Specification Testing: Choice of Residential Telephone Services Case

Market Segmentation

Files to use with BIOGEME:
Model files: *MNL_low_inc_tel.mod*, *MNL_med_inc_tel.mod*, *MNL_high_inc_tel.mod*, *MNL_socio_econ_tel.mod*
Data file: *tel.dat*

In this case we test if there is a taste variation accross market segments. We define different segments based on income and divide the population into three income groups. We estimate separate models for each income group using the same model specification, namely *MNL_socio_eco_tel.mod* used in the Multinomial Logit case study, and compare the estimation results with a model based on the complete dataset. The results in terms of final log-likelihood are summarized in table 1.

The null hypothesis is of no taste variations across the market segments, that is

$$H_0 \ : \ \beta_{HI} = \beta_{MI} = \beta_{LI}.$$

Performing a likelihood ratio test,

$$\begin{aligned} LR &= -2(L_N(\hat{\beta}) - \sum_{g=1}^{G} L_{N_g}(\widehat{\beta^g})) \\ &= -2(-467.804 + 120.103 + 297.99 + 46.668) = 6.086 \\ \chi^2_{0.95,13} &= 22.36, \end{aligned}$$

we can conclude that the null hypothesis cannot be rejected, that is market segmentation on income does not exist.

Model	Definition	Log-likelihood	Nb. of Coefficients
Low Income	$\text{Income} < 10000$	-120.103	6
Medium Income	$10000 < \text{Income} < 40000$	-297.99	7
High Income	$\text{Income} > 40000$	-46.668	7
Pooled Data Restricted Model	All	-468.791	7

Table 1: Results for the market segmentation test

IIA test estimation				
Variable number	Variable name	Coefficient estimate	Robust standard error	Robust *t statistic*
1	ASC_{BM}	0.366	0.336	1.077
2	ASC_{LF}	-0.207	0.388	-0.532
3	ASC_{EF}	-0.357	0.843	-0.423
4	ASC_{MA}	0.165	0.530	0.312
5	β_{userF}	0.409	0.108	3.792
6	β_{M}	-0.156	0.037	-4.178
7	β_{F}	-0.105	0.022	-4.785
8	β_{iia}	1.832	0.628	2.919

Summary statistics
Number of observations $= 434$
$\mathcal{L}(0) = -560.25$
$\mathcal{L}(\hat{\beta}) = -463.068$
$\bar{\rho}^2 = 0.173$

Table 2: Estimation results for the IIA test

McFadden IIA Test

Files to use with BIOGEME:
Model file: *mnl_tel.mod, iia_tel.mod*
Data file: *tel.dat, iia_tel.dat*
Excel worksheet: *iia.xls*

For the telephone dataset it is probable that there are common unobserved attributes between the measured options (alternatives *BM* and *SM*) and we can perform the McFadden IIA test to check this. That is, first we estimate a MNL model (*mnl_tel.mod*) on the full data set *tel.dat*. We use the estimated values of the parameters to compute in Excel the choice probabilities for each observation (individual) and for each alternative. As discussed above, we suppose in this case that the subset of alternatives suspected to be correlated is given by $\widehat{\mathrm{C}} = \{\mathrm{BM,SM}\}$. We then compute in Excel the two corresponding auxiliary variables for each observation of the data file to get the file *iia.xls*, which we export in the Text format file *iia_tel.dat*. Now we specify a new model (*iia_tel.mod*) which includes the auxiliary variables in the utility functions associated with the alternatives SM and BM. Finally, we estimate the model on this new data file and obtain the results reported in table 2.

The value of the parameter β_{iia} is significantly different from 0 at 95% level of confidence. This indicates that IIA assumption is not valid for BM and SM alternatives. Similarly, we can check for IIA violation among other potential groups of alternatives (e.g. flat options). In presence of such correlations, GEV models like the Nested Logit are more appropriate.

