

# CE 264 Problem Set 3: Specification, Estimation, and Testing of Multinomial Choice Models

Franklin Zhao (3033030808)

Junzhe Shi (3033030938)

22 Feb. 2018

## Part 1 Model Development

### Question 1:

The utility equations are shown in Equation (1)(2)(3). Note that variables are in the same order as shown in Table 1.

$$U_{Train} = \beta_1 \times X_1 + \beta_2 \times X_2 + \beta_3 \times X_3 + \beta_5 \times X_5 \quad (1)$$

$$U_{Swissmetro} = \beta_1 \times X_1 + \beta_3 \times X_3 + \beta_4 \times X_4 + \beta_5 \times X_5 + \beta_6 \times X_6 + \beta_7 \times X_7 + \beta_8 \times X_8 \quad (2)$$

$$U_{Car} = \beta_0 + \beta_1 \times X_1 \quad (3)$$

Table 1: Variables for our “best” model

Variables description	Parameter estimation	Standard error	T-stats	P-values	Utility equation
Alternative specific constant (Car)	-0.6292	0.081	-7.727	0.000	(3)- $\beta_0$
Travel Time, units:hrs (Train and Car)	-0.8615	0.028	-30.894	0.000	(1)(3)- $X_1$
Travel Time, units:hrs (Swissmetro)	-0.7713	0.037	-20.634	0.000	(2)- $X_1$
Headway, units:hrs (Train)	-0.3619	0.047	-7.697	0.000	(1)- $X_2$
Travel Cost, units:CHF (Train)	-0.0172	0.001	-24.068	0.000	(1)- $X_3$
Travel Cost, units:CHF (Swissmetro)	-0.0097	0.000	-23.695	0.000	(2)- $X_3$
Traveler's income threshold, 0=more than 1000 CHF per year, 1=otherwise	-0.7056	0.096	-7.367	0.000	(2)- $X_4$
Measure of luggage, 0=none, 1=one piece, 3=several pieces (Train)	0.5064	0.057	8.852	0.000	(1)- $X_5$
Measure of luggage, 0=none, 1=one piece, 3=several pieces (Swissmetro)	0.1592	0.045	3.542	0.000	(2)- $X_5$
First class traveler, 0=no, 1=yes	0.2241	0.048	4.715	0.000	(2)- $X_6$
Who paid for the ticket, 0=not known, 1=self, 2=employer, 3 = half-half	0.3447	0.034	10.172	0.000	(2)- $X_7$
Whether the ticket is free, 0=no, 1=yes	1.6154	0.171	9.424	0.000	(2)- $X_8$

The regression results are shown in Table 2.

Table 2: Regression results

<b>Dep. Variable:</b>	CHOICE	<b>No. Observations:</b>	10,719
<b>Model:</b>	Multinomial Logit Model	<b>Df Residuals:</b>	10,707
<b>Method:</b>	MLE	<b>Df Model:</b>	12
<b>Date:</b>	Thu, 22 Feb 2018	<b>Pseudo R-squ.:</b>	0.253
<b>Time:</b>	14:31:08	<b>Pseudo R-bar-squ.:</b>	0.252
<b>AIC:</b>	16,600.296	<b>Log-Likelihood:</b>	-8,288.148
<b>BIC:</b>	16,687.653	<b>LL-Null:</b>	-11,093.627

### Question 2:

Apart from the variables in the final model, we expected that “travel purpose” and the “dummy variables” for train, Swissmetro and car should also be appeared in the model, where the coefficient of “travel purpose” should be a negative number if we put it into the train utility equation. However, when adding “travel purpose”, the coefficient turned out to be really small, and the final log-likelihood went down again. As for the “dummy variables”, the coefficients of which were not large either, and they also had negative impact on the final log-likelihood. Also, they did not pass the t-stats test, so we had to ignore them.

### Question 3:

Yes, it does generally. For example, people are likely to choose the mode with less travel time and cost. Also, people love to choose the mode if the ticket is free. If the decision maker had more luggages, he/she would prefer to choose the train mode. Finally, low-income people are more likely to choose cheap modes generally.

### Question 4:

From Table 1 we can find out that “whether free ticket” variable has the most significant impact on people’s mode choice. It really makes sense. Who does not prefer a free ticket? Also, travel time and cost, income, and luggage variables, the coefficient absolute value of which are all above 0.5, also have great impacts on the mode choice. It also makes sense since again, people are likely to choose the mode with less travel time and cost. Train mode is preferred if people have more luggages.

