Computer Lab II
Biogeme &
Specifying Models

Evanthia Kazagli, Anna Fernandez Antolin & Matthieu de Lapparent

Transport and Mobility Laboratory
School of Architecture, Civil and Environmental Engineering
École Polytechnique Fédérale de Lausanne

September 22, 2015

Administrative info

- The schedule of the lectures has been slightly modified. See:

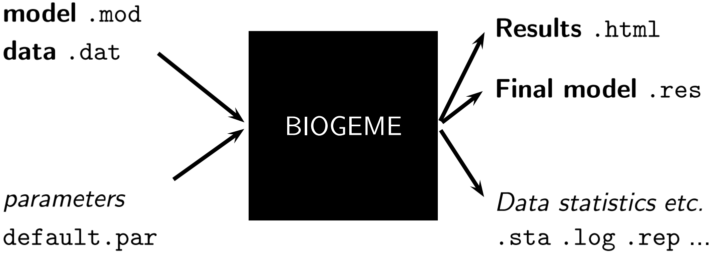
  http://transp-or.epfl.ch/courses/dca2015/schedule2015.php

- The assignment that will be submitted on the 27^{th} **of November** will be maximum 4 (2 double-sided) pages long, and is compulsory in order to be able participate in the final exam. More information will follow.

Today

1. Closer look at BIOGEME
2. Specifying, estimating and interpreting models

How does BIOGEME work?



model `.mod`

data `.dat`

parameters

`default.par`

BIOGEME

Results `.html`

Final model `.res`

Data statistics etc.

`.sta .log .rep` ...

BIOGEME - Data file

- File extension `.dat`
- It contains the data, what we call observations.
- One observation per row.
- First row contains column (variable) names.
- Each row must contain a choice indicator.

- Example with the Netherlands transportation mode choice data: choice between car and train.

BIOGEME - Data file

netherlands.dat

```
id      choice  rail_cost   rail_time   car_cost   car_time
1       0       40          2.5         5          1.167
2       0       35          2.016       9          1.517
3       0       24          2.017       11.5       1.966
4       0       7.8         1.75        8.333      2
5       0       28          2.034       5          1.267
...
219     1       35          2.416       6.4        1.283
220     1       30          2.334       2.083      1.667
221     1       35.7        1.834       16.667     2.017
222     1       47          1.833       72         1.533
223     1       30          1.967       30         1.267
```

BIOGEME - Data file

netherlands.dat

```
id      choice  rail_cost   rail_time   car_cost   car_time
1       0       40          2.5         5          1.167
2       0       35          2.016       9          1.517
3       0       24          2.017       11.5       1.966
4       0       7.8         1.75        8.333      2
5       0       28          2.034       5          1.267
...
219     1       35          2.416       6.4        1.283
220     1       30          2.334       2.083      1.667
221     1       35.7        1.834       16.667     2.017
222     1       47          1.833       72         1.533
223     1       30          1.967       30         1.267
```

Unique identifier of observations

BIOGEME - Data file

netherlands.dat

id	choice	rail_cost	rail_time	car_cost	car_time
1	0	40	2.5	5	1.167
2	0	35	2.016	9	1.517
3	0	24	2.017	11.5	1.966
4	0	7.8	1.75	8.333	2
5	0	28	2.034	5	1.267
...					
219	1	35	2.416	6.4	1.283
220	1	30	2.334	2.083	1.667
221	1	35.7	1.834	16.667	2.017
222	1	47	1.833	72	1.533
223	1	30	1.967	30	1.267

Choice indicator, 0: car and 1: train

BIOGEME - Model file

- File extension `.mod`
- Must be consistent with data file.
- Contains deterministic utility specifications, model type etc.
- The model file contains different `[Sections]` describing different elements of the model specification.

