EPFL
ENAC TRANSP-OR
Prof. M. Bierlaire

Mathematical Modeling of Behavior
Fall 2017

Airline itinerary case

1 Nonlinear specifications

The models studied previously were specified with linear-in-parameter formulations of the deterministic parts of the utilities (i.e. parameters that remain constant throughout the whole range of the values of each variable). However, in some cases, non-linear specifications may be more justified. In this section, we test three different nonlinear specifications of the deterministic utility functions: a piecewise linear specification of the time parameter of the non-stop itinerary, a power series method and Box-Cox transformation.

The base model that we will consider in this section is the one you developed during the previous session with alternative-specific coefficients for the travel time (*MNL_airline_specific.py*). The deterministic utilities for this model are the following:

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

1.1 Piecewise Linear Approximation

File to develop using the airline dataset:
Model file: *MNL_airline_piecewise.py*

In this first example, we want to test the hypothesis that the value of the travel time parameter for the non-stop itinerary alternative assumes different values for different ranges of values of the variable itself. We split the range of values for the total travel time of alternative 1 $\text{Total\_TT}_1 \in [0.67, 6.35]$ (`TripTimeHours_1` in the data) into three different intervals:

- $\text{Total\_TT}_1\_1 \in [0, 2]$
- $\text{Total\_TT}_1\_2 \in]2, 3]$
- $\text{Total\_TT}_1\_3 > 3$

The systematic utility expression for the non-stop alternative is the following:

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

To model the three intervals we need to define in PythonBiogeme three accumulative variables to represent the total travel time. We use the specification that has been seen in the lecture. More precisely,

$$\text{Total\_TT}_1\text{\_1} = \begin{cases} TripTimeHours_1 & \text{if } TripTimeHours_1 < 2 \\ 2 & \text{if } TripTimeHours_1 \geq 2 \end{cases} \tag{1}$$

$$= \min(TripTimeHours_1, 2) \tag{2}$$

$$\text{Total\_TT}_1\text{\_2} = \begin{cases} 0 & \text{if } TripTimeHours_1 < 2 \\ TripTimeHours_2 & \text{if } 2 \leq TripTimeHours_1 < 3 \\ 1 & \text{if } TripTimeHours_1 > 3 \end{cases} \tag{3}$$

$$= \max(0, \min(TripTimeHours_1 - 2, 1)) \tag{4}$$

$$\text{Total\_TT}_1\text{\_3} = \begin{cases} 0 & \text{if } TripTimeHours_1 < 3 \\ TripTimeHours_1 - 3 & \text{if } TripTimeHours_1 \geq 2 \end{cases} \tag{5}$$

$$= \max(0, TripTimeHours_1 - 3) \tag{6}$$

It is easy to see that the previous specification represents the total travel time. For instance, consider an individual with a travel time of $TripTimeHours_1 = 2.5$. In this case, the three variables will be calculated as follows:

1. $\text{Total\_TT}_1\text{\_1} = \min(TripTimeHours_1, 2) = \min(2.5, 2) = 2$
2. $\text{Total\_TT}_1\text{\_2} = \max(0, \min(TripTimeHours_1 - 2, 1))) = \max(0, \min(0.5, 1)) = \max(0, 0.5) = 0.5$
3. $\text{Total\_TT}_1\text{\_3} = \max(0, TripTimeHours_1 - 3) = \max(0, -0.5) = 0$

Thus, the original value of the travel time ($TripTimeHours_1$) is decomposed into the three variables ($TripTimeHours_1 = \text{Total\_TT}_1\text{\_1} + \text{Total\_TT}_1\text{\_2} + \text{Total\_TT}_1\text{\_3}$). You can easily define these three variables in PythonBiogeme with the instruction `DefineVariable` by using the formulations (2), (4) and (6).

The estimation results for this specification are shown in Table 1. All time coefficients related to the piecewise linear expression are negative. The coefficient associated with short trips (shorter than 2 hours) is the largest in absolute value, meaning that the same increase of travel time penalizes the utility of the non-stop alternative more if the trip is shorter than 2 hours than if is longer than 2 hours. Similarly, the coefficient associated with trips with an intermediate duration (between 2 and 3 hours) penalizes more the utility of the non-stop alternative than if the trip lasts longer than 3 hours.
We perform a likelihood ratio test where the restricted model is the one with linear travel time for the non-stop alternative (*MNL_airline_specific.py*) and the unrestricted model is the piecewise linear specification (*MNL_airline_piecewise.py*). The null hypothesis is given as follows:

$$H_0 : \beta_{Total\_TT_1\_1} = \beta_{Total\_TT_1\_2} = \beta_{Total\_TT_1\_3}$$

The statistic for the likelihood ratio test is the following:

$$-2(-2320.447 + 2315.041) = 10.812$$

Since $\chi^2_{0.95,2} = 5.99$, we can reject the null hypothesis of a linear travel time for the non-stop alternative at a 95% level of confidence.

