EPFL
ENAC TRANSP-OR
Prof. M. Bierlaire

Mathematical Modeling of Behavior
Fall 2016

EXERCISE SESSION 12

This session focuses on *mixture models* and *latent variables models*. Since Monte-Carlo integration is normally used when estimating mixtures choice models and latent variables models, we refer the student to http://biogeme.epfl.ch/documentation/montecarlo.pdf for complete explanations.

1 Mixture models

In order to have precised estimates for the random parameters, a high number of draws is usually considered. For the sake of convenience, we provide in the following subsections the specifications for different types of models and the estimates for a small number of draws (100 in this case, and 10 for mixed GEV model). You can try to run the model with a higher number of draws to identify the changes on the estimates, likelihood function, etc.

Try to analyse the given specification for each type of model:

1. What are the underlying assumptions?
2. Is the model correctly specified?
3. What conclusions can you draw from the estimation results?

It would also be interesting to compare these results with the results of models estimated during the previous lab sessions. Consult the Pythonbiogeme user manual for details on how different random distributions are specified.

For this type of models we characterize four different specifications by considering the Swissmetro data. The *.py* files are provided.

1.1 Heteroskedastic Model

Files to use with Biogeme:
Model file: *01_Mixture_Heteroskedastic.py*
Data file: *swissmetro.dat*

In this first model specification we assume that the ASCs are randomly distributed with mean $\bar{\alpha}_{car}$ and $\bar{\alpha}_{SM}$ and standard deviation σ_{car} and σ_{SM}. Below, we provide the utility expressions.

The normalization is with respect to the train alternative. The estimation results are reported in Table 1.

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

			Robust		
Parameter		Coeff.	Asympt.		
number	Description	estimate	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

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

Table 1: Heteroskedastic specification (with 100 draws)

1.2 Error Component Model

Files to use with Biogeme:
Model file: *02_Mixture_ErrorComp_01.py, 03_Mixture_ErrorComp_02.py*
Data file: *swissmetro.dat*

We present two different specifications of error component models. Below, we provide the systematic utility expressions. The train and SM modes share the random term ζ_{rail}, which is assumed to be normally distributed $\zeta_{rail} \sim N(m_{rail}, \sigma^2_{rail})$. We estimate the standard deviation σ_{rail} of this error component, while the mean m_{rail} is fixed to zero. The estimation results are reported in Table 2.

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

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 2: First Error component specification (with 100 draws)

In the second model we use a more complex error structure. The specification is presented below and the estimation results are reported in Table 3.

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

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	CLASSIC_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 3: Second Error component specification (with 100 draws)

1.3 Random Coefficients

Files to use with Biogeme:
Model file: *04_Mixture_RandomComp.py*
Data file: *swissmetro.dat*

In this specification the unknown parameters are assumed to be randomly distributed over the population. The utility expressions are shown below. The estimation results are reported in Table 4.

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

Different distributions

We hereby describe two examples for a specification of a random coefficient model where the parameters are log-normally and Johnson's Sb distributed, accordingly. Recall that, a variable X is log normally distributed if $y = \ln(X)$ is normally distributed. In the case of Johnson's

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 4: Random coefficient specification (with 100 draws)

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.

Try to implement these two versions for the parameter β_{time} by adapting the provided *.py* file.

1.4 Mixed GEV Models

Files to use with Biogeme:
Model file: *05_Mixture_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.

$$
\begin{aligned}
V_{car} &= ASC_{car} + \beta_{car\_time} CAR\_TT + \beta_{cost} CAR\_CO \\
V_{train} &= \beta_{train\_time} TRAIN\_TT + \beta_{cost} TRAIN\_CO + \beta_{he} TRAIN\_HE \\
& \quad + \beta_{ga} GA + \beta_{age} AGE \\
V_{SM} &= ASC_{SM} + \beta_{SM\_time} SM\_TT + \beta_{cost} SM\_CO + \beta_{he} SM\_HE \\
& \quad + \beta_{ga} GA + \beta_{seats} SEATS
\end{aligned}
$$

We specify a nest composed of the alternatives *car* and *train* representing standard transportation modes, while the Swissmetro alternative represents the technological innovation. We further assume a generic cost parameter and three randomly distributed alternative-specific time parameters. Normal distributions are used for the random coefficients, that is,

$$
\begin{aligned}
\beta_{car\_time} &\sim N(m_{car\_time}, \sigma^2_{car\_time}) \\
\beta_{train\_time} &\sim N(m_{train\_time}, \sigma^2_{train\_time}) \\
\beta_{SM\_time} &\sim N(m_{SM\_time}, \sigma^2_{SM\_time}).
\end{aligned}
$$

The estimation results are reported in Table 5.

2 Latent variables models

For this part you will use the documents provided in http://biogeme.epfl.ch/examples_latent.html. In particular, the tutorial on how to use PythonBiogeme to estimate latent variable models and one of the provided examples.

Before looking at the estimation results of the model:

1. Go through `05latentChoiceFull.py` in detail.
2. Which is the attitude considered?
3. Which are the indicators considered?
4. Write down the utility functions considered.
5. Write down the structural equations considered.
6. Write down the measurement equation considered.
7. Draw the graphical representation of this model.

Now look at the estimation results shown in Table 6. Interpret the parameters one by one.

