Missing data and data imputation with the Swiss Household Panel

André Berchtold

LIVES, LINES, Université de Lausanne

FORS SHP workshop – June 12-14, 2018

UNIL | Université de Lausanne

Outline

1. **WHAT ARE MISSING DATA ?**

2. **HOW TO TREAT MISSING DATA ?**

3. **LONGITUDINAL DATA, CAUSALITY, & ETHICS**

Definitions

- **Missing data** : Data whose collect was planned, but for which we do not have values
- **Partial missing data** : Only a part of the information is missing for a particular subject, ie for a subset of variables (= **item non-response**)
- **Complete missing data** : All information is missing for a particular subject in a given wave (= **unit non-response**)
- **Attrition** : Decrease in the available sample size of a longitudinal study, because some subjects have only missing data after a given wave

Working in the unknown

- Missing data are a very complicated field
- Some situations are (still ?) impossible to identify
- Even the best solutions to missing data can generate errors, and we cannot always identify these errors
- Developing field / Work in progress

Outline

1 WHAT ARE MISSING DATA ?
- Examples from SHP
- Consequences of missing data
- Classification of missing data

2 HOW TO TREAT MISSING DATA ?
- Basic notions
- Simple imputation
- Multiple imputation
- Some good questions about imputation

3 LONGITUDINAL DATA, CAUSALITY, & ETHICS
- Specificities of longitudinal data
- Experiments
- Missing data and ethics

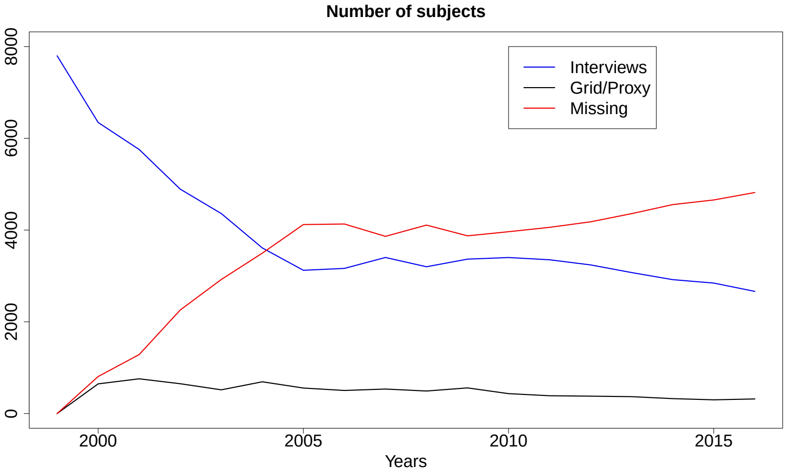
Example dataset

- File : Data_DM_FORS
- Available in SPSS and Stata formats. For R, you can load the SPSS file using the *foreign* library
- All individuals interviewed in 1999 (SHP I, n=7799)
- Data from wave 1 (1999) to wave 18 (2016)
- Subset of 900 variables covering many domains

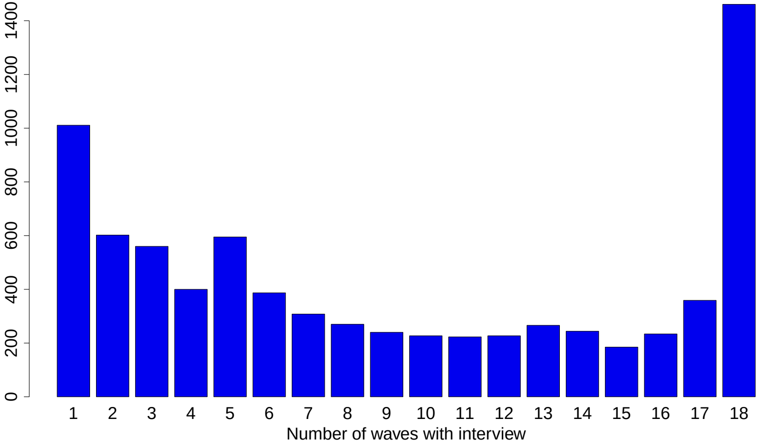
Unit non-response in our sample

- We begin by exploring unit non-response accross waves
- **... but why is it important ?**

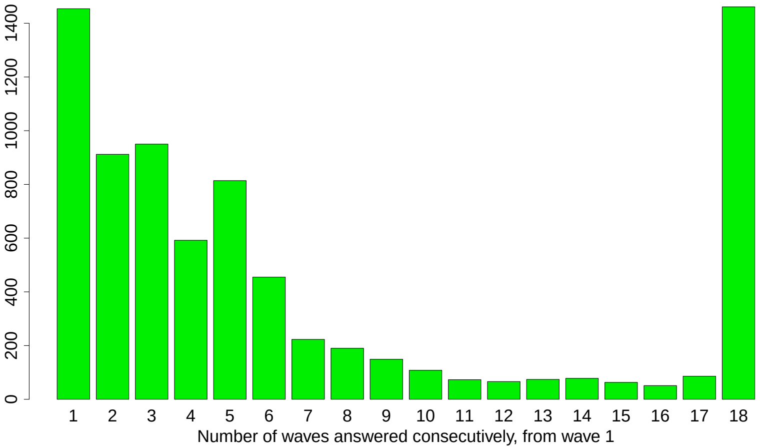
Number of answers, by wave



Number of waves answered, by subject



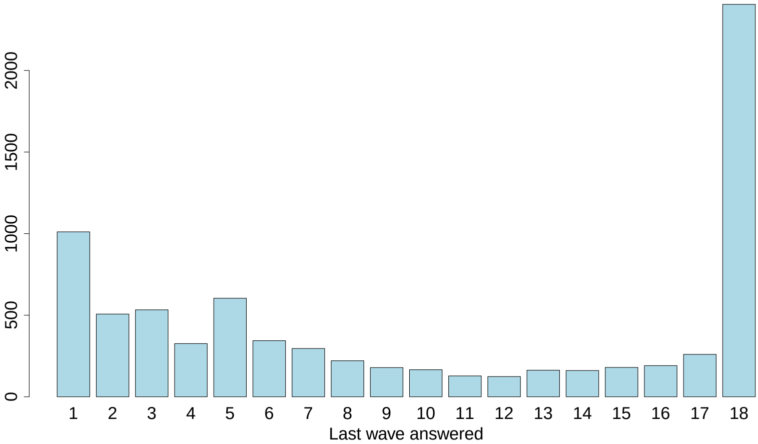
Number of consecutive answers, from wave 1