BIOGEME - Model file

- How can we write the following deterministic utility functions in BIOGEME?

$$\begin{aligned} V_{\text{car}} &= \text{ASC}_{\text{car}} + \beta_{\text{time}}\text{time}_{\text{car}} + \beta_{\text{cost}}\text{cost}_{\text{car}} \\ V_{\text{rail}} &= \beta_{\text{time}}\text{time}_{\text{rail}} + \beta_{\text{cost}}\text{cost}_{\text{rail}} \end{aligned}$$

BIOGEME - Model file

```
[Choice]
choice

[Beta]
// Name      DefaultValue LowerBound UpperBound  status
ASC_CAR      0.0          -100.0     100.0         0
ASC_RAIL     0.0          -100.0     100.0         1
BETA_COST    0.0          -100.0     100.0         0
BETA_TIME    0.0          -100.0     100.0         0

[Utilities]
//Id Name Avail linear-in-parameter expression
0    Car  one   ASC_CAR * one + BETA_COST * car_cost +
                BETA_TIME * car_time
1    Rail one   ASC_RAIL * one + BETA_COST * rail_cost +
                BETA_TIME * rail_time
```

BIOGEME - Model file

```
[Choice]
choice

[Beta]
// Name       DefaultValue LowerBound UpperBound  status
ASC_CAR       0.0          -100.0     100.0         0
ASC_RAIL      0.0          -100.0     100.0         1
BETA_COST     0.0          -100.0     100.0         0
BETA_TIME     0.0          -100.0     100.0         0

[Utilities]
//Id Name Avail linear-in-parameter expression
0    Car  one   ASC_CAR * one + BETA_COST * car_cost +
                BETA_TIME * car_time
1    Rail one   ASC_RAIL * one + BETA_COST * rail_cost +
                BETA_TIME * rail_time
```

BIOGEME - Model file

```
[Choice]
choice

[Beta]
// Name       DefaultValue LowerBound UpperBound  status
ASC_CAR       0.0          -100.0     100.0         0
ASC_RAIL      0.0          -100.0     100.0         1
BETA_COST     0.0          -100.0     100.0         0
BETA_TIME     0.0          -100.0     100.0         0

[Utilities]
//Id Name Avail linear-in-parameter expression
0    Car  one   ASC_CAR * one + BETA_COST * car_cost +
                BETA_TIME * car_time
1    Rail one   ASC_RAIL * one + BETA_COST * rail_cost +
                BETA_TIME * rail_time
```

BIOGEME - Model file

```
[Choice]
choice
```

What is one?

Which is the type of model?

```
[Beta]
// Name       DefaultValue LowerBound UpperBound  status
ASC_CAR       0.0          -100.0     100.0         0
ASC_RAIL      0.0          -100.0     100.0         1
BETA_COST     0.0          -100.0     100.0         0
BETA_TIME     0.0          -100.0     100.0         0

[Utilities]
//Id Name Avail linear-in-parameter expression
0    Car  one   ASC_CAR * one + BETA_COST * car_cost +
                BETA_TIME * car_time
1    Rail one   ASC_RAIL * one + BETA_COST * rail_cost +
                BETA_TIME * rail_time
```

BIOGEME - Model file

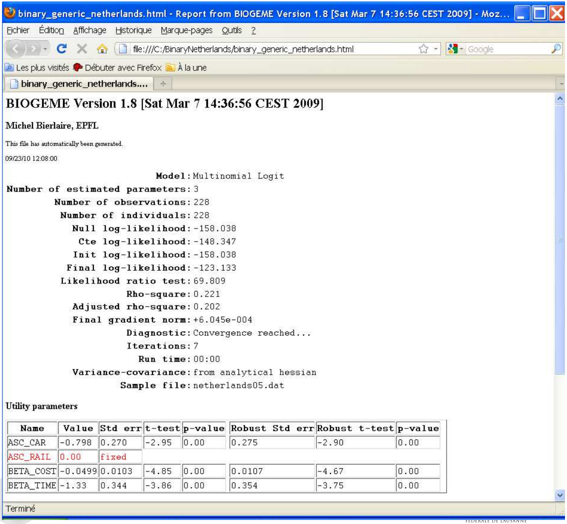
```
[Expressions]
// Define here arithmetic expressions for name that are not directly
// available from the data
one = 1

[Model]
// Currently, only $MNL (multinomial logit), $NL (nested logit), $CNL
// (cross-nested logit) and $NGEV (Network GEV model) are valid keywords
//
$MNL
```

Model and Data Files

- How to read and modify model files?
- How to read data files?
  - GNU Emacs, TextEdit (Mac) or Wordpad (Windows)
  - Notepad (Windows) should not be used!

