Computer Lab III
Summary

Evanthia Kazagli

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

March 3, 2015

Today

- Summary of what you've learnt so far:
  - Types of coefficients (generic, specific, socioeconomic).
  - Tests (likelihood ratio test, t-test).
- Help: dealing with missing data.
- You will work on lab 3 exercise.

Data set: Mode choice in Switzerland (Optima)

- Data set "optimaTOT3_valid.dat" on the website.
- Description of the data and variables available on the website:
  - General description.
  - List of variables.

OPTIMA dataset

- Objective: study mode choice in low density areas of switzerland.
- Data collection: Revealed Preferences (RP).
- Survey type: Mail survey.
- Context: Loop trips = cyclical sequence of trips.
- Dataset with 2265 observations including:
  - Trip features;
  - Socioeconomic variables.

Types of parameters

In linear formulation of the 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 between the alternatives.
  - Alternative specific:
    - Different coefficients between utility functions.
    - Capture the marginal utility specific to an alternative.
- Alternative-specific socioeconomic:
  - Reflect differences in preference as a function of the characteristics of the decision-maker.

Tests

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

- t-test
- Likelihood ratio test

Tests: t-test

- Goal: test whether a particular parameter in the model differs from some known constant, often zero.
- Valid only asymptotically (since we work with nonlinear models).
- t-test > 1.96 means significant parameter (95% confidence interval).

Tests: Likelihood ratio test

- Goal: compare different specifications (i.e., models).
- Restricted model (e.g., some βs $= 0$) (null hypothesis) vs unrestricted model.
- Number of degrees of freedom: difference between the number of estimated coefficients in the restricted and unrestricted models.
- χ^2 test with this number of d.o.f.: $-2(\mathcal{L}(\hat{\beta}_{restricted}) - \mathcal{L}(\hat{\beta}_{unrestricted}))$

Tests: Likelihood ratio test (cont.)

1. Calculate the degrees of freedom: difference between the number of estimated coefficients in the restricted (df_r)and unrestricted (df_u) models.
2. Calculate the value of the test statistic: $-2(\mathcal{L}(\hat{\beta}_{restricted}) - \mathcal{L}(\hat{\beta}_{unrestricted}))$
3. Look up in a table the value of the χ^2 you are interested in $\chi^2_{0.95,(df_r - df_u)}$
4. If $-2(\mathcal{L}(\hat{\beta}_{restricted}) - \mathcal{L}(\hat{\beta}_{unrestricted})) > \chi^2_{0.95,(df_r - df_u)}$ we can reject the null hypothesis. Therefore the unrestricted model is better.
5. 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).

Dealing with missing data

- Section `[Exclude]` tells BIOGEME to NOT consider some observations.
- **Example** of `binary_generic_boeing.mod`

```
[Exclude] ArrivalTimeHours_1 == -1 || BestAlternative_3
```

  1. Excludes missing data (-1) for variable ArrivalTimeHours_1
  2. Excludes alternative BestAlternative_3 (1 Stop with 2 different airlines)

- `[Exclude]` needs to be used in the Optima case study to exclude soft modes and only consider choice between public transportation and car for your assignment (binary logit model).

Dealing with missing data

- **Example**: if you want to use the gender variable (q17_gender).
- **Solution 1**
  - Exclude missing data (-1 and 99) from **the whole data set**
    → `[Exclude] q17_gender == 99 || q17_gender == -1`

Dealing with missing data

- **Example**: if you want to use the gender variable (q17_gender).
- **Solution 2 (better)**
  - Measure taste heterogeneity between men and women by introducing a term for missing data in the utility.
  - In section `[Expressions]` define:
    - `MissingGender = ( ( q17_Gender == -1 ) + ( q17_Gender == 99 ) ) > 0`
  - In section `[Utilities]` specify:
    - `+ Male_Opt2 * Male + MDGender * MissingGender`