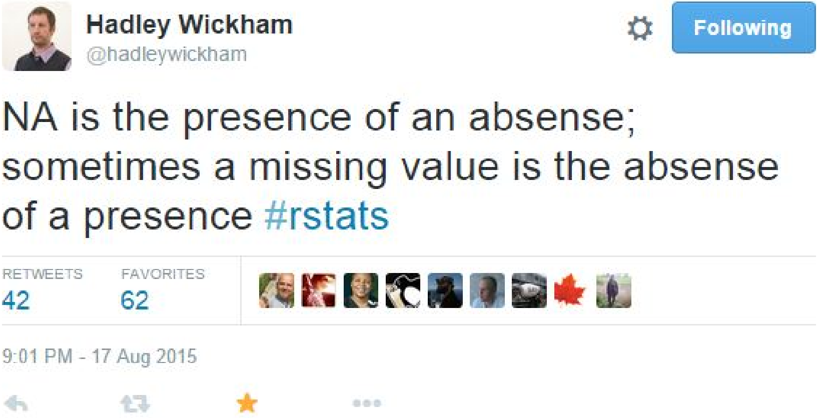
Missing Data

- Data sets often have missing values. Missing data is a problem, in particular, with multivariate modeling.
- If the analyst must discard an entire record because the value for one variable is missing, valuable information is lost.
- Open Question: Is it better to find a way keep the record, adjust for the missing value(s), and let the analysis proceed? (Some would say no)



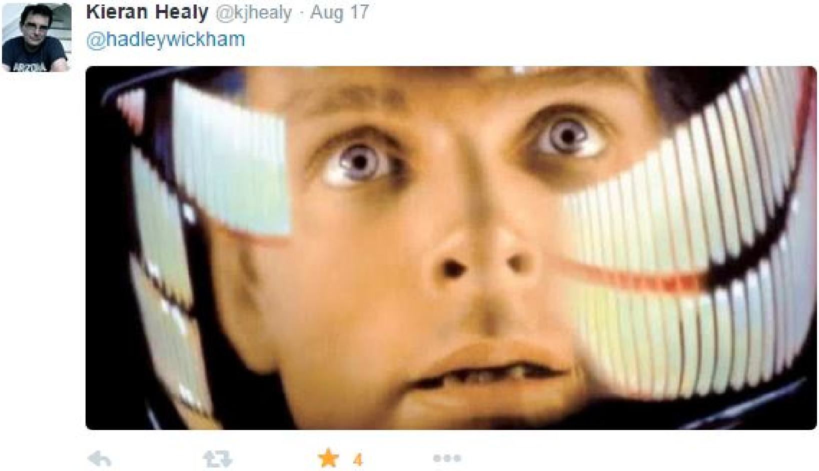
Hadley Wickham
@hadleywickham

NA is the presence of an absense; sometimes a missing value is the absense of a presence #rstats

RETWEETS 42

FAVORITES 62

9:01 PM - 17 Aug 2015



Kieran Healy @kjhealy · Aug 17

@hadleywickham

Missing Data

Introduction

- Missing data is a common problem in all kinds of research.
- The way you deal with it depends on how much data is missing, the kind of missing data (single items, a full questionnaire, a measurement wave), and why it is missing, i.e. the reasons that the data are missing.
- Handling missing data is an important step in several phases of a scientific study.

Dealing with Missing Data

- Missing values means reduced sample size and loss of data.
- The less data collected, the less data that can be analyzed, and reducing the data that can be analyzed reduces statistical power, which is the ability to detect real relationships in the data.
- Missing values may also indicate bias in the data. If the missing values are non-random, then the study is not accurately measuring the intended constructs.
- The results of your study may have been different if the missing data was not missing.

Missing Values in R

The NA Symbol

- In R, missing values are represented by the symbol NA (not available) .
- Impossible values (e.g., dividing by zero) are represented by the symbol NaN (not a number).

Excluding Missing Values from Analyses

Arithmetic functions on missing values yield missing values.

```
x <- c(1,2,NA,3)
mean(x) # returns NA
mean(x, na.rm=TRUE) # returns 2
```

- The function `complete.cases()` returns a logical vector indicating which cases are complete.

```
# list rows of data that
# have missing values
mydata[!complete.cases(mydata),]
```

- The function `na.omit()` returns the object with listwise deletion of missing values.

```
# create new dataset without missing data
newdata <- na.omit(mydata)
```

Why do missing values occur?

- Missing values are either **random** or **non-random**.
- Random missing values may occur because the subject inadvertently did not answer some questions.
- The study may be overly complex and/or long, or the subject may be tired and/or not paying attention, and miss the question.
- Random missing values may also occur through data entry mistakes.

Types of Missing Data

- Missing At Random
- Missing Completely At Random
- Missing Not At Random

Types of Missing Data

Missing Completely At Random

- There are several reasons why data may be missing.
- They may be missing because equipment malfunctioned, the weather was terrible, people got sick, or the data were not entered correctly.
- Here the data are **missing completely at random (MCAR)**.

