EPFL
ENAC TRANSP-OR
Prof. M. Bierlaire

Mathematical Modeling of Behavior
Fall 2020

Airline itinerary case

1 Test of Non-Nested Hypotheses: Cox test

Files to use with PandasBiogeme (provided):

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

In discrete choice analysis, we often perform tests based on 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 hypotheses.

The Cox test is described in detail in the course. Assume that we want to test a model M_1 against another model M_2 (and no model is 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 two likelihood ratio tests. 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 included linearly, 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`):

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

where the cost is *linear.*

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

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

where the cost is *logarithmic.*

We now define the composite model M_C (`MNL_airline_composite`), characterized by the following systematic utilities. Note that it includes both the linear term and the logarithmic term for the cost:

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

Table 1 summarizes the differences between the various models and Tables 2, 3, and 4 show the estimation results for models M_1, M_2 and M_C, respectively.

We can now apply the likelihood ratio test for M_1 against M_C. The null hypothesis is:

$$H_0^1 : \beta_{LogFare} = 0$$

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

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

The result of this first test is that we can reject the null hypothesis H_0^1: 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

$$H_0^2 : \beta_{Fare} = 0.$$

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_0^2 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.

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 1: Summary of the different model specifications

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	-24.0	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

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 2: 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	-10.5	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	-4.85	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 3: 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	-9.72	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

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 4: Estimation results for the model M_C

2 Market segmentation

Files to use with PandasBiogeme:

Provided model notebook:	*MNL_airline_specific.ipynb*,
Model notebooks to develop:	*MNL_airline_male.ipynb*, *MNL_airline_female.ipynb*, *MNL_airline_GenderNA.ipynb*,
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 number of 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 `database.remove` command from PandasBiogeme). We obtain the values shown in Table 5. 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 5: 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 $3 \cdot 9$ 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.

no / th / rk / mpp / jp / mw