AH2170 Transportation Data Collection and Analysis

Case study III

September 2018

1 Introduction

In this project we will use data from a travel survey conducted in Sweden. The data will be used to estimate a Multinomial Logit (MNL) model. You will get to choose a model specification and perform t-tests and likelihood ratio tests on the parameters.

2 Data

The data consist of 4000 observations of individiduals that has chosen one out of 6 modes of transportation for a trip. The data consist of attributes of the 6 available modes as well as the index of the choose alternative. The columns of the data are explained in Table 1 and the index of the chosen mode is explained in Table 2.

3 Exploratory data analysis and descriptive statistic

Now, using the variables and data provided by the survey:

1. Explore the data and report average values, standard deviations, and ranges of values for the variables.
2. Develop a-priori hypotheses about variables you think are the most important in explaining the mode choices of the residents. Make sure you include

Column index	Column name	Description
1	mode	chosen mode of transport
2	car_cost	cost of using car
3	car_time	travel time with car as driver
4	car_ok	dummy for availability of car
5	pass_cost	cost of travelling as passenger
6	pass_time	travel time when passenger
7	pass_ok	dummy for availability of passenger
8	bus_cost	cost of using bus
9	bus_g_time	walking time to bus stop
10	bus_w_time	waiting time for bus
11	bus_time	travel time with bus
12	bus_ok	dummy for availability of bus
13	train_cost	cost of travelling with train
14	train_w _time	wait time for train
15	train_g _time	walking time to train station
16	train_time	travel time with train
17	train_ok	dummy for availability of train
18	walk_time	travel time if walking
19	walk_ok	dummy for availability of walk
20	bike_time	travel time with bike
21	bike_ok	dummy for availability of bike
22	id	id of observation

Table 1: Explanations of columns in data

Value	Chosen mode
1	car driver
2	car passenger
3	bus
4	train
5	walk
6	bike

Table 2: Explanation of index of mode of transport.

a good set of potential variables and do not limit yourself. You may also consider transformations of variables. Discuss at least 4 variables.

3. If appropriate plot the histograms of the variables you consider to be important ones.

4. Find the correlation coefficients (if applicable) among the variables that you consider most important.

4 Estimation and model specification

1. Estimate (using MATLAB) and present your best two model specifications, and interpret the estimation results of the two models.

2. Comment on which of the two models is better and why.

In answering the above questions keep in mind the following:

- Think carefully about a priori hypotheses and expectations regarding the important factors affecting mode choice in this case.
- Explain your process of obtaining the final models clearly.
- Interpret statistical results and tests.
- Select the best model based on a priori hypotheses, statistics, and causal relationships.

4.1 Analysis of estimation result

1. Select two parameters that you have estimated, one with a large and one with a small t-value. Plot for one parameter at a time how the log-likelihood changes when the value of that parameter varies around its estimated value and including zero. Keep all other parameters fixed at their estimated.

2. Compare the t-test and log-likelihood ratio test for the two parameters.

3. Calculate the aggregate sample demand for the alternative modes with the estimated parameters.

4. Plot how the aggregate sample demand for respectively modes changes with a change in one of the attributes, e.g., when travel time with bus changes.

A Maximum likelihood estimation using MATLAB

In this section we will describe the steps involved in estimating a MNL model using MATLAB.

You first need to load the data stored in the CSV file into either a table or a matrix. This can be done using one of the following commands:

```
X =  readtable('ModeChoice.csv'); % Load as table, so that X.car_time gives
                                  % vector of travel time with car
```

or

```
X =  dlmread('ModeChoice.csv',';',2); % Gives a matrix with all attributes.
                                      % X(3,:) gives vector with
                                      % travel times with car
```

Next you need to create a function for the probability of selecting respectively mode. The function should accept the table `X` and a parameter vector `theta` as an input and return a matrix `P` which is $N_{observations} \times N_{alternatives}$ large, i.e, contains the probability for each alternative for each observation.

```
function [ P ] = ModeProbability( X, theta )
%MODEPROBABILITY Calculates MNL probabilities of respectively
%                mode for all observations in the sample
%   INPUT
%       X     N_obs x N_att - Table with attrbiutes
%       theta N_param x 1   - Parameter vector
%   OUTPUT
%       P     N_obs x N_alt  - Probability for all alternatives for all
%                              observations
```

The log-likelihood which should be maximised can now be obtained by summing $\log(p_{\text{obs}})$ for the observed alternative for each observation. This requires you to first select the chosen alternatives from the matrix P. The index of the chosen alternatives are given in the column named 'mode'. There are multiple ways to select the chosen alternatives from the matrix `P`. If mode was a vector containing the index of the chosen alternative for all observations, one way to obtain the probability that someone who was observed to choose 3 would choose 3 according to the model would be to use:

```
p3If3 = P(mode==3,3);
```

The log-likelihood can also be calculated using a for-loop or using the function `sub2ind`.

```
function [ ll ] = ModeLogLikelihood( X,theta )
%MODELIKELHOOD caluclates Loglikelihood of observations given attributes
%and choices in x and parameters theta
```

```
%  1) Call ModeProbability to calculate matrix of probabilities for all modes
%  2) obtain choosen alternatives from input table
%  3) Select choosen alternatives from matrix P
%  4) sum log(p)
```

Once the choice probabilities are known it is time to maximize the log-likelihood. This can be done using the function `fminunc` which minimze a function. Create a function handle to the function that should be maximized, in this case `ModeLogLikelihood` using `f=@(x)-1*ModeLogLikelihood(X,x)`. Now `f(theta)` gives you the negative log-likelihood for the parameter vector `theta`.

B How to write the report

You are to present the result in the form of a technical report. The report should at most contain 2000 words and follow the IMRAD structure (Introduction, Method, Results and Analysis/Discussion).

You are required to reference any sources correctly. Use an appropriate number of digits in any result you present. All figures should be numbered in sequence. Each figure should have a description, placed below the figure. Tables should also be numbered consecutively. Each table must have a description placed above it. Figures, tables and their respective descriptions have to be centred on the page.