CE264 – Spring 2017
Problem Set 3: Specification, Estimation and Testing of Multinomial Choice Models

Due: Tuesday, February 20, 2017

In Part 1 of this assignment, you will estimate a Multinomial Logit (MNL) model using real world data. In Part 2, you will be asked to reframe your research question as a multinomial choice. Part 3 of this assignment consists of two supplemental problems, one where you are to assess whether several different model specifications are estimable, and one where you will estimate a simple MNL model by hand (without using PyLogit).

The problem set is to be done in groups of 2 or 3 currently enrolled CE264 students with one solution being turned in by the group (i.e. choose one group member to upload the solution to bCourses). You must list the student ID number (SID) of each team member on homework submissions.

Part 1 (50 Points): Model Development

Objectives

The objective here is to further familiarize yourselves with discrete choice modeling, by specifying and estimating a Multinomial Logit (MNL) models. The modeling portion of this assignment is analogous to the binary choice assignment, but now you will work through the process in the case of more alternatives.

Dataset: Swissmetro

This dataset consists of survey data collected on the train between St. Gallen and Geneva, Switzerland, during March 1998. Each respondent provided information on their trip, including the trip purpose, amount of luggage carried, etc., as well as various socio-economic characteristics. Attribute of the alternatives were also collected and provided in the dataset. The purpose of collecting this dataset was to analyze the impact of the modal innovation in transportation, represented by the Swissmetro, a revolutionary maglev underground system, against the usual transportation modes represented by car and train. This dataset is further detailed in Appendix B.

Appendices

There are two appendices to help you with the assignment, each of which you should read.
A. Theoretical reminders
B. The Swissmetro Dataset

Report Content

Your report should be **no more than 3 pages** (no appendices) and include:

Model Presentation

1. Present the estimation results of your best model specification. This should be a cleaned up version of the PyLogit output that can be clearly understood by the reader (e.g., understandable description of the explanatory variables). The table should follow the guidelines specified in Problem Set 2 and posted again at the bottom of the page.

Discussion and analysis of your final model:

2. What were your a priori expectations for the final model in terms of the variables that you thought would be significant and the expected signs of the coefficients for those variables.
3. Does the model match your a priori expectations? Why or why not?
4. What heterogeneity have you captured or not captured?
5. What issues did you come across?
6. How did you use the statistical tests to arrive at your final model?
7. Describe at least one test on whether any of the cost (or time, if relevant) attributes should be generic or alternative-specific. Based on this test, discuss whether the use of generic or alternative-specific coefficients is justified for improving the model.

Attach a printout of the PyLogit script for the best model specification (this won't count towards the 3 page limit on the model report).

Don't forget parts 2 and 3 to the problem set.

Table Guidelines:

1. For each variable in the model, include the units and what alternative's utility equation the variable belongs to.
2. Use actual variable names (not the shorthand used in one's computer code). For example, use "travel time" instead of "tt".
3. Include the parameter estimates, t-stats, and p-values.
4. The report table should present everything that needs to be known about the model (clear variable names, choice indicator defined, dummy variables defined, clear base category(ies) for categorical variables, etc.). The table should convey all required information to understand and assess your "best" model.

Part 2 (10 Points): Research Project

In Problem Set 2, we asked you to frame your research problem in terms of a binary choice. Now we'd like you to extend it to a multinomial choice, identifying the dependent variable of interest, each of the alternatives, and all independent explanatory variables that you believe influence the choice. Specify the utility of each of the alternatives as some function of the explanatory variables, accompanied by a very brief description of your a priori expectations regarding each of the model parameters. As before, show how you are addressing the research question using the model specification.

For example, in Problem Set 2 your research problem may have been, "What is the impact of real time bus arrival information (e.g. NextBus) on transit ridership?" The way that the question was addressed was by reframing the problem as a binary choice where the dependent variable was: does an individual choose transit or not. The framework can now be extended to a multinomial model of travel mode choice, where the dependent variable is the travel mode chosen by an individual, and the alternatives are auto, transit, bike, walk and skateboard. Possible explanatory variables that influence an individual's decision may include attributes of the alternatives (travel times and costs), characteristics of the decision-maker (student at Cal), features of the trip (terrain), etc. In picking a functional form for the utility of each alternative, we employ the linear-in-parameters specification:

$$U_{mode}=\beta_1 x_{1,mode}+\beta_2 x_{2,mode}+...$$

where the x's represent the explanatory variables and the β's represent the model parameters.

