Airline itinerary choice (boeing case study)

September 25, 2011

Model Specification with Generic Attributes

Files to use with BIOGEME:

Model file:binary_generic_boeing.mod

Data file: boeing.dat

We assume the choice variable (dependent variable) includes following alternatives: Option 1 and Option 2. In this first model, we assume leg room, fare, schedule delays (early and late) are the factors influencing the choice. We also assume that the coefficients of travel time variables are generic, i.e., they do not vary between alternatives. The expression of the utilities for this simple model can be expressed as:

$$
\begin{aligned}
V_1 &= \beta_{Fare} * Opt1\_FARE + \beta_{Total\_TT} * TripTimeHours\_1 \\
&\quad + \beta_{Legroom} * Opt1\_Legroom + \beta_{SchedDE} * Opt1\_SchedDelayEarly \\
&\quad + \beta_{SchedDL} * Opt1\_SchedDelayLate \\
V_2 &= Constant2 \\
&\quad + \beta_{Fare} * Opt2\_FARE + \beta_{Total\_TT} * TripTimeHours\_2 \\
&\quad + \beta_{Legroom} * Opt2\_Legroom + \beta_{SchedDE} * Opt2\_SchedDelayEarly \\
&\quad + \beta_{SchedDL} * Opt2\_SchedDelayLate,
\end{aligned}
$$

where fare is coded as $Opt1\_FARE$ and $Opt2\_FARE$ in the unit of 100$, in order to reduce numerical issues; the schedule delay is categorized into early and late as variables:

- $Opt1\_SchedDelayEarly$,
- $Opt1\_SchedDelayLate$,
- $Opt2\_SchedDelayEarly$ and
- $Opt2\_SchedDelayLate$;

the leg room is coded as a continuous variable in inch unit. These variables are coded in the "[Expressions]" section of the model file.

The estimation results are reported in Table 1. The results indicate that all the rest being equal, the first option without stop is preferred. All the estimated coefficients are significantly different from zero. The signs of the time coefficient β_{Total_TT} and the fare coefficient β_{Fare} are negative, as expected, meaning that the utility of an alternative decreases with increase in travel time and fare. The signs of the schedule delay coefficients are both negative, indicating that people don't like delays. The positive sign of the leg room indicates that people like seats with bigger space.

Table 1: Binary model with generic attributes

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	$Constant2$	-1.41	0.176	-8.02	0.00
2	β_{Fare}	-1.83	0.104	-17.65	0.00
3	$\beta_{Legroom}$	0.115	0.0179	6.41	0.00
4	$\beta_{SchedDE}$	-0.111	0.0213	-5.23	0.00
5	$\beta_{SchedDL}$	-0.118	0.0189	-6.25	0.00
6	β_{Total_TT}	-0.236	0.0966	-2.44	0.01

Summary statistics

Number of observations = 3093

$$\mathcal{L}(0) = -2143.904$$
$$\mathcal{L}(c) = -1505.814$$
$$\mathcal{L}(\hat{\beta}) = -1171.504$$
$$-2[\mathcal{L}(0) - \mathcal{L}(\hat{\beta})] = 1944.800$$
$$\rho^2 = 0.454$$
$$\bar{\rho}^2 = 0.451$$

MNL with Alternative-Specific Attributes

Files to use with BIOGEME:

Model file:binary_specific_boeing.mod

Data file: boeing.dat

In this second specification we relax the hypothesis of generic coefficients. To illustrate this idea, two different time coefficients are introduced for two alternatives. The corresponding utility functions are reported below:

$$
\begin{aligned}
V_1 &= \beta_{Fare} * Opt1\_FARE + \beta_{Total\_TT1} * TripTimeHours\_1 \\
&\quad + \beta_{Legroom} * Opt1\_Legroom + \beta_{SchedDE} * Opt1\_SchedDelayEarly \\
&\quad + \beta_{SchedDL} * Opt1\_SchedDelayLate \\
V_2 &= Constant2 \\
&\quad + \beta_{Fare} * Opt2\_FARE + \beta_{Total\_TT2} * TripTimeHours\_2 \\
&\quad + \beta_{Legroom} * Opt2\_Legroom \\
&\quad + \beta_{SchedDE} * Opt2\_SchedDelayEarly \\
&\quad + \beta_{SchedDL} * Opt2\_SchedDelayLate,
\end{aligned}
$$

The estimation results are reported in Table 2. In this case, both time coefficients for the two options are estimated. Both their signs are negative, as expected. The absolute value of β_{Total_TT1} is larger, meaning that people are more sensitive to time in case of non-stop flights. The interpretation for other parameters remains the same.