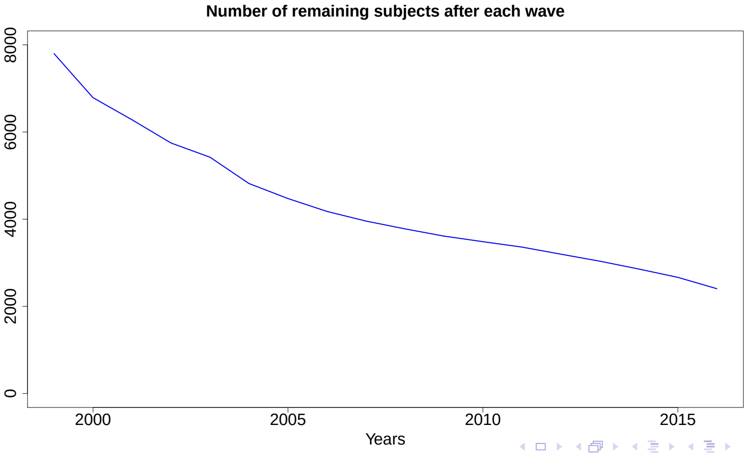
Attrition

- Attrition is a (definitive) diminution of the original sample size in a longitudinal study
- In some situations, impossible to know whether a subject without answer at wave *t* will answer again in the future
- Attrition rate is certain only after the end of a study

Last wave answered



Remaining subjects after each wave (as if SHP ended in 2016)



Causes of unit non-response

- Subjects leaving a longitudinal survey at some point (attrition)
- Impossibility to contact some subjects included in the study
- Answers from subjects included in the study, but who are not part of the population of interest
- Participant who do not meet the inclusion criteria anymore (leaving Switzerland for instance)
- Death of a participant
- ...

Item non-response in our sample

- In a second step, we look at item non-response
- **... but why is it important (and/or different from unit non-response) ?**

The case of 2010 (wave 12)

- Number of variables in the dataset : 46
- Number of answers :
  - individual questionnaire : 3401
  - proxy questionnaire : 40
  - grid : 394
- Number of complete data (value available for all 46 variables) : 0
- Data from proxy or grid, OK. But what about individual questionnaires ?
- $\Longrightarrow$ We consider the subsample of 3401 subjects having answered the individual questionnaire

Missing data per variable in 2010

```
    0    1    2    3    5    9   37  376  769
   17    1    2    3    1    1    1    1    1

       934 1155 1156 1165 1178 1181 1195 1238
         2    1    1    1    1    1    1    1

      1279 1586 2221 2247 2253 3401
         1    1    1    1    1    4
```

Variables without valid answers

- 4 variables without valid answers
- X10C05, X10C06, X10I04, X10I05
- Questions about health and income
- Coded as "inapplicable"

Logical vs real missing data

- In many situations, we do not have answers because we did not ask the question ! !
  $\Longrightarrow$ MD caused by logical skips
- This is less trivial than it could appear ...
- Different possibilities :
  - The variable exists in the database, because it corresponds to a question asked only to some respondents (SHP II for instance)
  - The variable was not asked in function of a previous answer (e.g. the age of the first child is not asked if the subject answered previously that he/she has no child)
  - The interviewer forgot to ask the question
  - ...

Number of MD (except completely missing)

```
IDHOUS10   STATUS10     SEX10      AGE10   RELARP10   COHAST10
       0          0     0          0          0        934

IDSPOU10   CIVSTA10   OWNKID10   EDUCAT10   EDCAT10    EDUGR10
     934          0          0          0         0          0

  EDGR10   EDYEAR10  OCCUPA10   NAT_1_10    WSTAT10   NOGA2M10
       0          0         1          0          0       1279

TR1MAJ10   I10PTOTG     I10WYG    WP10T1S  WP10LP1S     P10D29
    1181        376       1238          0         0          3

  P10C01     P10C11     P10C15     P10W04     P10W39     P10W42
       0         37        769       2247       1165       2253

  P10W43     P10W46    P10W216    P10W228    P10W92     P10I01
    2221       1195       1156       1155      1178          3

  P10N35     P10P01     P10A06     P10A01     P10A09     P10A15
    1586          3          9          5          2          2
```

Very few MD

- OCCUPA10 (actual occupation, from grid) : 1 MD
  - no answer (1)
- P10P01 (interest in politics) : 3 MD
  - does not know (3)
- P10A06 (satisfaction with leisure activities) : 9 MD
  - no answer (3)
  - does not know (6)
- ...

More MD

- P10C11 (number of doctor consultations, last 12 months) : 37 MD
  - no answer (6)
  - does not know (31)
- I10PTOTG (yearly total personal income, gross) : 376 MD
  - other error (34)
  - no personal income (131)
  - no answer (153)
  - does not know (58)
- ...

