Discrete Choice Case Study
Part 1: Multinomial Logit Models

1 Objectives of Part 1 of the Case Study

The topic of this case study is the Multinomial Logit Model. Different specifications are introduced using a stepwise modelling strategy, which increases the complexity by adding different variables at each step. The reference software is PandasBiogeme . The objectives of this case study can be summarized as follows:

- Specification and estimation of a first MNL model making use of generic attributes;
- Become familiar with MNL including alternative-specific parameters;
- Introduce you to generic vs specific tests techniques;
- Illustration of the market segmentation concept and relative testing.

1.1 Theoretical Reminder

Multinomial logit 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:

$$\begin{aligned} P_n(i|C) &= Pr(U_{in} \geq U_{jn}, \forall j \in C_n) \\ &= Pr(V_{in} + \epsilon_{in} \geq V_{jn} + \epsilon_{jn}, \forall j \in C_n) \\ &= Pr[V_{in} + \epsilon_{in} \geq max(V_{jn} + \epsilon_{jn}] \end{aligned}$$

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}:

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

where $A_{in} = 1$ if alternative i is available to decision maker n and 0 otherwise. More theoretical details can be found in Ben-Akiva and Lerman (1985) and Train (2003).

1.2 Datasets

The Swissmetro dataset. This dataset consists of survey data collected on the train between St. Gallen and Geneva, Switzerland, during March 1998. The respondents provided information in order to analyze the impact of the modal innovation in transportation, represented by the Swissmetro, a revolutionary mag-lev underground system, against the usual transport modes represented by car and train.

The residential telephone services dataset. A household survey was conducted in 1984 for a telephone company among 434 households in Pennsylvania. The dataset involves choices among five calling plans and consists of various alternative-specific and socio-economic variables. It was originally used to develop a model system to predict residential telephone demand (Train and Ben-Akiva (1987)).

1.3 Your Tasks

The procedure you should follow to work on the case study can be summarized as follows:

- select a dataset of interest;
- formulate hypothesis: selection of the explanatory variables and how they affect the utilities (generic and/or alternative-specific), inclusion of socio-economic variables, etc.;

- perform the estimation of the related model. Give an interpretation of the obtained results and check if they are plausible, performing different tests on the current specification;
- try to change the model specification, improving the final log likelihood value, keeping coefficient estimates which are coherent with the behavioral hypothesis.

1.4 Report Content

Your final report should contain:

- the presentation of your best model specification;
- the results of some tests used to arrive at your best model specification;
- a discussion of the estimated parameter values;
- the PandasBiogeme report file (*.html) for all discussed models (submitted electronically in a zip file).

Swissmetro Case

Model Specification with Generic Attributes

Files to use with Biogeme:
Model file: *MNL_SM_generic.py*
Data file: *swissmetro.dat*

The dataset consists of survey data collected on the trains between St. Gallen and Geneva in Switzerland. The idea is to analyze the impact of modal innovation in transportation, represented by the Swissmetro, against the more classic types of transport modes. The choice variable consists of three alternatives: train, Swissmetro and car (for car owners). In this first model specification, we assume that travel time, cost and headway of public transportation modes influence the utility functions. We also assume that the coefficients of the explanatory variables are generic, that is, they do not vary over the alternatives. The corresponding expressions of the utilities are defined as follows:

$$\begin{aligned} V_{car} &= \mathrm{ASC}_{\mathrm{car}} + \beta_{\mathrm{time}}\mathrm{CAR\_TT} + \beta_{\mathrm{cost}}\mathrm{CAR\_CO} \\ V_{train} &= \beta_{\mathrm{time}}\mathrm{TRAIN\_TT} + \beta_{\mathrm{cost}}\mathrm{TRAIN\_COST} + \beta_{\mathrm{he}}\mathrm{TRAIN\_HE} \\ V_{SM} &= \mathrm{ASC}_{\mathrm{SM}} + \beta_{\mathrm{time}}\mathrm{SM\_TT} + \beta_{\mathrm{cost}}\mathrm{SM\_COST} + \beta_{\mathrm{he}}\mathrm{SM\_HE} \end{aligned}$$

where CAR_TT is the car travel time, CAR_CO is the car cost, TRAIN_TT is the train travel time, TRAIN_COST is the train cost (considering the ownership of Swiss annual season ticket, GA), TRAIN_HE is train headway (in minutes), SM_TT is the Swissmetro travel time, SM_COST is the Swissmetro cost (considering the ownership of GA), and SM_HE is the Swissmetro headway.

The estimation results are shown in Table 1. For estimation purposes, we have normalized the alternative specific constant of train to zero. The estimated values for the alternative specific constants $\mathrm{ASC}_{\mathrm{car}}$ and $\mathrm{ASC}_{\mathrm{SM}}$ show that, all else being equal, there is a preference in the choice of car and Swissmetro with respect to train. Moreover, the higher value of $\mathrm{ASC}_{\mathrm{SM}}$ shows a greater preference for Swissmetro compared to car. As expected, both the travel time and cost coefficients have negative signs. The higher the travel

Logit model with generic attributes

Parameter number	Parameter name	Parameter estimate	Robust standard error	Robust *t statistic*
1	$\mathrm{ASC}_{\mathrm{car}}$	0.189	0.0798	2.37
2	$\mathrm{ASC}_{\mathrm{SM}}$	0.451	0.0932	4.84
3	β_{cost}	-0.0108	0.000682	-15.90
4	β_{he}	-0.00535	0.000983	-5.45
5	β_{time}	-0.0128	0.00104	-12.23

Summary statistics
Number of observations = 6768
$\mathcal{L}(0) = -6964.663$
$\mathcal{L}(\hat{\beta}) = -5315.386$
$\bar{\rho}^2 = 0.236$

Table 1: Logit model with generic attributes

time or the cost of an alternative, the lower the related utility. The negative estimate of the headway coefficient β_{he} indicates that the higher the headway, the lower the frequency of service, and thus the lower the utility.

