R PACKAGES TO STUDY MISSING AND LONGITUDINAL DATA

Susana Rafaela Guimarães Martins

10th October 2019

VI XORNADA DE USUARIOS DE R EN GALICIA

Overview

1. Introduction

2. Motivation

3. Methodology

4. Results

5. Conclusions

Overview

1. Introduction
2. Motivation
3. Methodology
4. Results
5. Conclusions

Introduction

- The application of statistics in the most diverse areas is a reality that we face on a daily basis. Data collection can be single or regular, which implies studies of different natures. If data is collected regularly, we can have a longitudinal study.
- Longitudinal studies are of great interest to medicine, sports and other areas, and they often present a problem that conditions them - the lack of data. The cause for this lack may be diverse, and it can cause constraints on the conclusions drawn.
- Nowadays, some techniques that allow the existence of missing data to be filled have been explored. However, the study of missing data in longitudinal databases is not yet explored. This way, the aim of this work is to state the progress of computational tools, namely the R-software.

Introduction

- The Department of Human Motricity of Instituto Politécnico de Viana do Castelo has started a study about morphofunctional study of the children from Viana do Castelo. This study supports my PhD work.
- The main objective of that study was to normatively characterize the variables of morphological growth and physical fitness throughout development in the county's juvenile population, to evaluate the adequacy of the profiles displayed by children and young people from Viana do Castelo according to health prevention criteria.
- This study had a longitudinal data and it had missing values.

Introduction

- One of the objectives of my study is to estimate the transition probabilities between different IOTF categories, as well as to understand their relationship with other longitudinally recorded variables, especially the physical variables.
- The IOTF is an index of cataloging individuals relative to weight. There are several categories: thinness, normal, overweight and obesity.
- The principal problem of data are missing values that in the most of the cases are "missing individual".

Overview

1. Introduction

2. Motivation

3. Methodology

4. Results

5. Conclusions

Motivation

- Data collection took place annually between 1997 and 2000 inclusive, and it was collected again in 2006. This is equivalent to 6, 7, 8, 9 and 15 years old.
- The data collected refers essentially to two types of individuals' characteristics: morphological characteristics and physical fitness characteristics.

Variables to morphological characteristics

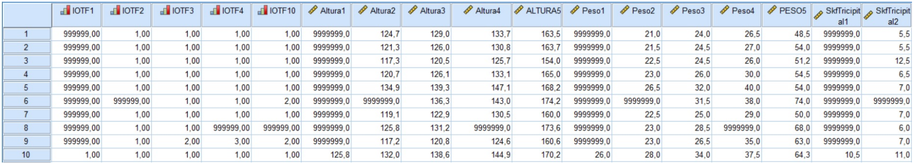
Fat percentage, body mass index, International obesity task force and sum of triceps adipose fold and subscapular adipose fold (Sum_SKF),

Variables to physical fitness characteristics

Maximum bar suspension time, sit and reach, 4x 10 meter agility run, long jump without preparatory run, 60-second sit-ups, 50 meter speed race and 20 meter reciprocating endurance race.

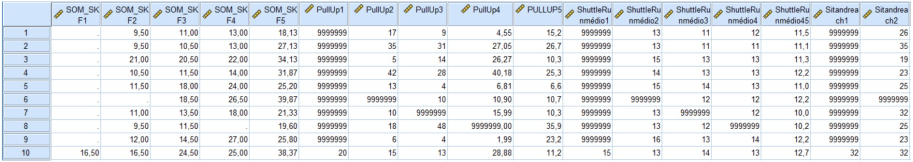
Motivation

The principal problem of data are missing values. In the most of the cases the missing values are missing on moment to all variables of individual.