Many MD

- IDSPOU10 (identification number of partner or spouse) : 934 MD
    - inapplicable (934)
- NOGA2M10 (current main job, nomenclature) : 1279 MD
    - inapplicable (1154)
    - no answer (125)
- P10W04 (seeking job, last for weeks) : 2247 MD
    - inapplicable (2247)
- ...

Causes of item non-response

- Intentional non-response to some questions
- Questions not asked because they were not relevant (logical skip)
- Error in the design of the questionnaire (e.g. questions not asked because of a wrong filter)
- Questions not asked in a short form of a questionnaire
- Removal of outliers
- ...

$\Longrightarrow$ **In some cases, the cause of MD is clearly identified (e.g. logical skip), in other cases it is not obvious (e.g. intentional non-response)**

The simple remedy

THE BEST REMEDY (PROVEN ! !) AGAINST MISSING DATA IS ...

NOT HAVING MISSING DATA !

- Sounds like a joke, but this is true
- Much attention and effort should be paid to prevent missing data :
  - questionnaire design
  - sampling method
  - incentives
  - accurate treatment of data
  - matching of databases
  - ...

Outline

1. WHAT ARE MISSING DATA ?
   - Examples from SHP
   - **Consequences of missing data**
   - Classification of missing data
2. HOW TO TREAT MISSING DATA ?
   - Basic notions
   - Simple imputation
   - Multiple imputation
   - Some good questions about imputation
3. LONGITUDINAL DATA, CAUSALITY, & ETHICS
   - Specificities of longitudinal data
   - Experiments
   - Missing data and ethics

Example : Pearson correlation (1)

- 4 variables : AGE10 (age), OWNKID10 (number of children), TR1MAJ10 (Treiman job prestige scale), I10PTOTG (yearly total income)

```
> cor(D10[c(4,9,19,20)])

         AGE10 OWNKID10 TR1MAJ10 I10PTOTG
AGE10     1.00     0.22       NA       NA
OWNKID10  0.22     1.00       NA       NA
TR1MAJ10    NA       NA        1       NA
I10PTOTG    NA       NA       NA        1
```

Example : Pearson correlation (2)

```
> cor(D10[c(4,9,19,20)],use="complete.obs")

          AGE10 OWNKID10 TR1MAJ10 I10PTOTG
AGE10     1.000    0.281   -0.048    0.058
OWNKID10  0.281    1.000   -0.074   -0.037
TR1MAJ10 -0.048   -0.074    1.000    0.197
I10PTOTG  0.058   -0.037    0.197    1.000
```

```
> cor(D10[c(4,9,19,20)],use="pairwise.complete.obs")

          AGE10 OWNKID10 TR1MAJ10 I10PTOTG
AGE10     1.000    0.218   -0.050   -0.096
OWNKID10  0.218    1.000   -0.084   -0.052
TR1MAJ10 -0.050   -0.084    1.000    0.197
I10PTOTG -0.096   -0.052    0.197    1.000
```

Example : Linear regression for I10PTOTG (1)

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    -23353      13932   -1.68  0.09387 .
D10$AGE10         802        226    3.56  0.00039 ***
D10$OWNKID10    -3732       1889   -1.98  0.04830 *
D10$TR1MAJ10     1644        180    9.11  < 2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 1e+05 on 2040 degrees of freedom
  (1357 observations deleted due to missingness)
Multiple R-squared:  0.0453,Adjusted R-squared:  0.0439
F-statistic: 32.2 on 3 and 2040 DF,  p-value: <2e-16
```

Example : Linear regression for I10PTOTG (2)

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    106734       6547   16.30  < 2e-16 ***
D10$AGE10        -559        118   -4.75  2.2e-06 ***
D10$OWNKID10    -2141       1314   -1.63      0.1
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 90800 on 3022 degrees of freedom
  (376 observations deleted due to missingness)
Multiple R-squared:  0.0101,Adjusted R-squared:  0.0094
F-statistic: 15.3 on 2 and 3022 DF,  p-value: 2.34e-07
```

Consequences of missing data

- Less data to compute statistics
  $\longrightarrow$ less statistical power
- Different number of data points at each wave of a longitudinal study or for each variable
  $\longrightarrow$ statistics computed on different subsets of the data
  $\longrightarrow$ difficult to compare results
- Possible bias of point estimates
- Possible underestimation of the variability of results
  $\longrightarrow$ too high probability of rejecting null hypotheses
- Impossibility to follow the individual trajectories of subjects in longitudinal surveys
- ...

Outline

1. WHAT ARE MISSING DATA ?
   - Examples from SHP
   - Consequences of missing data
   - Classification of missing data
2. HOW TO TREAT MISSING DATA ?
   - Basic notions
   - Simple imputation
   - Multiple imputation
   - Some good questions about imputation
3. LONGITUDINAL DATA, CAUSALITY, & ETHICS
   - Specificities of longitudinal data
   - Experiments
   - Missing data and ethics

Three types of missing data (1)

- Classification due to Rubin (1976)
- Let

$$Y = Y_o + Y_m$$

  denotes the complete dataset with Y_o the observed part of the data and Y_m the missing part
- Let R be the indicator matrix of missing data
- Three different kind of missing data are defined in function of the relation between Y_o, Y_m and R

Three types of missing data (2)

- **Missing Completely At Random (MCAR)** : Missing data are a random sample of the observations

$$P(R|Y) = P(R)$$

- **Missing At Random (MAR)** : The probability of missing depends on other variables (of the database)

$$P(R|Y) = P(R|Y_o)$$

- **Missing Not At Random (MNAR)** : The probability of missing depends on the missing values themselves

$$P(R|Y) = P(R|Y_m) \quad \text{or} \quad P(R|Y) = P(R|Y_m + Y_o)$$

Example : MCAR

$$Y = \left(\begin{array}{cc} \textbf{Nationality} & \textbf{Age} \\ \text{Swiss} & 20 \\ \text{Swiss} & 50 \\ \text{Swiss} & 20 \\ \text{Swiss} & 50 \\ \text{German} & 20 \\ \text{German} & 50 \\ \text{German} & 20 \\ \text{German} & 50 \end{array}\right) \qquad R = \left(\begin{array}{cc} \textbf{Nationality} & \textbf{Age} \\ 0 & 1 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{array}\right)$$