Test of Non-Nested Hypothesis

In discrete choice models, we often perform tests based on the so-called nested hypothesis, 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 in which we aim to

compare models which are not nested, meaning that one model cannot be obtained as a restricted version of another one. 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

Files to use with BIOGEME:
Model files: *base_tel.mod*, *M1_tel.mod*, *M2_tel.mod*, *MC_tel.mod*
Data file: *tel.dat*

In this section we show a practical application of such a kind of test. We start from a base model specification (M_b) from which two extensions M_1 and M_2 plus a third one (M_C), including both M_1 and M_2, are generated. The determininistic part of the utility functions for each of these specifications are:

1. M_b

$$\begin{aligned} V_{BM} &= ASC_{BM} + \beta_1 cost_{BM} \\ V_{SM} &= \beta_1 cost_{SM} \\ V_{LF} &= ASC_{LF} + \beta_1 cost_{LF} \\ V_{EF} &= ASC_{EF} + \beta_1 cost_{EF} \\ V_{MF} &= ASC_{MF} + \beta_1 cost_{MF} \end{aligned}$$

2. M_1

$$\begin{aligned} V_{BM} &= ASC_{BM} + \beta_M cost_{BM} \\ V_{SM} &= \beta_M cost_{SM} \\ V_{LF} &= ASC_{LF} + \beta_F cost_{LF} \\ V_{EF} &= ASC_{EF} + \beta_F cost_{EF} \\ V_{MF} &= ASC_{MF} + \beta_F cost_{MF} \end{aligned}$$

Model	Nb. of parameters	Log-likelihood	$\bar{\rho}^2$
Base (M_b)	5	-482.719	0.129
Extension 1 (M_1)	6	-476.04	0.139
Extension 2 (M_2)	6	-471.151	0.148
Composite Model (M_C)	7	-467.804	0.152

Table 3: Results from the non-nested hypothesis test

3. M_2

$$\begin{aligned} V_{BM} &= ASC_{BM} + \beta_1 cost_{BM} \\ V_{SM} &= \beta_1 cost_{SM} \\ V_{LF} &= ASC_{LF} + \beta_1 cost_{LF} + \beta_{user} users \\ V_{EF} &= ASC_{EF} + \beta_1 cost_{EF} + \beta_{user} users \\ V_{MF} &= ASC_{MF} + \beta_1 cost_{MF} + \beta_{user} users \end{aligned}$$

4. M_c

$$\begin{aligned} V_{BM} &= ASC_{BM} + \beta_M cost_{BM} \\ V_{SM} &= \beta_M cost_{SM} \\ V_{LF} &= ASC_{LF} + \beta_F cost_{LF} + \beta_{user} users \\ V_{EF} &= ASC_{EF} + \beta_F cost_{EF} + \beta_{user} users \\ V_{MF} &= ASC_{MF} + \beta_F cost_{MF} + \beta_{user} users \end{aligned}$$

The estimation results of the different models are summarized in table 3.

We first compare the M_1 model specification against the composite model M_C by means of a likelihood ratio test: $H_0 : \beta_{user} = 0$
$-2(L(\beta_{M1}) - L(\beta_{MC})) = -2(-476.04 + 467.804) = 16.472$
$\chi^2_{0.95,1} = 3.841 < 16.472$ We can therefore reject the null hypothesis of generic coefficients. We then compare M_2 against M_C: $H_0 : \beta_M = \beta_F$
$-2(L(\widehat{\beta_{M2}}) - L(\widehat{\beta_{MC}})) = -2(-471.151 + 467.804) = 6.694$

$\chi^2_{0.95,1} = 3.841 < 6.694$ We can therefore reject the null hypothesis of not including socio-economic variables.

The limitation of this approach is clearly shown in this example where we can reject the null hypothesis in both cases. Thus we reach the conclusion that both M_1 and M_2 are rejected.