### Question 5:

The biggest issue we came accross was that if we put two alternative specific constants (ASC), one of them would always fail the t-stats test, no matter train, Swissmetro, or Car. We thought it should be correct that two ASCs are in the model, but it turned out things did not work out

(p-value was about 0.5 so must be ignored). Other issues include encountered singular matrices problems when choosing inappropriate variables, and tried really hard to improve the log-likelihood.

### **Question 6:**

We used t-stats test to decide whether we should include a variable or not, even if the coefficients of which is large. Taking the 5% significant value, for p-values that are less than 0.05, we would reject the null hypothesis and include the corresponding variable; for p-values that are really large (i.e., 0.15 or more), we would definitely exclude those variables no matter how significant their impacts are; for p-values that are slightly exceed the 0.05 cut-off value, we would think about it considering their impacts on the general mode choice.

### **Question 7:**

For example, the “travel cost” variable should be alternative-specific. The test was simple, just set the variable to be generic and alternative-specific separately, and execute the code to see the result of the log-likelihood. When we implemented this, it turned out that the log-likelihood is much higher when we set the “travel cost” to be an alternative-specific variable. Together with other tests, we concluded that the use of some alternative-specific coefficients is justified for improving the model.

## PyLogit Script for the best model specification:

```
1 ## Our Best Model
2 alt_varying_variables = {u'travel_time': dict([(1, 'TRAIN_TT'),
3                                               (2, 'SM_TT'),
4                                               (3, 'CAR_TT')]),
5                          u'travel_cost': dict([(1, 'TRAIN_CO'),
6                                               (2, 'SM_CO'),
7                                               (3, 'CAR_CO')]),
8                          u'headway': dict([(1, 'TRAIN_FR'),
9                                             (2, 'SM_FR')]),
10                         u'INCOME': dict([(2, 'INCOME')]),
11                         u'seat_configuration': dict([(2, "SM_SEATS")]),
12                         u'LUGGAGE': dict([(2, 'LUGGAGE')]),
13                         u'GA': dict([(2, 'GA')]),
14                         u'WHO': dict([(2, 'WHO')]),
15                         u'FIRST': dict([(2, 'FIRST')])
16                      }
17
18 # Modify the "travel cost" variable by consider the "GA" factor
19 # ONLY consider FREE or NOT in the "TICKET" variable
20 # Set the income threshold to 1K CHF since we think it doesn't makes much difference
   when one's income is above the thre
21 long_swiss_metro["travel_cost_modified"] = (long_swiss_metro['GA']==0) *
   long_swiss_metro['travel_cost']
22 long_swiss_metro["free_ticket"] = long_swiss_metro['TICKET'] == 8
23 long_swiss_metro["free_ticket"] = long_swiss_metro["free_ticket"] * 1
24 long_swiss_metro["income_threshold"] = long_swiss_metro['INCOME'] > 0
25 long_swiss_metro["income_threshold"] = long_swiss_metro['income_threshold'] * 1
26
27 basic_specification = OrderedDict()
28 basic_names = OrderedDict()
29
30 basic_specification["intercept"] = [3]
31 basic_names["intercept"] = ['ASC Car']
32
33 basic_specification["travel_time_hrs"] = [[1, 3], 2]
34 basic_names["travel_time_hrs"] = ['Travel Time, units:hrs (Train and Car)',
35                                   'Travel Time, units:hrs (Swissmetro)']
36
37 basic_specification["headway_hrs"] = [1]
38 basic_names["headway_hrs"] = ["Headway, units:hrs (Train)"]
39
40
41 basic_specification["travel_cost_modified"] = [1, 2]
42 basic_names["travel_cost_modified"] = ["Travel Cost, units:CHF (Train)",
43                                       "Travel Cost, units:CHF (Swissmetro)"]
44
45 basic_specification["income_threshold"] = [2]
```

```

46 basic_names["income_threshold"] = ["Traveler's income threshold, 0=more than 1000
    CHF per year, 1=otherwise"]
47
48 basic_specification["LUGGAGE"] = [1, 2]
49 basic_names["LUGGAGE"] = ["Measure of luggage, 0=none, 1=one piece, 3=several pieces
    (Train)",
50                             "Measure of luggage, 0=none, 1=one piece, 3=several pieces
    (Swissmetro)"]
51
52 basic_specification["FIRST"] = [2]
53 basic_names["FIRST"] = ['First class traveler, 0=no, 1=yes']
54
55 basic_specification["WHO"] = [2]
56 basic_names["WHO"] = ['Who paid for the ticket, 0=not known, 1=self, 2=employer, 3 =
    half-half']
57
58 basic_specification["free_ticket"] = [2]
59 basic_names["free_ticket"] = ['Whether the ticket is free, 0=no, 1=yes']
60
61 swissmetro_mnl = pl.create_choice_model(data=long-swiss-metro,
62                                         alt_id_col=custom_alt_id,
63                                         obs_id_col=obs_id_column,
64                                         choice_col=choice_column,
65                                         specification=basic_specification,
66                                         model_type="MNL",
67                                         names=basic_names)
68
69 # Specify the initial values and method for the optimization.
70 swissmetro_mnl.fit_mle(np.zeros(12))
71 bestModel = swissmetro_mnl.get_statsmodels_summary()
72 bestModel
73
74 bestModel.tables[1].as_latex_tabular()
75 bestModel.tables[0].as_latex_tabular()

