Computer Lab 3
Data and missing values

Anna Fernandez Antolin & Evanthia Kazagli & Matthieu de Lapparent

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

September 29, 2015

Outline

Datasets

The datasets that will be used during the course are the following:

- Transportation Mode Choice in the Netherlands
- Residential Telephone Services
- Swissmetro
- Airline Itinerary Choice (Boeing)
- Mode choice in Switzerland (Optima)

Datasets

To do:

1. Download the dataset and the dataset description that can be found in the website.
2. Read carefully the dataset descriptions.
3. Open the dataset with a text editor and copy its contents in Excel.
4. Go through all the columns to make sure that you understand what they represent.
5. Choose randomly a few observations (each row corresponds to one observation) and study them in detail. Does the choice of this respondent correspond to your expectations? Are there any missing values? How are they coded?

Example

From the Netherlands dataset respondent with `id==6`:

- `car_walk_time == -1`
- `car_parking_fee == -1`
- `seat_status == -1`

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)

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`

Today's plan

- Look in detail into the different datasets;
- Identify missing values and understand why something needs to be done about them;
- Try to code the different approaches to deal with missing values.