Example : MAR

$$Y = \left(\begin{array}{cc} \textbf{Nationality} & \textbf{Age} \\ \text{Swiss} & 20 \\ \text{Swiss} & 50 \\ \text{Swiss} & 20 \\ \text{Swiss} & 50 \\ \text{German} & 20 \\ \text{German} & 50 \\ \text{German} & 20 \\ \text{German} & 50 \end{array} \right) \qquad R = \left(\begin{array}{cc} \textbf{Nationality} & \textbf{Age} \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{array} \right)$$

Example : MNAR

$$Y = \left(\begin{array}{cc} \textbf{Nationality} & \textbf{Age} \\ \text{Swiss} & 20 \\ \text{Swiss} & 50 \\ \text{Swiss} & 20 \\ \text{Swiss} & 50 \\ \text{German} & 20 \\ \text{German} & 50 \\ \text{German} & 20 \\ \text{German} & 50 \end{array}\right) \qquad R = \left(\begin{array}{cc} \textbf{Nationality} & \textbf{Age} \\ 0 & 1 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 0 \end{array}\right)$$

Ignorable vs non-ignorable MD

- Missing data are sometimes classified as *ignorable* and *non-ignorable*
- This is related to the possible impact of MD on statistical results
- Basically, MCAR are ignorable, and MAR & MNAR are non-ignorable
- When MD are not ignorable, the MD mechanism should be accounted for during statistical analyses

How to determine the type of missing data ?

- How to work with something that does not exist ? ? ?
- What should be tested ?
- What can be tested ?
- ... ideas ?
- *Remark : Of course, in a same database, we can have MD of different types*

Tests based on the mean (1)

- Hypotheses :

$$\begin{aligned} H_0 &: \text{MCAR} \\ H_1 &: \text{not MCAR} \end{aligned}$$

- The principle is to check whether the distributions of other variables are different when the data on the variable of interest are missing or not
- If the distributions are different, then the missing data are not completely random
- In practice, each variable with DM divides the sample in two parts (with and without MD), and the equality of the mean of other variables is tested between the two subsamples

Tests based on the mean (2)

- Dixon (1988) : individual Student t-test for each variable
- Little (1988) : global test based on the log-likelihood
- These tests consider only the mean
- Applicable on numerical data only
- Other test : Park & Davis (1993) : Extension of Little test to longitudinal categorical data

Tests based on the mean and covariance

- Jamshidian & Jalal (2010)
- Simultaneous test of normality and homogeneity of covariances
- If homogeneity rejected, then MCAR rejected
- Problem : the rejection of H_0 can also imply that normality (and not homogeneity) is rejected
- A second, non-parametric, test must be performed on the covariances after rejection of the first test
- ... quite complex to use in practice
- Applicable on numerical data only

Availability of tests

- Little :
  - R : LittleMCAR from library BaylorEdPsych
  - Stata : user written function mcartest
  - SPSS : in the Missing Value Analysis dialog box (tick the EM box)
- Jamshidian :
  - R : TestMCARNormality from library MissMech

Regression based test

- Rouzinov & Berchtold (2016-2018, in development)

1. Regression on the observed part of the data :

$$X^A_{1,obs} = f(X^A_{2,obs}, X^A_{3,obs}, ..., X^A_{k,obs}) \Longrightarrow \beta^A$$

2. Predictions for both the observed and missing parts of X_1 :

$$\hat{X}^A_{1,obs} = f(\hat{\beta}^A, X^A_{2,obs}, X^A_{3,obs}, ..., X^A_{k,obs})$$

$$\hat{X}^B_{1,mis} = f(\hat{\beta}^A, X^B_{2,obs}, X^B_{3,obs}, ..., X^B_{k,obs})$$

3. Comparison of the distributions of $\hat{X}^A_{1,obs}$ and $\hat{X}^B_{1,mis}$
Equality $\Longrightarrow$ MCAR

To summarize about tests

- No method available to test between all 3 types of MD
- Available methods generally designed for numerical data $\Longrightarrow$ What about categorical data ?
- Tests have only small power and can give contradictory results ...

CONFUSED ?

Tips & good practices

- The more you can understand about your MD, the better !
- Begin by testing each variable with MD separately
- Always check whether MD were caused by logical skips or are "real"

Outline

1 WHAT ARE MISSING DATA ?

- Examples from SHP
- Consequences of missing data
- Classification of missing data

2 HOW TO TREAT MISSING DATA ?

- **Basic notions**
- Simple imputation
- Multiple imputation
- Some good questions about imputation

3 LONGITUDINAL DATA, CAUSALITY, & ETHICS

- Specificities of longitudinal data
- Experiments
- Missing data and ethics

Four main approaches

- Ignoring
- Weighting
- Likelihood-based estimation
- Imputing

The prehistory : ignoring MD

- Only available data are analyzed. Missing data are simply discarded :
  - listwise deletion : all subjects with at least one MD are suppressed from all analyses
  - pairwise deletion : subjects with MD are suppressed only when variables with MD are used
- Should only be used with MCAR, ... but not optimal even in this case
- Otherwise : biased results

Somewhat rough ...

- ... but this is the default method in many statistical softwares (and the preferred choice of many social sciences researchers ...)