Model Specification with Alternative Specific Attributes

Files to use with Biogeme:
Model file: *MNL_SM_specific.py*
Data file: *swissmetro.dat*

In this second model, we relax the hypothesis of generic coefficients. To illustrate this idea, we use three different cost coefficients, one for each alternative. The corresponding utility functions are

$$
\begin{aligned}
V_{car} &= \mathrm{ASC_{car}} + \beta_{\mathrm{time}}\mathrm{CAR\_TT} + \beta_{\mathrm{car\_cost}}\mathrm{CAR\_CO} \\
V_{train} &= \beta_{\mathrm{time}}\mathrm{TRAIN\_TT} + \beta_{\mathrm{train\_cost}}\mathrm{TRAIN\_COST} + \beta_{\mathrm{he}}\mathrm{TRAIN\_HE} \\
V_{SM} &= \mathrm{ASC_{SM}} + \beta_{\mathrm{time}}\mathrm{SM\_TT} + \beta_{\mathrm{SM\_cost}}\mathrm{SM\_COST} + \beta_{\mathrm{he}}\mathrm{SM\_HE}.
\end{aligned}
$$

Logit model with alternative specific travel cost

Parameter number	Parameter name	Parameter estimate	Robust standard error	Robust *t statistic*
1	$\mathrm{ASC_{car}}$	-0.971	0.134	-7.22
2	$\mathrm{ASC_{SM}}$	-0.444	0.102	-4.34
3	$\beta_{\mathrm{car\_cost}}$	-0.00949	0.00116	-8.21
4	β_{he}	-0.00542	0.00101	-5.36
5	$\beta_{\mathrm{SM\_cost}}$	-0.0109	0.000703	-15.49
6	β_{time}	-0.0111	0.00120	-9.26
7	$\beta_{\mathrm{train\_cost}}$	-0.0293	0.00169	-17.32

Summary statistics

Number of observations = 6768
$\mathcal{L}(0) = -6964.663$
$\mathcal{L}(\hat{\beta}) = -5068.559$
$\bar{\rho}^2 = 0.271$

Table 2: Logit model with alternative-specific cost attributes

The estimation results for this model specification are shown in Table 2. The results show the significance of the alternative-specific cost coefficients. The influence of the cost is different, showing a larger negative impact on the train alternative with respect to car and Swissmetro. In this model, the ASC's are negative implying a preference, with all the rest constant, for the train alternative. These results are different from those of the previous model where ASC_{car} and ASC_{SM} were positive and significant. The larger negative value of ASC_{car} implies that this alternative is more negatively perceived with respect to train than the Swissmetro alternative. Considering that the deterministic utilities are very simple, only including three explanatory variables, the alternative specific constants can capture various effects. Their signs and magnitudes should therefore be further investigated.

Generic vs. Specific Test

To test whether a coefficient should be generic or alternative-specific, we use the likelihood ratio test (see pages 28 and 164-167 in Ben-Akiva and Lerman (1985)). We compare the log likelihood functions of the restricted and unrestricted models of interest. The restricted model includes generic travel cost coefficients over the three alternatives, and the unrestricted model includes alternative-specific travel cost coefficients. Hence, the null hypothesis is

$$H_0 : \beta_{\text{car\_cost}} = \beta_{\text{train\_cost}} = \beta_{\text{SM\_cost}}$$

and the test statistic for the null hypothesis is given by

$$-2(\mathcal{L}_R - \mathcal{L}_U)$$

which is asymptotically distributed as χ^2 with $df = K_U - K_R$ degrees of freedom, where K_U and K_R are the numbers of estimated parameters in the unrestricted and restricted models, respectively. We reject the null hypothesis that the restrictions are true if

$$-2(\mathcal{L}_R - \mathcal{L}_U) > \chi^2_{((1-\alpha),df)}$$

where α is the level of significance. In this specific case, using $\alpha = 0.05$ yields

$$-2(-5315.386 + 5068.559) = 493.654 > 5.991$$

We can therefore reject the null hypothesis and conclude that the travel cost coefficient should be alternative-specific.

Model Specification with Socio-Economic Characteristics

Files to use with Biogeme:
Model file: *MNL_SM_socioec.py*
Data file: *swissmetro.dat*

To capture the average of the differences between the individuals in the sample, we make use of socio-economic characteristics. These types of variables do not change over the choice set and are individual specific. In this example, we add two variables to the model: a dummy variable (SENIOR) for senior people (age above 65) and a dummy variable that captures the effect of the Swiss annual season ticket for train (GA). A few observations, where the variable AGE is unknown (coded as 6), are removed from the estimation. The deterministic utilities are:

$$
\begin{aligned}
V_{\text{car}} = {} & \text{ASC}_{\text{car}} + \beta_{\text{time}}\text{CAR\_TT} + \beta_{\text{car\_cost}}\text{CAR\_CO} + \beta_{\text{senior}}\text{SENIOR} \\
V_{\text{train}} = {} & \beta_{\text{time}}\text{TRAIN\_TT} + \beta_{\text{train\_cost}}\text{TRAIN\_COST} + \beta_{\text{he}}\text{TRAIN\_HE} + \\
& \beta_{\text{ga}}\text{GA} \\
V_{\text{SM}} = {} & \text{ASC}_{\text{SM}} + \beta_{\text{time}}\text{SM\_TT} + \beta_{\text{SM\_cost}}\text{SM\_COST} + \beta_{\text{he}}\text{SM\_HE} + \\
& \beta_{\text{senior}}\text{SENIOR} + \beta_{\text{ga}}\text{GA}
\end{aligned}
$$

The estimation results for this model are shown in Table 3. The coefficients of the socio-economic variables have been estimated and are significantly different from zero at a 95% confidence level. The negative sign of the age coefficient (referring to SENIOR dummy variable) reflects a preference of older individuals for the train alternative. It seems a reasonable conclusion, dictated probably by safety reasons with respect to the car choice and a kind of "inertia" with respect to the modal innovation represented by the Swissmetro alternative. The coefficient related to the ownership of the Swiss annual season ticket (GA) is positive, as expected. It reflects a preference for the SM and train alternative with respect to car, given that the traveler possesses a season ticket. Finally, the interpretation of the alternative specific constants is similar to that of the previous model specification.