As before, say we hypothesize that real-time arrival information influences transit ridership through its effect on the disutility of waiting time. Let x_1 be the waiting time when real-time arrival information is available, and zero otherwise; and similarly, let x_2 be the waiting time when real-time arrival information is NOT available, and zero otherwise. The hypothesis test will still boil down to whether β_1 is statistically different from β_2 or not. However, one of the benefits of reframing the question as a multinomial choice problem is that we are controlling for the effects of the level-of-service of other modes on transit ridership.

The specific questions to be answered in this part of the assignment are:

1. What are the dependent variables of interest in the multinomial version of your problem? In other words, what are the possible alternatives?
2. What are all of the independent variables that you think influence the choice amongst these alternatives?
3. What would the utility specifications be for each of your alternatives?
4. Briefly, what are your a priori expectations regarding each of your model parameters?
5. How is your research question addressed by your proposed model specification?

This part of the report should not be more than 3 pages.

Part 3A (20 Points): Supplemental Problem 1

You are considering estimating a logit mode choice model for three different alternatives: carpool, drive alone, and public transit. Below are five different possible model specifications. For each specification, determine whether the coefficients of the model are in fact estimable, and explain how you arrived at your conclusion.

Specification 1

$V_{carpool} = ASC_{carpool} + \beta_{tt} TT_{carpool} + \beta_{downtown} DCITY + \beta_{suburb} DSUBURB$

$V_{drivealone} = ASC_{drivealone} + \beta_{tt} TT_{drivealone}$

$V_{transit} = \beta_{tt} TT_{transit}$

where the $ASCs$ are alternative -specific constants and the β's are model coefficients , both of which need to be estimated . TT denotes travel time for each alternative , $DCITY$ is a dummy variable equal to 1 if an individual works in downtown , and $DSUBURB$ is a dummy variable equal to 1 if an individual does not work in downtown.

In general, for models with linear-in-parameters utility specifications, it helps to rewrite the utilities in the form of a specification table, shown below for Specification 1:

	$ASC_{carpool}$	$ASC_{drive\ alone}$	β_{tt}	$\beta_{downtown}$	β_{suburb}
Carpool	1	0	Carpool travel time	Dummy variable equal to 1 if individual works in downtown	Dummy variable equal to 1 if individual works in suburbs
Drive alone	0	1	Drive alone travel time	0	0
Transit	0	0	Transit travel time	0	0

where the rows of the table correspond to each of the three alternatives and the columns correspond to each of the five model parameters. The entry in any of the interior cells of the table represents the variable that the parameter corresponding to a particular column is multiplied by in the utility of the alternative corresponding to that row. These tables are clearer to read and quicker to check for identification . In writing your solution , convert each of the other four specifications into a table such as the one shown above.

Specification 2

$V_{carpool}=ASC_{carpool}+\beta_{tt}TT_{carpool}+\beta_{downtown,carpool}DCITY$

$V_{drivealone}=ASC_{drivealone}+\beta_{tt}TT_{drivealone}+\beta_{downtown,drivealone}DCITY$

$V_{transit}=\beta_{tt}TT_{transit}+\beta_{suburb,transit}DSUBURB$

where the variables hold the same definition as before.

Specification 3

$V_{carpool}=ASC_{carpool}+\beta_{tt}TT_{carpool}$

$V_{drivealone}=ASC_{drivealone}+\beta_{tt}TT_{drivealone}+\beta_{downtown,drivealone}DCITY$

$V_{transit}=\beta_{tt}TT_{transit}+\beta_{suburb,transit}DSUBURB$

where the variables hold the same definition as before.

Specification 4

$V_{carpool}=ASC_{carpool}+\beta_{tt}TT_{carpool}+\beta_{downtown,carpool}DCITY$

$+\beta_{numAuto,carpool}A$

$V_{drivealone}=ASC_{drivealone}+\beta_{tt}TT_{drivealone}+\beta_{downtown,drivealone}DCITY$

$+\beta_{numAuto,drivealone}A$

$V_{transit}=\beta_{tt}TT_{transit}$

where all the variables hold the same definition as before, and Arepresents the number of cars owned by an individual.