Weighting

- Applicable mainly to unit-missing data
- This is what the Swiss Household Panel does for attrition from wave to wave
- The idea is to modify the respective importance of each individual during the statistical analyses, in order to have a sample keeping a constant structure (sex, age, ...) through time
- With weights, results computed from different waves with different sample sizes can still be compared

Likelihood-based estimation (1)

- Several variants do exist : multi-group approach, Full Information Maximum Likelihood (FIML), EM, ...
- Basic idea : use all available information from all data to estimate parameters of interest, without explicitly imputing missing values
- For instance, if a strong correlation exists between two variables, one having missing data, then an information about the values of MD on the other variable can be obtained

Likelihood-based estimation (2)

- Multi-group approach : The full sample is split into several subgroups and the likelihood is computed separately from each subgroup. More information can then be extracted from the data, since the pattern of MD can be different in each subgroup
- FIML : Same idea, but pushed further : the likelihood is computed separately for each observation

Likelihood-based estimation (3)

- These methods generally suppose that data follow a multivariate-normal distribution
- Moreover, they are not available for all kind of models
- Quite simple to use (much simpler than multiple imputation for instance) and provide accurate results, but not for all statistical models and data
- In practice, when hypotheses are met, results are similar to results obtained with multiple imputation
- See e.g. Enders (2001) for an introduction

Imputation

- Imputation is the process of replacing missing values by likely ones
- Many approaches, from very rough to very sophisticated
- Can be based on the variable with missing data itself and/or on additional information
- Simple or multiple imputation

Outline

1. WHAT ARE MISSING DATA ?
   - Examples from SHP
   - Consequences of missing data
   - Classification of missing data
2. HOW TO TREAT MISSING DATA ?
   - Basic notions
   - **Simple imputation**
   - Multiple imputation
   - Some good questions about imputation
3. LONGITUDINAL DATA, CAUSALITY, & ETHICS
   - Specificities of longitudinal data
   - Experiments
   - Missing data and ethics

Basic idea

- Each missing data is replaced by a single imputed value
- Many mechanisms are available for the imputation (mean, mode, median, hot deck, probability distribution of observed values, regression model, ...)
- The choice of a specific mechanism should depend on our knowledge of the dataset and of the missing data (generating mechanism, ...)

General warning

- Imputed data are rarely precise at the individual level !
- *If a continuous variable has MCAR data, replacing missing data by the average of available data will results in an unbiased estimation of the mean, but of course at the individual level, almost all imputed values will be false*
- Imputed data should be used at the aggregated level only, to estimate characteristics of the population
- Even at the aggregated level, results can be biased !

The constant approach (1)

- For a given variable, all MD are replaced by a same value
- This value can be based on our knowledge of the data, but generally it is computed from the variable itself
- Knowledge of the data : We know from another study that people not answering to this question have a specific behavior or value
- Computed from the variable : mean, median, mode

The constant approach (2)

- Advantage : very easy to use
- Drawbacks :
  - Reinforced the central tendency of the variable (or another value of the distribution)
  - Limit the dispersion, hence the variance, of the variable
  - Very unrealistic in most cases

The constant approach (3)

- **Zero imputation**
- In multiple-choice questions, zero imputation consists in imputing a zero value (meaning that the event did not occur) in case of missing data
- *Do you smoke cigarettes ? Yes, No.*
- People not smoking may not answer because they are not concerned by the question
- Zero imputation impute them as *No*

The constant approach (4)

- In the case of categorical variables, missing data are sometimes considered as an **additional modality** of the variable
- This is not a true imputation, since we do not try to find the true value of the MD
- The idea behind this practice is that MD convey a specific information, ie respondents wanted to tell us something through the fact of not answering
- In practice, working with this additional modality can be complicated

The random approach

- In the random approach, a different value determined from a random distribution is imputed for each missing value
- *Hot deck* : a value taken from the same dataset is used
- *Cold deck* : a value taken from another dataset is used
- Easiest solution : computing the distribution of the variable with MD and randomly selecting one value

Matching

- For a given subject (the receiver) having a missing data on a specific variable, the closest other subject (the donor) is selected in function of variables without MD
- The value of the donor is then used as imputation value

The non-random approach

- In the non-random approach, a specific imputation value is computed for each MD on the basis of a set of explanatory variables
- Standard solutions : regression models, predictive mean matching
- Advantages :
  - Coherence between imputed values and other variables
  - Variability better preserved
- Drawback :
  - A good imputation model must be defined ↔ explanatory variables must exist

Outline

1 WHAT ARE MISSING DATA ?
- Examples from SHP
- Consequences of missing data
- Classification of missing data

2 HOW TO TREAT MISSING DATA ?
- Basic notions
- Simple imputation
- **Multiple imputation**
- Some good questions about imputation

3 LONGITUDINAL DATA, CAUSALITY, & ETHICS
- Specificities of longitudinal data
- Experiments
- Missing data and ethics

Main problems with simple imputation

1. The inherent variability of the non-observed true data is often underestimated by the imputed values
2. Results can also be systematically biased

$\Longrightarrow$ One modern solution : multiple imputation (Rubin, 1987 ; Schafer, 1999)

Principle of Multiple Imputation

- Each missing value is replaced by $m > 1$ imputed values instead of only one
- The advantage is to preserve the variability of the data
- Accurate results could be obtained with m as small as 3 or 5, but modern authors recommend to use more (Bodner, 2008)
- In practice, several datasets (replications) of imputed values are created. Statistical models are then computed independently on each dataset, and these intermediary results are combined into a final result
- Different imputation techniques can be used to generate the m replications, the only requirement being to be able to impute different values in each replication

