Advanced EEE Homework

Bin Cheng

January 2022

Question 1: Descriptive Statistics

Mean Statistics

First, we conclude the accuracy of the dataset by the fact that mean of all major characteristics of students remain the same across two time periods.

Second, we find that 90.5% of students go to high school and 58.2% decided to attend college in the next period. And 77.8% of students successfully earned their high school degree at the end of first period.

	(t=1)	(t=2)
	Mean	Mean
male	0.495541	0.495541
ability_math	0.044397	0.044397
ability_language	0.024622	0.024622
parentcollege	0.278790	0.278790
school	0.905196	0.581815
degree	0.777821	0.777821
dist		8.519775
N	5158	5158

Table 1: Mean

Correlation Statistics: Period 1

We find that male students tend to avoid schooling and be more likely to drop out without a degree. In addition, having excellent performances in math and language will improve the chance of gaining a degree. Also, with parents attending college in the past, students are more likely to attend high school and successfully earn a degree at the end.

			(t=1)			
	male	ability_math	ability_language	parentcollege	school	degree
male	1.000					
ability_math	0.012	1.000				
ability_language	-0.076	0.735	1.000			
parentcollege	-0.013	0.310	0.337	1.000		
school	-0.132	0.288	0.308	0.151	1.000	
degree	-0.137	0.445	0.473	0.244	0.606	1.000

Table 2: Correlation

Correlation Statistics: Period 2

The main insights from first period still hold, but about attend college not high school as before. And we find that distance to the closest higher education institution affect negatively students' incentive to attend college.

	t=2						
	male	ability_math	ability_language	parentcollege	school	degree	dist
male	1.000						
ability_math	0.012	1.000					
ability_language	-0.076	0.735	1.000				
parentcollege	-0.013	0.310	0.337	1.000			
school	-0.155	0.535	0.562	0.332	1.000		
degree	-0.137	0.445	0.473	0.244	0.630	1.000	
dist	0.060	0.016	0.013	-0.075	-0.025	-0.017	1.000

Table 3: Correlation

Simple Linear Regression

To give a more systematic view of previous observations, we run a naive linear regression of schooling decision and degree obtaining on major characteristics.

It turns out that a female with parents attending college before, and excellent in math and language will be more likely to attend high school and earn degree at the end. If additionally, such a female living close to closest higher education institution, the chance that she will attend college increase as well.

	HighSchool	HighSchoolDegree	College
male	-0.0701***	-0.0506***	-0.0967***
	(0.00772)	(0.00928)	(0.0121)
ability_math	0.0481***	0.0782***	0.129***
	(0.00641)	(0.00779)	(0.0104)
ability_language	0.0552***	0.0907***	0.128***
	(0.00615)	(0.00744)	(0.00991)
parentcollege	0.0292**	0.0505***	0.109***
	(0.00912)	(0.0106)	(0.0132)
dist			-0.00164
			(0.00116)
_cons	0.928***	0.847***	0.703***
	(0.00595)	(0.00703)	(0.0134)
N	5158	4669	4012

Standard errors in parentheses

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

Table 4: Simple Linear Regression

Question 2: Static Model

Question 2.a: Assumption and Step

We write the utility function of a student as follows:

$$U_{ijt} = u_{ijt} + \epsilon_{ijt} = s_{it}\alpha_j + \epsilon_{ijt}, \quad j \in J = \{0, 1\} \tag{1}$$

where s_{it} represent the student characteristics at period t.

In period 1, s_{i1} only include identity-related characteristics "male, ability_math, ability_language, parentcollege". But in period 2, s_{i2} also include other characteristics as "dist". $J = \{0, 1\}$ is interpreted as the choice set of "attend school" and "drop out".

We assume ϵ_{ijt} to be i, i, d draws from extreme value distribution, called taste shock. And we further impose rationality on observed choices in data as follows:

$$d_{it} = \arg\max_{j \in \{0,1\}} \{u_{ijt} + \epsilon_{ijt}\} \tag{2}$$

which, with extreme value assumption, leads to

$$Pr(d_{it} = 1|s_{it}) = \frac{exp(u_{i1t})}{\sum_{j'=0,1} exp(u_{ij't})} \tag{3}$$

We also normalize the utility of "drop out" to be $u_{i0t} = 0$ and identify the difference in parameters later. In other words, $\alpha = \alpha_1 - \alpha_0$ will be identified and we write $u_{i1t} = s_{it}\alpha$.

