Assignment DEEQA EEE: Discrete Choice Models

Young Kwang Kim, Wenxuan Xu

September 2020

1 Descriptive Evidence

Consider the cross-correlation between the dependent variable, *school*, and the observed student characteristic variables. All the comments in this section is based on correlation **not** causality between variables.

First of all, consider period 1; students decide whether to attend high school. From Table 1, male students tends to drop out from high school more than female students. In Table 2 we can see that students with at least one of parents with a degree tend to attend high school more than those without a degree. Lastly, it is clear from Table 3 that there is a positive correlation between the ability scores and school. In other words, the higher ability score, the higher chance of attend high school.

Table 1: Gender, $t = 1$

Variables	school	male
school	1.000	
male	-0.132	1.000

Table 2: College parent, $t = 1$

Variables	school	parentcollege
school	1.000	
parentcollege	0.151	1.000

Table 3: Ability scores, $t = 1$

Variables	school	math	Variables	school	language
school	1.000		school	1.000	
math	0.288	1.000	language	0.308	1.000

Now we look at period 2; conditional on high school graduation, students choose to continue study at collage or not. Overall, the correlations exhibits the similar patterns as it is in period 1. Attending college seems to be less common among men than women shown in Table 4. An interesting difference is that the correlations are much higher for *parentcollege* and ability scores shown in Table 5 and Table 6 compare to Table 2 and Table 3 in period 1. It may imply that the parents' education background as well as student's ability influence more heavily on decision to attend college than to attend high school. Finally in Table 5, the distance is also negatively correlated with the dependent variable in period 2.

Table 4: Gender, $t = 2$

Variables	school	male
school	1.000	
male	-0.100	1.000

Table 5: College parent & Distance $t = 2$

Variables	school	parentcollege	Variables	school	dist
school	1.000		school	1.000	
parentcollege	0.246	1.000	dist	-0.020	1.000

Table 6: Ability scores, $t = 2$

Variables	school	math	Variables	school	language
school	1.000		school	1.000	
math	0.420	1.000	language	0.437	1.000

2 Static Models of Choice

2.1 a,b)

In this section, Both Q2.a and Q2.b are discussed together and the index for period is suppressed. Let $J = \{0, 1\}$ be the alternative set where $j = 1$ indicates "attending school" and $j = 0$ indicates "not attending school". We assume that the utility function is specified as follows:

$$u_{ijt} = x_i\beta + \epsilon_{ijt} \tag{1}$$

where u_{ijt} is the utility of student i obtained from alternative j in period t, x_i is the vector of observed student characteristics, and ϵ_{ijt} captures idiosyncratic taste shock term. In Question 2, we assume that ϵ_{ijt} is drawn from *iid* extreme value distribution and students only differ in their taste shock (homogenous preferences : $\beta_i = \beta, \forall i$). This allow us to employ a simple multinomial logit estimation. Here we include x_is $male, parentcollege, ability\_math, ability\_language$, and *dist* is included iff $period = 2$. We assume that the graduation of high school can be awarded to students who attended high school only.

To predict the schooling decision in each period and the probability to get a high school degree, a simple multinomial (binary) logit model is used by *mlogit* in Stata and the dependent variables are *school* in each period and *degree*. The utility of not attending school is set as zero in each period, hence $u_{i0t} = \epsilon_{i0t}$.

2.2 c)

The estimation result is summarized in Table 7. All estimators are statistically significant except for distance (significant at 20% level) in column 2. The sign of parameters coincide with our expectation from Q1: (1) Only *male* and *dist* are negatively correlated with the dependent variable. (2) The size of ability parameters are larger in period 2 than in period 1. To summarize, students with higher abilities or college parents tend to prefer to go to school more than those with lower abilities or no college parents. Male students tend to prefer going to school less than female students.

Column 3 of Table 7 shows the estimation result for the probability of getting a high school degree. Again, the parameter for male is the only negative one, suggesting that male students are less likely to complete high school than female.

Table 7: Schooling decisions and high school graduation

	$(t = 1)$ school	$(t = 2)$ school	(high) degree
attend			
male	-0.949***	-0.634***	-0.481***
	(0.111)	(0.0866)	(0.0971)
parentcollege	0.958***	0.970***	1.050***
	(0.190)	(0.110)	(0.160)
ability_math	0.525***	0.795***	0.645***
	(0.0807)	(0.0743)	(0.0750)
ability_language	0.706***	0.939***	0.900***
	(0.0812)	(0.0754)	(0.0764)
dist		-0.0107	
		(0.00839)	
_cons	3.097***	1.126***	2.217***
	(0.104)	(0.0955)	(0.0812)
N	5158	4012	4669

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

2.3 d)

Recall that the utility function is imposed as (1), and let β_p, β_d be the parameter of $parentcollege, dist$ respectively. As the scale of utility does not matter in our estimation, we can normalize the utility in terms of kilometers by dividing β_d on the bothside of (1). Thus, β_p/β_d can be interpreted as the extra utility gain in kilometers by having college parents. The empirical counterpart is $\hat{\beta_p}/\hat{\beta_d}$ and the value is 90.45 km[1]. This means that not having college parents is equivalent to being 90.45 km further away from the nearest college.

