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_airline_male.mod*, *SpecTest_airline_female.mod*, *SpecTest_airline_GenderNA.mod*, *SpecTest_airline_full.mod*,
Data file: *airline.dat*

In this example, we test if there exists 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 the number of observations for each segment is equal to the total number of 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 estimate the same model for each gender group separately. Note that we make use of the `[Exclude]` section in the model specification file to define the observations which 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:

$$
\begin{aligned}
V_1 &= ASC_1 + \beta_{Fare} \cdot Fare_1 + \beta_{Legroom} \cdot Legroom_1 + \beta_{Total\_TT} \cdot Total\_TT_1 \\
&\quad + \beta_{SchedDE} \cdot Opt1\_SchedDelayEarly + \beta_{SchedDL} \cdot Opt1\_SchedDelayLate \\
V_2 &= ASC_2 + \beta_{Fare} \cdot Fare_2 + \beta_{Legroom} \cdot Legroom_2 + \beta_{Total\_TT} \cdot Total\_TT_2 \\
&\quad + \beta_{SchedDE} \cdot Opt2\_SchedDelayEarly + \beta_{SchedDL} \cdot Opt2\_SchedDelayLate \\
V_3 &= ASC_3 + \beta_{Fare} \cdot Fare_3 + \beta_{Legroom} \cdot Legroom_3 + \beta_{Total\_TT} \cdot Total\_TT_3 \\
&\quad + \beta_{SchedDE} \cdot Opt3\_SchedDelayEarly + \beta_{SchedDL} \cdot Opt3\_SchedDelayLate
\end{aligned}
$$

Let us remark that one of the three alternative specific constants ASC_1, ASC_2 and ASC_3 must be set to 1 for normalization purposes.

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 assumes 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_airline_full.mod,*
	SpecTest_airline_IIA.mod
Data file:	*airline.dat*

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

```
Biogeme SpecTest_airline_full_bis airline.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_airline_full_bis.res*. We need to rename it as a *.mod* file: *SpecTest_airline_full_bis_res.mod* in order to apply it on the same data file, using BioSim:

```
biosim SpecTest_airline_full_bis_res airline.dat
```

The original *.dat* file and the *SpecTest_airline_full_bis_res.enu* file need to be merged in order to create a new data file that contains both the original model variables and the auxiliary variables. This step is performed using BIOMERGE:

```
biomerge airline.dat SpecTest_airline_full_bis_res.enu
```

The merged data file is stored into a file named *biomergeOutput.lis*. We rename this file as *SpecTest_airline_IIATest.dat*. Now we specify a new model

(*SpecTest_airline_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 the original data file and *SpecTest_airline_full_res.enu*, using the following command:

```
Biogeme SpecTest_airline_IIA SpecTest_airline_IIATest.dat
```

The estimation results are shown in Table 2.

Logit model for IIA test for itineraries 1 and 2				
Parameter number	Parameter name	Parameter estimate	Robust standard error	Robust *t statistic*
1	ASC_2	-1.51	0.211	-7.14
2	ASC_3	-1.65	0.194	-8.51
3	β_{Fare}	-0.0198	0.00104	-18.94
4	β_{Legroom}	0.232	0.0281	8.24
5	β_{SchedDE}	-0.143	0.0168	-8.49
6	β_{SchedDL}	-0.107	0.0145	-7.40
7	$\beta_{\text{Total\_TT}_1}$	-0.341	0.0744	-4.58
8	$\beta_{\text{Total\_TT}_2}$	-0.304	0.0700	-4.34
9	$\beta_{\text{Total\_TT}_3}$	-0.312	0.00111	-4.65
10	β_{IIA}	-0.0489	0.0714	-4.37

Summary statistics
Number of observations = 3609
$\mathcal{L}(0) = -3964.892$
$\mathcal{L}(\hat{\beta}) = -2320.155$
$\bar{\rho}^2 = 0.412$

Table 2: Logit model for IIA test

In the IIA Test, we are interested in the value of the t-statistic for the coefficient related to the auxiliary variables. 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. It would mean that alternatives 1 and 2 share some unobserved attributes.

However Table 2 shows that parameter β_{IIA} is not significantly different from 0. Hence we cannot conclude that the IIA property does not hold. The calibration of more complex models such as Generalized Extreme Value (GEV) models which capture correlation between alternative sharing some common characteristics might not be justified in this case. We can hence keep the logit specification.