EPFL
ENAC TRANSP-OR
Prof. M. Bierlaire

Mathematical Modeling of Behavior
Fall 2017

MIXTURE MODELS

1 Base model

Files to use with Biogeme:
Model file: *Base_Model.py*
Data file: *swissmetro.dat*

The utility specifications of the base model are the following:

$$\begin{aligned} V_{car} &= ASC_{car} + \beta_{time} CAR\_TT + \beta_{cost} CAR\_CO \\ V_{train} &= \beta_{time} TRAIN\_TT + \beta_{cost} TRAIN\_CO + \beta_{he} TRAIN\_HE \\ V_{SM} &= ASC_{SM} + \beta_{time} SM\_TT + \beta_{cost} SM\_CO + \beta_{he} SM\_HE \end{aligned}$$

The estimates of the parameters are included in Table 1.

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	ASC_CAR	0.189	0.0798	2.37	0.02
2	ASC_SM	0.451	0.0932	4.84	0.00
3	BETA_COST	-0.0108	0.000682	-15.90	0.00
4	BETA_HE	-0.00535	0.000983	-5.45	0.00
5	BETA_TIME	-0.0128	0.00104	-12.23	0.00

Summary statistics
Number of observations = 6768
Number of excluded observations = 3960
Number of estimated parameters = 5

$$\begin{aligned} \mathcal{L}(\beta_0) &= -6964.663 \\ \mathcal{L}(\hat{\beta}) &= -5315.386 \\ -2[\mathcal{L}(\beta_0) - \mathcal{L}(\hat{\beta})] &= 3298.553 \\ \rho^2 &= 0.237 \\ \bar{\rho}^2 &= 0.236 \end{aligned}$$

Table 1: Estimates for the parameters of the base model

2 Heteroskedastic model

Files to use with Biogeme:
Model file: *Mixture_Heteroskedastic.py*
Data file: *swissmetro.dat*

Remark: In order to have accurate estimates for the random parameters, a high number of draws is usually considered. For the sake of convenience, in this section and the following ones we provide the specifications for different types of models and the estimates for a small number of draws (100 in all the cases except for the mixed GEV model, in which 50 draws are considered). You can try to run the model with a higher number of draws to identify the changes on the estimates, likelihood function, etc.

In this model specification we assume that ASC_{car} and ASC_{SM} are normally distributed with mean $\bar{\alpha}_{car}$ and $\bar{\alpha}_{SM}$ and standard deviation σ_{car} and σ_{SM}, respectively. The estimation results are reported in Table 2.

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	ASC_CAR_mean	0.248	0.111	2.24	0.03
2	ASC_CAR_std	-0.0501	0.0779	-0.64	0.52
3	ASC_SM_mean	0.917	0.198	4.62	0.00
4	ASC_SM_std	-3.25	0.427	-7.61	0.00
5	BETA_COST	-0.0178	0.00159	-11.20	0.00
6	BETA_HE	-0.00780	0.00137	-5.69	0.00
7	BETA_TIME	-0.0170	0.00206	-8.29	0.00

Summary statistics
Number of observations = 6768
Number of excluded observations = 3960
Number of estimated parameters = 7

$\mathcal{L}(\beta_0)$	=	-6964.663
$\mathcal{L}(\hat{\beta})$	=	-5239.062
$-2[\mathcal{L}(\beta_0) - \mathcal{L}(\hat{\beta})]$	=	3451.202
ρ^2	=	0.248
$\bar{\rho}^2$	=	0.247

Table 2: Estimates of the parameters for the heteroskedastic specification (with 100 draws)

We perform a likelihood ratio test in order to test if this model is better than the base model. The restricted model is the base model, since it assumes that the standard deviation is equal to 0 (i.e., $\sigma_{car} = 0$ and $\sigma_{SM} = 0$), which implies that the ASCs are directly the mean values (i.e., $ASC_{car} = \bar{\alpha}_{car}$ and $ASC_{SM} = \bar{\alpha}_{SM}$). Thus, the unrestricted model is the heteroskedastic model. The null hypothesis is given as follows:

$$H_0 : \sigma_{car} = \sigma_{SM} = 0.$$

The statistic for the likelihood ratio test is the following:

$$-2(-5315.386 + 5239.062) = 152.648,$$

which states that we can reject the null hypothesis since $\chi^2_{0.95,2} = 5.99$ at a 95% level of confidence.