Specification 5

$V_{carpool}=ASC_{carpool}+\beta_{tt}TT_{carpool}+\beta_{downtown,carpool}DCITY$

$V_{drivealone}=ASC_{drivealone}+\beta_{tt}TT_{drivealone}+\beta_{downtown,drivealone}DCITY$

$V_{transit}=\beta_{tt}TT_{transit}+\beta_{numAuto,transit}ATRANSIT$

where all variables hold the same definition as before; and $ATRANSIT$represents the number of cars owned by an individual if the individual chose transit, and zero otherwise.

Part 3B (20 Points): Supplemental Problem 2

Suppose we have the following information from a sample of 450 laptop owners at UC Berkeley:

Type of Laptop	*Number of Observations*
HP	97
Mac	213
Dell	140

Given these limited data, you hypothesize that people choose laptops based on a simple multinomial model where the systematic utility of each alternative is a constant term; i.e.:

$$P_n(\text{type of laptop} = i) = \frac{e^{\alpha_i}}{\sum_j e^{\alpha_j}}$$

1. Formulate the log-likelihood function for this model.
2. Determine analytically the maximum likelihood estimator for the α's and calculate these estimates empirically using the given data. (NOTE: While there are three constants (alpha1, alpha2, and alpha3), only two of these are estimable. Therefore you should fix one of them (any one) to zero and solve for the maximum likelihood estimates for the other two.)
3. Estimate the asymptotic standard error of the estimates in question 2. (HINT: This is where you have to use the equation that is the negative of the inverse of the Hessian as defined by the Cramer-Rao lower bound.)

Appendix A: Theoretical Reminder

The Multinomial Logit Model

MNL models arise when the available choice set for the decision maker includes more than two alternatives. Writing the utility of alternative i as perceived by individual n following the usual notation:

$$U_{in} = V_{in} + \epsilon_{in}$$

The general random utility formulation gives us the following choice probability for alternative i and individual n:

$$\begin{aligned} P_n(i|C) &= P\left(U_{in} \geq U_{jn}, \forall j \in C_n\right) \\ &= P\left(V_{in} + \epsilon_{in} \geq V_{jn} + \epsilon_{jn}, \forall j \in C_n\right) \\ &= P\left(V_{in} + \epsilon_{in} \geq \max_{j \in C_n}(V_{jn} + \epsilon_{jn})\right) \end{aligned}$$

Assuming that all the disturbances ϵ_{in} are independent and identically Gumbel distributed with location parameter $\eta = 0$ and scale parameter $\mu > 0$, we derive the closed form solution for the choice probability P_{in}:

$$P_n(i|C) = \frac{e^{\mu V_{in}}}{\sum_{j \in C_n} e^{\mu V_{jn}}} = \frac{A_{in} e^{\mu V_{in}}}{\sum_{j \in C_n} A_{in} e^{\mu V_{jn}}}$$

, where $A_{in} = 1$ if alternative i is available to individual n, and 0 otherwise. More theoretical details can be found in Ben-Akiva and Lerman (1985) and Train (2009).

References

Ben-Akiva, M., and Lerman, S. (1985), "**Discrete choice analysis: Theory and application to travel demand**," Vol. 9, The MIT press.

Train, K. E. (2009), "**Discrete choice models with simulation**," Cambridge University Press, Cambridge.

Appendix B: The Case of Swissmetro

Context

Innovation in the market for intercity passenger transportation is a difficult enterprise, as the existing modes – private car, coach, rail, as well as long-distance air services – continue to innovate in their own right by offering new combinations of speeds, services, prices and technologies. Consider for example high-speed rail links between major centers or direct regional jet services between smaller countries. The Swissmetro SA in Geneva is promoting such an innovation: a maglev underground system operating at speeds up to 500 km/h in partial vacuum connecting the major Swiss conurbations, in particular along the Mittelland corridor (St. Gallen, Zurich, Bern, Lausanne and Geneva).

Data

The Swissmetro is a true innovation. It is therefore not appropriate to base forecasts of its impact on observations of existing revealed preferences (RP) data. It is necessary to obtain data from surveys of hypothetical markets/situations, which include the innovation, to assess the impact. Survey data was collected on rail-based travels, interviewing 470 respondents. Due to data problems, only 441 are used here. Nine stated choice situations were generated for each of the 441 respondents, offering three alternatives: rail, Swissmetro and car (only for car owners).

