EPFL
ENAC TRANSP-OR
Prof. M. Bierlaire

Mathematical Modeling of Behavior
Fall 2017

Airline itinerary case

1 Model Specification with Generic Attributes

Files to use with Biogeme:

Model file:	*MNL_airline_generic.py*
Data file:	*airline.dat*

The choice set consists of the following three alternatives:

1. a non-stop flight,
2. a flight with one stop on the same airline,
3. a flight with one stop and a change of airline.

We define the deterministic part of the utility for the household by including the alternative specific constants (ASCs) and five attributes, namely fare (in the unit of 100$, in order to reduce numerical issues), legroom, total travel time (Total_TT), early and late schedule delays (SchedDE and SchedDL), with their respective generic coefficients β_{Fare}, β_{Legroom}, $\beta_{\text{Total\_TT}}$, β_{SchedDE} and β_{SchedDL}:

$$\begin{aligned}
V_1 &= \text{ASC}_1 + \beta_{\text{Fare}} \cdot \text{Fare}_1 + \beta_{\text{Legroom}} \cdot \text{Legroom}_1 + \beta_{\text{Total\_TT}} \cdot \text{Total\_TT}_1 \\
&\quad + \beta_{\text{SchedDE}} \cdot \text{SchedDE}_1 + \beta_{\text{SchedDL}} \cdot \text{SchedDL}_1 \\
V_2 &= \text{ASC}_2 + \beta_{\text{Fare}} \cdot \text{Fare}_2 + \beta_{\text{Legroom}} \cdot \text{Legroom}_2 + \beta_{\text{Total\_TT}} \cdot \text{Total\_TT}_2 \\
&\quad + \beta_{\text{SchedDE}} \cdot \text{SchedDE}_2 + \beta_{\text{SchedDL}} \cdot \text{SchedDL}_2 \\
V_3 &= \text{ASC}_3 + \beta_{\text{Fare}} \cdot \text{Fare}_3 + \beta_{\text{Legroom}} \cdot \text{Legroom}_3 + \beta_{\text{Total\_TT}} \cdot \text{Total\_TT}_3 \\
&\quad + \beta_{\text{SchedDE}} \cdot \text{SchedDE}_3 + \beta_{\text{SchedDL}} \cdot \text{SchedDL}_3
\end{aligned}$$

One of the alternative specific constants (arbitrarily ASC_1) is normalized to zero for identification. The corresponding alternative is the reference alternative for the ASCs. This is important for the interpretation we will perform in the next paragraphs.

The results are presented in Table 1. Note that we have excluded observations for which the arrival time record is missing by including the following expression into the code:

Generic MNL estimation				
Parameter	Parameter	Parameter	Robust	Robust
number	name	estimate	standard error	*t statistic*
1	ASC_2	-1.31	0.126	-10.36
2	ASC_3	-1.54	0.126	-12.15
3	β_{Fare}	-0.0194	0.000796	-24.42
4	β_{Legroom}	0.225	0.0266	8.45
5	β_{SchedDE}	-0.139	0.0163	-8.55
6	β_{SchedDL}	-0.104	0.0137	-7.59
7	$\beta_{\text{Total\_TT}}$	-0.300	0.0670	-4.48
Summary statistics				
Number of observations = 3609				
$\mathcal{L}(0) = -3964.892$				
$\mathcal{L}(\hat{\beta}) = -2321.153$				
$\bar{\rho}^2 = 0.413$				

Table 1: Logit model with generic attributes

```
BIOGEME_OBJECT.EXCLUDE = ArrivalTimeHours_1 == -1
```

Given our specification, and everything being equal, an ASC with negative sign indicates a lower utility level for the corresponding alternative compared to the normalized one (i.e., the first one). As it can be observed in Table 1, this is the case for both other alternatives (ASC_2 and ASC_3 are negative and statistically significant). It means that alternative 1 is preferred to alternatives 2 and 3, i.e., alternative without stop is preferred to alternatives with stops all other things being equal.