```

Log-likelihood at zero: -11,093.6273

Initial Log-likelihood: -11,093.6273

Estimation Time for Point Estimation: 0.20 seconds.

Final log-likelihood: -8,288.1096

## Part 2 Research Project

### Question 1:

First let us frame the research problem in terms of a multinomial choice, and be specific about the transportation mode. Let us say, whether driving costs at Berkeley will influence a Berkeley's graduate student's choice on driving, taking public transit or riding a bike to campus everyday, assuming the distance is fair enough and there is no other transportation mode (e.g., no BART, and too far to walk). Then the dependent variables will be  $\{B, C, T\}$ , where  $B$  is to choose riding a bike,  $C$  is to choose driving a car, and  $T$  is to choose taking public transit. Hence, we now change the problem from a binary choice problem to a multinomial choice problem.

### Question 2:

#### Variables specification:

**X<sub>1</sub>:** parking fee, which can be specified as money spent on parking every month.

**X<sub>2</sub>:** fuel cost, similarly can be specified as money spent on car fuel every month.

**X<sub>3</sub>:** bike cost (money spent on the bike).

**X<sub>4</sub>:** public transit cost, specified as money spent in taking public transit every month.

**X<sub>5</sub>:** maintenance cost. The cost for car maintenance every month (parking and fuel are not included).

**X<sub>6</sub>:** time cost. The average time spent getting to the campus everyday.

**X<sub>7</sub>:** traffic condition, which can be specified as the average traffic density of the regular route to campus.

**X<sub>8</sub>:** hills, which can be specified as the length of the road that has a slope angle greater than 10°.

**X<sub>9</sub>:** characteristics, which can be specified as the level of patience (0–5; “5” is the most patient).

**X<sub>10</sub>:** financial status (0–5; “5” has the best financial status).

**X<sub>11</sub>:** physical condition (0–5; “5” has the best physical condition).

### Question 3:

The utilities of the three alternatives are shown in Equation (4)

$$\begin{aligned} U_B &= \beta_B + \beta_3 X_3 + \beta_6 X_6 + \beta_8 X_8 + \beta_9 X_9 \\ U_C &= \beta_1 X_1 + \beta_2 X_2 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_{10} X_{10} + \beta_{11} X_{11} \\ U_T &= \beta_T + \beta_4 X_4 + \beta_6 X_6 + \beta_7 X_7 \end{aligned} \tag{4}$$

where  $U_B$ ,  $U_C$ , and  $U_T$  are utilities of bike, car, and transit;  $X$ s are the independent variables specified in Question 2;  $\beta$ s are the corresponding coefficients (model parameters).

**Question 4:**

$\beta_{\mathbf{B}}$  and  $\beta_{\mathbf{T}}$  are alternative specific constants, which capture the difference of two utilities when all else are equal.

$\beta_{\mathbf{1}}$  is a negative number.

$\beta_{\mathbf{2}}$  is a negative number.

$\beta_{\mathbf{3}}$  is a negative number.

$\beta_{\mathbf{4}}$  is a negative number.

$\beta_{\mathbf{5}}$  is a negative number.

$\beta_{\mathbf{6}}$  is a negative number.

$\beta_{\mathbf{7}}$  is a negative number.

$\beta_{\mathbf{8}}$  is a negative number.

$\beta_{\mathbf{9}}$  is a positive number.

$\beta_{\mathbf{10}}$  is a positive number.

