Tutorial: Discrete choice analysis
Masaryk University, Brno

Prepared by Stefanie Peer and Paul Koster

November 30, 2016

1 Introduction

Discrete choice analysis is widely applied in transport analysis and the logit model of McFadden (1974) has been the workhorse model for many applied transport studies. For example, it has been used to study the choice of transport mode or to value non-market goods such as travel time and traffic safety. In this tutorial you will develop some practical skills that enable you to analyze discrete choice data. For this assignment you need to estimate the value of travel time using data obtained from a stated preference experiment among car peak commuters. Please do not hesitate to ask questions during the exercises.

2 Trading travel time and money

The data is collected using a stated choice experiment that aims at estimating a monetary value attached to reductions in travel time (in short: value of time). Respondents make 6 choices each, where they need to trade off travel costs and travel time. An example of a trade-off is given below.

Suppose that these are the only two existing alternatives to travel from home to work. Indicate which one you prefer.

Clearly there is a trade-off between the faster and more expensive (Alternative 2) and the slower and cheaper alternative (Alternative 1)[1]. Therefore, if a respondent chooses

[1]That is also how the dataset to be used in the exercises is structured

	Alternative 1	Alternative 2
Travel costs (in Euro)	6	8
Travel time (in minutes)	40	30

alternative 1 or 2, we learn something about his trading procedure of money and travel time. In order to make this more explicit, we first look for the trade-off value of travel time (VOTT)[2] for which the respondent is indifferent between choice alternatives 1 and 2. This value is usually referred to as the bid-value. Suppose that Alternative 1 is always the slowest alternative. Then the bid-value is given by Eq. 1

$$bid = \frac{-(C_1 - C_2)}{T_1 - T_2}, \tag{1}$$

where C_1 and C_2 are the travel costs of alternative 1 and alternative 2 respectively. T_1 and T_2 are the travel times of alternatives 1 and 2 respectively. The bid-value of the example is therefore $-(6-8)/(40-30) = 2/10$ Euro per minute or 12 Euro per hour. Now suppose a respondent chooses Alternative 1 (i.e. the slower and cheaper alternative). Then we learn that the VOTT of this respondent is lower than (or equal to) 6 Euro/hour. Similarly a choice for alternative 2 implies that the respondents VOTT is larger than 12 Euro. Rather than looking at the trade-offs in each choice situation we would prefer to analyze the preferences of the whole sample of respondents. Therefore we proceed with a model that is able to analyze large datasets.

3 Binary logit analysis

The logit model can account for errors in decision making and other factors that are unobserved by the researcher and influence the choice of a respondent. We start by writing down the utility function for alternative $j = 1, 2$ and for choice $n = 1, \ldots, N$:

$$U_{jn} = V_{jn}[\beta; T_{jn}; C_{jn}] + \epsilon_{jn} \tag{2}$$

The utility function consist of two components: A systematic component given by $V_{jn}[.]$[3], and a random component ϵ_{jn}. The systematic component is a function of the travel time and travel cost (T_{jn} and C_{jn}) and the sensitivity to changes in these variables indicated (i.e. the coefficients to be estimated) by β. If we assume that ϵ_{jn} is logistically distributed, the probability that alternative $j = 1, 2$ is chosen is given by:

$$P_j = \frac{\exp V_{jn}[.]}{\exp V_{1n}[.] + \exp V_{2n}[.]} \tag{3}$$

The beauty of Eq. 3 is that the formula has a simple closed form expression. The value of travel time is then given as ratio of the marginal utilities and is given by Eq. 4:

$$VOTT = \frac{\partial V[.]/\partial T}{\partial V[.]/\partial C} \tag{4}$$

The ratio indicates the willingness to pay for reducing travel time by one unit. It is of key importance to understand that the VOTT is a ratio of marginal utilities. When

[2]Sometimes also abbreviated by VOT (value of time) or VOTTS (value of travel time savings).

[3]For convenience $V_{jn}[.]$is used, but of course V_{jn} is a function of several variables.

nonlinear formulations of the systematic utility are used, this ratio may be more complicated. Throughout this exercise we assume that a linear systematic utility applies such as in Equation 5:

$$V = \beta_C C + \beta_T T \tag{5}$$

Then the VOTT is given by:

$$VOTT = \beta_T / \beta_C, \tag{6}$$

which is the ratio of the sensitivities to changes in travel time and travel costs. More general specifications of Eqs. 5 and 6 will be discussed in the following exercises.

