Computer Lab II
Biogeme &
Binary Logit Model Estimation

Evanthia Kazagli, Anna Fernandez Antolin & Antonin Danalet

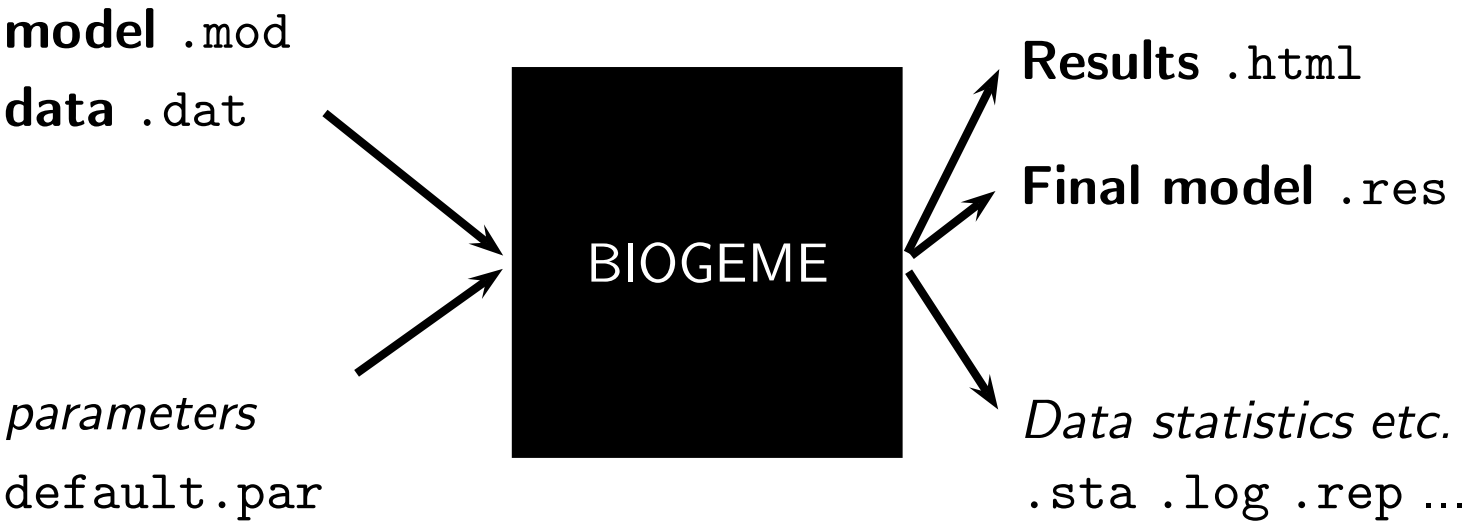
Transport and Mobility Laboratory
School of Architecture, Civil and Environmental Engineering
École Polytechnique Fédérale de Lausanne

September 23, 2014

Today

- Further introduction to BIOGEME
- Estimation of Binary Logit models

How does BIOGEME work?



BIOGEME - Data file

- File extension `.dat`
- First row contains column (variable) names.
- One observation per row.
- 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
...	Unique identifier of observations				
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

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
```

TRANSP-OR

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

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
```

What is one?

Which is the type of model?

