Airline Itinerary Choice (Boeing)

The topic of this case study is the testing of different hypotheses regarding both model specifications and structures. The objectives can be summarized as follows:

- Illustration of the market segmentation concept and related testing.
- Explanation of the McFadden IIA test to test the assumption of independence between alternatives.
- Testing of non-nested hypotheses using the composite model test.
- Testing of non-linear specifications using the piecewise linear approximation, the power series expansion and the Box-Cox transformation methods.

Market Segmentation

Files to use with BIOGEME:
Model files: *SpecTest_Boeing_male.mod*,
SpecTest_Boeing_female.mod,
SpecTest_Boeing_GenderNA.mod,
SpecTest_Boeing_full.mod,
Data file: *boeing.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:

$$\mathsf{N}_{\mathsf{Male}} + \mathsf{N}_{\mathsf{Female}} + \mathsf{N}_{\mathsf{NA}} = \mathsf{N}$$

We estimate a model on the full data set. Then we run the same model for each gender group separately. Note that we make use of the `[Exclude]` section in the model specification file to define which observations should be excluded for the estimation. We obtain the values shown in Table 1. The expressions of the utility

functions are the same for all models:

$$\begin{aligned}
V_1 &= ASC_1 + \beta_{Fare} \cdot Fare_1 + \beta_{Legroom} \cdot Legroom_1 + \beta_{Total_TT} \cdot Total_TT_1 \\
&\quad + \beta_{SchedDE} \cdot SchedDE_1 + \beta_{SchedDL} \cdot SchedDL_1 \\
V_2 &= ASC_2 + \beta_{Fare} \cdot Fare_2 + \beta_{Legroom} \cdot Legroom_2 + \beta_{Total_TT} \cdot Total_TT_2 \\
&\quad + \beta_{SchedDE} \cdot SchedDE_2 + \beta_{SchedDL} \cdot SchedDL_2 \\
V_3 &= ASC_3 + \beta_{Fare} \cdot Fare_3 + \beta_{Legroom} \cdot Legroom_3 + \beta_{Total_TT} \cdot Total_TT_3 \\
&\quad + \beta_{SchedDE} \cdot SchedDE_3 + \beta_{SchedDL} \cdot SchedDL_3
\end{aligned}$$

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

Table 1: 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 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) 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.

McFadden IIA Test

Files to use with BIOGEME:
Model files: *SpecTest_Boeing_socioec.mod, SpecTest_Boeing_IIA.mod*
Data files: *boeing.dat, boeing_exclude.dat*

In this survey, the choice is made between three flights, two of which are with the same company. It is possible that there are common unobserved attributes between the two tickets with the same company. It would seem logical to expect some kind of *relationship* between the traditional alternatives. They are maybe correlated. In order to test this assumption, we perform the McFadden IIA test. First we estimate an MNL model (*SpecTest_Boeing_socioec.mod*) on the full data set *boeing.dat*. The specification file *SpecTest_Boeing_socioec.mod* contains a section describing the correlation we want to test. The corresponding BIOGEME snapshot is shown in figure 1. Alternative 1 corresponds to the flight without stops, and alternative 2 to the same company but with one stop.

```
biogeme SpecTest_Boeing_socioec boeing.dat
```

```
[IIATest]
C12     1 2
```

Figure 1: BIOGEME snapshot: IIATest section

By defining the section *[IIATest]* in the orginal *.mod* file, auxiliary variables are automatically computed for each observation, and reported in the *.enu* output file. Biogeme also produces a file containing the specification of the estimated model, in the same format as the model specification file, *SpecTest_Boeing_socioec.res*. We need to rename it as a .mod file: *SpecTest_Boeing_socioec_res.mod* in order to apply it on the same data file, using BIOSIM:

biosim SpecTest_Boeing_socioec_res boeing.dat

The original *.dat* file and the *SpecTest_Boeing_socioec_res.enu* file need to be merged in order to create a new data file, *data_DAT_ENU_merged.dat*. Note that this merged data file could not contain the same number of observations than boeing.dat because observations are excluded in the first estimation (it's not the case in this example but it could happen very often). Now we specify a new model (*SpecTest_Boeing_IIA.mod*) which includes the auxiliary variables in the utility functions associated with alternatives 1 and 2. Finally, we estimate this model on the new data file created by merging.

```
biogeme SpecTest_Boeing_IIA data_DAT_ENU_merged.dat
```

The focus in this test is not related to the sign of the estimated IIA parameter. What is important is the value of the t-statistic for such a coefficient. If β_{IIA} is significantly different from 0 at a 95% level of confidence, this indicates that the IIA property does not hold for alternatives 1 and 2. In this case, it means alternatives 1 and 2 might share some unobserved attributes. Therefore we validate our hypothesis that the IIA property does not hold. This kind of correlation can be captured with Generalized Extreme Value (GEV) models. They allow for partially relaxing the IIA assumption. This topic will be part of the case study where we introduce the Generalized Extreme Value models.