Logit model with socio-economic variables

Parameter number	Parameter name	Parameter estimate	Robust standard error	Robust *t statistic*
1	ASC_{car}	-0.608	0.143	-4.24
2	ASC_{SM}	-0.135	0.106	-1.26
3	$\beta_{\text{car\_cost}}$	-0.00936	0.00117	-8.02
4	β_{he}	-0.00586	0.00106	-5.55
5	$\beta_{\text{SM\_cost}}$	-0.0104	0.000744	-14.02
6	β_{time}	-0.0111	0.00121	-9.20
7	$\beta_{\text{train\_cost}}$	-0.0268	0.00176	-15.24
8	β_{senior}	-1.88	0.109	-17.31
9	β_{ga}	0.557	0.191	2.91

Summary statistics

Number of observations $= 6759$

$\mathcal{L}(0) = -6958.425$

$\mathcal{L}(\hat{\beta}) = -4927.167$

$\bar{\rho}^2 = 0.291$

Table 3: Logit model with socio-economic variables

Choice of Residential Telephone Services Case

Model Specification with Generic Attributes

Files to use with Biogeme:
Model file: *MNL_Tel_generic.py*
Data file: *telephone.dat*

In this example, we model the household's choice of service option for local telephone services. The choice variable (dependent variable) includes the following alternatives: budget measured (BM), standard measured (SM), local flat(LF), extended flat(EF) and metro flat(MF). In this first model, we assume that the cost of the calling plan is the only factor influencing the choice of the calling plan. We also assume that the coefficients of the explanatory variables are generic, i.e., they do not vary among the alternatives. The expressions of the utilities for this simple model can be written as:

$$\begin{aligned} V_{\mathrm{BM}} &= \mathrm{ASC}_{\mathrm{BM}} + \beta_{\mathrm{cost}} \ln(\mathrm{cost\_BM}) \\ V_{\mathrm{SM}} &= \beta_{\mathrm{cost}} \ln(\mathrm{cost\_SM}) \\ V_{\mathrm{LF}} &= \mathrm{ASC}_{\mathrm{LF}} + \beta_{\mathrm{cost}} \ln(\mathrm{cost\_LF}) \\ V_{\mathrm{EF}} &= \mathrm{ASC}_{\mathrm{EF}} + \beta_{\mathrm{cost}} \ln(\mathrm{cost\_EF}) \\ V_{\mathrm{MF}} &= \mathrm{ASC}_{\mathrm{MF}} + \beta_{\mathrm{cost}} \ln(\mathrm{cost\_MF}). \end{aligned}$$

Here we have included the natural logarithm of the cost in order to better capture differences in cost among alternatives.

The estimation results are shown in Table 4. The results indicate that all the rest being equal, the budget measured (BM) alternative is the least desired alternative and the metro area flat (MF) is the most preferred alternative. The alternative specific constant for the extended flat (EF) alternative is not significantly different from zero, as shown by the related t-statistic value. The sign of the cost coefficient is negative, as expected, meaning that the utility of an alternative decreases with increase in cost.

Model Specification with Alternative-Specific Attributes

Files to use with Biogeme:
Model file: *MNL_Tel_specific.py*
Data file: *telephone.dat*

In this second specification, we relax the hypothesis of generic coefficients. To illustrate this idea, two different cost coefficients are introduced, one for the flat alternatives and the other for the measured alternatives. The corresponding utility functions are shown below:

$$
\begin{aligned}
V_{\mathrm{BM}} &= \mathrm{ASC}_{\mathrm{BM}} + \beta_{\mathrm{M\_cost}} \ln(\mathrm{cost\_BM}) \\
V_{\mathrm{SM}} &= \beta_{\mathrm{M\_cost}} \ln(\mathrm{cost\_SM}) \\
V_{\mathrm{LF}} &= \mathrm{ASC}_{\mathrm{LF}} + \beta_{\mathrm{F\_cost}} \ln(\mathrm{cost\_LF}) \\
V_{\mathrm{EF}} &= \mathrm{ASC}_{\mathrm{EF}} + \beta_{\mathrm{F\_cost}} \ln(\mathrm{cost\_EF}) \\
V_{\mathrm{MF}} &= \mathrm{ASC}_{\mathrm{MF}} + \beta_{\mathrm{F\_cost}} \ln(\mathrm{cost\_MF})
\end{aligned}
$$

The estimation results are shown in Table 5. In this case, both cost coefficients for flat and measured alternatives are estimated. Both their signs are negative, as expected, and the larger absolute value of $\beta_{\mathrm{M\_cost}}$ indicates that people are more sensitive to cost in case of measured alternatives. The value and the sign of the budget measured alternative specific constant still indicates that this option is the least desired, all the rest remaining constant. The other values of the ASC's for the flat options are not significant.

Generic vs. Specific Test

The likelihood ratio test (see pages 28 and 164-167 in Ben-Akiva and Lerman (1985)) can be used to test a generic versus an alternative-specific model specification. The likelihood ratio test statistic for the null hypothesis of generic attributes is

$$-2(L(\hat{\beta}_R) - L(\hat{\beta}_U))$$

where R and U denote the restricted (generic) and unrestricted (alternative-specific) models, respectively. It is χ^2 distributed with the number of degrees of freedom equal to the number of restrictions $(\mathsf{K}_\mathsf{U} - \mathsf{K}_\mathsf{R})$, where K_U and K_R are the numbers of estimated coefficients in the unrestricted and restricted models, respectively. In this case, $-2(-477.557 + 476.608) = 1.898$. Since $\chi^2_{0.95,1} = 3.841$ at a 95% level of confidence, we can conclude that the null hypothesis of a generic cost coefficient cannot be rejected. The restricted model should therefore be preferred.