A similar method for relevant car trips with a household or telephone survey was deemed impractical. The sample was therefore constructed using license plate observations on the motorways in the corridor by means of video recorders. A total of 1059 relevant license plates were recorded during September 1997. The central Swiss car license agency had agreed to sending up to 10,000 owners of these cars a survey pack. Until April 1998, 9658 letters had been mailed, of which 1758 were returned. A total of 1070 people filled in the survey completely and were willing to participate in the second stated preference (SP) survey, which was generated using the same approach used for the rail interviews. 750 usable SP surveys were returned from the license plate based survey.

Variables and Descriptive Statistics

The variables of the dataset are described in Tables 1, 2 and 3. A more detailed description of the dataset as well as the data collection procedure can be found in Bierlaire et al. (2001).

References

Bierlaire, M., Axhausen, K., and Abay, G. (2001), "**The acceptance of modal innovation: The case of Swissmetro**," *Proceedings of the 1st Swiss Transportation Research Conference*.

GROUP	Different groups in the population
SURVEY	Survey performed in train (0) or car (1)
SP	It is fixed to 1 for all observations
ID	Respondent identifier
PURPOSE	Travel purpose 1 = commuter, 2 = shopping, 3 = business, 4 = leisure, 5 = return from work, 6 = return from shopping, 7 = return from business, 8 = return from leisure, 9 = other
FIRST	First class traveler 0 = no, 1 = yes
TICKET	Travel ticket 0 = none, 1 = two way with half-price card, 2 = one way with half-price card, 3 = two way normal prices, 4 = one way normal price, 5 = half day, 6 = annual season ticket, 7 = annual season ticket, junior or senior, 8 = free travel after 7 PM, 9 = group ticker, 10 = other
WHO	Who paid for the ticket 0 = not known, 1 = self, 2 = employer, 3 = half-half
LUGGAGE	Measure of luggage 0 = none, 1 = one piece, 3 = several pieces
AGE	Categorical variable 1 = age≤24, 2 = 24<age≤39, 3 = 39<age≤54, 4 = 54<age≤65, 5 = age>65, 6 = not known
MALE	Traveler's gender 0 = female, 1 = male
INCOME	Traveler's income per year (thousand CHF) 0 or 1 = under 50, 2 = between 50 and 100, 3 = over 100, 4 = not known
GA	Possession of Swiss annual season ticket for the rail system and most local public transit as well 1 = individual owns a GA, 0 = otherwise

Table 1: Description of variables associated with individual characteristics

ORIGIN	Trip origin (a number corresponding to a canton, see Table 3)
DEST	Trip destination (a number corresponding to a canton, see Table 3)
TRAIN_AV	Train availability dummy
CAR_AV	Car availability dummy
SM_AV	Swissmetro availability dummy
TRAIN_TT	Train travel time (minutes) door-to-door travel times making assumptions about car-base distances (1.25*crow-flight distance)
TRAIN_CO	Train cost (CHF) without considering GA
TRAIN_FR	Train headways (minutes)
SM_TT	Swissmetro travel time (minutes) speeds of 500 km/h were assumed
SM_CO	Swissmetro cost (CHF) calculated at the current relevant rail fare, without considering GA, multiplied by a factor of 1.2 to reflect higher speeds
SM_FR	Swissmetro headways (minutes)
SM_SEATS	Swissmetro seat configuration 1 = airline seats, 0 = otherwise
CAR_TT	Car travel time (minutes)
CAR_CO	Car cost (CHF) assuming an average cost per kilometer of 1.2 CHF
CHOICE	Stated choice 1 = train, 2 = Swissmetro, 3 = car

Table 2: Description of variables associated with alternative specific characteristics

1	Zurich
2	Bern
3	Lucerne
4	Uri
5	Schwyz
6	Obwalden
7	Nidwalden
8	Glarus
9	Zug
10	Freiburg
11	Solothurn
12	Basel-Stadt
13	Basel-Land
14	Schaffhausen
15	Appenzell Ausserrhoden
16	Appenzell Interhoden
17	St. Gallen
18	Graubünden
19	Aargau
20	Thurgau
21	Ticino
22	Vaud
23	Valais
24	Neuchâtel
25	Geneva
26	Jura

Table 3: Coding of cantons