The $\bar{\rho}^2$ test

In this procedure, the adjusted likelihood ratio index $\bar{\rho}^2$ is used as a goodness of fit measure. This can be calculated using the formula:

$$\bar{\rho}^2 = 1 - \frac{L(\hat{\beta}) - K}{L(0)}$$

BIOGEME automatically reports the value of $\bar{\rho}^2$. In this case:

$$\bar{\rho}^2_{M1} = 0.139$$
$$\bar{\rho}^2_{M2} = 0.148$$

Following the theoretical reminder, we verify the inequality:

$$\Pr(\bar{\rho_2}^2 - \bar{\rho_1}^2 > z) \leq \Phi\{-[-2Nzln(J) + (K_2 - K_1)]^{\frac{1}{2}}\}$$

Here we have the difference in $\bar{\rho}^2$ equal to $z = 0.009$, the number of observations $N = 434$, and the difference in the number of parameters between M_1 and M_2 is zero. So, we have:

$$\Pr(\bar{\rho_2}^2 - \bar{\rho_1}^2 > z) \leq \Phi\{-3.545\}$$
$$\Pr(\bar{\rho_2}^2 - \bar{\rho_1}^2 > z) \leq 0.0002$$

This indicates that the probability that such a difference would be exceeded is 0.0002. So, we can choose the model with the higher $\bar{\rho}^2$.

Tests of Non-Linear Specifications

In the previous case study the examples dealt with linear cost coefficients (coefficients that remain constant throughout the whole range of the values of each variable). However, in some cases non-linear specifications may be more justified (e.g. sensitivity to cost may not be the same in all cost ranges). In this section we apply three different procedures we have described in the theoretical reminder concerning the tests for non-linear specifications. We have used the MNL model with alternative specific cost coefficients as the base model.

Piecewise Linear Approximation

Files to use with BIOGEME:
Model file: *piecewise_tel.mod*
Data file: *tel.dat*

In the first model we assume that the coefficient of measured cost assume different values for different ranges of the cost variable. The full range of values for the measured cost variable is \$3.28 to \$435.5. To test the piecewise linear hypothesis the following variables are generated:

$$\begin{aligned} \mathrm{cost}_{i1} &= \min\{\mathrm{cost}_i, 10\} \\ \mathrm{cost}_{i2} &= \max\{0, \min\{\mathrm{cost}_i - 10, 40\}\} \\ \mathrm{cost}_{i3} &= \max\{0, \mathrm{cost}_i - 50\} \end{aligned}$$

with $i = \{BM, SM\}$ and cost_i is the cost of the alternative i in the original dataset. We are practically splitting the range of values for cost_i (which is $\mathrm{cost}_i \in [3.28, 435.5]$, expressed in minutes) into three different intervals: $\mathrm{cost}_{i1} \in [0, 10]$, $\mathrm{cost}_{i2} \in [10, 50]$ and $\mathrm{cost}_{i3} \in [50, 435.5]$. As mentioned in the theoretical reminder section, the selection of these ranges is based on a priori hypothesis of the user behavior and distribution of cost in the

```
    [Expressions]
    // Define here arithmetic expressions for name
    // that are not directly available from the data
cost11 =min(cost1 ,10)
cost12 =max(0,min(cost1 - 10 ,40))
cost13 =max(0,cost1 - 50)
cost21 =min(cost2 ,10)
cost22 =max(0,min(cost2 - 10 ,40))
cost23 =max(0,cost2 - 50)
```

Figure 1: BIOGEME snapshot for the piecewise linear approximation

observed sample. The reader is encouraged to experiment different ranges. An extract from the BIOGEME model file to code the ranges of costs is presented in figure 1.