Model Specification with Socio-Economic Characteristics

Files to use with Biogeme:
Model file: *MNL_Tel_socioec.py*
Data file: *telephone.dat*

The previous two models only include variables that are attributes of the alternatives. We now introduce a socio-economic characteristic, namely the number of users in the household (*users*), in the utility of the flat options. It should be noted that the socio-economic variables do not vary among the alternatives and are individual specific. The utility functions can be written now as follows:

$$\begin{aligned}
V_{\mathrm{BM}} &= \mathrm{ASC}_{\mathrm{BM}} + \beta_{\mathrm{M\_cost}} \ln(\mathrm{cost\_BM}) \\
V_{\mathrm{SM}} &= \beta_{\mathrm{M\_cost}} \ln(\mathrm{cost\_SM}) \\
V_{\mathrm{LF}} &= \mathrm{ASC}_{\mathrm{LF}} + \beta_{\mathrm{F\_cost}} \ln(\mathrm{cost\_LF}) + \beta_{\mathrm{users}}\mathrm{users} \\
V_{\mathrm{EF}} &= \mathrm{ASC}_{\mathrm{EF}} + \beta_{\mathrm{F\_cost}} \ln(\mathrm{cost\_EF}) + \beta_{\mathrm{users}}\mathrm{users} \\
V_{\mathrm{MF}} &= \mathrm{ASC}_{\mathrm{MF}} + \beta_{\mathrm{F\_cost}} \ln(\mathrm{cost\_MF}) + \beta_{\mathrm{users}}\mathrm{users}
\end{aligned}$$

The estimation results are shown in Table 6. The coefficient of the *users* variable is statistically significantly different from zero and indicates that people have higher preference towards flat options if the number of users is higher (as expected). The interpretation of the other coefficients remains the same as in the previous model specifications.

Logit model with generic attributes				
Parameter number	Parameter name	Parameter estimate	Robust standard error	Robust *t statistic*
1	$\mathrm{ASC}_{\mathrm{BM}}$	-0.721	0.152	-4.76
2	$\mathrm{ASC}_{\mathrm{LF}}$	1.20	0.159	7.56
3	$\mathrm{ASC}_{\mathrm{EF}}$	1.00	0.703	1.42
4	$\mathrm{ASC}_{\mathrm{MF}}$	1.74	0.267	6.51
5	β_{cost}	-2.03	0.212	-9.55

Summary statistics

Number of observations = 434

$\mathcal{L}(0) = -560.250$

$\mathcal{L}(\hat{\beta}) = -477.557$

$\bar{\rho}^2 = 0.139$

Table 4: Logit model with generic attributes

Logit model with alternative specific attributes				
Parameter number	Parameter name	Parameter estimate	Robust standard error	Robust *t statistic*
1	$\mathrm{ASC}_{\mathrm{BM}}$	-0.748	0.155	-4.82
2	$\mathrm{ASC}_{\mathrm{LF}}$	0.154	0.691	0.22
3	$\mathrm{ASC}_{\mathrm{EF}}$	-0.0925	1.00	-0.09
4	$\mathrm{ASC}_{\mathrm{MF}}$	0.479	0.817	0.59
5	$\beta_{\mathrm{M\_cost}}$	-2.16	0.243	-8.90
6	$\beta_{\mathrm{F\_cost}}$	-1.71	0.273	-6.25

Summary statistics

Number of observations = 434

$\mathcal{L}(0) = -560.250$

$\mathcal{L}(\hat{\beta}) = -476.608$

$\bar{\rho}^2 = 0.139$

Table 5: Logit model with alternative-specific attributes

Logit model with socio-economic characteristics

Parameter	Parameter	Parameter	Robust	Robust
number	name	estimate	standard error	*t statistic*
1	ASC_{BM}	-0.731	0.153	-4.77
2	ASC_{LF}	-0.0871	0.700	-0.12
3	ASC_{EF}	-0.319	1.02	-0.31
4	ASC_{MF}	0.274	0.830	0.33
5	β_{users}	0.394	0.108	3.63
6	$\beta_{\text{M\_cost}}$	-1.96	0.246	-7.96
7	$\beta_{\text{F\_cost}}$	-1.79	0.286	-6.25

Summary statistics

Number of observations = 434

$\mathcal{L}(0) = -560.250$

$\mathcal{L}(\hat{\beta}) = -468.791$

$\bar{\rho}^2 = 0.151$

Table 6: Logit model with socio-economic characteristics

Discrete Choice Case Study
Part 2: Forecasting

2 Objectives of Part 2 of the Case Study

This part of the case study deals with forecasting population market shares for different policy scenarios using the MNL model that you estimated in Part 1 of the case study. The reference software is PandasBiogeme. You will also need a spreadsheet application, such as Excel.

The estimated coefficients of a discrete choice model can be used to calculate the choice probability of each alternative for each observation in the sample. In forecasting, however, we are interested in the aggregate market shares for the entire population, or for different segments. It could also be interesting to know how these aggregate market shares are affected by a change in an independent variable.

In this case study, you will learn to aggregate the individual probabilities to obtain market shares, and to test the effects of different alternate scenarios on the market shares.