Types of Missing Data

Missing Completely At Random

- When we say that data are missing completely at random, we mean that the probability that an observation (X_i) is missing is unrelated to the value of X_i or to the value of any other variables.
- Thus data on family income would not be considered MCAR if people with low incomes were less likely to report their family income than people with higher incomes.

Types of Missing Data

Missing At Random

- Often data are not missing completely at random, but they may be classifiable as **missing at random (MAR)**.
  (MAR is not really a good name for this condition because most people would take it to be synonymous with MCAR, which it is not. However, the name has stuck.)

Types of Missing Data

Missing At Random

- For data to be missing completely at random, the probability that X_i is missing is unrelated to the value of X_i or other variables in the analysis.
- But the data can be considered as missing at random if the data meet the requirement that missing-ness does not depend on the value of X_i after controlling for another variable.

Types of Missing Data

- MCAR : Completely at Random throughout the data
- MAR : Randomly Occuring within variables, but more likely to happen with some variables than other.

Types of Missing Data

Missing Not at Random

- If data are not MCAR or MAR then they are classed as Missing Not at Random (MNAR).
- MNAR data is data that is missing for a specific reason (ie. the value of the variable that's missing is related to the reason it's missing)

Types of Missing Data

Missing Not at Random

- When we have data that are MNAR we have a problem.
- The only way to obtain an unbiased estimate of parameters is to model missingness.
- In other words we would need to write a model that accounts for the missing data.
- That model could then be incorporated into a more complex model for estimating missing values.

- Non-random missing values may occur because the subject purposefully did not answer some questions.
- The question may be confusing, so many subjects do not answer the question.
- Also, the question may not provide appropriate answer choices, such as **no opinion** or **not applicable**, so the subject chooses not to answer the question.

Dealing with Missing Data

There are three approaches to dealing with missing data.

Option 1 Continue With the Incomplete Data

Option 2 Casewise deletion

Option 3 Imputation

Option 1 : Continue With the Incomplete Data

- The first option is to leave the data as is, with the missing values in place.
- This is the most frequent approach, for a few reasons. First, the number of missing values are typically small. Second, missing values are typically non-random.
- Third, even if there are a few missing values on individual items, you typically create composites of the items by averaging them together into one new variable, and this composite variable will not have missing values because it is an average of the existing data.

Option 2 : Case-Wise Deletion

- The next option is to delete cases with missing values.
- For every missing value in the dataset, you can delete the subjects with those missing values.
- Thus, you are left with complete data for all subjects.

Option 2 : Case-Wise Deletion

- The disadvantage to this approach is you reduce the sample size of your data (sometimes most of it).
- If you have a large dataset, then it may not be a big disadvantage because you have enough subjects even after you delete the cases with missing values.

Option 2 : Case-Wise Deletion

- Another disadvantage to this approach is that the subjects with missing values may be different than the subjects without missing values (i.e MNAR), so you have a non-representative sample after removing the cases with missing values.

Option 3 : Imputation

- The last option is to replace the missing values, called **imputation**.
- There is little agreement about whether or not to conduct **imputation**.
- There is some agreement, however, in which type of imputation to conduct.
- You typically do NOT conduct **Mean substitution** or **Regression substitution**.

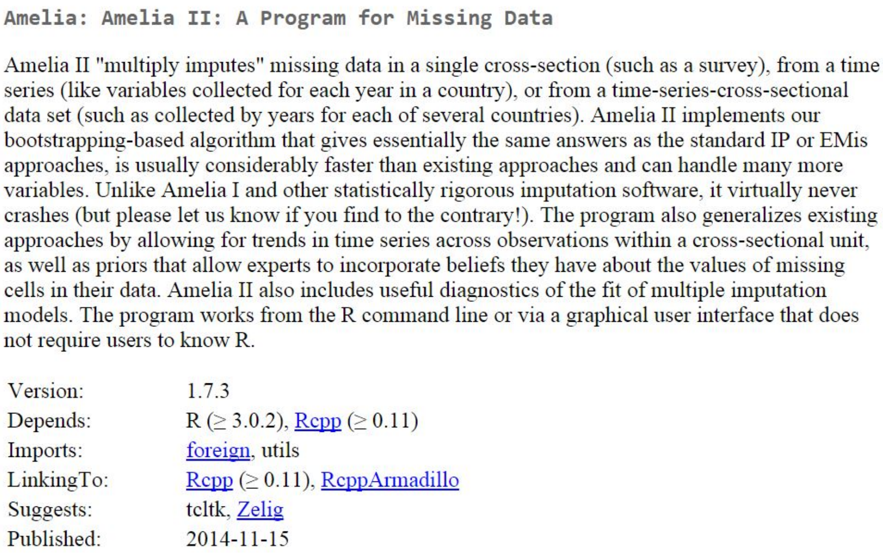
Option 3 : Imputation

- **Mean substitution** is replacing the missing value with the mean of the variable.
- **Regression substitution** uses regression analysis to replace the missing value. Regression analysis is designed to predict one variable based upon another variable, so it can be used to predict the missing value based upon the subjects answer to another variable.