BIOGEME - Model file

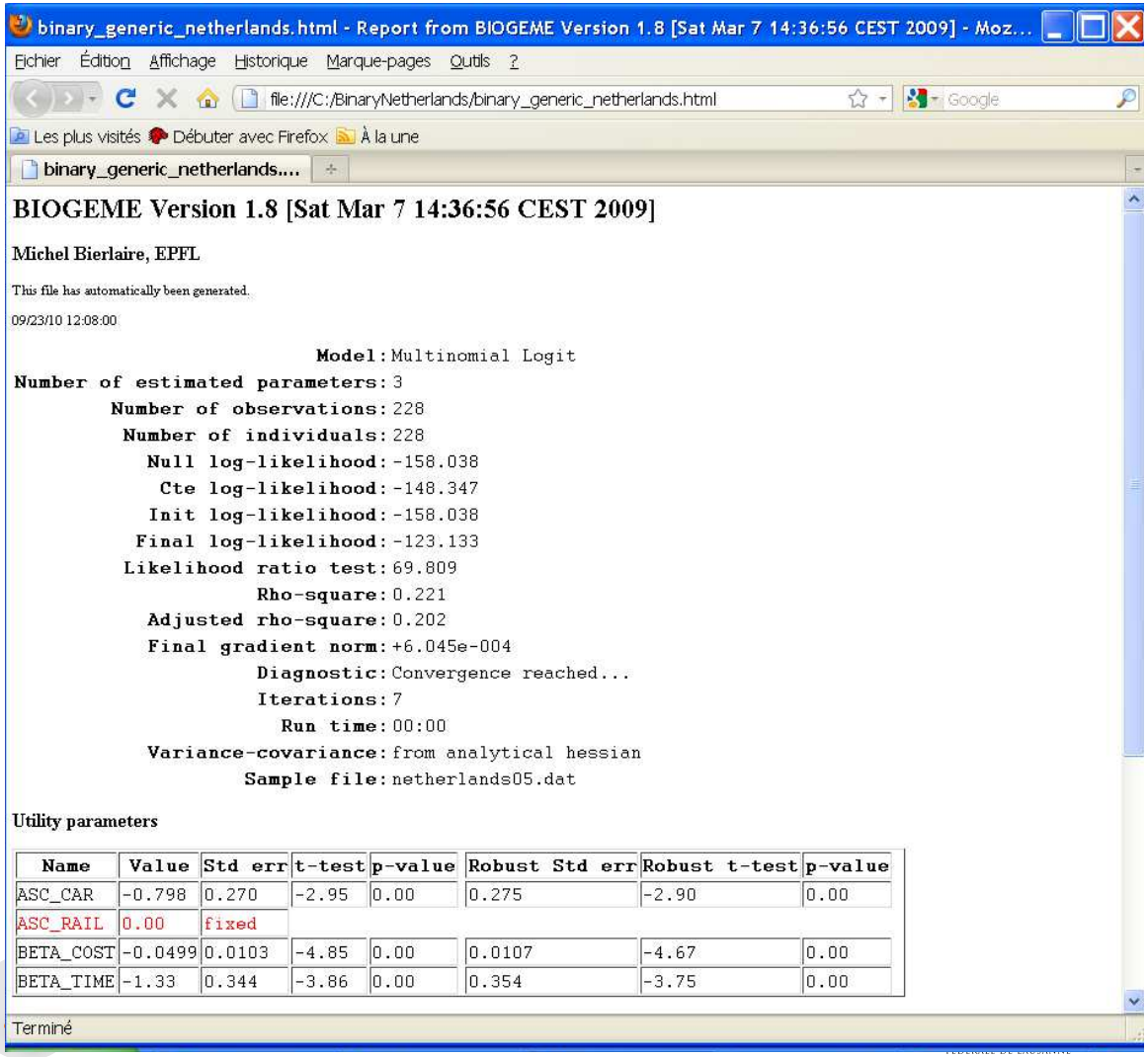
```
[Expressions]
// Define here arithmetic expressions for name that are not directly
// available from the data
one = 1

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

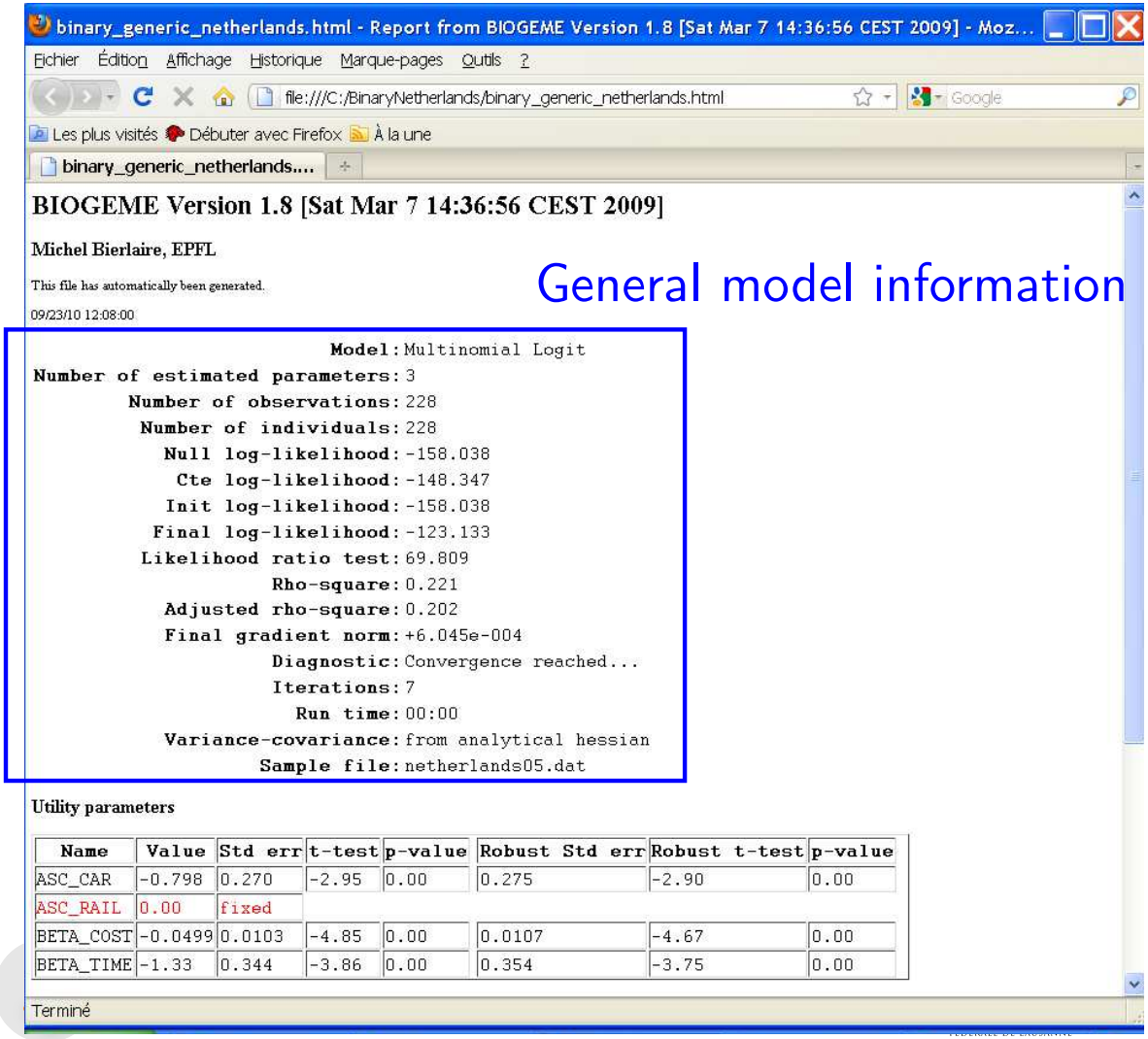
Model and Data Files

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