Test of Non-Nested Hypotheses

Files to use with BIOGEME:
Model files: *SpecTest_Boeing_full_LogFare.mod* (M_1),
SpecTest_Boeing_full.mod (M_2),
SpecTest_Boeing_full_C.mod (M_C)
Data file: *boeing.dat*

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 which are not nested, meaning that 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 for each of the restricted models against the composite model.

Composite Model Test

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:

- One of the two models is rejected. Then we keep the one that is not rejected.

- Both models are rejected. Then better models should be developed. The composite model could be used as a new basis for future specifications.
- Both models are accepted. Then we choose the model with the higher $\bar{\rho}^2$ index.

We show next the expressions of the utility functions used for the three different models M_1, M_2 and M_C. M_1 has the following systematic utilities:

$$\begin{aligned}
V_1 &= ASC_1 + \beta_{Fare} \cdot Fare_1 + \beta_{Legroom} \cdot Legroom_1 + \beta_{Total_TT} \cdot Total_TT_1 \\
&\quad + \beta_{SchedDE} \cdot SchedDE_1 + \beta_{SchedDL} \cdot SchedDL_1 \\
V_2 &= ASC_2 + \beta_{Fare} \cdot Fare_2 + \beta_{Legroom} \cdot Legroom_2 + \beta_{Total_TT} \cdot Total_TT_2 \\
&\quad + \beta_{SchedDE} \cdot SchedDE_2 + \beta_{SchedDL} \cdot SchedDL_2 \\
V_3 &= ASC_3 + \beta_{Fare} \cdot Fare_3 + \beta_{Legroom} \cdot Legroom_3 + \beta_{Total_TT} \cdot Total_TT_3 \\
&\quad + \beta_{SchedDE} \cdot SchedDE_3 + \beta_{SchedDL} \cdot SchedDL_3
\end{aligned}$$

where the cost related coefficients are *linear*. The systematic utilities of M_2 are:

$$\begin{aligned}
V_1 &= ASC_1 + \beta_{LogFare} \cdot \log(Fare_1) + \beta_{Legroom} \cdot Legroom_1 + \beta_{Total_TT} \cdot Total_TT_1 \\
&\quad + \beta_{SchedDE} \cdot SchedDE_1 + \beta_{SchedDL} \cdot SchedDL_1 \\
V_2 &= ASC_2 + \beta_{LogFare} \cdot \log(Fare_2) + \beta_{Legroom} \cdot Legroom_2 + \beta_{Total_TT} \cdot Total_TT_2 \\
&\quad + \beta_{SchedDE} \cdot SchedDE_2 + \beta_{SchedDL} \cdot SchedDL_2 \\
V_3 &= ASC_3 + \beta_{LogFare} \cdot \log(Fare_3) + \beta_{Legroom} \cdot Legroom_3 + \beta_{Total_TT} \cdot Total_TT_3 \\
&\quad + \beta_{SchedDE} \cdot SchedDE_3 + \beta_{SchedDL} \cdot SchedDL_3
\end{aligned}$$

where the cost related coefficients are *logarithmic*. We now define the composite model M_C with the following systematic utilities:

$$\begin{aligned}
V_1 &= ASC_1 + \beta_{Fare} \cdot Fare_1 + \beta_{LogFare} \cdot \log(Fare_1) + \beta_{Legroom} \cdot Legroom_1 \\
&\quad + \beta_{Total_TT} \cdot Total_TT_1 + \beta_{SchedDE} \cdot SchedDE_1 + \beta_{SchedDL} \cdot SchedDL_1 \\
V_2 &= ASC_2 + \beta_{Fare} \cdot Fare_1 + \beta_{LogFare} \cdot \log(Fare_2) + \beta_{Legroom} \cdot Legroom_2 \\
&\quad + \beta_{Total_TT} \cdot Total_TT_2 + \beta_{SchedDE} \cdot SchedDE_2 + \beta_{SchedDL} \cdot SchedDL_2 \\
V_3 &= ASC_3 + \beta_{Fare} \cdot Fare_1 + \beta_{LogFare} \cdot \log(Fare_3) + \beta_{Legroom} \cdot Legroom_3 \\
&\quad + \beta_{Total_TT} \cdot Total_TT_3 + \beta_{SchedDE} \cdot SchedDE_3 + \beta_{SchedDL} \cdot SchedDL_3
\end{aligned}$$

In Table 2, we summarize the differences between the various models, and we show in Tables 3, 4 and 5 the estimation results for the M_1, M_2 and M_C models, respectively.