BIOGEME - Output - Netherlands dataset


binary_generic_netherlands.html - Report from BIOGEME Version 1.8 [Sat Mar 7 14:36:56 CEST 2009] - Moz...
Fichier Édition Affichage Historique Marque-pages Outils ?
file:///C:/BinaryNetherlands/binary_generic_netherlands.html
Google
Les plus visités Débuter avec Firefox À la une
binary_generic_netherlands....
Terminé

BIOGEME Version 1.8 [Sat Mar 7 14:36:56 CEST 2009]

Michel Bierlaire, EPFL

This file has automatically been generated.

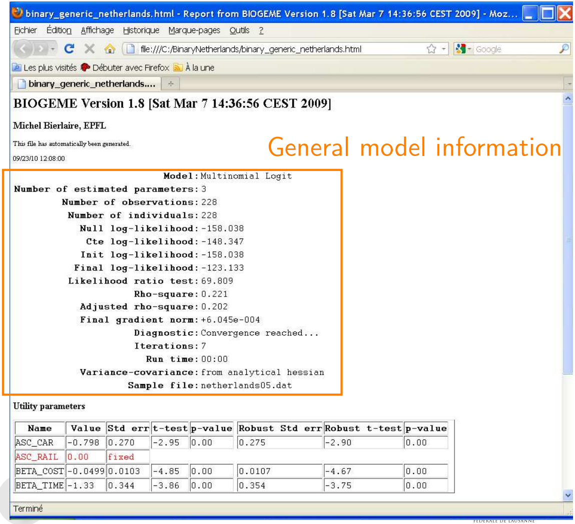
09/23/10 12:08:00

```
                       Model: Multinomial Logit
Number of estimated parameters: 3
        Number of observations: 228
         Number of individuals: 228
           Null log-likelihood: -158.038
            Cte log-likelihood: -148.347
           Init log-likelihood: -158.038
          Final log-likelihood: -123.133
         Likelihood ratio test: 69.809
                   Rho-square: 0.221
          Adjusted rho-square: 0.202
          Final gradient norm: +6.045e-004
                   Diagnostic: Convergence reached...
                   Iterations: 7
                     Run time: 00:00
          Variance-covariance: from analytical hessian
                  Sample file: netherlands05.dat
```

Utility parameters

Name	Value	Std err	t-test	p-value	Robust Std err	Robust t-test	p-value
ASC_CAR	-0.798	0.270	-2.95	0.00	0.275	-2.90	0.00
ASC_RAIL	0.00	fixed					
BETA_COST	-0.0499	0.0103	-4.85	0.00	0.0107	-4.67	0.00
BETA_TIME	-1.33	0.344	-3.86	0.00	0.354	-3.75	0.00

BIOGEME - Output


binary_generic_netherlands.html - Report from BIOGEME Version 1.8 [Sat Mar 7 14:36:56 CEST 2009] - Moz...
Fichier Édition Affichage Historique Marque-pages Outils ?
file:///C:/BinaryNetherlands/binary_generic_netherlands.html
Google
Les plus visités Débuter avec Firefox À la une
binary_generic_netherlands....
Terminé

BIOGEME Version 1.8 [Sat Mar 7 14:36:56 CEST 2009]

Michel Bierlaire, EPFL

This file has automatically been generated.

