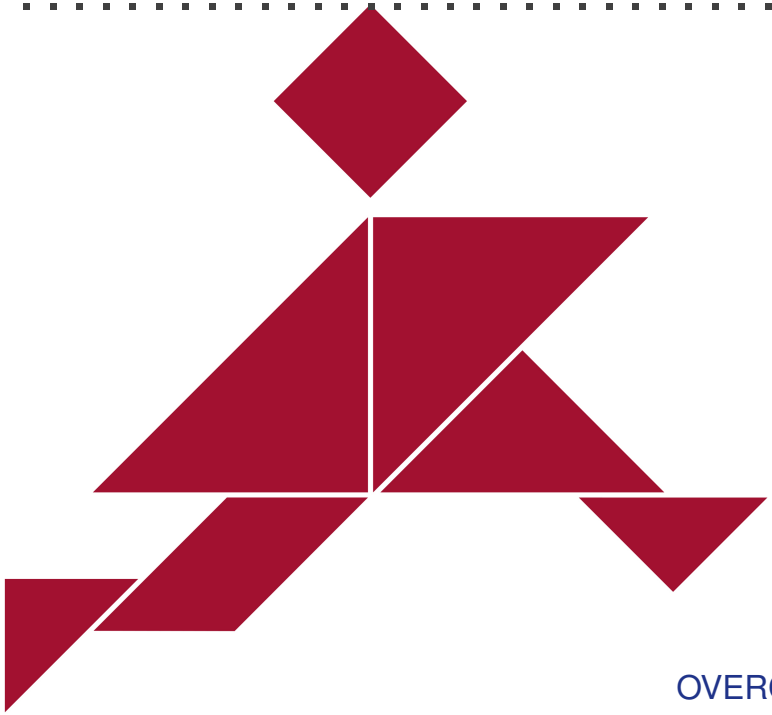
FNSNF Swiss National Science Foundation

The National Centers of Competence in Research (NCCR)
are a research instrument
of the Swiss National Science Foundation

Unil UNIL | Université de Lausanne

UNIVERSITÉ DE GENÈVE

MISSING DATA: AN OVERVIEW OF MODERN METHODS



Serguei Rouzinov

HSE, Moscow, 27.12.2016

LIVES
Swiss National Centre of Competence in Research

OVERCOMING VULNERABILITY: LIFE COURSE PERSPECTIVE

- 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


MISSING DATA
MISSING DATA EVERYWHERE!
nerator.net

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

obs/var	V_{i1}	V_{i2}	V_{i3}	V_{i4}	obs/var	V_{1t_j}	V_{2t_j}	V_{3t_j}	V_{4t_j}
1	x	x	x	x	1	x	.	x	x
2	x	x	x	.	2	.	x	.	x
3	x	x	.	.	3	x	x	.	.
4	x	.	.	.	4	.	x	x	.

TABLE – Monotone and arbitrary pattern

INTRODUCTION(2)

- 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

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

TREATMENTS OF MD

- 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 effcient 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. $Y = aX1 + bX2 + cX3$
    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

TREATMENTS OF MD (3)

- 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

- Allison P. D., *Missing Data*. Sage University Papers Series on Quantitative Applications in Social Sciences, 07-136. Thousand Oaks, CA : Sage ; 2010

- Berchtold A., and Surís J.-C. Imputation of Repeatedly-observed Multinomial Variables in Longitudinal Surveys. *Communications in Statistics - Simulation and Computation*. DOI : 10.1080/03610918.2015.1082588 ; 2015

- Fitzmaurice G., Davidian M., Verbeke G., and Molenberghs G. *Longitudinal Data Analysis*. CRC Press ; 2009

BIBLIOGRAPHY(2)

- Glasser M. Linear Regression Analysis with Missing Observations among the Independent Variables. *Journal of the American Statistical Association* Volume 59, Issue 307, 1964

- Jamshidian M., and Yuan K.-H. Examining Missing Data mechanisms via homogeneity of parameters, homogeneity of distributions, and multivariate normality. *WIREs Comput Stat* 2014, 6 :56-73. Doi :10.1002/wics.1287

BIBLIOGRAPHY(3)

- Kromrey J. D., and Hines C. V. Nonrandomly Missing Data in Multiple Regression : An Empirical Comparison of Common Missing-Data Treatments. *Educational and Psychological Measurement* vol. 54 no. 3 (1994) 573-593

- Little R. J. A. Test of Missing Completely at Random for Multivariate Data with Missing Values. *Journal of the American Statistical Association Vol*. 83, No. 404 (1988), pp. 1198-1202.

BIBLIOGRAPHY(4)

- Van Buuren S., Oudshoorn C. G. M. Multivariate Imputation by Chained Equations : MICE V1.0 User's manual. *TNO Report PG/VGZ/00.038*. TNO Preventie en Gezondheid : Leiden, 2000. Available from : http ://www.multiple-imputation.com/

- White I. R., Royston P., and Wood A. M. Multiple imputation using chained equations : Issues and guidance for practice. *Statistics in Medicine* Volume 30, Issue 4, pages 377–399, 20 February 2011

BIBLIOGRAPHY(5)

- Wothke W. Nonpositive definite matrices in structural modeling. In K.A. Bollen & J.S. Long (Eds.). *Testing structural equation models* (pp. 256-293), Newbury Park, CA : Sage ; 1993