4 Exercise I: Exploratory analysis

Applied analysis always should always with a first exploratory analysis of the data. This provides insides in the quality of the data and some general trends. For this analysis you should use Excel and the dataset *brno.xlsx*. Here you find data for 1357 respondents making 6 choices each. The columns in the dataset indicate the variables. In the tab *varnames* you find a dictionary of the variables. Read this carefully before you proceed.

Ex. 1. Make summary statistics (minimum, average, maximum) of the reference travel time, income for the whole sample and for males and females separately.[4]

Ex. 2. How many times is the fastest alternative chosen? How many times is the slowest alternative chosen?

Ex. 3. What is the *probability* that the fastest alternative is chosen? What is the probability that the slowest alternative is chosen?

Ex. 4. Are there any dominant alternatives (hence, alternatives that are both cheaper AND faster) included in the dataset?

5 Exercise II: Logit analysis

For this analysis you should use Biogeme and the dataset *brno.dat*. Biogeme is open source software developed by Michel Bierlaire (Bierlaire, 2003). Open the file *VOTT_BL.mod*.[5] Read the file carefully in order to understand the structure. What are the parameters to be estimated? Where is the utility specified? How does the file relate to the dataset? If you understand the file, you can proceed. If you have any questions please ask!

[4] The excel function *averageif* is useful in this context.

[5] Right click on the file and use open with and choose *Wordpad*.

Ex. 5. Estimate a logit model with the linear specification of utility of Eq. 5 using the Biogeme model file *VOTT_BL.mod*.[6] Do the coefficients have the expected sign? Are the coefficients significantly different from 0? Calculate the VOTT in Euro/hour using Eq. 5.

Ex. 6. Try to use the [Exclude] section in the file *VOTT_BL.mod* to exclude some data. For example to only consider those with a lower income, use: income > 4 . Several statements can be combined using the OR operator ||, for example: (statement 1) || (statement 2)

6 Exercise III: Covariates

It may well be that the VOTT depends on the income of the respondent, because respondents with a high income are likely having a lower sensitivity to changes in travel costs. Therefore we extend the utility function of Eq. 5 to incorporate an interaction effect of the variable INC in your dataset:

$$V = \beta_C C + \beta_{C,INC} C\ INC + \beta_T T \tag{7}$$

Ex. 7. What is the interpretation of $\beta_{C,INC}$? Derive the formula for the VOTT as in Eq. 6.

Ex. 8. Adjust the Biogeme model file *VOTT_BL.mod* to incorporate the interaction effect of travel costs and income. Store the new modfile as *VOTT_BL_covars.mod*. First, calculate the variable C_INC in the [Expressions] section. Second, use Eq. 7 and change the utility function in the [Utilities] section. Third, add the new variable $\beta_{C,INC}$ in the [Beta] section.

Ex. 9. Estimate the model and interpret your results.[7] What is the VOTT at the average income level ? Compare this average to your findings at exercise 6. What is the maximum and minimum VOTT?

Ex. 10. Test if males and females have different marginal utilities of travel time (e.g. different β_T). First, write down the extension of the utility function of 7. Second, adjust the Biogeme model file and save it. Third, run the model and interpret your results. Derive the VOTT at minimum, average and maximum income levels for males and females separately.

[6] This can be done in the following way: 1) Go to `biogeme.epfl.ch`, go to downloads and click on *(Executables for) Windows*. Go to the download folder, choose *gui* and open inside this folder *guibiogeme*. You should then see a graphical user interface for Biogeme. 2) To specify the model click on the *Select File* button (the highest one) and select the model file (in this case *VOTT_BL.mod*). 3) To specify the data click on the *Select File* button (the lower one) and select the datafile *brno.dat*. 4) To estimate the model, click on *estimate*. The model will be estimated. 5) If the model is finished, click *Display File* to see the results. You can also open the html file that is created in the directory you are working in.

[7] You can use the income variable as if it was continuous. Thus, you do not have to take into account that in fact it is an ordinal variable.