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	ASC_2	-2.32	0.411	-5.65	0.00
2	ASC_3	-2.55	0.438	-5.83	0.00
3	β_{Fare}	-0.0193	0.000799	-24.10	0.00
4	$\beta_{Legroom}$	0.227	0.0267	8.51	0.00
5	$\beta_{SchedDE}$	-0.140	0.0165	-8.47	0.00
6	$\beta_{SchedDL}$	-0.105	0.0137	-7.64	0.00
7	$\beta_{Total\_TT_1\_1}$	-0.824	0.238	-3.46	0.00
8	$\beta_{Total\_TT_1\_2}$	-0.444	0.188	-2.36	0.02
9	$\beta_{Total\_TT_1\_3}$	-0.229	0.0889	-2.57	0.01
10	β_{Total_TT2}	-0.301	0.0701	-4.29	0.00
11	β_{Total_TT3}	-0.301	0.0701	-4.29	0.00

Summary statistics

Number of observations = 3609

Number of excluded observations = 0

Number of estimated parameters = 11

$$\begin{aligned} \mathcal{L}(\beta_0) &= -3964.892 \\ \mathcal{L}(\hat{\beta}) &= -2315.041 \\ -2[\mathcal{L}(\beta_0) - \mathcal{L}(\hat{\beta})] &= 3299.701 \\ \rho^2 &= 0.416 \\ \bar{\rho}^2 &= 0.413 \end{aligned}$$

Table 1: Airline itinerary piecewise linear model

1.2 The Power Series Expansion

File to develop using the airline dataset:
Model file: MNL_airline_powerseries.py

We introduce here a power series expansion for the travel time of the non-stop itinerary. Other polynomial expressions could be tried as well, but in the following example, we only specify a squared term.

The specification of the model presented in this section is the same as the one in *MNL_airline_specific.py* except for the alternative relative to the non-stop itinerary. The latter is given as follows:

$$\begin{aligned} V_1 = \; & \text{ASC}_1 + \beta_{\text{Fare}} \cdot \text{Fare}_1 + \beta_{\text{Legroom}} \cdot \text{Legroom}_1 + \beta_{\text{Total\_TT}_1\_1} \cdot \text{Total\_TT}_1\_1 \\ & + \beta_{\text{Total\_TT}_1\_sq} \cdot \text{Total\_TT}_1\_sq + \beta_{\text{SchedDE}} \cdot \text{SchedDE}_1 + \beta_{\text{SchedDL}} \cdot \text{SchedDL}_1 \end{aligned}$$

In order to define the squared term of Total_TT_1 in PythonBiogeme, we add the following instruction to define it as a variable:

```
TripTimeHours_1_sq = DefineVariable('TripTimeHours_1_sq', TripTimeHours_1 ** 2)
```

The estimation results for this specification are shown in Table 2. The estimated parameter associated with the linear term of the power series expansion is negative while the estimated

parameter associated with the squared term is positive. However, for reasonable travel times, the cumulative effect of the travel time variable on the utility is still negative, as the coefficient associated with the power series term is much smaller in absolute value.

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	ASC_2	-2.21	0.298	-7.42	0.00
2	ASC_3	-2.43	0.312	-7.78	0.00
3	β_{Fare}	-0.0193	0.000800	-24.11	0.00
4	$\beta_{Legroom}$	0.227	0.0267	8.51	0.00
5	$\beta_{SchedDE}$	-0.139	0.0165	-8.46	0.00
6	$\beta_{SchedDL}$	-0.105	0.0137	-7.63	0.00
7	$\beta_{Total\_TT_1}$	-0.870	0.172	-5.05	0.00
8	$\beta_{Total\_TT_1\_sq}$	0.0745	0.0220	3.38	0.00
9	$\beta_{Total\_TT_2}$	-0.301	0.0701	-4.30	0.00
10	$\beta_{Total\_TT_3}$	-0.302	0.0701	-4.31	0.00