2.4 e,f)

To predict the counterfactual college enrolment, we used the estimated primitives of period 2 in Section 2.c. For both counterfactuals, there is no effect on high school enrolment or high

[1] This assumes that the normalized coefficient of distance is -1.

school graduation rate due to the static nature of the choices. The result of estimations (by *invlogit*) are shown in Table 8. The first row of Table 8 is the actual conditional enrolment rate in our data, the second is the predicted conditional enrolment rate using our model. The second column shows the percentage difference in rate of enrolment between the predicted value and the counterfactual. The enrolment rate in Counterfactual 1 can be estimated by setting $dist = 0, \forall i$.

In the case of Counterfactual 2, as only 50% of students with a high school degree can enter college (randomly) upon choosing to apply, the expected utility of a student entering a college in period 2 is $0.5u_{i12} + 0.5u_{i01} = 0.5u_{i12}$. This is equivalent to the situation where the utility of entering a college is halved for all high school graduates in period 2. In general, we run the counterfactuals assuming that the errors that determine decisions will be redrawn for a population of students that have the same characteristics as the data.

Table 8: Static Model: College Enrolment Rate

	College Enrolment	Change (%)
Actual	0.748	-
Predicted	0.748	-
Counter 1	0.760	1.6
Counter 2 (Incentives)	0.663	-11.4
Counter 2 (Incentives + Random Entry)	0.332	-55.6

The result is not surprising at all. Since the distance is an obstacle for going to college, we can see that the enrolment rate increases by 1.6% in Counterfactual 1. On the other hand, the rate in Counterfactual 2 decreases by 11.4% as the utility of entering college is halved, and by 55.6% when we take into account both the incentive changes and the mechanical effect of the policy change.

3 Dynamic Model

All our estimation results for question 3-5 are presented in section 6.

3.1 a, b) Model Setup

We posit the following 2-period model. In period 1, a student decides whether or not to go to high school or drop out based on his 1st period utility for high school vs an alternative (for which we normalize its utility to 0) and the expected value of the option of going to college.

$$u_{i11} + \beta V_2 + \epsilon_{i11} > u_{i01} + \epsilon_{i01} \qquad \text{(Period 1)}$$

The 2nd period value function here depends on the probability of graduating high school and of entering college, p_{hi} and p_c respectively, and maximum of the two period 2 alternatives, going to college or not.

$$V_2 = p_{hi} * \max\{p_c * v_{i12} + ((1 - p_c) * v_{i02}), v_{i02}\}$$

In the 2nd period, a high school graduate chooses to go to college if

$$u_{i12} + \epsilon_{i12} > u_{i02} + \epsilon_{i02} \qquad \text{(Period 2)}$$

We assume that between all stages, the errors have *iid* extreme value distributions. In addition, since the model terminates in period 2, we have $v_{ij2} = u_{ij2}$, and we normalize $u_{i02} = 0$

Here, from the logsum formula, we have the following expression for the value function

$$\begin{aligned} V_2 &= p_{hi}(\gamma + \ln \sum_{j'} exp(p_{j'} u_{ij'2})) \\ &= p_{hi}(\gamma + \ln(1 + exp(p_c * u_{i12}))) \end{aligned}$$

with the normalization of $u_{i02} = 0$. With this, we can estimate the model by two stages. We start with the 2nd period, where we assume that the deterministic component of the utility of going to college is

$$u_{i12} = X_i * \theta_2$$

and $u_{i02} = 0$. From there, we go back to the 1st stage and form the 1st period choice of going to college as

$$u_{i11} + \beta p_{hi}(\gamma + \ln(1 + exp(p_c * X_i * \theta 2))) + \epsilon_{i11} > \epsilon_{i01}$$

with $u_{i11} = X_i * \theta_1$. We estimate this equation in one step using the maximum likelihood estimator, taking p_{hi} from the static model and $p_c = 1$ for the baseline case.

4 2 Period CCP

Applying the CCP to the 2nd period, we get the value function in period 2 as

$$V_2 = p_{hi}(\gamma - \ln(Pr(Dropout_i)))$$

where we will estimate the dropout probability of an individual with a logit model,

$$Pr(Dropout_i) = \frac{1}{1 + exp(\theta_2)}.$$

Next, the student decides to go to high school in period 1 if

$$X_i * \theta_1 + \beta p_{hi}(\gamma - \ln(\frac{1}{1 + exp(X_i * \theta_2)})) + \epsilon_{i11} > \epsilon_{i01}$$

from which we can estimate the period 1 high school utility parameters. This is very much the same as the structural model, but in two steps instead of one. Note that the utility of dropout is normalized to zero for all period.