2.1 Theoretical Reminder

For forecasting, it is necessary to aggregate the individual probabilities to obtain market shares. Aggregation can be done by the method of sample enumeration.” This approach assumes that the random sample that makes up the data is representative of the entire population. The aggregate proportion of the sample choosing each alternative can therefore be used as an estimate of the population market shares. More specifically, if $P(i|x_n)$ is the probability of individual n choosing alternative i given the set of attributes x_n, and there are N_s individuals in the sample, then the proportion choosing alternative i can be estimated from the sample as follows:

$$\hat{W}(i) = \frac{1}{N_s} \sum_{i=1}^{N_s} P(i|x_n) \qquad (1)$$

where

$$P(i|x_n) = \frac{e^{V_{in}}}{\sum_{j \in C_n} e^{V_{jn}}}$$

$\hat{W}(i)$ is then used as an estimate of the proportion $W(i)$ of the population choosing alternative i. If the sample is divided into market segments, the above formula applies to each market segment. See page 146 of Ben-Akiva and Lerman (1985) for details.

2.2 Datasets

Either the Swissmetro mode choice dataset or the residential telephone choice dataset can be used for the case study. The detailed descriptions of the datasets are presented in Appendix A.

2.3 Your Tasks

- Select your best model specification from Part 1
- Use your best model for forecasting different policy alternatives under different market segmentations.

Some sample policy alternatives are provided below for each of the datasets. You should use a spreadsheet application to complete this section. In each case, examine and comment on the change in market share of each alternative for the population as a whole and for the market segments defined in each section. If the forecasts obtained for a particular scenario do not appear to be plausible, then it might be necessary to re-specify the model and re-estimate parameters. As mentioned earlier, a good starting specification would be to use your best specification from Part 1.

2.4 Report Content

Your final report should contain:

- Presentation of your best model specification

- Forecasts for different policy scenarios for different market segmentation

- The PandasBiogeme reports (*.html) and spreadsheets for all discussed models.

2.5 Swissmetro Case

Forecasting the Effect of Change in Swissmetro Cost

Files to use with Biogeme:
Model files: *MNL_SM_socioec.py*
MNL_SM_socioec_Low_simul.py
MNL_SM_socioec_Med_simul.py
MNL_SM_socioec_Hi_simul.py
MNL_SM_socioec_Low_forecast.py
MNL_SM_socioec_Med_forecast.py
MNL_SM_socioec_Hi_forecast.py
Data file: *swissmetro.dat*

In this case study, we forecast the effects of change in Swissmetro costs across different market segments. (See Chapter 6 in Ben-Akiva and Lerman (1985)) for details on forecasting techniques.) Suppose that we know that market segmentation exists on income. We can then consider three markets, namely, low income, medium income and high income that are defined as follows

- Low Income: under $50,000 (INCOME = 0 or 1)
- Medium Income: between $50,000 and $100,000 (INCOME = 2)
- High Income: Over $100,000 (INCOME= 3).

We use the model *MNL_SM_socioec.py* from Part 1 of this case study. The procedure used for forecasting market shares is the following

- Estimate the model with Biogeme.
- Copy the output from the *..._param.py* file (e.g., *MNL_SM_socioec_param.py*), which is generated during the estimation, and paste the output into *MNL_Tel_socioec.py*. Name the new file as *MNL_Tel_socioec_simul.py*. Do these for each segment you create by excluding the other segments from the file. The files *MNL_SM_socioec_Low_simul.py*, *MNL_SM_socioec_Med_simul.py* and *MNL_SM_socioec_Hi_simul.py* are provided in the folder that contains the files relative to this case study.

- Compute the predicted probabilities with Biogeme, based on the estimated model. A simulation file (e.g., provided *MNL_SM_socioec_Low_simul.py*) is used for each market segment to compute these predicted probabilities. Run the simulation files with Biogeme. Notice that the simulation file includes *BIOGEME_OBJECT.SIMULATE* line of code that computes the outputs of interest.

- Biogeme can also be used to compute elasticities, as the expression *Derive* automatically generates the analytical derivatives of the log likelihood function.

- The output of the simulation file contains the observations and their corresponding probabilities. For each market segment, you can also find the market share of each alternative (average of the probabilities of the alternative). The *Exclude* function is applied according to the market segment we study.

We would like to investigate the cost influence on the market shares of Swissmetro. We therefore increase the cost for the Swissmetro by 20% and we forecast the market shares after this change. We modify the simulation files to take into account the cost policy in the following way:

```
# Expressions
SM_COST = DefineVariable('SM_COST', 1.2 * SM_CO * ( GA  ==  0 ))
```

The probabilities and the market shares are computed analogously as for the base case. The results for the base case and the new cost scenario are given in Table 7. We can note a decrease in the market shares of Swissmetro for all market segments. However, it is not an important decrease which indicates that travelers are not very sensitive to cost changes for this new transportation mode (see elasticities in the output file generated from the simulation file).

Figure 1 shows the market shares of the Swissmetro alternative for the low and high income segments as a function of changes in Swissmetro cost. We can see that surprisingly the sensitivity to cost is higher for the high income group than for the low income group. This might indicate that a different model specification should be attempted (for example, one that includes income as an explanatory variable). We can also note that surprisingly the

Swissmetro alternative has a higher market share for the low income group than for the high income group. This could be due to the SP data collection where the price for Swissmetro may not have been high enough to capture the differences between these groups.

	Base case			Forecast		
	Low INC	**Med INC**	**Hi INC**	**Low INC**	**Med INC**	**Hi INC**
CAR	14	28	32	16	31	36
TRAIN	23	12	9	24	13	10
SM	62	60	60	60	56	54

Table 7: Market Shares (percent) for increased cost of Swissmetro

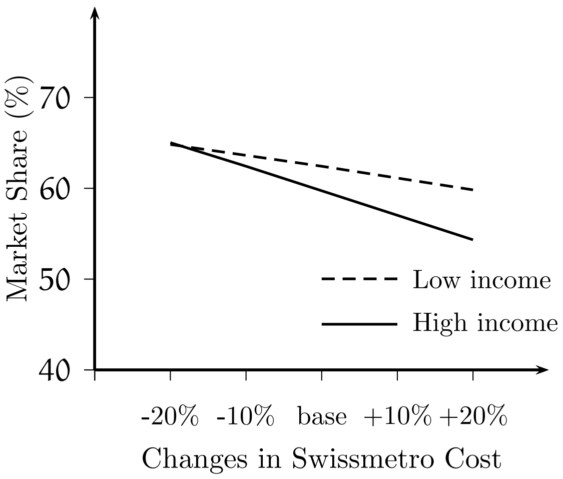

Figure 1: Swissmetro: Market Shares for Low and High Income Segments

It would also be interesting to investigate the impact on the market shares for the following two policy scenarios:

- The Swissmetro SA has decided to provide a 20% discount to youths (age < 24) and 50% discount to elderly (age > 65) when using Swissmetro. To compensate for the lost revenue, the company considers increasing the general Swissmetro fare uniformly by 10%.