The parameter related to leg room has a positive sign and it is significantly different from zero. It implies that more room for legs increases the utility of the alternative. For other parameters, like fare, delays and travel time, the sign is negative. It means that all these factors have a negative impact on utility: they make the alternative less likely to be chosen.

2 Model Specification with Alternative-Specific Coefficients

File to develop using the same dataset as before:
Model file: MNL_airline_specific.py

Next we present a model (unrestricted) with alternative-specific travel time coefficients and we compare it with the (restricted) model with generic coefficients presented in the previous section. We carry out a statistical test (likelihood ratio test) to assess if one specification is significantly better than the other. We perform the analysis on the coefficient of the travel time. The deterministic utilities for this model with alternative-specific travel times are:

$$
\begin{aligned}
V_1 &= \text{ASC}_1 + \beta_{\text{Fare}} \cdot \text{Fare}_1 + \beta_{\text{Legroom}} \cdot \text{Legroom}_1 + \beta_{\text{Total\_TT\_1}} \cdot \text{Total\_TT}_1 \\
&\quad + \beta_{\text{SchedDE}} \cdot \text{SchedDE}_1 + \beta_{\text{SchedDL}} \cdot \text{SchedDL}_1 \\
V_2 &= \text{ASC}_2 + \beta_{\text{Fare}} \cdot \text{Fare}_2 + \beta_{\text{Legroom}} \cdot \text{Legroom}_2 + \beta_{\text{Total\_TT\_2}} \cdot \text{Total\_TT}_2 \\
&\quad + \beta_{\text{SchedDE}} \cdot \text{SchedDE}_2 + \beta_{\text{SchedDL}} \cdot \text{SchedDL}_2 \\
V_3 &= \text{ASC}_3 + \beta_{\text{Fare}} \cdot \text{Fare}_3 + \beta_{\text{Legroom}} \cdot \text{Legroom}_3 + \beta_{\text{Total\_TT\_3}} \cdot \text{Total\_TT}_3 \\
&\quad + \beta_{\text{SchedDE}} \cdot \text{SchedDE}_3 + \beta_{\text{SchedDL}} \cdot \text{SchedDL}_3
\end{aligned}
$$

Note that instead of only $\beta_{\text{Total\_TT}}$, we have now $\beta_{\text{Total\_TT\_1}}, \beta_{\text{Total\_TT\_2}}$ and $\beta_{\text{Total\_TT\_3}}$. The results for the unrestricted model are reported in Table 2.

Alternative-specific MNL estimation				
Parameter number	Parameter name	Parameter estimate	Robust standard error	Robust *t statistic*
1	ASC_2	-1.43	0.183	-7.81
2	ASC_3	-1.64	0.192	-8.53
3	β_{Fare}	-0.0193	0.000802	-24.05
4	β_{Legroom}	0.226	0.0267	8.45
5	β_{SchedDE}	-0.139	0.0163	-8.53
6	β_{SchedDL}	-0.104	0.0137	-7.59
7	$\beta_{\text{Total\_TT}_1}$	-0.332	0.0735	-4.52
8	$\beta_{\text{Total\_TT}_2}$	-0.299	0.0696	-4.29
9	$\beta_{\text{Total\_TT}_3}$	-0.302	0.0699	-4.32
Summary statistics				
Number of observations = 3609				
$\mathcal{L}(0) = -3964.892$				
$\mathcal{L}(\hat{\beta}) = -2320.447$				
$\bar{\rho}^2 = 0.412$				

Table 2: Logit model with alternative-specific travel-time attributes

Generic vs Specific Test Under the null hypothesis:

$$H_0 : \beta_{\text{Total\_TT\_1}} = \beta_{\text{Total\_TT\_2}} = \beta_{\text{Total\_TT\_3}}$$

We reject null hypothesis (generic travel time coefficient) if :

$$-2(L_R - L_U) > \chi_{((1-\alpha),df)}$$

Next we describe the standard steps to perform the test:

1. L_R and L_U represent the log-likelihood for both the restricted and the unrestricted models:
$$\begin{aligned} L_R &= -2321.153 \\ L_U &= -2320.447 \end{aligned}$$
2. The degree of freedom is given by the difference in the number of estimated parameters between the models:
$$df = K_U - K_R = 9 - 7 = 2$$
3. $-2(L_R - L_U) = -2(-2321.153 + 2320.447) = 1.412$
4. The critical value for $\chi_{(0.95,2)}$ is 5.99.
5. We conclude that we cannot reject the null hypothesis H_0 and we keep the generic coefficient.