7 Exercise IV: Unobserved heterogeneity - cross-sectional mixed logit

For this exercise you should again use Biogeme and the dataset *brno.dat*. The standard logit model assumes that unobserved effects are captured by the error term. For example: education may have an effect on the VOTT, but if we do not measure education (e.g. the variable is not in our data), the effect of education is *unobserved*. This effect will therefore end up in the random part of the utility. Ignoring unobserved heterogeneity may lead to biases in the logit estimates and therefore it is important to control for these effects. The common workhorse model to do so is the mixed logit model. The mixed logit model estimates a distribution of preferences instead of a single VOTT, assuming that the distribution of preferences is continuous. An alternative are latent class (or: discrete mixture) models that assume a distribution with discrete masspoints (classes), whereby the researcher determines the number of classes. Because of limited time we will only consider the mixed logit model here.

Assume a linear in parameters utility function as in Eq. 5. The choice probability of Eq. 3 then will change to:

$$P_{jn} = \int \frac{\exp(\beta_C C_{jn} + \beta_T T_{jn})}{\exp(\beta_C C_{1n} + \beta_T T_{1n}) + \exp(\beta_C C_{2n} + \beta_T T_{2n})} f[\beta_C] d\beta_C \tag{8}$$

The logit probability is a continuous mixture of logit probabilities, where the mixture is governed by the probability distribution $f[\beta_C]$. A mixture can be viewed as a weighted average of logit probabilities. We are interested in the distribution $f[\beta_C]$, which provides the distribution of the cost coefficient in the sample and try to estimate it. Meanwhile, the time coefficient is still assumed homogenous across choices. Throughout the analysis, a normal distribution is assumed.

Ex. 11. Use the Biogeme model file *VOTT_CSLC.mod* to estimate the model and interpret your results (running the model may take some time).[8]

Ex. 12. Again try to include the interaction variables income and gender. Compare your results with the result of exercise I and II.

8 Exercise V: Unobserved heterogeneity - panel mixed logit

Skip this exercise if you are short on time, as the time required to estimate the model is quite long.

For this exercise you should use Biogeme and again the dataset *brno.dat*. The cross sectional mixed logit model (as estimated in the previous section) assumes that the error

[8] If a bug results due to numerical problems: lower the number of Draws in the section [Draws] and try to estimate the model. Use the results of the estimation that works and gradually increase the number of draws.

term over a series over choices is uncorrelated and therefore analyzes the probability of an isolated choice. In other words, it does not take into account that each person makes 6 choices rather than just 1, and that the preferences across these 6 choices will be similar. The panel mixed logit model goes one step further and analyzes the probability that an individual makes a sequence of choices. Assume a linear in parameters utility function as in Eq. 5. The choice probability of a sequence of choices made by individual i is given by:

$$P_{ji} = \int \left[\prod_{z=1}^{Z(i)} \frac{\exp(\beta_C C_{jiz} + \beta_T T_{jiz})}{\exp(\beta_C C_{1n} + \beta_T T_{1iz}) + \exp(\beta_C C_{2n} + \beta_T T_{2iz})} \right] f[\beta_C] d\beta_C, \tag{9}$$

where $Z(i)$ is the number of choices of the number of choices of individual i (in our case 6 for each individual). The key difference with Eq. 8 is that instead of analyzing each choice separately, we now analyze the probability of a *sequence of choices*, which is given by the product of the choice probabilities of each separate choice. Therefore Eq. 9 integrates over the product of a sequence of choices, where β_C is kept constant over the series of choices of an individual.

Ex. 13. Use the Biogeme model file *VOTT_PMXL.mod* to estimate the model and interpret your results (running the model may take some time). Furthermore, compare your results with the estimates of Exercise 12.

9 Exercise VI: Advanced examples from the Biogeme website

Go to `http://biogeme.epfl.ch/swissmetro/examples.html` and download the dataset concerning the Swissmetro. First check what the dataset is about.

Ex. 14. Run some of the advanced models shown on the site using the Swissmetro dataset (e.g. (cross-) nested logit). Become acquainted with the logic of the models and make some amendments (e.g. add some exclude statements or interaction terms). See in which way the results change.

Ex. 15. Check the pythonbiogeme version of the models, and try to understand their structure and syntax.