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

Models used for the composite test		
Model	Parameters	Description
M_1	9	two ASC's, one generic cost *linear* coefficient, three alternative specific time coefficients and three generic coefficients
M_2	9	two ASC's, one generic cost *logarithmic* coefficient, three alternative specific time coefficients and three generic coefficients
M_C	10	two ASC's, one generic cost *logarithmic* coefficient, one generic cost *logarithmic* coefficient, three alternative specific time coefficients and three generic coefficients

Table 2: Summary of the different model specifications

Parameter number	Parameter name	Parameter estimate	Robust standard 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	-4.31	0.00

Summary statistics

Number of observations = 3609

$\mathcal{L}(0) = -3964.892$

$\mathcal{L}(\hat{\beta}) = -2320.447$

$\bar{\rho}^2 = 0.412$

Table 3: Estimation results for the M_1 model

Parameter number	Parameter name	Parameter estimate	Robust standard 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	Fare	-8.54	0.305	-28.02	0.00
4	Legroom	0.219	0.0261	8.38	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

$\mathcal{L}(0) = -3964.892$

$\mathcal{L}(\hat{\beta}) = -2283.103$

$\bar{\rho}^2 = 0.422$

Table 4: Estimation results for the M_2 model

Parameter number	Parameter name	Parameter estimate	Robust standard error	t-stat	p-value
1	Constant2	-1.69	0.193	-8.74	0.00
2	Constant3	-1.94	0.199	-9.72	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

$\mathcal{L}(0) = -3964.892$

$\mathcal{L}(\hat{\beta}) = -2271.656$

$\bar{\rho}^2 = 0.425$

Table 5: Estimation results for the M_C model

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

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

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

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

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

In this case, the likelihood ratio test with $K = 2$ 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 logartimic model is also rejected. Since both models are rejected, better models should be developed: we cannot keep the composite models with two cost-related coefficients, it doesn't make sense. If both models were accepted, we would choose the one with the higher $\bar{\rho}^2$ index.

Tests of Non-Linear Specifications

Files to use with BIOGEME:
Model files: *SpecTest_boeing_piecewise.mod*,
SpecTest_boeing_powerseries.mod,
SpecTest_boeing_boxcox.mod
Data file: *boeing.dat*

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 non-linear 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.

Piecewise Linear Approximation

In this first example, we want to test the hypothesis that the value of the travel time related 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 travel time $\mathsf{TripTimeHours}_1 \in [0.67, 6.35]$ (expressed in hours) into three different intervals: $\mathsf{TripTimeHours}_{1_1} \in [0, 2]$, $\mathsf{TripTimeHours}_{1_2} \in]2, 3]$, $\mathsf{TripTimeHours}_{1_3} > 3$. Figure 2 displays the corresponding BIOGEME code.

```
[Expressions]
TripTimeHours_1_1 = min( TripTimeHours_1 , 2)
TripTimeHours_1_2 = max(0,min( TripTimeHours_1 - 2, 1))
TripTimeHours_1_3 = max(0,TripTimeHours_1 - 3)
```

Figure 2: BIOGEME snapshot concerning the piecewise variables definition

The systematic utility expressions used in this model are given as follows:

