MISSING DATA: AN OVERVIEW OF MODERN METHODS

Serguei Rouzinov
HSE, Moscow, 27.12.2016
Education

- Bachelor of Science in Economics (Université de Genève)
- Master of Science in Statistics (Université de Neuchâtel)
- PhD student (Université de Lausanne) in Mathematics Applied to Human and Social Sciences
- Thesis Advisor: Professor André Berchtold
- Mail: Serguei.Rouzinov@unil.ch
OVERVIEW

- Introduction
- Classification of missing data (MD)
- Testing the MD mechanism
- Treatments of MD
- Bibliography
GUIDELINE

Introduction

Classification of MD

Testing the MD mechanism

Treatments of MD

Bibliography
INTRODUCTION

- **Definition of MD**
  - Missing data are data whose collect was planned, but not realized

- **Examples**: longitudinal and transversal cases

<table>
<thead>
<tr>
<th>obs/var</th>
<th>( V_{i1} )</th>
<th>( V_{i2} )</th>
<th>( V_{i3} )</th>
<th>( V_{i4} )</th>
<th>obs/var</th>
<th>( V_{1t} )</th>
<th>( V_{2t} )</th>
<th>( V_{3t} )</th>
<th>( V_{4t} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>1</td>
<td>x</td>
<td>.</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>2</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>.</td>
<td>2</td>
<td>.</td>
<td>x</td>
<td>.</td>
<td>x</td>
</tr>
<tr>
<td>3</td>
<td>x</td>
<td>x</td>
<td>.</td>
<td>.</td>
<td>3</td>
<td>x</td>
<td>x</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>4</td>
<td>x</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>4</td>
<td>.</td>
<td>x</td>
<td>x</td>
<td>.</td>
</tr>
</tbody>
</table>

**Table** – Monotone and arbitrary pattern
Why handle MD?
- Example of the Russian median salary

How to handle MD?
- Test the MD mechanism

Apply a method
1. listwise and pairwise deletion
2. single and multiple imputation
3. others methods
GUIDELINE

Introduction

Classification of MD

Testing the MD mechanism

Treatments of MD

Bibliography
CLASSIFICATION OF MD

Classification of Rubin (1972)

- Missing completely at random (MCAR)
  \[ Pr(Y_{mis}|Y, X) = Pr(Y_{mis}) \]

- Missing at random (MAR)
  \[ Pr(Y_{mis}|X, Y) = Pr(Y_{mis}|X, Y_{obs}) \]

- Missing not at random (MNAR)
  \[ Pr(Y_{mis}|X, Y) = Pr(Y) \]

Huge impact on the method to be used for handling missing data

Mixture of mechanisms
GUIDELINE

Introduction

Classification of MD

Testing the MD mechanism

Treatments of MD

Bibliography
TESTING THE MD MECHANISM

- Difficult to test something that does not exist
- First step: determine if MD is MCAR or not
- Easiest: sample groups difference by using a \( t\)-test
  - problematic when there are more than 2 variables
- Little (1988)
- Jamshidian & Yuan (2014)
TESTING THE MD MECHANISM

- Limits
  - Developed for transversal cases
  - Not applicable to some type of data (non-numerical)
  - Not robust to a change of distributions
  - Not applicable to MNAR
GUIDELINE

Introduction

Classification of MD

Testing the MD mechanism

Treatments of MD

Bibliography
Deletion methods

- Listwise deletion or complete-case analysis
  1. Definition
  2. Only for MCAR
  3. MCAR is unrealistic in longitudinal studies (Kromrey & Hines, 1994)

- Pairwise deletion or available-case analysis
  1. Definition
  2. Generally more efficient than listwise for MCAR
  3. Relatively big bias for MAR and MNAR
  4. Not well implemented in usual software
  5. Variance-covariance matrix could be not positive definite in relatively small samples (Wothke, 1993)
  6. Not efficient in longitudinal contexts (Fitzmaurice et al., 2009)
TREATMENTS OF MD (2)

- **Imputation methods**
  - **Simple imputation**
    1. Imputation by the mean, median, mode, (Haitovsky, 1968)
    2. Hot deck and cold deck imputations
    3. Regression imputation
Simple regression imputation

1. \( Y = aX_1 + bX_2 + cX_3 \)
2. \( Y = [Y_{obs}, Y_{mis}] \) is the dependent variable \( X_1, X_2, X_3 \) are the independent variables
3. \( a, b, c \) are coefficients
4. biased coefficients and standard errors (Allison, 2010)
5. use of random effect in order to increase the variance of the data
Imputation methods

- Multiple imputation
  1. One of the most efficient technique for dealing with MD
  2. Impute several times each MD
  3. Better reproduce the true but unknown variability of the unobserved values
  4. For instance by using several single imputation...
  5. ... or by using multiple imputation by chained equations (for all types of data, Van Buuren et. al, 2000 ; White, et al., 2011)
Thank you for your attention!

Any questions?!
GUIDELINE

Introduction

Classification of MD

Testing the MD mechanism

Treatments of MD

Bibliography
BIBLIOGRAPHY


