Computer Lab IV

Summary

Evanthia Kazagli

evanthia.kazagli@epfl.ch

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Today

- Summary of what you've learnt so far:
  - Types of variables (generic, specific, socioeconomic)
  - Tests (likelihood ratio test, t-test)
- Help: dealing with missing data.
- You'll work on lab 4 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

Types of explanatory variables

In linear formulation of utility function, βs are called coefficients or parameters. Different kinds:

- 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 functions of 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 freedom:
  $-2(\mathcal{L}(\hat{\beta}_{unrestricted}) - (\hat{\beta}_{restricted}))$

Interpretation

- Is the coefficient significant?
- Sign
  - Coefficients are expected to have a behavioral meaning: a negative coefficient means lower utility when the variable is high, and higher utility when the variable is low, e.g. travel time, cost.
  - The other way around: same interpretation.

Dealing with missing data

- Section `[Exclude]` tells BIOGEME not to consider some observations.
- **Example** of `binary_generic_boeing.mod`
  - `[Exclude] ArrivalTimeHours_1 == -1 || BestAlternative_3`
  - Excludes missing data (-1) for variable ArrivalTimeHours_1
  - Excludes alternative BestAlternative_3 (1 Stop with 2 different airlines)
- The same needs to be done for the Optima case study: exclude soft modes, and keep public transportation and cars if you want to estimate a binary choice model only for the motorised modes.

Dealing with missing data (cont.)

- **Example**: if want to use gender variable (q17_gender)
- **Solution 1**
  - Exclude missing data (-1 and 99) from **whole data set**
  - `[Exclude] ArrivalTimeHours_1 == -1 || BestAlternative_3 || q17_gender == 99 || q17_gender == -1`

Dealing with missing data (cont.)

- **Example**: if want to use gender variable (q17_gender)
- **Solution 2 (BETTER)**
  - Measure taste heterogeneity between men and women by introducing a term for missing data in utility
  - `[Exclude]` section identical
    - `[Exclude] ArrivalTimeHours_1 == -1 || BestAlternative_3`
  - In section `[Expressions]` define:
    - `MissingGender = ((q17_Gender == -1) + (q17_Gender == 99)) > 0`
  - In section `[Utilities]` specify:
    - `+ Male_Opt2 * Male + MDGender * MissingGender`