MI estimator

- Let θ be a parameter to be estimated
- From each of the m replicated datasets, we obtain an estimation $\hat{\theta}_i$
- The MI estimator of θ is then

$$\hat{\theta}_{MI} = \frac{\sum_{i=1}^{m} \hat{\theta}_i}{m}$$

- The variance of the MI estimator is obtained as a combination of the variance of each $\hat{\theta}_i$ and the variance between the $\hat{\theta}_i$. If $\hat{V}_i$ is the variance of $\hat{\theta}_i$, then

$$\hat{V}_{\hat{\theta}_{MI}} = \frac{\sum_{i=1}^{m} \hat{V}_i}{m} + \left(1 + \frac{1}{m}\right) \frac{1}{m-1} \sum_{i=1}^{m} (\hat{\theta}_i - \hat{\theta}_{MI})^2$$

Chained equations (1)

- Chained equations (aka Fully Conditional Specification) is an imputation principle due, among others, to Van Buuren, Boshuizen & Knook (1999) :
  1. Regression models are defined to explain each variable with missing values
  2. Missing values are first replaced by random values
  3. Each regression model is then used in turn to impute missing values
  4. The algorithm iterates several times through all regression models, missing values being each time replaced by the value imputed during the previous iteration
  5. Imputations of the last iteration are replaced by the closest values really observed in the dataset
- Repeating the whole process *m* times leads to *m* different imputed datasets

Chained equations (2)

- Chained equations are available in the R package *mice* (Van Buuren & Groothuis-Oudshoorn, 2011).
- This method was also implemented in Stata under the name *ice* and was then integrated as a standard component of Stata.

Advantages

- Missing data on different variables can be imputed simultaneously
- Independent variables in the regression models can also have missing data
- Different regression models (linear, logistic, multinomial, ...) can be used simultaneously for different kind of variables (continuous, dichotomous, multinomial, ...)
- By default, all variables are used in all regression models, but it is also possible to specify a particular model for each variable with missing data
- The order of imputation of the different variables can be chosen

Outline

1 WHAT ARE MISSING DATA ?
- Examples from SHP
- Consequences of missing data
- Classification of missing data

2 HOW TO TREAT MISSING DATA ?
- Basic notions
- Simple imputation
- Multiple imputation
- **Some good questions about imputation**

3 LONGITUDINAL DATA, CAUSALITY, & ETHICS
- Specificities of longitudinal data
- Experiments
- Missing data and ethics

"Exact imputation"

- Sometimes, the true value of a MD can be found, for instance by matching the data with a different datafile
- When MD were caused by logical skips, the true value can also sometimes be found
- In such cases, it is of course beneficial to replace the MD with its true value
- This is not a real imputation, and there are no drawbacks

Logical skip

- Logical skips are very specific MD, because they are intentional
- Should we impute these MD ?
- It depends ! !
    - If we know the true value (e.g. number of children) $\Longrightarrow$ IMPUTE
    - If we used a short version of the questionnaire $\Longrightarrow$ POSSIBLE TO IMPUTE
    - Otherwise $\Longrightarrow$ NO REASON TO IMPUTE

Multi-item scales

- Average of the available items
  $\Longrightarrow$ Often used, but theoretical properties unknown
- Two possibilities for imputation :
  - total score
  - individual items
- More accurate results are obtained when imputing the items rather than the total score (Eekhout et al., 2014)
- Especially true when the number of missing data is high

Which values can be used as imputation values ?

- Some imputation methods can produce imputation values that were not observed in the sample, or that are not possible at all
  - *never observed income value*
  - *non-integer or negative number of doctor visits*
- If we want to prevent such values, we can replace the imputation value by the closest observed (or possible) value or category

Separate or simultaneous imputation

- When several variables have missing data, the trend now is clearly to consider all these variables simultaneously during the imputation step
- Chained equations is an example of algorithm able to treat all MD in one step
- "Simultaneously" refers to the fact that at the end of the procedure, all MD are inputed. The process itself is more of the iterative kind
- True simultaneous imputation could also be used

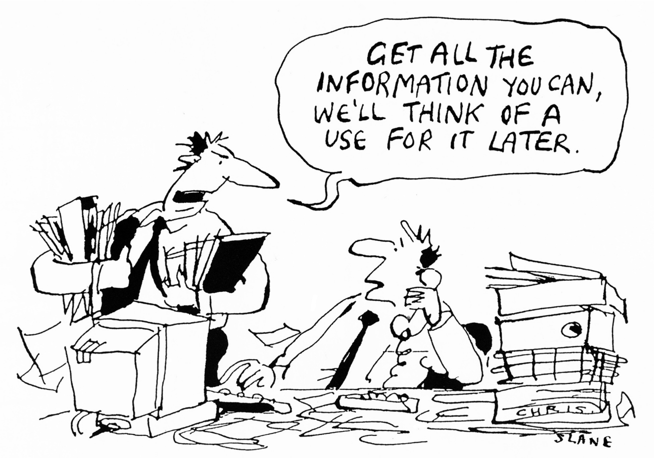
Independent and dependent variables

- All MD can be imputed, whether the variable will be used as dependent or independent in statistical analyses
- However, some authors suggest that, after imputation, cases with MD on the dependent variable should not be used in the statistical model (e.g. von Hippel, 2007)
- The argument is that in the case of MAR, the MD of the dependent variable Y do not provide information on the regression of independent variable on Y
- OK, but
  - von Hippel considered only the case of multiple imputation, not simple imputation
  - Maybe not true when MD are not MAR (and certainly not true in the case of MNAR)

Meaning of variables

- Which variables can be imputed ?
$\Longrightarrow$ *socio-demographic, income, health, sport practice, psychological behavior, ...*
- Theoretically, all variables can be imputed
- BUT not to impute is better than a wrong imputation
$\Longrightarrow$ impute only when a good imputation model do exist

Which variables in the imputation model ? (1)