3 Error Component Model

Files to develop from the file `Base_Model.py` in Biogeme:
Model file: *Error_Component_01.py*
Error_Component_02.py
Data file: *swissmetro.dat*

We present two different specifications of error component models. In the first specification, the train and SM modes share the random term ζ_{rail}, which is assumed to be normally distributed $\zeta_{rail} \sim N(m_{rail}, \sigma^2_{rail})$. This error component model attempts to capture the correlation between the train and Swissmetro alternatives. They are both rail-based transportation modes, so the hypothesis is that they share unobserved attributes. A similar idea could be implemented by means of a nested logit model. The systematic utility expressions are the following:

$$\begin{aligned} V_{car} &= ASC_{car} + \beta_{time} CAR\_TT + \beta_{cost} CAR\_CO \\ V_{train} &= \beta_{time} TRAIN\_TT + \beta_{cost} TRAIN\_CO + \beta_{he} TRAIN\_HE + \zeta_{rail} \\ V_{SM} &= ASC_{SM} + \beta_{time} SM\_TT + \beta_{cost} SM\_CO + \beta_{he} SM\_HE + \zeta_{rail} \end{aligned}$$

We estimate the standard deviation σ_{rail} of this error component, while the mean m_{rail} is fixed to zero for identification reasons. Indeed, it cannot be estimated as its value is *contained* in the associated ASC. The estimation results are reported in Table 3.

We perform a likelihood ratio test in order to test if this model is better than the base model. The null hypothesis is given as follows:

$$H_0 : \sigma_{rail} = 0.$$

The statistic for the likelihood ratio test is the following:

$$-2(-5315.386 + 5315.385) = 0.002,$$

which states that we cannot reject the null hypothesis since $\chi^2_{0.95,1} = 3.84$ at a 95% level of confidence.

In the second specification, we use a more complex error structure. The idea is that train and SM are correlated, being both rail-based transportation modes but also that train and car are correlated representing existing transportation modes as opposed to the more innovative Swissmetro. A similar correlation pattern could be specified by means of a cross-nested logit model where the SM alternative belongs to a *rail* nest, the car mode belongs to an *existing* nest and

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	ASC_CAR	0.189	0.0798	2.37	0.02
2	ASC_SM	0.451	0.0932	4.84	0.00
3	BETA_COST	-0.0108	0.000682	-15.90	0.00
4	BETA_HE	-0.00535	0.000983	-5.45	0.00
5	BETA_TIME	-0.0128	0.00104	-12.23	0.00
6	RAIL_std	-0.00677	0.0114	-0.59	0.55

Summary statistics

Number of observations = 6768

Number of excluded observations = 3960

Number of estimated parameters = 6

$$\begin{aligned} \mathcal{L}(\beta_0) &= -6964.663 \\ \mathcal{L}(\hat{\beta}) &= -5315.385 \\ -2[\mathcal{L}(\beta_0) - \mathcal{L}(\hat{\beta})] &= 3298.556 \\ \rho^2 &= 0.237 \\ \bar{\rho}^2 &= 0.236 \end{aligned}$$

Table 3: Estimates of the parameters for the first error component specification (with 100 draws)

the train alternative is assigned with certain degrees of membership to both rail and existing nests.

The utility specifications are the following:

$$\begin{aligned} V_{car} &= ASC_{car} + \beta_{time} CAR\_TT + \beta_{cost} CAR\_CO + \zeta_{existing} \\ V_{train} &= \beta_{time} TRAIN\_TT + \beta_{cost} TRAIN\_CO + \beta_{he} TRAIN\_HE + \zeta_{rail} + \zeta_{existing} \\ V_{SM} &= ASC_{SM} + \beta_{time} SM\_TT + \beta_{cost} SM\_CO + \beta_{he} SM\_HE + \zeta_{rail} \end{aligned}$$

As before, the random terms are supposed to be normally distributed $\zeta_{rail} \sim N(m_{rail}, \sigma^2_{rail})$ and $\zeta_{existing} \sim N(m_{existing}, \sigma^2_{existing})$. The standard deviations, σ_{rail} and $\sigma_{existing}$ are estimated, while the means m_{rail} and $m_{existing}$ are fixed to zero. The estimates of the parameters can be found in Table 4.