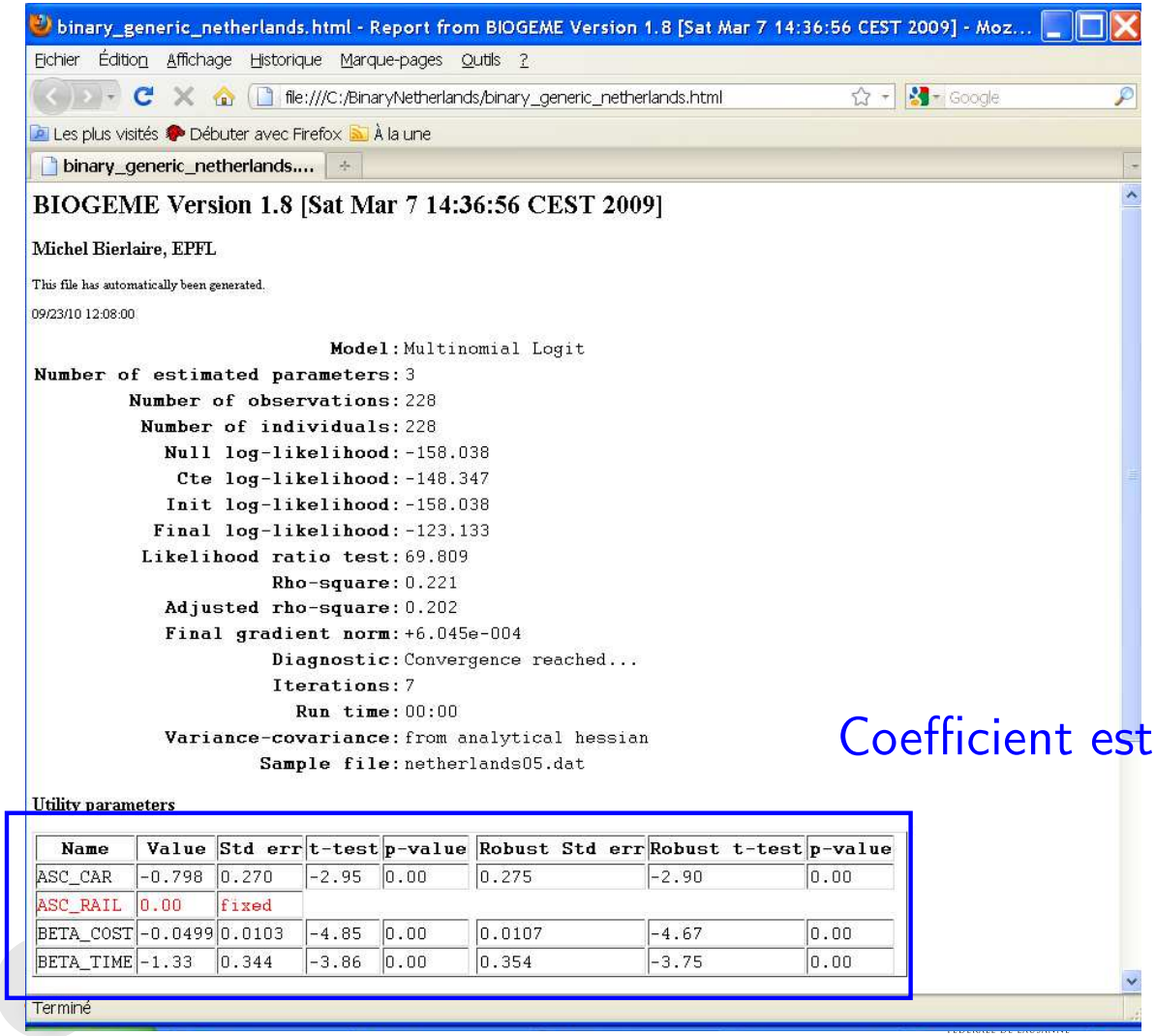
BIOGEME - Results - Netherlands dataset



BIOGEME - Results



BIOGEME - Results



Coefficient estimates

Today

- Further introduction to BIOGEME
- **Estimation of Binary Logit 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 as an assignment (next lab session).

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.

Today's plan

1. Independent work on 2-3 Case Studies
   - choose a case study;
   - estimate a model;
   - interpret the results.
2. Group work
   - gather in groups;
   - generate .mod file (base);
   - test an idea/ hypothesis.

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).