09/23/10 12:08:00

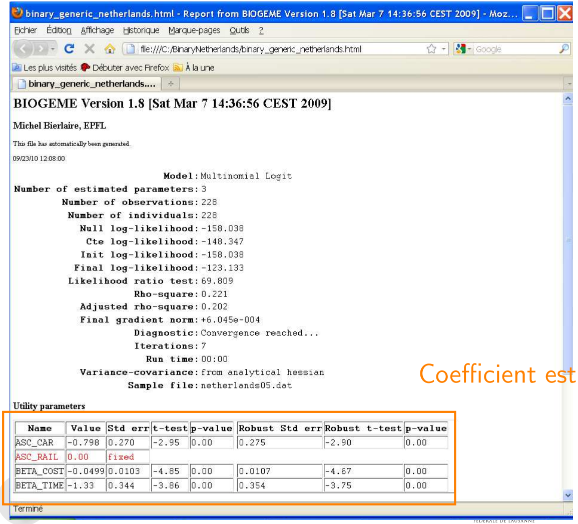
General model information

```
                       Model: Multinomial Logit
Number of estimated parameters: 3
      Number of observations: 228
       Number of individuals: 228
         Null log-likelihood: -158.038
          Cte log-likelihood: -148.347
         Init log-likelihood: -158.038
        Final log-likelihood: -123.133
       Likelihood ratio test: 69.809
                  Rho-square: 0.221
         Adjusted rho-square: 0.202
         Final gradient norm: +6.045e-004
                  Diagnostic: Convergence reached...
                  Iterations: 7
                    Run time: 00:00
         Variance-covariance: from analytical hessian
                 Sample file: netherlands05.dat
```

Utility parameters

Name	Value	Std err	t-test	p-value	Robust Std err	Robust t-test	p-value
ASC_CAR	-0.798	0.270	-2.95	0.00	0.275	-2.90	0.00
ASC_RAIL	0.00	fixed					
BETA_COST	-0.0499	0.0103	-4.85	0.00	0.0107	-4.67	0.00
BETA_TIME	-1.33	0.344	-3.86	0.00	0.354	-3.75	0.00

BIOGEME - Output


binary_generic_netherlands.html - Report from BIOGEME Version 1.8 [Sat Mar 7 14:36:56 CEST 2009] - Moz...
Fichier Édition Affichage Historique Marque-pages Outils ?
file:///C:/BinaryNetherlands/binary_generic_netherlands.html
Google
Les plus visités Débuter avec Firefox À la une
binary_generic_netherlands....
Terminé

BIOGEME Version 1.8 [Sat Mar 7 14:36:56 CEST 2009]

Michel Bierlaire, EPFL

This file has automatically been generated.

09/23/10 12:08:00

```
                      Model: Multinomial Logit
Number of estimated parameters: 3
        Number of observations: 228
         Number of individuals: 228
           Null log-likelihood: -158.038
            Cte log-likelihood: -148.347
           Init log-likelihood: -158.038
          Final log-likelihood: -123.133
         Likelihood ratio test: 69.809
                    Rho-square: 0.221
           Adjusted rho-square: 0.202
           Final gradient norm: +6.045e-004
                    Diagnostic: Convergence reached...
                    Iterations: 7
                      Run time: 00:00
           Variance-covariance: from analytical hessian
                   Sample file: netherlands05.dat
```

Utility parameters

Name	Value	Std err	t-test	p-value	Robust Std err	Robust t-test	p-value
ASC_CAR	-0.798	0.270	-2.95	0.00	0.275	-2.90	0.00
ASC_RAIL	0.00	fixed					
BETA_COST	-0.0499	0.0103	-4.85	0.00	0.0107	-4.67	0.00
BETA_TIME	-1.33	0.344	-3.86	0.00	0.354	-3.75	0.00

Coefficient estimates

Today

1. Further introduction to BIOGEME
2. Specifying, estimating and interpreting models

Binary Logit Case Study

- Available datasets:
  - Airline itinerary choice (Boeing)
  - Choice-Lab marketing
  - Mode choice in Netherlands
  - Residential Telephone Services
  - Mode choice in Switzerland (Optima)
- Descriptions available on the course webpage.
- Optima dataset does not contain `.mod` files. A specification has to be proposed for the assignment.