Summary statistics
Number of observations = 3609
Number of excluded observations = 0
Number of estimated parameters = 10

$$\begin{aligned} \mathcal{L}(\beta_0) &= -3964.892 \\ \mathcal{L}(\hat{\beta}) &= -2314.435 \\ -2[\mathcal{L}(\beta_0) - \mathcal{L}(\hat{\beta})] &= 3300.914 \\ \rho^2 &= 0.416 \\ \bar{\rho}^2 &= 0.414 \end{aligned}$$

Table 2: Airline itinerary power series model

To see if the power series specification is better than the linear one, we perform a likelihood ratio test. Here, the restricted model is the one with linear travel time for the non-stop alternative (*MNL_airline_specific.py*) and the unrestricted model is the one with the power series expansion (*MNL_airline_powerseries.py*). The null hypothesis is given by:

$$H_0 : \beta_{Total\_TT_1\_sq} = 0$$

The statistic for the likelihood ratio test is given as follows:

$$-2(-2320.447 + 2314.435) = 12.024$$

Since $\chi^2_{0.95,1} = 3.841$, we can reject the null hypothesis of a linear travel time for the non-stop alternative at a 95% level of confidence.

1.3 The Box-Cox Transformation

File to develop using the airline dataset:
Model file: *MNL_airline_boxcox.py*

In this section, we specify a Box-Cox transformation, which is a non-linear transformation of a variable that also depends on an unknown parameter λ. Precisely, a Box-Cox transformation of a variable x is given as follows:

$$\frac{x^\lambda - 1}{\lambda}, \text{ where } x \geq 0. \tag{7}$$

We apply this transformation to the travel time variable for the non-stop itinerary. The utilities are the same as the previous models, apart from the one relative to the non-stop itinerary, which we report below:

$$\begin{aligned} V_1 &= \text{ASC}_1 + \beta_{\text{Fare}} \cdot \text{Fare}_1 + \beta_{\text{Legroom}} \cdot \text{Legroom}_1 + \beta_{\text{Total\_TT}_1} \cdot \frac{Total\_TT_1^\lambda - 1}{\lambda} \\ &\quad + \beta_{\text{SchedDE}} \cdot \text{SchedDE}_1 + \beta_{\text{SchedDL}} \cdot \text{SchedDL}_1 \end{aligned}$$

Let us note that in this specification, we have one more unknown parameter, λ. In PythonBiogeme, we define this parameter together with the other parameters of the model:

```
LAMBDA = Beta('LAMBDA',1,-10000,10000,0)
```

Moreover, the expression (7) for the travel time of alternative 1 is coded as follows:

```
( ( ( TripTimeHours_1 ** LAMBDA ) - 1 ) / LAMBDA )
```

The results relative to the model including the Box-Cox transformation are shown in Table 3. Let us remark that the Box-Cox transformation reduces to a linear function as a special case when the parameter λ is equal to 1. The estimate of λ is significantly different from 1 at a 95 % level of confidence, with a t-test equal to -3.36.

We perform a likelihood ratio test between the linear model (*MNL_airline_specific.py*) and the Box-Cox model (*MNL_airline_boxcox.py*). The null hypothesis is given by:

$$H_0 : \lambda = 1$$

The statistic of the likelihood ratio test for this null hypothesis is given as follows:

$$\begin{gathered} -2(-2320.447 + 2314.574) = 11.746 \\ \chi^2_{0.95,1} = 3.841 < 11.746 \end{gathered}$$

The null hypothesis of a linear specification is hence rejected at a 95 % level of confidence. Therefore, the Box-Cox transformation of the time is more adequate.

2 Test of Non-Nested Hypotheses: Cox test

Files to use with PythonBiogeme (provided):

Model files:	*MNL_airline_specific.py* (M_1)
	MNL_airline_log.py (M_2)
	MNL_airline_composite.py (M_C)
Data file:	*airline.dat*

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	ASC_2	-1.51	0.263	-5.77	0.00
2	ASC_3	-1.74	0.280	-6.22	0.00
3	β_{Fare}	-0.0193	0.000799	-24.12	0.00
4	λ	-0.139	0.338	-0.41	0.68
5	$\beta_{Legroom}$	0.227	0.0267	8.52	0.00
6	$\beta_{SchedDE}$	-0.140	0.0165	-8.47	0.00
7	$\beta_{SchedDL}$	-0.105	0.0137	-7.63	0.00
8	$\beta_{Total\_TT_1}$	-1.24	0.372	-3.34	0.00
9	$\beta_{Total\_TT_2}$	-0.306	0.0681	-4.49	0.00
10	$\beta_{Total\_TT_3}$	-0.306	0.0683	-4.48	0.00