GET ALL THE INFORMATION YOU CAN, WE'LL THINK OF A USE FOR IT LATER.
CHRIS
SLANE

Which variables in the imputation model ? (2)

- Most current advice : use all available variables, and at least all variables that will be used in the statistical model used to analyze the data
- But a variable unrelated with the variable to impute is useless ...
- Better to select variables in function of their predictive power regarding the variable to impute
- WARNING : longitudinal data are a special case

Which variables in the imputation model ? (3)

- Important to distinguish between
  - variables explaining the presence of missing data
  - variables explaining the (observed) values

$$\begin{pmatrix} \text{Swiss} & 20 \\ \text{Swiss} & 50 \\ \text{Swiss} & 20 \\ \text{Swiss} & 50 \\ \text{German} & 20 \\ \text{German} & 50 \\ \text{German} & 20 \\ \text{German} & 50 \end{pmatrix} \qquad \begin{pmatrix} \text{Swiss} & . \\ \text{Swiss} & . \\ \text{Swiss} & 20 \\ \text{Swiss} & 50 \\ \text{German} & 20 \\ \text{German} & 50 \\ \text{German} & 20 \\ \text{German} & 50 \end{pmatrix}$$

Which variables in the imputation model ? (4)

- Based on observed values, both variables are independent
- Based on missingness, nationality is a strong predictor of MD on age

	20	50
Swiss	50%	50%
German	50%	50%

	observed	*missing*
Swiss	50%	50%
German	100%	0%

Use of a MAR model for MCAR imputation

- Even if MD are MCAR, imputation values should be compatible with data observed on other variables
- Therefore, it is better to use a strong imputation model, similarly to MAR missing data
- Problem/question : Why is it important/useful to determine the type of MD, if in all cases we use an imputation model ?
- Ideas ?

Models for MNAR

- Under MNAR, the probability of missing depends on the missing values
- Therefore, it is necessary to model jointly the variable with missing values and the missingness process
- Two classical approaches (e.g. Enders, 2011) :
  1. selection models
  2. pattern mixture models
- Recent works suggest that MI could also be applicable (Galimard et al., 2016)
- Depends on very strict hypotheses
- Rarely used in practice
- Remember that MNAR is not really testable ...

Sensitivity analysis

- A sensitivity analysis should always be performed to evaluate the treatment aplied to missing data
- The idea is to evaluate the variability of final results in function of the treatments
- For instance, different imputation models can be used, and results compared, or different runs of the same imputation technique can be compared
- Possible problem : very different results achieved by different MD treatments ...

Outline

1. WHAT ARE MISSING DATA ?
   - Examples from SHP
   - Consequences of missing data
   - Classification of missing data
2. HOW TO TREAT MISSING DATA ?
   - Basic notions
   - Simple imputation
   - Multiple imputation
   - Some good questions about imputation
3. LONGITUDINAL DATA, CAUSALITY, & ETHICS
   - Specificities of longitudinal data
   - Experiments
   - Missing data and ethics

Time ordering

- Dependence between variables of different waves :
  - same variable observed through time
  - different variables
- Specific time order between variables
- One of the conditions to demonstrate causality

Causality

- Let A and B be to variables and suppose that A is the cause of B

$$A \overset{??}{\Longrightarrow} B$$

- To demonstrate this relationship, we must at least :
  1. Show that A and B are correlated
  2. Exclude all other possible causes of the observed relation between A and B
  3. Check that the cause, A, occured **before** the consequence, B

Specific imputation methods

- Last Observation Carried Forward (LOCF)
- Average of previous and next observations
- Linear inerpolation
- Regression on previous observations of the same variable
- ...
- More generally, we should exploit the correlation between waves to improve the quality of imputations

The "use all variables" advice

- Current advice : all variables (highly) correlated with the variable with MD should be incorporated into the imputation model
- In the longitudinal case, it is quite obvious that using variables from posterior waves (in addition to previous waves) will improve the imputation quality
- OK ... but what about causality ?
- The current trend in social sciences research is to collect longitudinal data, one of the final objective being to put into evidence causal relationships between events
- What could be the impact of imputation on causality if imputed data do not respect the temporal order ?

Outline

1 WHAT ARE MISSING DATA ?
- Examples from SHP
- Consequences of missing data
- Classification of missing data

2 HOW TO TREAT MISSING DATA ?
- Basic notions
- Simple imputation
- Multiple imputation
- Some good questions about imputation

3 LONGITUDINAL DATA, CAUSALITY, & ETHICS
- Specificities of longitudinal data
- **Experiments**
- Missing data and ethics

Imputation in the TREE data (1)

- Example taken from Berchtold & Surís (2017)
- We consider a sample of n=1999 subjects from the Transition from Education to Employment (TREE) cohort
- Seven waves from 2001 (T1) to 2007 (T7)
- Our variable of interest is smoking tobacco, with 5 modalities (from never to daily)
- The objective is to estimate a multinomial regression for smoking at T7
- Explanatory variables : smoking at T1, ..., T6
- Results are reported as Nagelkerke's R^2
- For the original data without missing, R^2=0.4935

Imputation in the TREE data (2)

- About 10% of missing data (MAR) were randomly generated on each of the seven variables
- Different multiple imputation procedures based on chained equations were used to impute the missing data
- Each time, the regression model for smoking at T7 was estimated
- The whole experiment was replicated 50 times with 50 different sets of missing values
- We also considered the case of 20% of missing data on each variable