5 Multi Period CCP

In this section, we answer what happens if the students' decisions continue after period 2, such that in every period that a student is in college, they have the option to exit to the labor market and never return. Here, we make use of the decision to exit the program being a terminal action.

Adopting the convention of the CCP, we have just as in the 2 period case, the value function associated with the choice to enter college or take the terminal action to dropout after a high school degree.

$$V_2 = p_{hi}(\gamma - \ln(\frac{1}{1 + exp(X_i * \theta_2)})).$$

This means that just as before, we estimate the following decision,

$$X_i * \theta_1 + \beta p_{hi}(\gamma - \ln(\frac{1}{1 + exp(X_i * \theta_2)})) + \epsilon_{i11} > \epsilon_{i01}.$$

The structural estimation, CCP with 2 periods and the multi period CCP estimates for the first stage coincide for this decision process. As a result, we do not need more data.

6 Results for Q3,4,5

In Table 9, we observe the updated period 1 estimates where students are now forward looking (estimates do not change in period 2 since there is no forward looking element in that period, and the result for Q3.c is the same as in Q2.c). We see that in comparison to the static model, there is a general trend towards zero for all coefficient on our parameters. This is due to the fact that the forward looking element of the utility function helps explain much of the variation in choices so that period 1 parameters do not matter as much.

Table 9: Schooling decision in period 1 : Dynamic model

	(1) deci_1
male	-0.691***
	(0.111)
parentcollege	0.315
	(0.191)
ability_math	0.236**
	(0.0803)
ability_language	0.340***
	(0.0816)
_cons	1.988***
	(0.108)
N	5158

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

In Table 10, 11 and 12, we explore the counterfactual experiments of Q2 in the dynamic model. Note that here, high school graduation rate and college enrolment rates are based on the whole population (unconditional) in period 1 and not conditional on attending high school or of graduating high school respectively.

In Table 10, we see that the dynamic model performs slightly worse than the static model in matching the actual high school enrolment in the data. However, this dynamic model in contrast to the static one, does respond to policy changes that affect period 2 choices since students are forward looking. As expected, we see that setting distance to zero increases the enrolment to high school while making college entry random (50% entry) decreases it.

In Table 11 we see that in this dynamic model, an element of selection on observables can be seen. The percentage change in counterfactual 1 for high school graduation rate is lower than

Table 10: Dynamic Model: High School Enrolment Rate

Period 1	Enrolment	Change (%)
Actual	.9051958	-
Static	.9051958	-
Predicted	.8567566	-
Counter 1	.8622052	.63595658
Counter 2 (Incentives)	.8548568	-.22174326

that of high school enrolment. This implies that when we take out distance as a negative factor on the utility of going to college, more people who were less likely to graduate high school have selected into high school. For the second counterfactual, we see that the negative response of high school graduation to random college entry is much greater than the effect in enrolment to high school. This too suggests a selection effect where who that value college most according to the model are those who are much more likely to be able to graduate high school.

Table 11: Dynamic Model: High School Graduation Rate

Period 1	Graduation	Change (%)
Actual	.7778209	-
Static	.7783271	-
Predicted	.7481094	-
Counter 1	.7521326	.53778231
Counter 2 (Incentives)	.7398048	-1.1100783

In Table 12, we present the counterfactuals of the static model and the dynamic model on unconditional period 1 college enrolment rates, with the percentage change in the counterfactuals being considered with respect to their static and dynamic counterparts. We see that although in counterfactual 1, distance going to zero increases college enrolment by more than the static case, in counterfactual 2, there only a slight difference between the static and the dynamic model. This is likely because the policy changes are on period 2 values and these are taken into account by the static model already. The additional effects on enrolment are only due to selection effects on the population of the people that are given the choice and is likely to only be second order.

Table 12: Dynamic Model: College Enrolment Rate

Period 2	Enrolment	Change (%)
Actual	.5818147	-
Static	.5848413	-
Static C1	.5922242	1.2623766
Static C2 (Incentives)	.5064415	-13.405312
Static C2' (Incentives + Random Entry)	.25322075	-56.702656
Predicted	.5714161	-
C1	.5827899	1.9904763
C2 (Incentives)	.4952484	-13.329623
C2' (Incentives + Random Entry)	.2476242	-56.664812

Finally, for completeness, we provide the estimted coefficients in the CCP model for Q4,5 below.

Table 13: Schooling decision in period 1 : CCP model

	(1) deci_1
male	-0.691***
	(0.111)
parentcollege	0.315
	(0.191)
ability_math	0.236**
	(0.0803)
ability_language	0.340***
	(0.0816)
_cons	1.988***
	(0.108)
N	5158

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$