3 Inclusion of Socio-Economic Characteristics

File to develop using the same dataset as before:
Model file: MNL_airline_socioecon.py

It is reasonable to assume that people make choices not only in relation to the attributes that characterize the alternatives but also depending on some personal characteristics or socioeconomic indicators. The availability of individual-specific information gives us the opportunity to model partly the heterogeneity present in the population. We modify the previous model by adding income (continuous income, *Cont_Income* in the airline dataset) of respondents into the utilities.

$$\begin{aligned} V_1 &= \text{ASC}_1 + \beta_{\text{Fare}} \cdot \text{Fare}_1 + \beta_{\text{Legroom}} \cdot \text{Legroom}_1 + \beta_{\text{Total\_TT}} \cdot \text{Total\_TT}_1 \\ &\quad + \beta_{\text{SchedDE}} \cdot \text{SchedDE}_1 + \beta_{\text{SchedDL}} \cdot \text{SchedDL}_1 + \beta_{\text{Inc}_1} \cdot \text{Income} \\ V_2 &= \text{ASC}_2 + \beta_{\text{Fare}} \cdot \text{Fare}_2 + \beta_{\text{Legroom}} \cdot \text{Legroom}_2 + \beta_{\text{Total\_TT}} \cdot \text{Total\_TT}_2 \\ &\quad + \beta_{\text{SchedDE}} \cdot \text{SchedDE}_2 + \beta_{\text{SchedDL}} \cdot \text{SchedDL}_2 + \beta_{\text{Inc}_2} \cdot \text{Income} \\ V_3 &= \text{ASC}_3 + \beta_{\text{Fare}} \cdot \text{Fare}_3 + \beta_{\text{Legroom}} \cdot \text{Legroom}_3 + \beta_{\text{Total\_TT}} \cdot \text{Total\_TT}_3 \\ &\quad + \beta_{\text{SchedDE}} \cdot \text{SchedDE}_3 + \beta_{\text{SchedDL}} \cdot \text{SchedDL}_3 + \beta_{\text{Inc}_3} \cdot \text{Income} \end{aligned}$$

Since the variable of the income does not vary between the alternatives and only differences in utilities matter, we need to normalize one alternative to zero. We interpret the estimated coefficients for the remaining alternatives with respect to the reference alternative, which arbitrarily is alternative 1. It is similar to what we did when specifying alternative specific constants.

We assume that the income of the respondent affects differently each alternative. Note that since the values of the fares are expressed in $ and the values for the income are expressed in 1000 $, the orders of magnitude of the associated parameters are different. One can avoid numerical issues by adapting the units (*e.g.* expressing the income in 10000 $ instead).

In this model, we need to deal with missing data for income. One solution is to exclude missing data (-1) from the data set by including the following instruction into the code, that tells Biogeme not to consider the observations whose values for *Cont_Income* are -1:

```
BIOGEME_OBJECT.EXCLUDE = (Cont_Income== -1) > 0
```

The estimation results of this model are reported in Table 4.

Socio-economic MNL estimation				
Parameter number	Parameter name	Parameter estimate	Robust standard error	Robust *t statistic*
1	ASC_2	-1.12	0.147	-7.59
2	ASC_3	-0.989	0.156	-6.35
3	β_{Fare}	-0.0196	0.000861	-22.72
4	β_{Income_2}	-0.00104	0.000665	-1.56
5	β_{Income_3}	-0.00462	0.000885	-5.22
6	β_{Legroom}	0.219	0.0287	7.64
8	β_{SchedDE}	-0.139	0.0173	-7.99
9	β_{SchedDL}	-0.0940	0.0146	-6.44
10	$\beta_{\text{Total\_TT}}$	-0.339	0.0719	-4.72
Summary statistics				
Number of observations = 3111				
$\mathcal{L}(0) = -3417.783$				
$\mathcal{L}(\hat{\beta}) = -2004.285$				
$\bar{\rho}^2 = 0.411$				