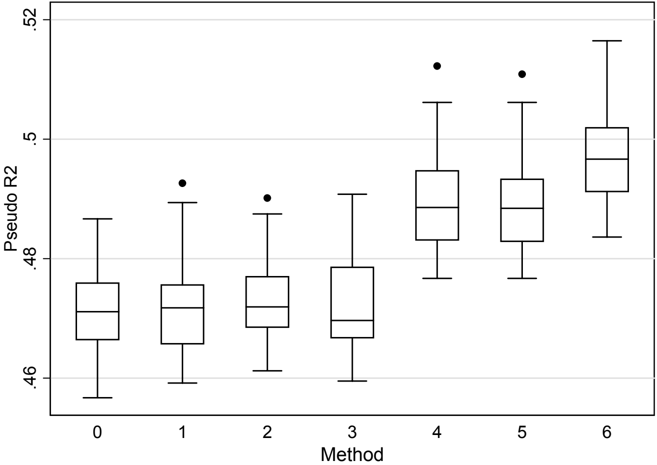
Imputation in the TREE data (3)

Wave	1	2	3	4	5	6	7
Subject A	O	O	O	O	O	O	O
Subject B	O	O	O	O	O	.	O
Subject C	O	O	.	O	.	.	.
Subject D	.	O	O	O	O	O	O

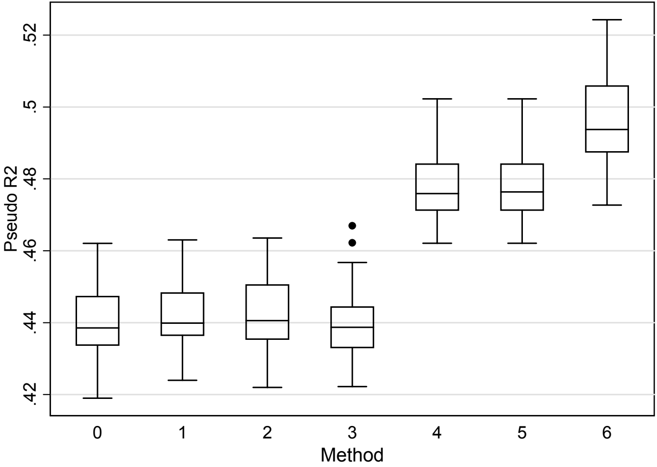
Imputation in the TREE data (4)

- Imputation models :
  0 Respect of temporality ; 6 covariates (age, gender, linguistic region, birth country, family wealth, mandatory school track)
  1 Same as 0, without age, gender, linguistic region
  2 Same as 0, without birth country, family wealth, mandatory school track
  3 Same as 0, with 4 additional covariates (reading level, family structure, index of cultural possessions, index of educative support provided by the family)
  4 Same as 0, but wave t+1 is also use to impute t ; no imputation of T1
  5 Same as 0, but wave t+1 is also use to impute t ; with imputation of T1
  5 Same as 0, but all waves are used to impute any other wave

Results with 10% of missing data



Results with 20% of missing data



What have we learned ?

- To preserve the relationships between data, we should respect the design of the study
- If data were collected in a specific order, then imputation should preserve this order
- On the other hand, more accurate imputed values can be obtained by using more information
- Remember that when using information, we should interèret results at the aggregated level only, not at the individual level

What imputation should do and not do

- Imputation should be used to complete datasets in the case of missing data
- Imputation should lead to more accurate results

BUT

- Imputation should not change the relations between variables
- Imputation should not dictate conclusions

Outline

1. WHAT ARE MISSING DATA ?
   - Examples from SHP
   - Consequences of missing data
   - Classification of missing data
2. HOW TO TREAT MISSING DATA ?
   - Basic notions
   - Simple imputation
   - Multiple imputation
   - Some good questions about imputation
3. **LONGITUDINAL DATA, CAUSALITY, & ETHICS**
   - Specificities of longitudinal data
   - Experiments
   - **Missing data and ethics**

Why speaking of ethics ?

- Missing data, and imputation in particular, have a clear relationship with ethics :
  - MD can be viewed as a missing part of the real phenomenon under study
  - Depending on the treatment method, MD can lead to biased results and incorrect conclusions
  - Imputation = "making up" data ?
- However, very few authors did seriously consider the relationship between MD and ethics
  $\Longrightarrow$ Enders & Gottschall (2011)

Ethics and data collection

- Remember : The best treatment method for MD is not having MD !
- Data collection step is essential, but where is the limit ?
  - Are incentive ethical ?
  - Is it ethical to ask many time for an answer (by phone, mail, email, ...) ?
  - ...
- On the other hand, is it ethical to give up a study, or to modify the design or the hypotheses, because (good) data cannot be obtained ?

Ethics and data analysis

- MD cannot just be ignored, even if there are few, because there is no threshold for "safe" MD
- The main issue is maybe not the number of missing data, but the MD mechanism
- First identify the mechanism, then use an adequate treatment
- No "all purposes ready-made solution" !
- There is maybe not a perfect solution, but there are many bad ones ! !

Ethics and imputation

- One often heard criticism against imputation : this is making up data
- Not so simple ; it strongly depends on the imputation model :
  - simple imputation tends to produce bias and to underestimate variance
  - multiple imputation (and likelihood based methods) do not have these issues
  - imputed values must not be analyzed at the individual level
  - imputed values are only a mean for computing accurate population level parameters

Ethics and results reporting

- At least three aspects of MD should be reported :
  1. The presence of MD in a study must always be acknowledged
  2. All treatments applied to MD must be described **and justified ! ! !**
  3. The impact of MD and of treatments on final results must be discussed
- Trade off between available space, readership, ...

Last words

- **“Are there three kinds of lies : lies, damned lies, and imputation ?”**
- Missing data are a problem for all statistical analyses
- A correct handling of missing data leads to an increase in the number of usable observations and in more accurate results
- Multiple imputation is now a standard way to handle missing data
- **Imputation is a process of data creation and this process must be strictly controlled and understood**