	IOTF1	IOTF2	IOTF3	IOTF4	IOTF10	Altura1	Altura2	Altura3	Altura4	ALTURA5	Peso1	Peso2	Peso3	Peso4	PESO5	SkfTricipital1	SkfTricipital2
1	999999,00	1,00	1,00	1,00	1,00	9999999,0	124,7	129,0	133,7	163,5	9999999,0	21,0	24,0	26,5	48,5	9999999,0	5,5
2	999999,00	1,00	1,00	1,00	1,00	9999999,0	121,3	126,0	130,8	163,7	9999999,0	21,5	24,5	27,0	54,0	9999999,0	5,5
3	999999,00	1,00	1,00	1,00	1,00	9999999,0	117,3	120,5	125,7	154,0	9999999,0	22,5	24,5	26,0	51,2	9999999,0	12,5
4	999999,00	1,00	1,00	1,00	1,00	9999999,0	120,7	126,1	133,1	165,0	9999999,0	23,0	26,0	30,0	54,5	9999999,0	6,5
5	999999,00	1,00	1,00	1,00	1,00	9999999,0	134,9	139,3	147,1	168,2	9999999,0	26,5	32,0	40,0	54,0	9999999,0	7,0
6	999999,00	999999,00	1,00	1,00	2,00	9999999,0	9999999,0	136,3	143,0	174,2	9999999,0	9999999,0	31,5	38,0	74,0	9999999,0	9999999,0
7	999999,00	1,00	1,00	1,00	1,00	9999999,0	119,1	122,9	130,5	160,0	9999999,0	22,5	25,0	29,0	50,0	9999999,0	7,0
8	999999,00	1,00	1,00	999999,00	999999,00	9999999,0	125,8	131,2	9999999,0	173,6	9999999,0	23,0	28,5	9999999,0	68,0	9999999,0	6,0
9	999999,00	1,00	2,00	3,00	2,00	9999999,0	117,2	120,8	124,6	160,6	9999999,0	23,0	26,5	35,0	63,0	9999999,0	7,0
10	1,00	1,00	1,00	1,00	1,00	125,8	132,0	138,6	144,9	170,2	26,0	28,0	34,0	37,5	64,3	10,5	11,0

Figure: Database image



	SOM_SKF1	SOM_SKF2	SOM_SKF3	SOM_SKF4	SOM_SKF5	PullUp1	PullUp2	PullUp3	PullUp4	PULLUP5	ShuttleRunmédio1	ShuttleRunmédio2	ShuttleRunmédio3	ShuttleRunmédio4	ShuttleRunmédio45	Sitandreach1	Sitandreach2
1	.	9,50	11,00	13,00	18,13	9999999	17	9	4,55	15,2	9999999	13	11	12	11,5	9999999	26
2	.	9,50	10,50	13,00	27,13	9999999	35	31	27,05	26,7	9999999	13	11	11	11,1	9999999	35
3	.	21,00	20,50	22,00	34,13	9999999	5	14	26,27	10,3	9999999	15	13	13	11,3	9999999	19
4	.	10,50	11,50	14,00	31,87	9999999	42	28	40,18	25,3	9999999	14	13	13	12,2	9999999	23
5	.	11,50	18,00	24,00	25,20	9999999	13	4	6,81	6,6	9999999	15	14	13	11,0	9999999	25
6	.	.	18,50	26,50	39,87	9999999	9999999	10	10,90	10,7	9999999	9999999	12	12	12,2	9999999	9999999
7	.	11,00	13,50	18,00	21,33	9999999	10	9999999	15,99	10,3	9999999	13	9999999	12	10,0	9999999	32
8	.	9,50	11,50	.	19,60	9999999	18	48	9999999,00	35,9	9999999	13	12	9999999	10,2	9999999	25
9	.	12,00	14,50	27,00	25,80	9999999	6	4	1,99	23,2	9999999	16	13	14	12,2	9999999	23
10	16,50	16,50	24,50	25,00	38,37	20	15	13	28,88	11,2	15	13	14	13	12,7	32	32

Figure: Database image

Motivation

- 229 children was available.
  101 (44.1%) were female and 128 (55.9%) were male.

