EPFL
ENAC TRANSP-OR
Prof. M. Bierlaire

Mathematical Modeling of Behavior
Fall 2020

Airline itinerary case

1 Model Specification with Generic Attributes

Files to use with Biogeme:

Model notebook:	*MNL_airline_generic.ipynb*
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, and
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.

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

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:

```
exclude = (ArrivalTimeHours_1==-1)
database.remove(exclude)
```

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 notebook: MNL_airline_specific.ipynb

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.

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:

$$df = K_U - K_R = 9 - 7 = 2$$

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

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 notebook: *MNL_airline_socioecon.ipynb*

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.

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

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:

```
exclude = (Cont_Income==-1)
database.remove(exclude)
```

The estimation results of this model are reported in Table 3.

File to develop using the same dataset as before:
Model notebook: *MNL_airline_socioecon_mi.ipynb*

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), database)
MissingIncome = DefineVariable(’MissingIncome’,
(Cont_Income == -1), database)
```

Socio-economic MNL estimation				
Parameter	Parameter	Parameter	Robust	Robust
number	name	estimate	standard error	*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
7	β_{SchedDE}	-0.139	0.0173	-7.99
8	β_{SchedDL}	-0.0940	0.0146	-6.44
9	$\beta_{\text{Total\_TT}}$	-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 - excluding missing data

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.

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

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.

4 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 i tinerary, a power series method, and a Box-Cox transformation.

For the exercises in this section you should modify the model specification with alternative-specific coefficients developed in Section 2 (*MNL_airline_specific.ipynb*). The deterministic utilities for this model are the following:

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

4.1 Piecewise Linear Approximation

File to develop using the airline dataset:
Model notebook: *MNL_airline_piecewise.ipynb*

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:

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

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

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

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

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

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

$$\text{Total\_TT}_1\text{\_3} = \begin{cases} 0 & \text{if } TripTimeHours_1 \leq 3 \\ TripTimeHours_1 - 3 & \text{if } TripTimeHours_1 > 3 \end{cases} \tag{3}$$

$$\tag{4}$$

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 as follows:

1. $\text{Total\_TT}_1\text{\_1} = 2$
2. $\text{Total\_TT}_1\text{\_2} = 0.5$
3. $\text{Total\_TT}_1\text{\_3} = 0$

Thus, the original value of the travel time ($TripTimeHours_1$) is decomposed into the three variables ($TripTimeHours_1 = \text{Total\_TT}_1\text{\_1} + \text{Total\_TT}_1\text{\_2} + \text{Total\_TT}_1\text{\_3}$).

The `piecewiseFormula` function in PandasBiogeme allows you to define the piecewise specification. It is imported using the following statement:

```
from biogeme.models import piecewiseFormula
```

This function will automatically create β parameters and Biogeme variables, given a list of thresholds and initial values for the parameters.

```
thresholds = [None, 2, 3, None]
init_Betas_TT1 = [0,0,0]
```

The piecewise transformation of the travel time variable can thus be included in the utility function by replacing the corresponding term ($\beta_{\text{Total\_TT\_1}} \cdot \text{Total\_TT}_1$) with the following expression:

```
piecewiseFormula(TripTimeHours_1, thresholds, init_Betas_TT1)
```

The estimation results for this specification are shown in Table 5. 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.ipynb*) and the unrestricted model is the piecewise linear specification (*MNL_airline_piecewise.ipynb*). 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.

4.2 The Power Series Expansion

File to develop using the airline dataset:
Model notebook: *MNL_airline_powerseries.ipynb*

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.ipynb* except for the alternative relative to the non-stop itinerary. The latter is given as follows:

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

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

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

Piecewise MNL estimation				
Parameter	Parameter	Parameter	Robust	Robust
number	name	estimate	standard error	*t statistic*
1	ASC_2	-2.32	0.411	-5.65
2	ASC_3	-2.55	0.438	-5.83
3	β_{Fare}	-0.0193	0.000799	-24.10
4	$\beta_{Legroom}$	0.227	0.0267	8.51
5	$\beta_{SchedDE}$	-0.140	0.0165	-8.47
6	$\beta_{SchedDL}$	-0.105	0.0137	-7.64
7	$\beta_{Total\_TT_1\_1}$	-0.824	0.238	-3.46
8	$\beta_{Total\_TT_1\_2}$	-0.444	0.188	-2.36
9	$\beta_{Total\_TT_1\_3}$	-0.229	0.0889	-2.57
10	β_{Total_TT2}	-0.300	0.0701	-4.29
11	β_{Total_TT3}	-0.301	0.0701	-4.29
Summary statistics				
Number of observations = 3609				
$\mathcal{L}(0) = -3964.892$				
$\mathcal{L}(\hat{\beta}) = -2315.041$				
$\bar{\rho}^2 = 0.413$				

Table 5: Airline itinerary piecewise linear model

The estimation results for this specification are shown in Table 6. 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.

Power series estimation				
Parameter number	Parameter name	Parameter estimate	Robust standard error	Robust *t statistic*
1	ASC_2	-2.21	0.298	-7.42
2	ASC_3	-2.43	0.312	-7.78
3	β_{Fare}	-0.0193	0.000800	-24.11
4	$\beta_{Legroom}$	0.227	0.0267	8.51
5	$\beta_{SchedDE}$	-0.139	0.0165	-8.46
6	$\beta_{SchedDL}$	-0.105	0.0137	-7.63
7	$\beta_{Total\_TT_1}$	-0.870	0.172	-5.05
8	$\beta_{Total\_TT_1\_sq}$	0.0745	0.0220	3.38
9	$\beta_{Total\_TT_2}$	-0.301	0.0701	-4.30
10	$\beta_{Total\_TT_3}$	-0.302	0.0701	-4.31
Summary statistics				
Number of observations = 3609				
$\mathcal{L}(0) = -3964.892$				
$\mathcal{L}(\hat{\beta}) = -2314.435$				
$\bar{\rho}^2 = 0.414$				

Table 6: Airline itinerary power series linear 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.ipynb*) and the unrestricted model is the one with the power series expansion (*MNL_airline_powerseries.ipynb*). 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.

4.3 The Box-Cox Transformation

File to develop using the airline dataset:
Model notebook: *MNL_airline_boxcox.ipynb*

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:

$$\frac{x^{\lambda}-1}{\lambda}, \text{ where } x \geq 0. \tag{5}$$

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:

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

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

```
LAMBDA  = Beta('LAMBDA',1,None,None,0)
```

Moreover, the expression (5) 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 7. 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 (be careful, the t-test provided by Biogeme is against 0, you need to calculate by hand the t-test against 1).

We can also perform a likelihood ratio test between the linear model (*MNL_airline_specific.ipynb*) and the Box-Cox model (*MNL_airline_boxcox.ipynb*). 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:

$$-2(-2320.447 + 2314.574) = 11.746$$

The null hypothesis of a linear specification is rejected at a 95 % level of confidence because $\chi^2_{0.95,1} = 3.841 < 11.746$. Therefore, the Box-Cox transformation of the time is more adequate.

Box-Cox estimation				
Parameter number	Parameter name	Parameter estimate	Robust standard error	Robust *t statistic*
1	ASC_2	-1.51	0.263	-5.77
2	ASC_3	-1.74	0.280	-6.22
3	β_{Fare}	-0.0193	0.000799	-24.12
4	λ	-0.139	0.338	-0.41
5	$\beta_{Legroom}$	0.227	0.0267	8.52
6	$\beta_{SchedDE}$	-0.140	0.0165	-8.47
7	$\beta_{SchedDL}$	-0.105	0.0137	-7.63
8	$\beta_{Total\_TT_1}$	-1.24	0.372	-3.34
9	$\beta_{Total\_TT_2}$	-0.306	0.0681	-4.49
10	$\beta_{Total\_TT_3}$	-0.306	0.0683	-4.48
Summary statistics				
Number of observations = 3609				
$\mathcal{L}(0) = -3964.892$				
$\mathcal{L}(\hat{\beta}) = -2314.574$				
$\bar{\rho}^2 = 0.414$				

Table 7: Airline itinerary Box-Cox model