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	ASC_CAR	-0.175	0.121	-1.44	0.15
2	ASC_SM	0.237	0.109	2.17	0.03
3	BETA_CAR_TIME_mean	-0.0119	0.000889	-13.40	0.00
4	BETA_CAR_TIME_std	0.00478	0.00125	3.81	0.00
5	BETA_COST	-0.00947	0.000803	-11.80	0.00
6	BETA_GA	0.949	0.143	6.66	0.00
7	BETA_HE	-0.00466	0.000857	-5.44	0.00
8	BETA_SEATS	-0.276	0.0994	-2.78	0.01
9	BETA_SENIOR	1.50	0.126	11.91	0.00
10	BETA_SM_TIME_mean	-0.0140	0.00122	-11.43	0.00
11	BETA_SM_TIME_std	-0.00700	0.00134	-5.24	0.00
12	BETA_TRAIN_TIME	-0.0146	0.000945	-15.44	0.00
13	BETA_TRAIN_TIME_std	-0.000142	0.000730	-0.19	0.85
14	Classic	1.90	0.175	10.83	0.00

Summary statistics

Number of observations = 6759

Number of excluded observations = 3969

Number of estimated parameters = 14

$$\begin{aligned}
\mathcal{L}(\beta_0) &= -6958.425 \\
\mathcal{L}(\hat{\beta}) &= -4970.224 \\
-2[\mathcal{L}(\beta_0) - \mathcal{L}(\hat{\beta})] &= 3976.402 \\
\rho^2 &= 0.286 \\
\bar{\rho}^2 &= 0.284
\end{aligned}$$

Table 5: Mixed Nested Logit estimation results (with 10 draws)

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	ASC_CAR	0.703	0.118	5.96	0.00
2	ASC_SM	0.261	0.345	0.76	0.45
3	BETA_COST_HWH	-1.43	0.341	-4.19	0.00
4	BETA_COST_OTHER	-0.526	0.161	-3.27	0.00
5	BETA_DIST	-1.41	0.386	-3.66	0.00
6	BETA_TIME_CAR_CL	-0.955	0.169	-5.65	0.00
7	BETA_TIME_CAR_REF	-9.50	1.94	-4.90	0.00
8	BETA_TIME_PT_CL	-0.456	0.143	-3.19	0.00
9	BETA_TIME_PT_REF	-3.23	0.839	-3.84	0.00
10	BETA_WAITING_TIME	-0.0205	0.00963	-2.13	0.03
11	B_Envir02_F1	-0.459	0.0308	-14.88	0.00
12	B_Envir03_F1	0.484	0.0316	15.32	0.00
13	B_Mobil11_F1	0.572	0.0419	13.65	0.00
14	B_Mobil14_F1	0.575	0.0350	16.42	0.00
15	B_Mobil16_F1	0.525	0.0425	12.36	0.00
16	B_Mobil17_F1	0.514	0.0420	12.25	0.00
17	INTER_Envir02	0.460	0.0308	14.92	0.00
18	INTER_Envir03	-0.367	0.0289	-12.69	0.00
19	INTER_Mobil11	0.418	0.0373	11.22	0.00
20	INTER_Mobil14	-0.173	0.0278	-6.21	0.00
21	INTER_Mobil16	0.147	0.0336	4.39	0.00
22	INTER_Mobil17	0.140	0.0329	4.24	0.00
23	SIGMA_STAR_Envir02	0.918	0.0344	26.64	0.00
24	SIGMA_STAR_Envir03	0.857	0.0352	24.34	0.00
25	SIGMA_STAR_Mobil11	0.895	0.0409	21.89	0.00
26	SIGMA_STAR_Mobil14	0.759	0.0333	22.81	0.00
27	SIGMA_STAR_Mobil16	0.873	0.0397	21.97	0.00
28	SIGMA_STAR_Mobil17	0.876	0.0392	22.36	0.00
29	coef_ContIncome_0_4000	0.146	0.0606	2.41	0.02
30	coef_ContIncome_10000_more	0.119	0.0365	3.25	0.00
31	coef_ContIncome_4000_6000	-0.279	0.114	-2.45	0.01
32	coef_ContIncome_6000_8000	0.321	0.137	2.34	0.02
33	coef_ContIncome_8000_10000	-0.666	0.157	-4.25	0.00
34	coef_age_65_more	0.0403	0.0748	0.54	0.59
35	coef_haveChildren	-0.0276	0.0563	-0.49	0.62
36	coef_haveGA	-0.745	0.0999	-7.46	0.00
37	coef_highEducation	-0.265	0.0670	-3.96	0.00
38	coef_individualHouse	-0.116	0.0560	-2.08	0.04
39	coef_intercept	0.373	0.169	2.21	0.03
40	coef_male	0.0776	0.0534	1.45	0.15
41	coef_moreThanOneBike	-0.365	0.0686	-5.32	0.00
42	coef_moreThanOneCar	0.711	0.0667	10.66	0.00
43	delta_1	0.328	0.0127	25.81	0.00
44	delta_2	0.989	0.0358	27.64	0.00
45	sigma_s	0.855	0.0549	15.57	0.00

Summary statistics

Number of observations = 1906

Number of excluded observations = 359

Number of estimated parameters = 45

$$\begin{aligned}\mathcal{L}(\beta\_0) &= -28534.376 \\ \mathcal{L}(\hat{\beta}) &= -18383.063 \\ -2[\mathcal{L}(\beta\_0) - \mathcal{L}(\hat{\beta})] &= 20302.626 \\ \rho^2 &= 0.356 \\ \bar{\rho}^2 &= 0.354\end{aligned}$$

Table 6: Estimation results of the latent variable model.