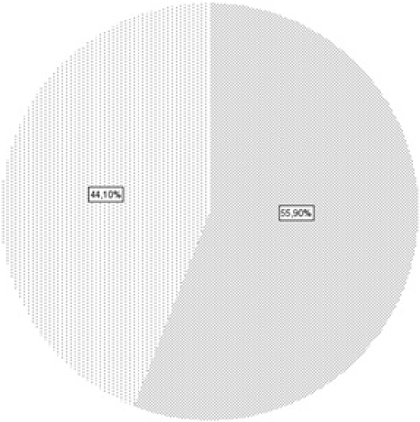


Figure: Children divided by gender

Motivation

Variable	6	7	8	9	15
Age	67	5	2	0	0
Fat	67	8	12	6	0
BMI	67	8	12	6	0
Sum_SK	67	8	12	6	0
IOTF	67	8	12	6	6
TSB	67	8	13	6	0
SR	67	9	13	6	0
SHR	67	8	13	7	0
SCP	67	9	13	6	0
ABD	67	8	13	6	1
C50	71	11	13	6	0
CVV	68	11	13	7	0

Table: Number of missing values of the girls over the years.

Motivation

Variable	6	7	8	9	15
Age	86	8	6	6	0
Fat	86	8	6	6	2
BMI	86	8	6	6	2
Sum_SK	86	8	8	6	2
IOTF	86	8	8	8	8
TSB	86	10	6	6	0
SR	86	9	6	6	0
SHR	86	11	7	6	0
SCP	86	12	6	6	0
ABD	86	10	6	6	1
C50	90	11	8	6	0
CVV	86	11	6	9	0

Table: Number of missing values of the boys over the years.

Overview

1. Introduction
2. Motivation
3. Methodology
4. Results
5. Conclusions

Methodology

- In order to have an overview of the methods of analysing missing values in longitudinal data, particularly in software resources, a global survey using the methodology traditionally known as the "Scoping Review" was made.
- For this, the reference base "Web of Science" and the keywords "R package missing longitudinal data" were used. We found 27 results of which only 11 were open access articles. From these articles we selected those that referred to at least one R package, leaving a total of 8 articles for analysis.
- For each article, a short summary highlighting the package used was elaborated, as well as its main functionality.

Overview

1. Introduction
2. Motivation
3. Methodology
4. Results
5. Conclusions

Results

Measuring the Impact of Nonignorable Missingness Using the R Package ISNI

The article presents a package that implements the ISNI - index of local sensitivity to nonignorability - methodology. This methodology assesses the dependence of inferences on the ignorability assumption by measuring its sensitivity to violation. The ISNI methodology has been little used due to the lack of software support for its application, which is now being addressed. The ISNI package implements the index of local sensitivity to nonignorability methodology and is a good tool for a systematic and efficient analysis of the reliability of empirical inferences that derive from incomplete data.

Results

Identifying patterns of item missing survey data using latent groups: an observational study

The article is based on a study on health and physical activity and its main objective is to understand if individuals can be grouped according to unanswered questions. The article does not explore any specific packages, but uses them to find results. The packages mentioned are MI and Rmixmod. MI is an imputation package that allows a global view of patterns in missing data. Rmixmod is a package that allows data modeling.

Results

Time-Course Gene Set Analysis for Longitudinal Gene Expression Data

The article introduces a new TcGSA package, which implements the method of the same name: Time-Course Gene Set Analysis. This package implements the extension of existing gene set analysis methods to longitudinal data.

Results

Accounting for Interactions and Complex Inter-Subject Dependency in Estimating Treatment Effect in Cluster-Randomized Trials with Missing Outcomes

The article focuses on the study of semiparametric methods for estimating the effects of randomly correlated clinical trial results. When results are randomly absent, weighted inverse probability methods are used to incorporate missing covariates. However, there are often interactions between covariates and treatment and there is no method that, alone, can produce consistent estimates. In this article, double robust estimator, DR, is proposed, which allows to provide correct estimates for missing data as well as allows a correctly specified result model. This estimator is implemented in the CRTgeeDR package.

Results

CRTgeeDR: an R Package for Doubly Robust Generalized Estimating Equations Estimations in Cluster Randomized Trials with Missing Data

The article presents the CRTgeeDR package, developed for cluster randomized clinical trials with missing data. This package has the advantage of addressing the failure of other software that generates biased estimates of the correlation structure when independence is not verified. When compared to an already existing geepack, this package has better results associated with the use of DR to analyse a binary result when using logistic regression.