Summary statistics

Number of observations = 3609

Number of excluded observations = 0

Number of estimated parameters = 10

$$\begin{aligned} \mathcal{L}(\beta_0) &= -3964.892 \\ \mathcal{L}(\hat{\beta}) &= -2314.574 \\ -2[\mathcal{L}(\beta_0) - \mathcal{L}(\hat{\beta})] &= 3300.636 \\ \rho^2 &= 0.416 \\ \bar{\rho}^2 &= 0.414 \end{aligned}$$

Table 3: Airline itinerary Box Cox model

In discrete choice analysis, we often perform tests based on the so-called nested hypotheses, which means that we specify two models such that the first one (the restricted model) is a special case of the second one (the unrestricted model). For this type of comparison, the classical likelihood ratio test can be applied. However, there are situations, such as non-linear specifications, in which we aim at comparing models that are not nested, i.e. one model cannot be obtained as a restricted version of the other. One way to compare two non-nested models is to build a composite model from which both models can be derived. We can thus perform two likelihood ratio tests, testing each of the restricted models against the composite model. This procedure is known as the Cox test of separate families of hypothesis.

The Cox test is described in detail in the slides of the course. Assume that we want to test a model M_1 against another model M_2 (and one model is not a restricted version of the other). We start by generating a composite model M_C such that both models M_1 and M_2 are restricted cases of M_C. We then test M_1 against M_C and M_2 against M_C using the likelihood ratio test. There are three possible outcomes of this test:

1. One of the two models is rejected. Then we keep the one that is not rejected.
2. Both models are rejected. Then better models should be developed. The composite model could be used as a new basis for future specifications.
3. Both models are accepted. Then we choose the model with the highest $\bar{\rho}^2$ index.

We present here the expressions of the utility functions used for three different models M_1, M_2 and M_C developed on the airline itinerary case study. In M_1 the fare is linearly included, in M_2 the logarithm of the fare is included and in M_C both terms are included.

M_1 has the following systematic utilities (`MNL_airline_specific.py`):

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

where the cost is *linear*.

The systematic utilities of M_2 are expressed as follows (`MNL_airline_log.py`):

$$\begin{aligned}
V_1 &= \text{ASC}_1 + \beta_{\text{LogFare}} \cdot \log(\text{Fare}_1) + \beta_{\text{Legroom}} \cdot \text{Legroom}_1 + \beta_{\text{Total_TT}_1} \cdot \text{Total_TT}_1 \\
&\quad + \beta_{\text{SchedDE}} \cdot \text{SchedDE}_1 + \beta_{\text{SchedDL}} \cdot \text{SchedDL}_1 \\
V_2 &= \text{ASC}_2 + \beta_{\text{LogFare}} \cdot \log(\text{Fare}_2) + \beta_{\text{Legroom}} \cdot \text{Legroom}_2 + \beta_{\text{Total_TT}_2} \cdot \text{Total_TT}_2 \\
&\quad + \beta_{\text{SchedDE}} \cdot \text{SchedDE}_2 + \beta_{\text{SchedDL}} \cdot \text{SchedDL}_2 \\
V_3 &= \text{ASC}_3 + \beta_{\text{LogFare}} \cdot \log(\text{Fare}_3) + \beta_{\text{Legroom}} \cdot \text{Legroom}_3 + \beta_{\text{Total_TT}_3} \cdot \text{Total_TT}_3 \\
&\quad + \beta_{\text{SchedDE}} \cdot \text{SchedDE}_3 + \beta_{\text{SchedDL}} \cdot \text{SchedDL}_3
\end{aligned}$$

where the cost is *logarithmic*.