How to go through the Case Studies

- Choose a dataset to work with (data descriptions are available on the course webpage).
- Copy the files related to the chosen dataset and case study from the course webpage.
- Go through the `.mod` files with the help of the descriptions.
- Run the `.mod` files with BIOGEME.
- Interpret the results and compare your interpretation with the one we have proposed.
- Develop other model specifications.

Course webpage

- http://transp-or.epfl.ch/
  → Teaching → Mathematical modeling of behavior → Laboratories
- BIOGEME software
  (including documentation and utilities)
- For each Case Study:
  - Data files;
  - Model specification files;
  - Possible interpretation of results.

Types of parameters

- In the linear formulation of utility functions, the βs are called coefficients or parameters. Different types:
  - Alternative specific constants (ASC).
  - Generic:
    - Appearing in all utility functions with equal coefficients.
    - Assume all choice makers have the same marginal utility among the alternatives.
  - Alternative specific:
    - Different coefficients among utility functions.
    - Capture the marginal utility specific to an alternative.
  - Alternative-specific socioeconomic:
    - Reflect differences in preference as functions of characteristics of the decision-maker.

Tests

Goal: test alternative specifications of the explanatory variables in the utility functions. Different tests:

- t-test
- Likelihood ratio test

t-test

- Goal: test whether a particular parameter in the model differs from some known constant –usually zero.
- Valid only asymptotically.
- t-test > 1.96 means significant parameter (95% confidence interval).

Likelihood ratio test (LRT)

- Goal: compare different specifications (i.e. models).
- Restricted model (e.g. some βs $= 0$ –null hypothesis) vs unrestricted model.
- Number of degrees of freedom (d.o.f.): difference between the number of estimated coefficients in the restricted and unrestricted model.
- χ^2 test with this number of d.o.f.: $-2(\mathcal{L}(\hat{\beta}_{unrestricted}) - (\hat{\beta}_{restricted}))$
- Find the LRT excel file in the Utilities tab on biogeme's official homepage.

Interpretation

- Is the coefficient significant?
- Are the signs reasonable?
  - Coefficients are expected to have a behavioral meaning, i.e. a negative coefficient means lower utility when the variable value increases, and higher utility when the variable value decreases (e.g. cost, travel time etc.).
  - The interpretation the other way around is the same (e.g. speed).

Specifying models: Recommended steps

- Formulate a-priori hypothesis:
  - Expectations and intuition regarding the explanatory variables that appear to be significant for mode choice.
- Specify a minimal model:
  - Start simple;
  - Include the main factors affecting the mode choice of (rational) travelers;
  - This will be your starting point.
- Continue adding and testing variables that improve the initial model in terms of *causality*, and *efficiency* with respect to what actually happened in the sample.

Evaluating models

The main indicators used to evaluate and compare the various models are summarised here:

- Informal tests:
  - *signs* and *relative magnitudes* of the parameters β values (under our a-priori expectations);
  - *trade-offs* among some attributes and ratios of pairs of parameters (e.g. reasonable value of time).
- Overall goodness of fit measure:
  - *adjusted rho-square* (likelihood ratio index): takes into account the different number of explanatory variables used in the models and normalizes for their effect $\rightarrow$ suitable to compare models with different number of independent variables. We check this value to have a first idea about which model might be better (among models of the same type), but it is not a statistical test.

Evaluating models (cont.)

- Statistical tests:
  - *t-test values*: statistically significant explanatory variables are denoted by t-statistic values remarkably higher/ lower than ± 2 (for a 95% level of confidence);
  - *final log-likelihood* for the full set of parameters: should be remarkably different from the ones in the naive approach (null log-likelihood and log-likelihood at constants); we ask for high values of likelihood ratio test $[-2(LL(0) - LL(\beta))]$ in order to have a model significantly different than the naive one.
- Test of entire models:
  - *likelihood ratio test* $[-2(LL(\hat{\beta}_R) - LL(\hat{\beta}_U))]$: used to test the null hypothesis that two models are equivalent, under the requirement that the one is the restricted version of the other. The likelihood ratio test is X^2 distributed, with degrees of freedom equal to $K_U - K_R$ (where K the number of parameters of the unrestricted and restricted model, respectively).