The utility functions are reported in equations 1.

$$
\begin{aligned}
V_{BM} &= ASC_{BM} + \beta_{M1}cost_{BM1} + \beta_{M2}cost_{BM2} + \beta_{M3}cost_{BM3} \\
V_{SM} &= \beta_{M1}cost_{SM1} + \beta_{M2}cost_{SM2} + \beta_{M3}cost_{SM3} \\
V_{LF} &= ASC_{LF} + \beta_{F}cost_{LF} \\
V_{EF} &= ASC_{EF} + \beta_{F}cost_{EF} \\
V_{MF} &= ASC_{MF} + \beta_{F}cost_{MF}
\end{aligned}
$$

The estimation results are reported in table 4.

The results indicate that the sensitivity to measured cost becomes less in the range $10 < cost_i < 50$ compared to the range $cost_i < 10$, but has a steep increase after that. This model has a better goodness-of-fit than the model with linear coefficients in general. To test whether or not the improvement in goodness-of-fit is statistically significant, we need to perform a likelihood ratio test between the two different specifications.

The null hypothesis in this case is

Piecewise linear approximation				
Variable number	Variable name	Coefficient estimate	Robust standard error	Robust *t statistic*
1	ASC_{BM}	-0.613	0.152	-4.028
2	ASC_{LF}	-0.631	-0.450	-1.263
3	ASC_{EF}	-0.843	0.869	-0.970
4	ASC_{MF}	-0.261	0.640	-0.409
5	β_{M1}	-0.294	0.066	-4.443
6	β_{M2}	-0.149	0.067	-2.234
7	β_{M3}	-4.264	0.777	-5.491
8	β_{F}	-0.105	0.022	-4.837
Summary statistics				
Number of observations $= 434$				
$\mathcal{L}(0) = -560.25$				
$\mathcal{L}(\hat{\beta}) = -474.703$				
$\bar{\rho}^2 = 0.152694$				

Table 4: Estimation results for the piecewise linear approximation

$$H_0 : \beta_{M1} = \beta_{M2} = \beta_{M3}$$

The χ^2 statistics for this null hypothesis is as follows:

$$-2(L(\beta_R) - L(\beta_U)) = -2(-476.04 + 474.703) = 2.674$$
$$\chi^2_{0.95,2} = 5.991 > 2.674$$

where the restricted model (R) is represented by the linear specification while the unrestricted model (U) corresponds to the piecewise linear specification. The improvement in goodness-of-fit due to the introduction of the piecewise linear specification is not significant and the null hypothesis that the cost coefficient is linear cannot be rejected.

The Power Series Expansion

Files to use with BIOGEME:
Model file: *power1_tel.mod*
Data file: *tel.dat*

In this test we relax the hypothesis of linear coefficients for measured options by assuming a second order power series (a squared term and a linear term). The corresponding systematic utility functions are

$$\begin{aligned}
V_{BM} &= ASC_{BM} + \beta_{M1}cost_{BM} + \beta_{M2}cost^2_{BM} \\
V_{SM} &= \beta_{M1}cost_{SM} + \beta_{M2}cost^2_{SM} \\
V_{LF} &= ASC_{LF} + \beta_F cost_{LF} \\
V_{EF} &= ASC_{EF} + \beta_F cost_{EF} \\
V_{MF} &= ASC_{MF} + \beta_F cost_{MF}.
\end{aligned}$$

From the estimation results presented in table 5 we can see that this model has in general a better goodness-of-fit than the model with linear coefficients. However, the coefficient of the squared term, though statistically

Power series estimation				
Variable number	Variable name	Coefficient estimate	Robust standard error	Robust *t statistic*
1	ASC_{BM}	-0.563	0.147	-3.831
2	ASC_{LF}	-0.162	0.369	-0.439
3	ASC_{EF}	-0.377	0.813	-0.464
4	ASC_{MF}	0.215	0.531	0.404
5	β_{M1}	-0.227	0.042	-5.319
6	β_{M2}	0.0005	0.0001	5.073
7	β_{F}	-0.106	0.021	-4.907

Summary statistics
Number of observations $= 434$
$\mathcal{L}(0) = -560.25$
$\mathcal{L}(\hat{\beta}) = -475.465$
$\bar{\rho}^2 = 0.139$