- The Swissmetro SA is considering an alternative option of making incremental investment in Swissmetro and initially starting with half the maglev trains they originally planned to purchase. To meet the growing demand, they are also considering doubling the frequency of the regular trains.

2.6 Choice of Residential Telephone Services Case

Forecasting the Effect of Change in Cost Across Market Segments

Files to use with Biogeme:

Model files:	*MNL_Tel_socioec.py*
	MNL_Tel_socioec_param.py
	MNL_Tel_socioec_simul.py
	MNL_Tel_socioec_simul2.py
Data file:	*telephone.dat*
Excel worksheet:	*telephone.xls*

In this case study, we forecast the effects of change in cost of alternatives across different market segments (See Chapter 6 in Ben-Akiva and Lerman (1985) for details on forecasting techniques.) Suppose that we know that market segmentation exists on income (Inc). We can then consider three markets, namely, low income, medium income and high income. We define these market segments as follows

- Low Income: under $20,000 (Inc = 1 or 2)
- Medium Income: Between $20,000 and $40,000 (Inc = 3 or 4)
- High Income: Over $40,000 (Inc = 5).

We use the model *MNL_Tel_socioec.py* from Part 1 of this case study. The procedure used for forecasting market shares is as follows:

- Estimate the model with Biogeme.
- Copy the output from *MNL_Tel_socioec_param.py*, which is generated during the estimation, and paste the output into *MNL_Tel_socioec.py*. Name the new file as *MNL_Tel_socioec_simul.py*.

- Introduce variables for the estimated probability of choosing each alternative; in the example code that is provided, the variables are named probBM, probSM, etc. Run the simulation file with Biogeme. Notice that the simulation file includes *BIOGEME_OBJECT.SIMULATE* line of code that computes the outputs of interest.

- Biogeme can also be used to compute elasticities, as the expression *Derive* automatically generates the analytical derivatives of the log likelihood function.

- Excel can be used for editing and processing the data and probabilities. For example, you can open the data file with Excel and paste the probabilities or market shares which are found in the output file into the Excel file.

We have provided an Excel file (*telephone.xls*) containing the observations and their corresponding probabilities. This file has also been used for computing market shares by averaging the alternative probabilities over each income-based market segment.

Assume that the telephone company in an effort to increase revenues considers raising the fixed costs for alternatives SM, LF, EF and MF by \$4, \$6, \$7 and \$11, respectively. We would like to forecast the market shares after this change. We modify the file *MNL_Tel_socioec_simul.py* to take into account the cost policy in the following way:

```
logcostBM  = DefineVariable('logcostBM', log(cost1))
logcostSM  = DefineVariable('logcostSM', log(cost2+4))
logcostLF  = DefineVariable('logcostLF', log(cost3+6))
logcostEF  = DefineVariable('logcostEF', log(cost4+7))
logcostMF  = DefineVariable('logcostMF', log(cost5+11))
```

We name this file *MNL_Tel_socioec_simul2.py*, and it is provided with this case study. We simulate again using Biogeme in order to obtain the alternative probabilities under this new scenario. The probabilities are again copied from the output file and pasted into the Excel file (*telephone.xls*), and the market shares by market segment can be computed in the same way as for the base case. The results for the base case and the new cost scenario are given in Table 8. The cost change does not result in important changes for the EF and MF alternatives. There is, however, an important increase for all market segments towards the BM alternative.

Figure 2 shows the market shares of the standard measure (SM) alternative for the low and high income segments as a function of changes in SM cost. Such figures can be created readily using Excel. We can see that the sensitivity to cost is about the same for the two market segments. The SM alternative has however a higher market share for the low income group than for the high income group.

	Base case			Forecast		
	Low INC	**Med INC**	**Hi INC**	**Low INC**	**Med INC**	**Hi INC**
BM	19	14	13	34	26	23
SM	30	28	23	22	21	18
LF	40	43	41	34	39	37
EF	0	1	2	0	1	2
MF	11	14	21	10	13	19

Table 8: Market Shares (percent)

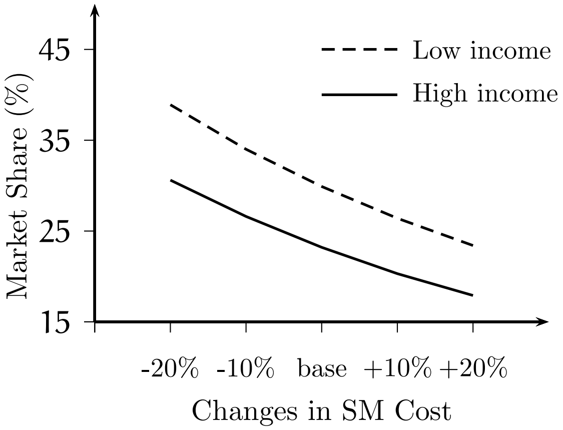

Figure 2: Market Shares for Low and High Income Segments, SM alternative

It would also be interesting to investigate the impact on the market shares for the following two policy scenarios:

- Due to legal restrictions, the telephone company is expected to subsidize the telephone costs of elderly households (a household with at least 1 household member older than 65 years) and low-income households (a household with annual household income less than \$20,000). The telephone company must provide a 50% discount to these households' telephone costs. To compensate for these losses in the revenues, the company considers increasing the telephone costs of all other households uniformly by 10%.