$\beta_{\mathbf{11}}$  is a positive number

**Question 5:**

For our research question, we can use the hypothesis testing. The null hypothesis will be that the driving costs will not influence a Berkeley's graduate student's mode choice, and the alternative hypothesis will be the opposite. Then we use t-stats method to test coefficients  $\beta_1$ ,  $\beta_2$ , and  $\beta_5$  (which corresponds to the driving cost variables  $X_1$ ,  $X_2$ , and  $X_5$ ). If we reject the null hypothesis, then the driving cost will increase the probability that a decision maker chooses riding a bike or transit rather than driving. If we fail to reject the null hypothesis, then driving cost may be an irrelevant factor, which has no impact on mode choice.

### Part 3A: Supplemental Problem1

#### Specification 1:

We cannot estimate the model, since  $1 - \text{DCITY} = \text{DSUBURB}$  which means they are linearly correlated.

#### Specification 2:

	$ASC_c$	$ASC_d$	$\beta_{tt}$	$\beta_{d,c}$	$\beta_{d,d}$	$\beta_{s,t}$
1 Carpool	1	0	$TT_1$	DCITY	0	0
2 Drive Alone	0	1	$TT_2$	0	DCITY	0
3 Transit	0	0	$TT_3$	0	0	DSUBURB

We cannot estimate the model, since  $1 - \text{DCITY} = \text{DSUBURB}$  where DCITY is a characteristic of decision makers. DCITY is restricted to  $J - 1 = 2$  alternatives.

#### Specification 3:

	$ASC_c$	$ASC_d$	$\beta_{tt}$	$\beta_{d,d}$	$\beta_{s,t}$
1 Carpool	1	0	$TT_1$	0	0
2 Drive Alone	0	1	$TT_2$	DCITY	0
3 Transit	0	0	$TT_3$	0	DSUBURB

We can estimate the model, DCITY is only restricted to  $J - 1 = 2$ .

#### Specification 4:

	$ASC_c$	$ASC_d$	$\beta_{tt}$	$\beta_{d,c}$	$\beta_{d,d}$	$\beta_{n,c}$	$\beta_{n,d}$
1 Carpool	1	0	$TT_1$	DCITY	0	A	0
2 Drive Alone	0	1	$TT_2$	0	DCITY	0	A
3 Transit	0	0	$TT_3$	0	0	0	0

We can estimate the model, both DCITY and A don't vary in different alternatives.

#### Specification 5:

	$ASC_c$	$ASC_d$	$\beta_{tt}$	$\beta_{d,c}$	$\beta_{d,d}$	$\beta_{n,t}$
1 Carpool	1	0	$TT_1$	DCITY	0	0
2 Drive Alone	0	1	$TT_2$	0	DCITY	0
3 Transit	0	0	$TT_3$	0	0	A when close transit



We cannot estimate the model. The ATRANSIT only be included when people choose transit. It means that  $\beta_{n,t}$  could be an arbitrarily large and positive number.

### Part 3B: Supplemental Problem 2

Let numbers of people choosing HP, Mac and Dell are  $N_1$ ,  $N_2$ , and  $N_3$ .  $N = N_1 + N_2 + N_3$

$$\begin{aligned}
 1. \quad L^*(\alpha) &= \left(\prod_{n=1}^{N_1} P_n(1)\right) \left(\prod_{n=1}^{N_2} P_n(2)\right) \left(\prod_{n=1}^{N_3} P_n(3)\right) \\
 &= \left(\frac{e^{\alpha_1}}{e^{\alpha_1} + e^{\alpha_2} + e^{\alpha_3}}\right)^{N_1} \left(\frac{e^{\alpha_2}}{e^{\alpha_1} + e^{\alpha_2} + e^{\alpha_3}}\right)^{N_2} \left(\frac{e^{\alpha_3}}{e^{\alpha_1} + e^{\alpha_2} + e^{\alpha_3}}\right)^{N_3} \\
 &= \frac{e^{\alpha_1 N_1 + \alpha_2 N_2 + \alpha_3 N_3}}{(e^{\alpha_1} + e^{\alpha_2} + e^{\alpha_3})^N}
 \end{aligned}$$

$$L(\alpha) = \ln(L^*(\alpha)) = 1N_1 + \alpha_2 N_2 + \alpha_3 N_3 - N \ln(e^{\alpha_1} + e^{\alpha_2} + e^{\alpha_3})$$