Question 2.b: Estimation

Inspired by the binary nature of choice set, we employ *logit* in *Stata* to estimate our parameters and run Logistic regression of dependent variables like "Attending High School", "Obtaining Degree" and "Attending College" on characteristics.

Question 2.c: Estimated Parameters

We summarize the results in Table 5, where only distance is not significant.

It is no surprise that the sign of parameters are consistent with our insight from correlation statistics and a simple linear regression. Again, only being male and distance to closest higher education institution have negative impact on schooling decisions and degree obtaining. In addition, having parents attending college before and excellent performances in math and language all improve the probability of attending school and obtaining a degree successfully.

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

Standard errors in parentheses

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

Table 5: Logistic Regression

Question 2.d: Impact

In period 2, we define the parameters of "parentcollege" and "dist" as α_{pc} and α_{dist}. Then we compute $|\alpha_{pc}/\alpha_{dist}|$, which is 90.45. And we interpret this ratio as the utility gain from having parents attending college

measured in kilometers, which means having parents attending college is equivalent to living 90.45 kilometers closer to the nearest higher education institution.

Question 2.e&f: Counterfactual Analysis

We used the estimated parameters in the third column of Table 5.

In addition, we define college enrollment rate as the product of high school attending rate, degree obtaining rate and college attending rate.

In counterfactual 1, we ignore the negative impact from distance and conclude that college enrollment rate increase by 0.9788% on average, which is not so significant as the increase is around 1.68% when measured in benchmark case. And the high school enrollment rate is unaffected and remain 90.5196%.

However, since we halve the probability of entering college in counterfactual 2, the college enrollment rate decrease enormously by 29.0976% on average, which is of course 50% change when measured in benchmark case. And the high school enrollment rate is also unaffected and remain 90.5196%.

	t=1
Benchmark Rate	0.581953
Counterfactual 1 Rate	0.591741
Counterfactual 2 Rate	0.290976
Difference in Counterfactual 1	0.009788
Difference in Counterfactual 2	-0.290976
N	5158

Table 6: College Enrollment Rate in Static Model

	t=1
Benchmark Rate	0.905196
Counterfactual 1 Rate	0.905196
Counterfactual 2 Rate	0.905196
Difference in Counterfactual 1	0.000000
Difference in Counterfactual 2	0.000000
N	5158

Table 7: High School Enrollment Rate in Static Model

Question 3: Dynamic Model

Question 3.a: Assumption and Step

With the same choice set $J = \{0, 1\}$, we define the conditional value function in period 1

$$v_{ij1} = u_{ij1} + \beta E[u^*_{i2} + \epsilon^*_{i2} | d_{it} = j] \tag{4}$$

where $u^*_{i2} + \epsilon^*_{i2}$ is the utility of the chosen alternative of i in period 2.

And we also write in period 2

$$v_{ij2} = u_{ij2} = \begin{cases} s_{it}\alpha, & \text{if } j = 1 \\ 0, & \text{if } j = 0 \end{cases} \tag{5}$$

Again, we assume ϵ_{ijt} are i, i, d draws from extreme value distribution and impose rationality on choice in data that

$$d_{it} = \arg\max_{j \in \{0,1\}} \{v_{ijt} + \epsilon_{ijt}\} \tag{6}$$

which, with extreme value assumption, leads to

$$Pr(d_{it} = 1 | s_{it}) = \frac{exp(v_{i1t})}{\sum_{j'=0,1} exp(v_{ij't})} \tag{7}$$

We specify $E[u^*_{i2} + \epsilon^*_{i2}|d_{it} = j]$ as follows

$$E[u^*_{i2} + \epsilon^*_{i2}|d_{it} = j] = Pr(degree = 1|s_{i1})\bar{V}_2(s_{i2}) + Pr(degree = 0|s_{i1})\bar{V}_2(s'_{i2}) \tag{8}$$

where

$$\bar{V}_2(s_{i2}) = \gamma + \ln \sum_{j'} exp(v_{ij'2}) \tag{9}$$

$$= \gamma + \ln \sum_{j'} exp(u_{ij'2}) \tag{10}$$

$$\bar{V}_2(s'_{i2}) = \gamma + \ln exp(0) \tag{11}$$

since without a degree, a student can't attend college and derive 0 utility.