We now define the composite model M_C with the following systematic utilities (`MNL_airline_composite.py`):

$$\begin{aligned}
V_1 &= \text{ASC}_1 + \beta_{\text{Fare}} \cdot \text{Fare}_1 + \beta_{\text{LogFare}} \cdot \log(\text{Fare}_1) + \beta_{\text{Legroom}} \cdot \text{Legroom}_1 + \beta_{\text{Total_TT}_1} \cdot \text{Total_TT}_1 \\
&\quad + \beta_{\text{SchedDE}} \cdot \text{SchedDE}_1 + \beta_{\text{SchedDL}} \cdot \text{SchedDL}_1 \\
V_2 &= \text{ASC}_2 + \beta_{\text{Fare}} \cdot \text{Fare}_2 + \beta_{\text{LogFare}} \cdot \log(\text{Fare}_2) + \beta_{\text{Legroom}} \cdot \text{Legroom}_2 + \beta_{\text{Total_TT}_2} \cdot \text{Total_TT}_2 \\
&\quad + \beta_{\text{SchedDE}} \cdot \text{SchedDE}_2 + \beta_{\text{SchedDL}} \cdot \text{SchedDL}_2 \\
V_3 &= \text{ASC}_3 + \beta_{\text{Fare}} \cdot \text{Fare}_3 + \beta_{\text{LogFare}} \cdot \log(\text{Fare}_3) + \beta_{\text{Legroom}} \cdot \text{Legroom}_3 + \beta_{\text{Total_TT}_3} \cdot \text{Total_TT}_3 \\
&\quad + \beta_{\text{SchedDE}} \cdot \text{SchedDE}_3 + \beta_{\text{SchedDL}} \cdot \text{SchedDL}_3
\end{aligned}$$

Table 4 summarizes the differences between the various models and Tables 5, 6 and 7 show the estimation results for models M_1, M_2 and M_C, respectively.

Now we can apply the likelihood ratio test for M_1 against M_C. In this case, the null hypothesis is:

$$H_0 : \beta_{LogFare} = 0$$

As usual, $-2(L(M_1) - L(M_C))$ is χ^2 distributed with $K = 1$ degrees of freedom. In this case, we have:

Models used for the Cox test		
Model	Parameters	Description
M_1	9	two ASCs, one generic cost *linear* coefficient, alternative specific time coefficients and three generic coefficients (for legroom, schedule delay – early departure, schedule delay – late departure)
M_2	9	two ASCs, one generic cost *logarithmic* coefficient, three alternative specific time coefficients and three generic coefficients (for legroom, schedule delay – early departure, schedule delay – late departure)
M_C	10	two ASCs, one generic cost *linear* coefficient, one generic cost *logarithmic* coefficient, three alternative specific time coefficients and three generic coefficients (for legroom, schedule delay – early departure, schedule delay – late departure)

Table 4: Summary of the different model specifications

$$-2(-2320.447 + 2271.656) = 97.582 > 3.84$$

The result of this first test is that we can reject the null hypothesis H_0: it means the composite model is better than M_1. The linear model is rejected. Applying the same test for M_2 against M_C, we have

$$H_1 : \beta_{Fare} = 0.$$

In this case, the likelihood ratio test with $K = 1$ degrees of freedom gives

$$-2(-2283.103 + 2271.656) = 22.894 > 3.84$$

and we can therefore reject the null hypothesis H_1 in this case as well. The logaritmic model is also rejected.

Since both models are rejected, better models should be developed: we cannot keep the composite model with two different cost-related coefficients since it does not have a behavioral interpretation. If both models had been accepted, we would choose the one with the highest $\bar{\rho}^2$ index.

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	Constant2	-1.43	0.183	-7.81	0.00
2	Constant3	-1.64	0.192	-8.53	0.00
3	Fare	-0.0193	0.000802	-24.05	0.00
4	Legroom	0.226	0.0267	8.45	0.00
5	SchedDE	-0.139	0.0163	-8.53	0.00
6	SchedDL	-0.104	0.0137	-7.59	0.00
7	Total_TT1	-0.332	0.0735	-4.52	0.00
8	Total_TT2	-0.299	0.0696	-4.29	0.00
9	Total_TT3	-0.302	0.0699	-4.32	0.00

Summary statistics

Number of observations = 3609

Number of excluded observations = 0

Number of estimated parameters = 9

$$\begin{aligned} \mathcal{L}(\beta_0) &= -3964.892 \\ \mathcal{L}(\hat{\beta}) &= -2320.447 \\ -2[\mathcal{L}(\beta_0) - \mathcal{L}(\hat{\beta})] &= 3288.889 \\ \rho^2 &= 0.415 \\ \bar{\rho}^2 &= 0.412 \end{aligned}$$

