HW3-YS

Yonghan Shi

3/4/2022

Exercise 1 Basic Statistics

- Number of students, schools, programs

student	school	program
340823	640	33

Note the number of school means the number of schools being applied by students (not that in the data of schools or that with students admitted).

- Number of choices (school, program) (Hint: Create a matrix of school, programs. Convert data from Wide to Long)

The number of choices is 2773.

- Number of students applying to at least one senior high schools in the same district to home (Suppose students live in the same district to their junior high schools)

The number of students applying to at least one senior high schools in the same district to home is 2.62145×10^5.

- Number of students each senior high school admitted

schoolcode	size
10101	398
10102	248
10103	443
10104	220
10105	346
10106	395

- The cutoff of senior high schools (the lowest score to be admitted)

schoolcode	cutoff
10101	284
10102	343
10103	316
10104	245
10105	260
10106	293

- The quality of senior high schools (the average score of students admitted)

schoolcode	quality
10101	320.2312
10102	394.1492
10103	353.8330
10104	296.9182
10105	351.2139
10106	340.1013

Exercise 2 Data

schoolcode	program	size	cutoff	quality	schoolname	sssdi
10101	General Arts	100	316	330.0900	EBENEZER SENIOR HIGH. SCHOOL, DANSOMAN	Accr
10101	Visual Arts	50	296	311.5400	EBENEZER SENIOR HIGH. SCHOOL, DANSOMAN	Accr
10101	Agriculture	49	288	310.1429	EBENEZER SENIOR HIGH. SCHOOL, DANSOMAN	Accr
10101	General Science	50	299	329.1000	EBENEZER SENIOR HIGH. SCHOOL, DANSOMAN	Accr
10101	Business	100	305	324.8600	EBENEZER SENIOR HIGH. SCHOOL, DANSOMAN	Accr
10101	Home Economics	49	284	300.5714	EBENEZER SENIOR HIGH. SCHOOL, DANSOMAN	Accr

Exercise 3 Distance

V1	schoolcode	program	choice_rank	dist
1	50112	Home Economics	1	8.813579
2	70102	General Arts	1	0.000000
3	50702	Business	1	0.000000
4	90501	Visual Arts	1	0.000000
5	51802	Home Economics	1	102.388006
6	10102	General Arts	1	121.565099

Exercise 4 Dimensionality Reduction

- Recode the schoolcode into its first three digits (substr). Call this new variable scode_rev.
- Recode the program variable into 4 categories: arts (general arts and visual arts), economics (business and home economics), science (general science) and others. Call this new variable pgm_rev.
- Create a new choice variable choice_rev.
- Recalculate the cutoff and the quality for each recoded choice.
- Consider the 20,000 highest score students.
- The rest of the assignment uses the recoded choices and the 20,000 highest score students.

V1	score	agey	male	scode_rev	pgm_rev	choice_rev	choice_rank	admitted	cutoff	quality
335624	469	15	0	301	Science	301 Science	1	TRUE	321	410.9369
335624	469	15	0	301	Economics	301 Economics	2	FALSE	202	348.1246
335624	469	15	0	501	Arts	501 Arts	3	FALSE	259	358.7018
335624	469	15	0	215	Economics	215 Economics	4	FALSE	190	270.7927
335624	469	15	0	104	Arts	104 Arts	5	FALSE	209	299.1473
335624	469	15	0	101	Arts	101 Arts	6	FALSE	243	340.0850

Exercise 5 First Model

- Propose a model specification. Write the Likelihood function.

Test score is the same for each students, so we use multinomial logit model.

```
> ## X and Y for the model
> X1 <- dat5 %>%
+   mutate(intercept = 1) %>%
+   select(intercept, score)
>
> X1 <- as.matrix(X1)
>
> Y1 <- dummy("choice_rev", data = dat5, sep="_")
>
> beta1 = solve(t(X1) %*% X1) %*% t(X1) %*% Y1
>
> # Likelihood Function for multinomial logit model
> mlogit.ll <- function(beta) {
+   B <- matrix(beta, nrow = 2, byrow = TRUE)
+   B[,1] <- 0
+   est <- exp(X1 %*% B)
+   Pr <- t(apply(est, 1, function(x) x / sum(x)))
+   Pr <- Pr*Y1
+   loglikelihood <- sum(log(rowSums(Pr)))
+   return(-loglikelihood)
+ }
```