Table 3: Logit model with socio-economic variables

File to develop using the same dataset as before:
Model file: *MNL_airline_socioecon_mi.py*

A second and better solution consists in defining another variable, called "MissingIncome" (MI). "MissingIncome" is equal to 1 if *Cont_Income* =-1. Still these missing values exist in the *Cont_Income* column. To separate their effect we further define:

```
Cont_Income_full = DefineVariable('Cont_Income_full', \
Cont_Income * (Cont_Income != -1) )
```

We do not exclude any observation any more. We just modify the utility functions as follows:

$$\begin{aligned}
V_1 &= \beta_{\text{Fare}}\text{Fare}_1 + \beta_{\text{Legroom}}\text{Legroom}_1 + \beta_{\text{Total\_TT}}\text{Total\_TT}_1 \\
&\quad +\beta_{\text{SchedDE}}\text{SchedDE}_1 + \beta_{\text{SchedDL}}\text{SchedDL}_1 \\
V_2 &= \text{ASC}_2 + \beta_{\text{Fare}}\text{Fare}_2 + \beta_{\text{Legroom}}\text{Legroom}_2 + \beta_{\text{Total\_TT}}\text{Total\_TT}_2 \\
&\quad +\beta_{\text{SchedDE}}\text{SchedDE}_2 + \beta_{\text{SchedDL}}\text{SchedDL}_2 + \beta_{\text{Inc}_2}\text{Cont\_Income\_full} \\
&\quad +\beta_{\text{MI}}\text{MissingIncome} \\
V_3 &= \text{ASC}_3 + \beta_{\text{Fare}}\text{Fare}_3 + \beta_{\text{Legroom}}\text{Legroom}_3 + \beta_{\text{Total\_TT}}\text{Total\_TT}_3 \\
&\quad +\beta_{\text{SchedDE}}\text{SchedDE}_3 + \beta_{\text{SchedDL}}\text{SchedDL}_3 + \beta_{\text{Inc}_3}\text{Cont\_Income\_full} \\
&\quad +\beta_{\text{MI}}\text{MissingIncome}
\end{aligned}$$

Note that this new term in the utility function can only appear in two of the three utility functions to be able to identify it. We choose arbitrarily to leave it out in V_1. The estimation results for the model with the variable "MissingIncome" are reported in table 4.

In both approaches we have specified two different β parameters associated with the attribute *Cont_Income*. β_{Inc} for alternative 1 has been normalized to zero. The two parameter estimates have negative signs, implying that the higher the income of the respondent, the lower the likelihood for choosing these two alternatives (with stops) compared to the first one (without stops). The parameter β_{MI} has no interpretation.

---

mbi/ ek/ afa /mpp

Generic logit model estimation				
Parameter	Parameter	Parameter	Robust	Robust
number	name	estimate	standard error	*t statistic*
1	ASC_2	-1.14	0.139	-8.16
2	ASC_3	-1.12	0.146	-7.65
3	β_{Fare}	-0.0198	0.000804	-24.60
4	β_{Inc_2}	-0.00133	0.000658	-2.02
5	β_{Inc_3}	-0.00424	0.000824	-5.14
6	β_{Legroom}	0.228	0.0267	8.53
7	β_{MI}	-0.399	0.137	-2.92
8	β_{SchedDE}	-0.139	0.0162	-8.53
9	β_{SchedDL}	-0.104	0.0138	-7.51
10	$\beta_{\text{Total\_TT}}$	-0.302	0.0670	-4.51
Summary statistics				
Number of observations = 3609				
$\mathcal{L}(0) = -3964.892$				
$\mathcal{L}(\hat{\beta}) = -2303.217$				
$\bar{\rho}^2 = 0.415$				

Table 4: Logit model with socio-economic variables and MissingIncome