Week 2: Stated preference analysis

Thomas de Graaff and Hans Koster

Version: January, 2025

About the tutorial

Intended Learning Outcome: Apply a logit analysis and discuss the results in the context of stated preference research.

Contents: Discrete choice analysis is widely applied in transport demand 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 (VoTT) using data obtained from a stated preference experiment amongst car commuters traveling during peak hours.

Questions?: Whenever you have problems with the assignments go back to the lecture slides, and the syllabus. For questions about `Stata` Google is your best friend (there is a large `Stata` community out there). If needed, you can ask for help by email (h.koster@vu.nl).

Important guidelines:

- Provide a separate `Stata` do-file (or `R` script) file so that your answers can be reproduced.
- The dataset to be used can be found on www.urbaneconomics.nl/aese.
- Feel free to write a report in a style that you deem useful. We will discuss the results collectively tomorrow.

1 Introduction: Trading travel time and money

1.1 Data

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 in this dataset make 6 choices each, where they are asked to trade off travel costs and travel time. An example of such 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.*

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

Clearly this choice provides 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 Alternative 1 or 2, we learn something about his trade-off between money and travel time. In order to make this more explicit, we first look for the trade-off value of travel time (VOTT) 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 Eqn. (1.1):

$$\text{bid} = \frac{-(C_1 - C_2)}{T_1 - T_2} \tag{1.1}$$

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) 12 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 for a sample of respondents. Therefore we proceed with a model that is able to analyze large datasets.

1.2 Binary logit analysis

The logit model allows for decision errors and other factors that are unobserved by the researcher that influence the choice of a respondent. We start by writing down the random utility function for Alternative $j = 1,2$, for choice $n = 1,\ldots,N$;

$$U_{jn} = V_{jn}[\beta; T_{jn}; C_{jn}] + \epsilon_{jn} \tag{1.2}$$

This utility function consist of two components: A systematic component given by $V_{jn}[.]$[2], 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 are indicated (i.e., the coefficients to

[1] This is also how the dataset to be used in the exercises is constructed.

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

be estimated) by the vector β. If we assume that ϵ_{jn} has a Gumbel distribution (or the overlapping Generalized Extreme value distribution), then the probability that alternative $j = 1,2$ is chosen is given by:

$$P_j = \frac{\exp V_{jn}[.]}{\exp V_{1n}[.] + \exp V_{2n}[.]} \tag{1.3}$$

The beauty of Eqn. (1.3) is that the formula for the choice probability has a simple closed-form expression. This is a key advantage of the logit model. For example, the probit model needs simulation or other techniques to approximate the choice probability. The value of travel time is then given as the ratio of the **marginal** utilities:

$$VOTT = \frac{\partial V[.]/\partial T}{\partial V[.]/\partial C} \tag{1.4}$$

This 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 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 given in Eqn. (1.5):

$$V = \beta_C C + \beta_T T, \tag{1.5}$$

where the coefficients β_C and β_T are expected to have negative signs. Then the VOTT is given by:

$$VOTT = \beta_T / \beta_C, \tag{1.6}$$

which is the ratio of the marginal change in utility due to a change in travel time and travel costs. More general specifications of Eqs. (1.5) and (1.6) will be discussed in the following exercises.

2 The assignment

2.1 Assignment I: Exploratory analysis

Applied analysis always should start with a first exploratory analysis of the data. This provides insides in the quality of the data, the representativeness of the sample and other general trends. For this analysis you should use `Stata` and the dataset `VOTTdat.xlsx`. Here you find data for 1,337 respondents making 6 choices each. The columns in the dataset indicate the variables. In the sheet `varnames` you find a dictionary of the variables. Read this dictionary carefully before you proceed.

1. The data you will use comes from a real stated preference experiment. Name and explain two potential biases that might arise with this particular experiment.
2. Import the excel file in `Stata` and derive summary statistics. What is the sample probability that the fastest alternative is chosen? What is the sample probability that the slowest alternative is chosen? In other words, is there a structure in the choice variable?
3. Are there any dominant alternatives (hence, alternatives that are both cheaper **and** faster) included in the dataset?

2.2 Assignment II: Logit analysis

For this analysis you should use `Stata` and the imported dataset `VOTTdat.xlsx`. Before you start estimating take another look at the data. Logit analysis in `Stata` is only possible when your dependent variable is binary (i.e., 0 or 1); is that the case here? And can the cost and time variable

directly be used when you would like to say someting about the *difference* in cost and time between the alternatives (see as well the lecture slides again and especially slide 11 with the formula $\Pr(B) = f(\text{traveltime}_B - \text{traveltime}_A)$. Transform the data thus accordingly.

1. Name two advantages of using a Logit model over using a linear probability model.
2. Now first estimate a logit model with only the difference in costs as an independent variable. Plot the probability of choosing the more costly alternative given the difference in costs.[3]
3. Now, estimate a logit model with the linear specification of the systematic part of utility of Eqn. (1.5). Do the coefficients have the expected sign? Are the coefficients significantly different from 0? Calculate the VOTT in Euro/hour using Eq. (1.5).
4. What does the constant mean in this estimation?
5. Explain what happens with the VOTT when you exclude everyone with high income (say income > 4)?

2.3 Assignment III: Covariates

It may well be that the VOTT depends on the income of the respondent, because respondents with a high income are expected to be less sensitive 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 as follows:

$$V = \beta_C C + \beta_{C,INC} C \times INC + \beta_T T, \tag{2.1}$$

1. What is the interpretatation of $\beta_{C,INC}$? Derive the formula for the VOTT as in Eqn. (1.6).
2. Estimate Eqn. (2.1) and interpret your results.
3. Calculate the choice probability for the various income groups and plot them. What do you infer?
4. Test whether males and females have different marginal utilities of travel time (e.g., different β_T)

2.4 Assignment IV: Unobserved heterogeneity—panel logit

The cross sectional logit model (as estimated in the previous section) assumes that the error 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 logit model goes one step further and analyzes the probability that an individual makes a sequence of choices.

1. Estimate a panel logit model (with fixed effects). Why do some individuals drop out of the estimation?
2. Compare your results with the non-fixed estimation and explain why they are (not that) different.

[3] The most straightforward way of estimating a logit is estimating a generalized linear model with the command `glm`. See for an example of plotting: https://www.econometrics-with-r.org/11.2-palr.html.