$$\begin{aligned}
 2. \quad \frac{\partial L}{\partial \alpha_1} &= 0 = N_1 - N \left( \frac{e^{\hat{\alpha}_1}}{e^{\hat{\alpha}_1} + e^{\hat{\alpha}_2} + e^{\hat{\alpha}_3}} \right) = N_1 - N \hat{P}_n(1) \\
 \frac{\partial L}{\partial \alpha_2} &= 0 = N_2 - N \left( \frac{e^{\hat{\alpha}_2}}{e^{\hat{\alpha}_1} + e^{\hat{\alpha}_2} + e^{\hat{\alpha}_3}} \right) = N_2 - N \hat{P}_n(2) \\
 \frac{\partial L}{\partial \alpha_3} &= 0 = N_3 - N \left( \frac{e^{\hat{\alpha}_3}}{e^{\hat{\alpha}_1} + e^{\hat{\alpha}_2} + e^{\hat{\alpha}_3}} \right) = N_3 - N \hat{P}_n(3)
 \end{aligned}$$

set  $\alpha_1 = 0$

$$\frac{P_n(1)}{P_n(2)} = \frac{N_1}{N_2} = \frac{e^{\hat{\alpha}_1}}{e^{\hat{\alpha}_2}} = \frac{1}{e^{\hat{\alpha}_2}} \Rightarrow \hat{\alpha}_2 = \ln\left(\frac{N_2}{N_1}\right) \approx 0.79$$

$$\frac{P_n(1)}{P_n(3)} = \frac{N_1}{N_3} = \frac{e^{\hat{\alpha}_1}}{e^{\hat{\alpha}_3}} = \frac{1}{e^{\hat{\alpha}_3}} \Rightarrow \hat{\alpha}_3 = \ln\left(\frac{N_3}{N_1}\right) \approx 0.37$$

$$\begin{aligned}
 3. \quad \text{Var}(\hat{\alpha}) &= -E(\nabla^2 L)^{-1} = - \begin{bmatrix} E\left(\frac{\partial^2 L}{\partial \alpha_2^2}\right) & E\left(\frac{\partial^2 L}{\partial \alpha_3 \partial \alpha_2}\right) \\ E\left(\frac{\partial^2 L}{\partial \alpha_3 \partial \alpha_2}\right) & E\left(\frac{\partial^2 L}{\partial \alpha_3^2}\right) \end{bmatrix}^{-1} \\
 E\left(\frac{\partial^2 L}{\partial \alpha_2^2}\right) &= E\left[-N \left( \frac{(e^{\alpha_1} + e^{\alpha_2} + e^{\alpha_3})e^{\alpha_2} - e^{\alpha_2}e^{\alpha_2}}{(e^{\alpha_1} + e^{\alpha_2} + e^{\alpha_3})^2} \right)\right] = E[-N(P_n(2) - P_n(2)P_n(2))] \\
 &= -N(P_n(2) - P_n(2)P_n(2)) = -N\left(\frac{N_2}{N} - \left(\frac{N_2}{N}\right)^2\right) = -112.18 \\
 E\left(\frac{\partial^2 L}{\partial \alpha_3^2}\right) &= E\left[-N \left( \frac{(e^{\alpha_1} + e^{\alpha_2} + e^{\alpha_3})e^{\alpha_3} - e^{\alpha_3}e^{\alpha_3}}{(e^{\alpha_1} + e^{\alpha_2} + e^{\alpha_3})^2} \right)\right] = E[-N(P_n(3) - P_n(3)P_n(3))] \\
 &= -N(P_n(3) - P_n(3)P_n(3)) = -N\left(\frac{N_3}{N} - \left(\frac{N_3}{N}\right)^2\right) = -96.44 \\
 E\left(\frac{\partial^2 L}{\partial \alpha_3 \partial \alpha_2}\right) &= E\left[-N \left( \frac{-e^{\alpha_2}e^{\alpha_3}}{(e^{\alpha_1} + e^{\alpha_2} + e^{\alpha_3})^2} \right)\right] = E[N(P_n(2)P_n(3))] = N(P_n(2)P_n(3)) \\
 &= N\left(\frac{N_2}{N} \frac{N_3}{N}\right) = 66.27
 \end{aligned}$$

$$\text{Var}(\hat{\alpha}) = - \begin{bmatrix} -112.18 & 66.27 \\ 66.27 & -96.44 \end{bmatrix}^{-1} = \begin{bmatrix} 0.0150 & 0.0103 \\ 0.0103 & 0.0175 \end{bmatrix}$$

The asymptotic variance of  $\hat{\alpha}_2$  is 0.0150, and standard error is 0.122.

The asymptotic variance of  $\hat{\alpha}_3$  is 0.0175, and standard error is 0.132.