Results

Joint models for predicting transplant-related mortality from quality of life data

The article aims to test whether health-related longitudinal measures related to life quality cause changes related to paediatric hematopoietic stem cell transplantation. In this article, where longitudinal data is analysed, we use the package JM, whose main use is modelling.

Results

kml and kml3d: R Packages to Cluster Longitudinal Data

The article is based on the presentation of two packages kml and kml3d. These packages provide an implementation of the k-means method and provide tools for longitudinal data analysis, such as path imputation methods, methods for defining k-means starting conditions, and quality criteria for choosing the best number of clusters. In addition, they allow to produce graphs on the trajectories of the variables, which can be unique or articulated.

Results

LMest: An R Package for Latent Markov Models for Longitudinal Categorical Data

The article is based on the presentation of the LMest package that allows the implementation of latent Markov models. These models are very useful in longitudinal data analysis, particularly when the response variables are categorical in nature. The presented package is appropriate to implement the traditional model, as well as some formulations to implement the model using covariates as well as missing data.

Overview

1 Introduction

2 Motivation

3 Methodology

4 Results

5 Conclusions

Conclusions

Package	Missing data	Longitudinal data
ISNI	X	–
MI	X	–
Rmixmod	–	–
TcGSA	–	X
CRTgeeDR	X	X
JM	–	X
Kml	X	X
kml3D	X	X
LMest	X	X

Table: Summary of presented packages.

- Are several packages available for missing data and longitudinal data studies.
- Only packages CRTgeeDR, Kml, kml3D and LMest are suitable for missing and longitudinal data.

Future Work

- Do research similar to this on other scientific bases.
- Do the experimentation of the identified packages and the verification of their suitability for the analysis of the data under study.
- Develop a guide for using existing packages.

Thank you for your attention!

References I

Adrian G Barnett, Paul McElwee, Andrea Nathan, Nicola W Burton, and Gavin Turrell.
Identifying patterns of item missing survey data using latent groups: an observational study.
BMJ Open, 7(10), 2017.

Boris P. Hejblum, Jason Skinner, and Rodolphe Thiébaut.
Time-course gene set analysis for longitudinal gene expression data.
PLOS Computational Biology, 11(6):1–21, 06 2015.

Christophe Genolini, Xavier Alacoque, Mariane Sentenac, and Catherine Arnaud.
kml and kml3d: R packages to cluster longitudinal data.
Journal of Statistical Software, Articles, 65(4):1–34, 2015.

References II

Francesco Bartolucci, Silvia Pandolfi, and Fulvia Pennoni.
Lmest: An r package for latent markov models for longitudinal categorical data.
Journal of Statistical Software, Articles, 81(4):1–38, 2017.

Hui Xie, Weihua Gao, Baodong Xing, Daniel F. Heitjan, Donald Hedeker, and Chengbo Yuan.
Measuring the impact of nonignorable missingness using the r package isni.
Computer Methods and Programs in Biomedicine, 164:207 – 220, 2018.

References III

Melanie Prague, Rui Wang, Alisa Stephens, Eric Tchetgen Tchetgen, and Victor DeGruttola.
Accounting for interactions and complex inter-subject dependency in estimating treatment effect in cluster-randomized trials with missing outcomes.
Biometrics, 72(4):1066–1077, 2016.

Melanie Prague, Rui Wang, and Victor De Gruttola.
CRTgeeDR: an R Package for Doubly Robust Generalized Estimating Equations Estimations in Cluster Randomized Trials with Missing Data.
The R Journal, 9(2):105–115, 2017.

References IV

Norma Terrin, Angie Mae Rodday, and Susan K. Parsons.
Joint models for predicting transplant-related mortality from quality of life data.
Quality of Life Research, 24(1):31–39, Jan 2015.

R PACKAGES TO STUDY MISSING AND LONGITUDINAL DATA

Susana Rafaela Guimarães Martins

10th October 2019

VI XORNADA DE USUARIOS DE R EN GALICIA