Table 5: Estimation results for the model M_1

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	Constant2	-1.82	0.194	-9.39	0.00
2	Constant3	-2.09	0.200	-10.46	0.00
3	Legroom	0.219	0.0261	8.38	0.00
4	LogFare	-8.54	0.305	-28.02	0.00
5	SchedDE	-0.142	0.0167	-8.50	0.00
6	SchedDL	-0.105	0.0139	-7.54	0.00
7	Total_TT1	-0.465	0.0729	-6.37	0.00
8	Total_TT2	-0.335	0.0690	-4.86	0.00
9	Total_TT3	-0.321	0.0692	-4.63	0.00

Summary statistics

Number of observations = 3609

Number of excluded observations = 0

Number of estimated parameters = 9

$$\begin{aligned} \mathcal{L}(\beta_0) &= -3964.892 \\ \mathcal{L}(\hat{\beta}) &= -2283.103 \\ -2[\mathcal{L}(\beta_0) - \mathcal{L}(\hat{\beta})] &= 3363.577 \\ \rho^2 &= 0.424 \\ \bar{\rho}^2 &= 0.422 \end{aligned}$$

Table 6: Estimation results for the model M_2

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	Constant2	-1.69	0.193	-8.74	0.00
2	Constant3	-1.94	0.199	-9.73	0.00
3	Fare	-0.00658	0.00154	-4.28	0.00
4	Legroom	0.223	0.0265	8.40	0.00
5	LogFare	-5.96	0.665	-8.96	0.00
6	SchedDE	-0.142	0.0167	-8.51	0.00
7	SchedDL	-0.106	0.0140	-7.57	0.00
8	Total_TT1	-0.415	0.0739	-5.62	0.00
9	Total_TT2	-0.324	0.0694	-4.67	0.00
10	Total_TT3	-0.316	0.0697	-4.53	0.00

Summary statistics

Number of observations = 3609

Number of excluded observations = 0

Number of estimated parameters = 10

$$\begin{aligned} \mathcal{L}(\beta_0) &= -3964.892 \\ \mathcal{L}(\hat{\beta}) &= -2271.656 \\ -2[\mathcal{L}(\beta_0) - \mathcal{L}(\hat{\beta})] &= 3386.472 \\ \rho^2 &= 0.427 \\ \bar{\rho}^2 &= 0.425 \end{aligned}$$

Table 7: Estimation results for the model M_C

3 Market segmentation

Files to use with PythonBiogeme (provided):

Model files: *MNL_airline_specific.py*,
MNL_airline_male.py
MNL_airline_female.py
MNL_airline_GenderNA.py

Data file: *airline.dat*

In this example, we test if there is a taste variation across market segments. The segmentation is made on the gender variable. We first create three market segments as follows: Male, Female, and no answer (NA). The sum of observations for each segment is equal to the total observations (N):

$$N_{Male} + N_{Female} + N_{NA} = N$$

We estimate a model on the full data set. Then we run the same model for each gender group separately. Note that each time we exclude the observations that do not belong to the considered segment (using the exclude command from PythonBiogeme). We obtain the values shown in

Table 8. The expressions of the utility functions are the same for all models:

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

Model	Log likelihood	Number of coefficients
Male	-1195.819	9
Female	-929.325	9
NA	-178.017	9
M1 model	-2320.447	9

Table 8: Values for the market segmentation test

The null hypothesis is of no taste variation across the market segments:

$$H_0 \ : \ \beta^{Male} = \beta^{Female} = \beta^{NA}$$

where $\beta^{segment}$ is the vector of coefficients of the market segment. Note that in the above equation Male, Female and NA refer to market segments and not to variables in the dataset.

The likelihood ratio test (with 27-9=18 degrees of freedom, where 27 corresponds to the 3*9 parameters of the three *segment* models and 9 to the number of parameters of the general model) yields:

$$
\begin{aligned}
LR &= -2\Big(\mathcal{L}_N(\hat{\beta}) - \big(\mathcal{L}_{N_{Male}}(\hat{\beta}^{Male}) + \mathcal{L}_{N_{Female})}(\hat{\beta}^{Female}) + \mathcal{L}_{N_{NA}}(\hat{\beta}^{NA})\big)\Big) \\
&= -2(-2320.447 + 1195.819 + 929.325 + 178.017) = 34.572
\end{aligned}
$$

$$\chi^2_{0.95,18} = 28.87$$

and we can therefore reject the null hypothesis at a 95% level of confidence: market segmentation on gender does exist.