$$
\begin{aligned}
V_1 &= \mathsf{ASC}_1 + \beta_{\mathsf{Fare}} \cdot \mathsf{Fare}_1 + \beta_{\mathsf{Legroom}} \cdot \mathsf{Legroom}_1 + \\
&\quad \beta_{\mathsf{SchedDE}} \cdot \mathsf{Opt1SchedDelayEarly} + \beta_{\mathsf{SchedDL}} \cdot \mathsf{Opt1SchedDelayLate} + \\
&\quad \beta_{\mathsf{Total\_TT}_1\_1} \cdot \mathsf{Total\_TT}_1\_1 + \beta_{\mathsf{Total\_TT}_1\_2} \cdot \mathsf{Total\_TT}_1\_2 + \\
&\quad \beta_{\mathsf{Total\_TT}_1\_3} \cdot \mathsf{Total\_TT}_1\_3 \\
V_2 &= \mathsf{ASC}_2 + \beta_{\mathsf{Fare}} \cdot \mathsf{Fare}_2 + \beta_{\mathsf{Legroom}} \cdot \mathsf{Legroom}_2 + \\
&\quad \beta_{\mathsf{SchedDE}} \cdot \mathsf{Opt2SchedDelayEarly} + \beta_{\mathsf{SchedDL}} \cdot \mathsf{Opt2SchedDelayLate} + \\
&\quad \beta_{\mathsf{Total\_TT}_2} \cdot \mathsf{Total\_TT}_2 \\
V_3 &= \mathsf{ASC}_3 + \beta_{\mathsf{Fare}} \cdot \mathsf{Fare}_3 + \beta_{\mathsf{Legroom}} \cdot \mathsf{Legroom}_3 + \\
&\quad \beta_{\mathsf{SchedDE}} \cdot \mathsf{Opt3SchedDelayEarly} + \beta_{\mathsf{SchedDL}} \cdot \mathsf{Opt3SchedDelayLate} + \\
&\quad \beta_{\mathsf{Total\_TT}_3} \cdot \mathsf{Total\_TT}_3
\end{aligned}
$$

The estimation results are shown in Table 6. All time coefficients related to the piecewise linear expression are negative. The coefficient associated with short trips (< 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.

Piecewise linear model: estimation results				
Parameter number	Parameter name	Coeff. estimate	Robust standard error	Robust t-stat
1	ASC_2	-2.33	0.412	-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	$\beta_{Total\_TT_1\_1}$	-0.825	0.238	-3.47
8	$\beta_{Total\_TT_1\_2}$	-0.443	0.188	-2.36
9	$\beta_{Total\_TT_1\_3}$	-0.229	0.0889	-2.57
10	$\beta_{Total\_TT_2}$	-0.300	0.0701	-4.29
11	$\beta_{Total\_TT_3}$	-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 6: Estimation results for the piecewise linear model

We perform a likelihood ratio test where the restricted model is the one with linear travel time for the non-stop alternative and the unrestricted model is the piecewise linear specification. 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.

The Power Series Expansion

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 squared term.

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

$$\begin{aligned} V_1 \;=\; & ASC_1 + \beta_{Fare} \cdot Fare_1 + \beta_{Legroom} \cdot Legroom_1 + \\ & \beta_{SchedDE} \cdot Opt1SchedDelayEarly + \beta_{SchedDL} \cdot Opt1SchedDelayLate + \\ & \beta_{Total\_TT_1} \cdot Total\_TT_1 + \beta_{Total\_TT_1\_sq} \cdot Total\_TT_1\_sq \end{aligned}$$

The estimation results for this specification are shown in Table 7. 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.

In order 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 and the unrestricted model is the one with the power series expansion. 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:

Power series model: estimation results				
Parameter number	Parameter name	Coeff. estimate	Robust standard error	Robust t-stat
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 7: Estimation results for the power series model

```
[GeneralizedUtilities]
1 Total_TT1 * ( ( ( TripTimeHours_1 ) ^ LAMBDA - 1 ) / LAMBDA )
```

Figure 3: BIOGEME snapshot of Box-Cox transformation

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

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.

The Box-Cox Transformation

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:

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

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:

$$\begin{aligned} V_1 = {} & ASC_1 + \beta_{Fare} \cdot Fare_1 + \beta_{Legroom} \cdot Legroom_1 + \\ & \beta_{SchedDE} \cdot Opt1SchedDelayEarly + \beta_{SchedDL} \cdot Opt1SchedDelayLate + \\ & \beta_{Total\_TT_1} \cdot \frac{Total\_TT_1^\lambda - 1}{\lambda} \end{aligned}$$

Let us note that in this specification, we have one more unknown parameter, λ. Figure 3 displays a BIOGEME snapshot from the model specification file.

The results relative to the model including the Box-Cox transformation are shown in Table 8.

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 and the Box-Cox model. 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:

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

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.

Box-Cox transformed model: estimation results				
Parameter number	Parameter name	Coeff. estimate	Robust standard error	Robust t-stat
1	Constant2	-1.51	0.263	-5.77
2	Constant3	-1.74	0.280	-6.22
3	Fare	-0.0193	0.000799	-24.12
4	LAMBDA	-0.139	0.338	-0.41
5	Legroom	0.227	0.0267	8.52
6	SchedDE	-0.140	0.0165	-8.47
7	SchedDL	-0.105	0.0137	-7.63
8	Total_TT1	-1.24	0.372	-3.34
9	Total_TT2	-0.306	0.0681	-4.49
10	Total_TT3	-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 8: Estimation results for the Box-Cox transformed model