Table 2: Binary model with alternative specific attributes

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	$Constant2$	-1.48	0.205	-7.22	0.00
2	β_{Fare}	-1.82	0.105	-17.27	0.00
3	$\beta_{Legroom}$	0.115	0.0179	6.41	0.00
4	$\beta_{SchedDE}$	-0.112	0.0214	-5.21	0.00
5	$\beta_{SchedDL}$	-0.118	0.0190	-6.24	0.00
6	β_{Total_TT1}	-0.257	0.104	-2.47	0.01
7	β_{Total_TT2}	-0.236	0.0967	-2.44	0.01

Summary statistics

Number of observations = 3093

$$
\begin{aligned}
\mathcal{L}(0) &= -2143.904 \\
\mathcal{L}(c) &= -1505.814 \\
\mathcal{L}(\hat{\beta}) &= -1171.318 \\
-2[\mathcal{L}(0) - \mathcal{L}(\hat{\beta})] &= 1945.172 \\
\rho^2 &= 0.454 \\
\bar{\rho}^2 &= 0.450
\end{aligned}
$$

Generic vs Specific Test

The likelihood ratio test can be used to test the generic vs. the alternative-specific model specifications. The likelihood ratio test statistic for the null hypothesis of generic attributes is

$$-2(L(\beta_R) - L(\beta_U)),$$

where R and U denote the restricted (generic) and unrestricted (alternative-specific) models, respectively. It is χ^2-distributed with the number of degrees of freedom equal to the number of restrictions ($K_U - K_R$), with K_U and K_R the numbers of estimated coefficients in the unrestricted and restricted models, respectively. In this case, $-2(-1171.504 + 1171.318) = 0.372$. Since $\chi^2_{0.90,1} = 2.71$ at 90% level of confidence, we can conclude that the null hypothesis of a generic time coefficient can not be rejected. So the model with alternative specific coefficient does not have a significant improvement in fit.

Inclusion of Socio-Economic Characteristics

Files to use with BIOGEME:

Model file:binary_socio_econ_boeing.mod

Data file: boeing.dat

The previous two models only include variables that are attribute of the alternatives. We now introduce a socio-economic characteristic, namely the gender of the respondent. $MALE$ is a dummy variable and is equal to 1 if the gender is male and zero if female. It should be noticed that the socio-economic variables do not vary among the alternatives (recall that only difference in the utilities matters), we have normalized alternative 2 to zero. However, this is an arbitrary normalization, as we could also have normalized alternative 1. The utility functions can be written now as follows:

$$\begin{aligned}
V_1 &= \beta_{Fare} * Opt1_FARE + \beta_{Total_TT} * TripTimeHours_1 \\
&\quad + \beta_{Legroom} * Opt1_Legroom + \beta_{SchedDE} * Opt1_SchedDelayEarly \\
&\quad + \beta_{SchedDL} * Opt1_SchedDelayLate \\
V_2 &= Constant2 \\
&\quad + \beta_{Male_{Opt2}} * Male + \beta_{Fare} * Opt2_FARE \\
&\quad + \beta_{Total_TT} * TripTimeHours_2 + \beta_{Legroom} * Opt2_Legroom \\
&\quad + \beta_{SchedDE} * Opt2_SchedDelayEarly \\
&\quad + \beta_{SchedDL} * Opt2_SchedDelayLate,
\end{aligned}$$

The estimation results are reported in Table 3. The coefficient of the β_{Male_Opt2} is not statistically significant different from zero and indicates that different genders have the same preferences on the two options. The interpretation of the other coefficients rest the same as the previous model specifications.

Table 3: Binary model with socio-economic characteristics

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	$Constant2$	-1.44	0.184	-7.86	0.00
2	β_{Fare}	-1.83	0.104	-17.66	0.00
3	$\beta_{Legroom}$	0.115	0.0179	6.41	0.00
4	β_{Male_Opt2}	0.0620	0.105	0.59	0.55
5	$\beta_{SchedDE}$	-0.111	0.0212	-5.22	0.00
6	$\beta_{SchedDL}$	-0.118	0.0189	-6.26	0.00
7	β_{Total_TT}	-0.234	0.0967	-2.42	0.02

Summary statistics

Number of observations = 3093

$$\mathcal{L}(0) = -2143.904$$
$$\mathcal{L}(c) = -1505.814$$
$$\mathcal{L}(\hat{\beta}) = -1171.329$$
$$-2[\mathcal{L}(0) - \mathcal{L}(\hat{\beta})] = 1945.150$$
$$\rho^2 = 0.454$$
$$\bar{\rho}^2 = 0.450$$