Question 3.b: Estimation

To estimate parameters with forward looking, we use (6) in period 1 and run a logistic regression of attending high school on characteristics s_{i1} and the discounted expected utility in period 2.

Question 3.c: Estimated Parameters

We summarize the results in the first column in Table 10, where only "parentcollege" is not significant.

The sign of parameters are still consistent with logistic regression in static model. But the scale of parameters decreased greatly. For example, the coefficient of "ability_math" is 0.525 in static model but only 0.236 in dynamic model.

This finding support the validity of forward looking behavior of students, since our expected utility in period 2 explain much variation and therefore decrease the magnitude of characteristics.

Question 3.d: Impact

As period 2 is the last period, our estimate is still the same as 90.45 in kilometers.

Question 3.e&f: Counterfactual Analysis

We used the estimated parameters in the first column of Table 10.

In counterfactual 1, we ignore the negative impact from distance and conclude that college enrollment rate increase by 1.0799% on average, which is not so significant as the increase is around 1.85% when measured in benchmark case. And the high school enrollment rate is affected by forward looking behavior and increase by 0.2052%.

However, since we halve the probability of entering college in counterfactual 2, the college enrollment rate decrease enormously by 29.0976% on average, which is of course 50% change when measured in benchmark case. And the high school enrollment rate is affected by forward looking behavior and decrease by 2.4698%.

	t=1
Benchmark Rate	0.582533
Counterfactual 1 Rate	0.593332
Counterfactual 2 Rate	0.283515
Difference in Counterfactual 1	0.010799
Difference in Counterfactual 2	-0.299018
N	5158

Table 8: College Enrollment Rate in Dynamic Model

	t=1
Benchmark Rate	0.905196
Counterfactual 1 Rate	0.907247
Counterfactual 2 Rate	0.880498
Difference in Counterfactual 1	0.002052
Difference in Counterfactual 2	-0.024698
N	5158

Table 9: High School Enrollment Rate in Dynamic Model

Question 4: CCP Estimation

We propose a new method to compute $\bar{V}_2(s_{i2}), \bar{V}_2(s'_{i2})$ as

$$\bar{V}_2(s_{i2}) = \gamma + v_{i02} - \ln Pr(d_{i2} = 0|s_{i2}) \tag{12}$$
$$\bar{V}_2(s'_{i2}) = \gamma + v_{i02} - \ln exp(0)/exp(0) \tag{13}$$

since without a degree, a student can only dropout from college and derive 0 utility.

In addition, since period 2 is the last period, we can use parameters estimated in static model without considering forward looking behaviors.

The parameters are summarized in the second column of Table 10.

	Structural Estimation	CCP Estimation
school		
male	-0.691***	-0.691***
	(0.111)	(0.111)
ability_math	0.236**	0.236**
	(0.0803)	(0.0803)
ability_language	0.340***	0.340***
	(0.0816)	(0.0816)
parentcollege	0.315	0.315
	(0.191)	(0.191)
_cons	1.440***	1.440***
	(0.108)	(0.108)
N	5158	5158

Standard errors in parentheses

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

Table 10: Parameters in Dynamic Model and CCP Estimation

Question 5: CCP Estimation with More Periods

Since leaving college is terminal, we still have the conditional value function in the terminal period where choices between attending college and dropout are made.

In other words, we can still write $\bar{V}_T(s_{iT})$ as

$$\bar{V}_T(s_{iT}) = \gamma + v_{i0T} - \ln Pr(d_{iT} = 0|s_{iT}) \tag{14}$$

where $Pr(d_{iT} = 0|s_{iT})$ is the same as $Pr(d_{i2} = 0|s_{i2})$ in our 2 period model.

So high school utility and estimates will be the same. And we don't need more data.