- Due to recession, the number of employed persons per household has reduced to half of the previous scenario and the telephone company has decided to provide a 20% discount for households that have no employed persons. To compensate for these losses in the revenues, the company considers increasing the telephone costs of households with at least one employed person by 10%.

A Datasets

A.1 Dataset1: Swissmetro Case

This dataset consists of survey data collected on the trains between St. Gallen and Geneva, Switzerland, during March 1998. The respondents provided information in order to analyze the impact of the modal innovation in transportation, represented by the Swissmetro, a revolutionary mag-lev underground system, against the usual transport modes represented by car and train.

Context

Innovation in the market for intercity passenger transportation is a difficult enterprise as the existing modes: private car, coach, rail as well as regional and 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 the major centers or direct regional jet services between smaller countries. The Swissmetro SA in Geneva is promoting such an innovation: a mag-lev 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 Collection

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 were 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 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 10529 relevant license plates were recorded during September 1997. The central Swiss car license agency had agreed to send up to 10000 owners of these cars a survey-pack. Until April 1998, 9658 letters were mailed, of which 1758 were returned. A total of 1070 persons filled in the survey completely and were willing to participate in the second 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 9 and 10, and the descriptive statistics are summarized in Table 11. A more detailed description of the data set as well as the data collection procedure is given in: M. Bierlaire and K.W. Axhausen and G. Abay, The acceptance of modal innovation: The case of swissmetro, in Proceedings of the 1st Swiss Transportation Research Conference, Ascona, Switzerland, March 13, 2001.

Variable	**Description**
GROUP	Different groups in the population. 2: current rail users, 3: current road users
SURVEY	Equivalent to GROUP but using different coding: 0: train users, 1: car users
SP	It is fixed to 1 (stated preference survey)
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 price, 4: One way normal price, 5: Half day, 6: Annual season ticket, 7: Annual season ticket Junior or Senior, 8: Free travel after 7pm card, 9: Group ticket, 10: Other
WHO	Who pays (0: unknown, 1: self, 2: employer, 3: half-half)
LUGGAGE	0: none, 1: one piece, 3: several pieces
AGE	It captures the age class of individuals. The age-class coding scheme is of the type: 1: age≤24, 2: 24<age≤39, 3: 39<age≤54, 4: 54<age≤ 65, 5: 65 <age, 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: unknown
GA	Variable capturing the effect of the Swiss annual season ticket for the rail system and most local public transport. It is 1 if the individual owns a GA, zero otherwise.
ORIGIN	Travel origin (a number corresponding to a Canton, see Table 12)

Table 9: Description of variables

Variable	**Description**
DEST	Travel destination (a number corresponding to a Canton, see Table 12)
TRAIN_AV	Train availability dummy
CAR_AV	Car availability dummy
SM_AV	SM availability dummy
TRAIN_TT	Train travel time [minutes]. Travel times are door-to-door making assumptions about car-based distances (1.25*crow-flight distance)
TRAIN_CO	Train cost [CHF]. If the traveler has a GA, this cost equals the cost of the annual ticket.
TRAIN_HE	Train headway [minutes] Example: If there are two trains per hour, the value of TRAIN_HE is 30.
SM_TT	SM travel time [minutes] considering the future Swissmetro speed of 500 km/h
SM_CO	SM cost [CHF] calculated at the current relevant rail fare, without considering GA, multiplied by a fixed factor (1.2) to reflect the higher speed.
SM_HE	SM headway [minutes] Example: If there are two Swissmetros per hour, the value of SM_HE is 30.
SM_SEATS	Seats configuration in the Swissmetro (dummy). Airline seats (1) or not (0).
CAR_TT	Car travel time [minutes]
CAR_CO	Car cost [CHF] considering a fixed average cost per kilometer (1.20 CHF/km)
CHOICE	Choice indicator. 0: unknown, 1: Train, 2: SM, 3: Car

Table 10: Description of variables

Variable	**Min**	**Max**	**Mean**	**St. Dev.**
GROUP	2	3	2.63	0.48
SURVEY	0	1	0.63	0.48
SP	1	1	1.00	0.00
ID	1	1192	596.50	344.12
PURPOSE	1	9	2.91	1.15
FIRST	0	1	0.47	0.50
TICKET	1	10	2.89	2.19
WHO	0	3	1.49	0.71
LUGGAGE	0	3	0.68	0.60
AGE	1	6	2.90	1.03
MALE	0	1	0.75	0.43
INCOME	0	4	2.33	0.94
GA	0	1	0.14	0.35
ORIGIN	1	25	13.32	10.14
DEST	1	26	10.80	9.75
TRAIN_AV	1	1	1.00	0.00
CAR_AV	0	1	0.84	0.36
SM_AV	1	1	1.00	0.00
TRAIN_TT	31	1049	166.63	77.35
TRAIN_CO	4	5040	514.34	1088.93
TRAIN_HE	30	120	70.10	37.43
SM_TT	8	796	87.47	53.55
SM_CO	6	6720	670.34	1441.59
SM_HE	10	30	20.02	8.16
SM_SEATS	0	1	0.12	0.32
CAR_TT	0	1560	123.80	88.71
CAR_CO	0	520	78.74	55.26
CHOICE	1	3	2.15	0.63

Table 11: Descriptive statistics

Number	**Canton**
1	ZH
2	BE
3	LU
4	UR
5	SZ
6	OW
7	NW
8	GL
9	ZG
10	FR
11	SO
12	BS
13	BL
14	Schaffhausen
15	AR
16	AI
17	SG
18	GR
19	AG
20	TH
21	TI
22	VD
23	VS
24	NE
25	GE
26	JU

Table 12: Coding of Cantons

A.2 Dataset 2: Choice of Residential Telephone Services Case

Context

Local telephone service typically involves the choice between flat (i.e., a fixed monthly charge for unlimited calls within a specified geographical area) and measured (i.e., a reduced fixed monthly charge for a limited number of calls and additional usage charges for additional calls) services. Various flat rate services differ by the size of the geographical area within which calling is provided at no extra charge, the monthly charge being higher for larger areas. Measured services differ with respect to the threshold number (or dollar value) of calls beyond which the customer is charged. The availability of each service may depend on the geographical location within the service area.