Table 5: Estimation results for the power series expansion

significant, has a very small value. It may be noted that the coefficient of the squared term is positive while the coefficient of the linear term is negative and the coefficient of the linear term is greater than that of the squared term. However, since the squared term is very small in magnitude, the total effect is expected to remain negative in the cost range. To test whether or not the improvement in goodness-of-fit is statistically significant, we need to do a likelihood ratio test. The null hypothesis in this case is:

$$H_0 : \beta_{M2} = 0$$

The χ^2 statistics for this null hypothesis is as follows:

$$-2(L(\widehat{\beta_R}) - L(\widehat{\beta_U})) = -2(-476.04 + 475.465) = 1.15$$
$$\chi^2_{0.95,1} = 3.841 > 1.15$$

where now the unrestricted model (U) corresponds to the power series specification. The improvement in goodness-of-fit due to the introduction of the second order power specification is not significant and the null hypothesis that the cost coefficient is linear cannot be rejected.

The Box-Cox Transformation

Files to use with BIOGEME:
Model file: *boxcox_tel.mod*
Data file: *tel.dat*

In this section we analyze the possibility to test for non-linear transformations of variables which are non linear in the unknown parameters. As explained in the theoretical reminder section, one such transformation is the Box-Cox expressed as

$$\frac{x^\lambda - 1}{\lambda}$$

where λ is a parameter that has to be estimated. We apply such a transformation to the measured cost variable. The utilities remain the same with

```
[Utilities]
// Id Name  Avail  linear-in-parameter expression
1     BM   avail1   ASC_1 * one
2     SM   avail2   ASC_2 * one
3     LF   avail3   ASC_3 * one + B2_FCOST * cost3
4     EF   avail4   ASC_4 * one + B2_FCOST * cost4
5     MF   avail5   ASC_5 * one + B2_FCOST * cost5

[GeneralizedUtilities]
1 B1_MCOST * ( ( ( cost1 )^ LAMBDA - 1)/LAMBDA )
2 B1_MCOST * ( ( ( cost2 )^ LAMBDA - 1)/LAMBDA )
```

Figure 2: BIOGEME snapshot for the Box-Cox transformation

the substitution of the measured cost variable with its Box-Cox transformation. The BIOGEME snapshot defining such a transformation is reported in figure 2.

The parameter λ is estimated along with the other parameters (its starting value needs to be specified as a value other than zero, since $\lambda \neq 0$).

The estimation results are reported in table 6.

The estimate of λ was not found to be significant. However, since there is an improvement of the likelihood value, a likelihood ratio test is performed. The null hypothesis is given by:

$$H_0 : \lambda = 1$$

The χ^2 statistics for this null hypothesis is as follows:

$$-2(L(\widehat{\beta_L}) - L(\widehat{\beta_{BC}})) = -2(-476.04 + 472.624) = 6.832$$
$$\chi^2_{0.95,1} = 3.841 < 6.832$$

Therefore, the null hypothesis can be rejected as the improvement in goodness-of-fit from introduction of the Box-Cox transformation is significant.

Box-Cox estimation				
Variable number	Variable name	Coefficient estimate	Robust standard error	Robust *t statistic*
1	ASC_{BM}	-0.695	0.166	-4.195
2	ASC_{LF}	-1.756	1.202	-1.460
3	ASC_{EF}	-1.977	1.386	-1.427
4	ASC_{MF}	-1.389	1.278	-1.087
5	β_F	-0.104	0.022	-4.832
6	β_M	-1.296	0.880	-1.473
7	λ	0.234	0.305	0.768

Summary statistics
Number of observations $= 434$
$\mathcal{L}(0) = -560.25$
$\mathcal{L}(\hat{\beta}) = -472.624$
$\bar{\rho}^2 = 0.143$

Table 6: Estimation results for the Box-Cox transformation