We perform a likelihood ratio test in order to test if this model is better than the base model. The null hypothesis is given as follows:

$$H_0 : \sigma_{rail} = \sigma_{existing} = 0.$$

The statistic for the likelihood ratio test is the following:

$$-2(-5315.386 + 5240.481) = 149.81,$$

which states that we can reject the null hypothesis since $\chi^2_{0.95,2} = 5.99$ at a 95% level of confidence.

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	ASC_CAR	0.252	0.110	2.29	0.02
2	ASC_SM	0.936	0.186	5.04	0.00
3	BETA_COST	-0.0177	0.00159	-11.18	0.00
4	BETA_HE	-0.00782	0.00137	-5.69	0.00
5	BETA_TIME	-0.0169	0.00199	-8.50	0.00
6	EXISTING_std	3.29	0.431	7.63	0.00
7	RAIL_std	-0.0155	0.0823	-0.19	0.85

Summary statistics

Number of observations = 6768

Number of excluded observations = 3960

Number of estimated parameters = 7

$$\begin{aligned} \mathcal{L}(\beta_0) &= -6964.663 \\ \mathcal{L}(\hat{\beta}) &= -5240.481 \\ -2[\mathcal{L}(\beta_0) - \mathcal{L}(\hat{\beta})] &= 3448.363 \\ \rho^2 &= 0.248 \\ \bar{\rho}^2 &= 0.247 \end{aligned}$$

Table 4: Estimates of the parameters for the second error component specification (with 100 draws)

4 Random Coefficients

Files to use with Biogeme:
Model file: *Random_Coefficients.py*
Data file: *swissmetro.dat*

In this specification the unknown parameters are assumed to be randomly distributed over the population. They capture the so called *taste variation* of individuals. In this case, the base model is modified by defining alternative-specific coefficients for the cost of all alternatives. The resulting utility specifications are the following:

$$\begin{aligned} V_{car} &= ASC_{car} + \beta_{time} CAR\_TT + \beta_{car\_cost} CAR\_CO \\ V_{train} &= \beta_{time} TRAIN\_TT + \beta_{train\_cost} TRAIN\_CO + \beta_{he} TRAIN\_HE \\ V_{SM} &= ASC_{SM} + \beta_{time} SM\_TT + \beta_{SM\_cost} SM\_CO + \beta_{he} SM\_HE \end{aligned}$$

The model in which the parameters β_{car_cost}, β_{train_cost}, β_{SM_cost} and β_{time} are assumed to be randomly distributed over the population is provided in the file `Random_Coefficients.py`.
Note that we have three alternative-specific coefficients for the cost variable, which are normally distributed, with means m_{car_cost}, m_{train_cost}, m_{SM_cost} and standard deviations σ_{car_cost}, σ_{train_cost}, σ_{SM_cost}, respectively. The coefficient related to headway is also assumed to be normally distributed over the population, with mean m_{he} and standard deviation σ_{he}. The estimation results are reported in Table 5.

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	ASC_CAR	-1.58	0.210	-7.51	0.00
2	ASC_SM	-1.03	0.158	-6.54	0.00
3	BETA_CAR_COST_mean	-0.0209	0.00395	-5.29	0.00
4	BETA_CAR_COST_std	0.0117	0.00293	3.98	0.00
5	BETA_HE_mean	-0.00737	0.00172	-4.29	0.00
6	BETA_HE_std	-0.00595	0.00352	-1.69	0.09
7	BETA_SM_COST_mean	-0.0187	0.00234	-7.98	0.00
8	BETA_SM_COST_std	-0.0109	0.00221	-4.95	0.00
9	BETA_TIME	-0.0139	0.00194	-7.16	0.00
10	BETA_TRAIN_COST_mean	-0.0659	0.00583	-11.30	0.00
11	BETA_TRAIN_COST_std	-0.0255	0.00299	-8.54	0.00

Summary statistics
Number of observations = 6768
Number of excluded observations = 3960
Number of estimated parameters = 11

$$\begin{aligned} \mathcal{L}(\beta_0) &= -6964.663 \\ \mathcal{L}(\hat{\beta}) &= -4967.484 \\ -2[\mathcal{L}(\beta_0) - \mathcal{L}(\hat{\beta})] &= 3994.359 \\ \rho^2 &= 0.287 \\ \bar{\rho}^2 &= 0.285 \end{aligned}$$

Table 5: Estimates of the parameters for the random coefficients specification (with 100 draws)