In developing a model of the residential demand for local telephone service, it is necessary to explicitly account for the inter-relationship between class of service choice and usage patterns. For example, expected usage patterns will influence the household's choice of service option since households with high usage levels typically could minimize their monthly bill for local telephone service by choosing some sort of flat rate service, while households with relatively low usage would be better off with a measured service. Given that a household has chosen a particular service option, usage patterns would be dependent to a certain extent upon the service option that is chosen since it determines the marginal price of calls. To accommodate these interrelationships, the model representing the household's choice of calling patterns and service options needs to include:

1. choice of the service option, which is modeled conditional upon the calling portfolio chosen by the household;
2. choice of the calling portfolio or the usage pattern as represented by the number and duration of calls by time of day and calling band.

This case study deals only with the first choice.

Data Collection

A household survey was conducted in 1984 for a telephone company among 434 households in Pennsylvania. The dataset involves choices among five calling plans and consists of various attributes and socio-economic characteristics. It was originally used to develop a model system to predict residential telephone demand. For more information, see: K. Train, D. McFadden and M. Ben-Akiva, The demand for local telephone service: a fully discrete model of residential calling patterns and service choices. In Rand Journal of Economics, 1987.

Variables and Descriptive Statistics

In the current application, five types of services are involved: two measured options and three flat options. The availability of these service options varies depending upon geographic location. Table 13 below lists the five service alternatives and their availability within the different service areas. Names and definitions of the variables are shown in Table 14. Some descriptive statistics of the dataset are summarized in Table 15.

Complications caused by very few respondents choosing alternative 4: If you examine the dataset, you see that only 3 of the respondents chose alternative 4 (extended area flat service). This implies that it is not possible to estimate numerous alternative specific coefficients for alternative 4. The intuition is that the dataset does not provide enough information on why people chose or did not choose alternative 4. If you try to estimate too many alternative specific coefficients for alternative 4, you get ”Singularity in the Hessian” error, and in order to estimate the model you have to reduce the number of coefficients specific to alternative 4. A practical solution to this problem is to use an “enriched sample” although such a sample is not available here. It is however not recommended to omit the observations for which the chosen alternative is 4 or combine alternative 4 with a different alternative.

		Availability		
Service option	Description	metro, suburban, some perimeter areas	other perimeter areas	non-metro areas
1. Budget measured	No fixed monthly charge; usage charges apply to each call made.	yes	yes	yes
2. Standard measured	A fixed monthly charge covers up to a specified dollar amount (greater than the fixed charge) of local calling, after which usage charges apply to each call made.	yes	yes	yes
3. Local flat	A greater monthly charge that may depend upon residential location; unlimited free calling within local calling area; usage charges apply to calls made outside local calling area.	yes	yes	yes
4. Extended area flat	A further increase in the fixed monthly charge to permit unlimited free calling within an extended area.	no	yes	no
5. Metro area flat	The greatest fixed monthly charge that permits unlimited free calling within the entire metropolitan area.	yes	yes	no

Table 13: Service options and their availability

Name	**Description**
age0 age1 age2 age3 age4 age5 age6 age7	number of household members under age 6 number of household members age 6-12 number of household members age 13-19 number of household members age 20-29 number of household members age 30-39 number of household members age 40-54 number of household members age 55-64 number of household members 65 and older
area	location of household residence 1=metro, 2=suburban, 3=perimeter with extended, 4=perimeter without extended, 5=non-metro
avail1, avail2, avail3, avail4, avail5	binary indicators of availability of each option. availX=0 if alternative X is not available to the household, availX=1 if alternative X is available to the household
choice	chosen alternative (dependent variable) 1=budget measured, 2=standard measured, 3=local flat, 4=extended flat, 5=metro flat
cost1, cost2, cost3, cost4, cost5	costX = monthly cost (in $) of alternative X.
employ	number of household members employed
inc	annual household income 1=under $10,000, 2=$10,000-20,000, 3=$20,000-30,000, 4=$30,000-40,000, 5=0ver $40,000
ones	ones = 1 for all observations
status	marital status 1=single, 2=married, 3=widowed, 4=divorced, 5=other
users	number of phone users in household

Table 14: Description of variables

	mean	max	min	stand dev	range
age0	0.21	4	0	0.53	4
age1	0.23	3	0	0.58	3
age2	0.24	4	0	0.67	4
age3	0.41	3	0	0.71	3
age4	0.44	2	0	0.73	2
age5	0.36	2	0	0.67	2
age6	0.31	3	0	0.61	3
age7	0.38	2	0	0.65	2
area	2.93	5	1	1.65	4
avail1	1.00	1	1	0.00	0
avail2	1.00	1	1	0.00	0
avail3	1.00	1	1	0.00	0
avail4	0.03	1	0	0.17	1
avail5	0.65	1	0	0.48	1
choice	2.65	5	1	1.17	4
cost1	11.73	433.5	3.28	24.13	430.22
cost2	11.49	432.8	5.78	23.90	427.02
cost3	14.82	435.5	7.03	23.56	428.47
cost4	62.19	433.03	10.48	117.88	422.55
cost5	27.48	38.28	23.28	4.17	15
employ	1.07	3	0	0.89	3
inc	2.53	5	1	1.28	4
ones	1.00	1	1	0.00	0
status	2.22	5	1	0.91	4
users	2.30	6	1	1.28	5

Table 15: Descriptive Statistics