- Estimate parameters and compute marginal effect of the proposed model.

```
## # weights:  735 (488 variable)
## initial  value 110014.161694
## iter  10 value 77097.978580
## iter  20 value 76713.535742
## iter  30 value 76712.716760
## iter  40 value 76711.705840
## iter  50 value 76710.424513
## iter  60 value 75793.642000
## iter  70 value 75594.520846
## iter  80 value 75593.556829
## iter  90 value 75479.355379
## iter 100 value 74236.055047
## final  value 74236.055047
## stopped after 100 iterations
```

Parameters:

	100 Economics	100 Others	100 Science	101 Arts	101 Economics	101 Others
(Intercept)	0.2924757	-0.0173560	0.5613559	1.5154205	3.3124952	-0.0929679
score	-0.0001586	-0.0033966	0.0013861	0.0066803	-0.0003778	0.0050557

Marginal Effect:

X100.Economics	X100.Others	X100.Science	X101.Arts	X101.Economics	X101.Others
0.0024458	0.0004239	0.0066062	0.1874971	0.1094368	0.0110936

Exercise 6

- Propose a model specification. Write the Likelihood function.

School Quality is the same for each choice, so we use conditional logit model.

```
> ## X and Y for the model
>
```

```
> X2 <- dat5%>%
+   select(choice_rev, quality) %>%
+   filter(!duplicated(choice_rev)) %>%
+   arrange(choice_rev)
>
> X2 <- t(data.frame(X2, row.names=1))
>
> X2f <- as.matrix(X2-X2[1])
>
> Y2 <- dummy("choice_rev", data = dat5, sep="_")
>
> beta2 <- runif(246,-1,1)
>
> # Likelihood Function for conditional logit model
> clogit.ll <- function(beta) {
+   B <- beta2[1]
+   A <- as.matrix(beta2[2:246])
+   A[1] <- 0
+   est <- exp(X2f * B + ( matrix(1,nrow = nrow(X2f * B),ncol = 1) %*% t(A)))
+   Pr <- t(apply(est, 1, function(x) x / sum(x)))
+   Pr <- matrix(rep(Pr,each=nrow(Y2)),nrow=nrow(Y2))
+   Pr <- Pr * Y2
+   loglikelihood <- sum(log(rowSums(Pr)))
+   return(-loglikelihood)
+ }
```

- Estimate parameters and compute marginal effect of the proposed model.

Parameters:

intercept	choice_rev_100 Economics	choice_rev_100 Others	choice_rev_100 Science	choice_rev_101 Arts	ch
-0.4991033	0.515345	0.646372	0.0197543	-0.9778607	

Marginal Effect:

```
##                    100 Arts 100 Economics    100 Others   100 Science
## 100 Arts      -1.924437e-08  3.046357e-13  4.255898e-09  2.820834e-22
## 100 Economics  3.046357e-13 -7.900611e-06  1.747250e-06  1.158088e-19
## 100 Others     4.255898e-09  1.747250e-06 -8.596696e-02  1.617900e-15
## 100 Science    2.820834e-22  1.158088e-19  1.617900e-15 -7.315842e-15
## 101 Arts       2.828124e-30  1.161080e-27  1.622081e-23  1.075125e-36
## 101 Economics  5.168569e-27  2.121945e-24  2.964453e-20  1.964857e-33
##                    101 Arts 101 Economics
## 100 Arts       2.828124e-30  5.168569e-27
## 100 Economics  1.161080e-27  2.121945e-24
## 100 Others     1.622081e-23  2.964453e-20
## 100 Science    1.075125e-36  1.964857e-33
## 101 Arts      -7.334747e-23  1.969934e-41
## 101 Economics  1.969934e-41 -1.340470e-19
```

Exercise 7 Counterfactual Simulations

- Explain and justify, which model (first or second model) you think is appropriate to conduct this exercise.

I think the first model is appropriate, as for the second model, moving "others" out won't affect the result much.

- Calculate choice probabilities under the appropriate model.

100 Arts	100 Economics	100 Others	100 Science	101 Arts	101 Economics
0.0008822	0.0011132	0.000241	0.0026102	0.0503289	0.0210007

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

```
## # weights:  585 (388 variable)
## initial  value 101779.437479
## iter  10 value 71752.801447
## iter  20 value 71403.721815
## iter  30 value 71402.644399
## iter  40 value 71402.413437
## iter  50 value 71384.166902
## iter  60 value 70538.845084
## iter  70 value 70327.649794
## iter  80 value 70324.252305
## iter  90 value 70323.683664
## iter 100 value 69815.594178
## final  value 69815.594178
## stopped after 100 iterations
```

100 Arts	100 Economics	100 Science	101 Arts	101 Economics	101 Science
0.0010146	0.001581	0.0021891	0.0456075	0.0119669	0.0610666

The choice probabilities become higher when excluding those choices.