We perform a likelihood ratio test in order to test if this model is better than the associated restricted model. Note that the restricted model is not the one included in `Base_Model.py`, but the modified version with alternative-specific parameters for the cost variable. The loglikelihood of this model is -5068.559. The null hypothesis is given as follows:

$$H_0 : \sigma_{car\_cost} = \sigma_{train\_cost} = \sigma_{SM\_cost} = \sigma_{he} = 0.$$

The statistic for the likelihood ratio test is the following:

$$-2(-5068.559 + 4967.484) = 202.15,$$

which states that we can reject the null hypothesis since $\chi^2_{0.95,4} = 9.49$ at a 95% level of confidence.

Different distributions You can use this file as a template to model different distributions:

1. We can assume that the parameter β_{time} is log-normally distributed. Recall that, a variable X is log normally distributed if $y = \ln(X)$ is normally distributed.

2. We can assume that the parameter β_{time} follows a Johnson's Sb distribution. In the case of Johnson's Sb distribution, the functional form is derived using a logit-like transformation of a Normal distribution, as defined in the following equation:

$$\xi = a + (b - a)\frac{e^{\zeta}}{e^{\zeta} + 1} \tag{1}$$

where $\zeta \sim N(\mu, \sigma^2)$. This distribution is very flexible; it is bounded between a and b and its shape can change from a very flat one to a bimodal, by changing the parameters of the normal variable. The estimation of four parameters (a, b, μ and σ) and a nonlinear specification are required, assuming as before, a generic time coefficient following such a distribution.

Remark: The computational time for these specifications can be high. For the sake of trying the given specification or the ones with different distributions you can reduce the number of draws.

5 Mixed GEV Models

Files to use with Biogeme:
Model file: *Mixed_GEV.py*
Data file: *swissmetro.dat*

In this example we capture the substitution patterns by means of a nested logit model, and we allow for some parameters to be randomly distributed over the population. This approach is very interesting because it allows us to formulate hypotheses about the partition of the unobserved heterogeneity. We consider the model developed in Section 4. One nest contains the existing transportation modes (rail and car) and the other nest is composed by the Swissmetro alternative (innovative). The estimation results are reported in Table 6. In this case the considered number of draws is 50.

We perform a likelihood ratio test in order to test if this model is better than the base model with alternative-specific coefficients for the cost variable. The null hypothesis is given as follows:

$$H_0 : \sigma_{car\_cost} = \sigma_{train\_cost} = \sigma_{SM\_cost} = \sigma_{he} = 0, \mu_{existing} = 1,$$

The statistic for the likelihood ratio test is the following:

$$-2(-5068.559 + 4970.153) = 196.812,$$

which states that we can reject the null hypothesis since $\chi^2_{0.95,5} = 11.07$ at a 95% level of confidence.

---

mbi/ ek/ afa /mpp

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	ASC_CAR	-1.26	0.163	-7.74	0.00
2	ASC_SM	-0.810	0.116	-7.00	0.00
3	BETA_CAR_COST_mean	-0.0131	0.00271	-4.83	0.00
4	BETA_CAR_COST_std	0.00588	0.00192	3.06	0.00
5	BETA_HE_mean	-0.00586	0.00116	-5.04	0.00
6	BETA_HE_std	0.000646	0.0103	0.06	0.95
7	BETA_SM_COST_mean	-0.0130	0.00152	-8.55	0.00
8	BETA_SM_COST_std	0.00566	0.00153	3.71	0.00
9	BETA_TIME	-0.0116	0.00134	-8.65	0.00
10	BETA_TRAIN_COST_mean	-0.0480	0.00658	-7.30	0.00
11	BETA_TRAIN_COST_std	-0.0195	0.00297	-6.59	0.00
12	Existing	1.39	0.236	5.88	0.00

Summary statistics

Number of observations = 6768

Number of excluded observations = 3960

Number of estimated parameters = 12

$$\begin{aligned} \mathcal{L}(\beta_0) &= -6964.663 \\ \mathcal{L}(\hat{\beta}) &= -4970.153 \\ -2[\mathcal{L}(\beta_0) - \mathcal{L}(\hat{\beta})] &= 3989.020 \\ \rho^2 &= 0.286 \\ \bar{\rho}^2 &= 0.285 \end{aligned}$$

Table 6: Estimates for the parameters in the mixed nested logit model (50 draws)