
5	301economics	4.510621e-02	4.675436e-02	1.036539
6	301arts	7.277288e-02	7.543194e-02	1.036539
7	211science	8.513219e-03	8.824285e-03	1.036539
8	301science	7.848763e-02	8.135550e-02	1.036539
9	213arts	3.866560e-03	4.007841e-03	1.036539
10	501arts	8.592854e-02	8.906830e-02	1.036539
11	204arts	7.337592e-03	7.605702e-03	1.036539
12	213economics	2.348058e-03	2.433854e-03	1.036539
13	201economics	1.369031e-02	1.419054e-02	1.036539
14	211arts	1.257179e-02	1.303115e-02	1.036539
15	303science	2.489798e-03	2.580773e-03	1.036539
16	201arts	1.985326e-02	2.057868e-02	1.036539
17	203economics	1.564917e-02	1.622098e-02	1.036539
18	101economics	3.159860e-02	3.275318e-02	1.036539
19	204economics	3.520373e-03	3.649005e-03	1.036539

First couple rows of compared choice probabilities.