The favored type of imputation is replacing the missing values using different estimation methods.

Some of these packages also have functions to explore patterns of missingness

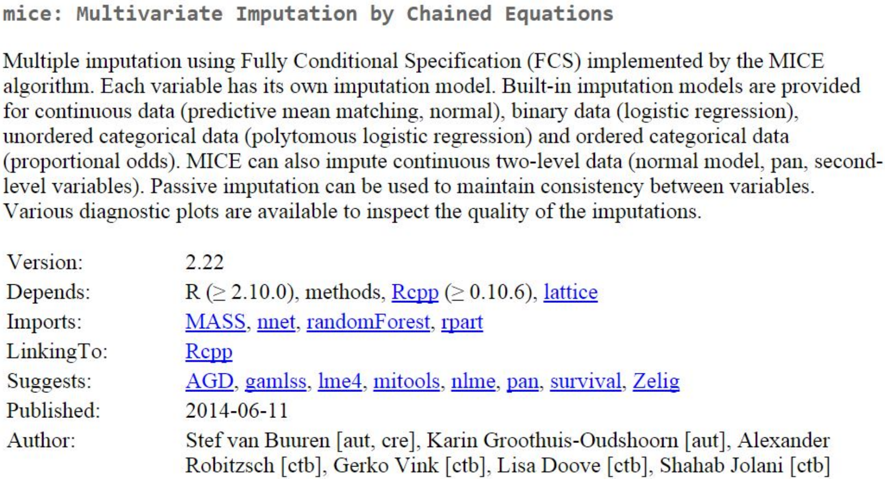
- **Amelia II:** A Program for Missing Data
- **Hmisc:** Harrell Miscellaneous
- **mi:** Missing Data Imputation and Model Checking
- **mitools:** Tools for multiple imputation of missing data



Amelia: Amelia II: A Program for Missing Data

Amelia II "multiply imputes" missing data in a single cross-section (such as a survey), from a time series (like variables collected for each year in a country), or from a time-series-cross-sectional data set (such as collected by years for each of several countries). Amelia II implements our bootstrapping-based algorithm that gives essentially the same answers as the standard IP or EMis approaches, is usually considerably faster than existing approaches and can handle many more variables. Unlike Amelia I and other statistically rigorous imputation software, it virtually never crashes (but please let us know if you find to the contrary!). The program also generalizes existing approaches by allowing for trends in time series across observations within a cross-sectional unit, as well as priors that allow experts to incorporate beliefs they have about the values of missing cells in their data. Amelia II also includes useful diagnostics of the fit of multiple imputation models. The program works from the R command line or via a graphical user interface that does not require users to know R.

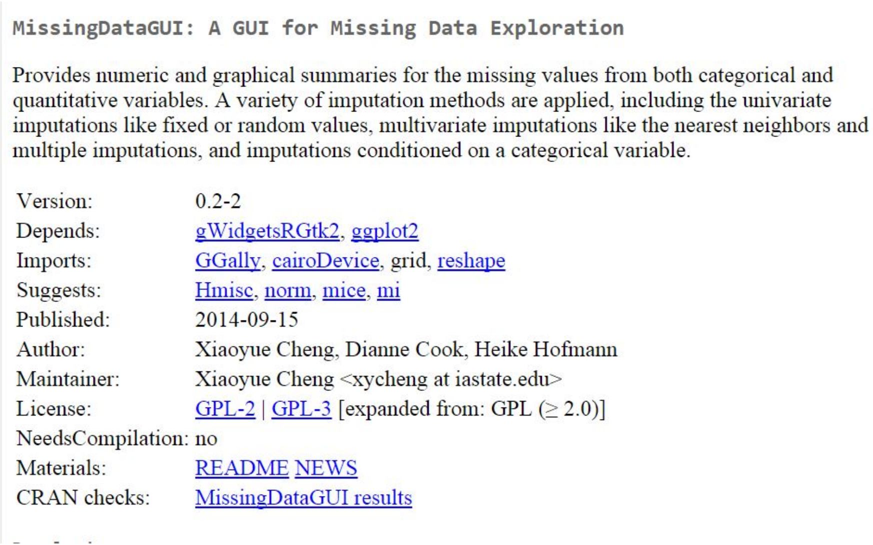
Version:	1.7.3
Depends:	R (≥ 3.0.2), Rcpp (≥ 0.11)
Imports:	foreign, utils
LinkingTo:	Rcpp (≥ 0.11), RcppArmadillo
Suggests:	tcltk, Zelig
Published:	2014-11-15



mice: Multivariate Imputation by Chained Equations

Multiple imputation using Fully Conditional Specification (FCS) implemented by the MICE algorithm. Each variable has its own imputation model. Built-in imputation models are provided for continuous data (predictive mean matching, normal), binary data (logistic regression), unordered categorical data (polytomous logistic regression) and ordered categorical data (proportional odds). MICE can also impute continuous two-level data (normal model, pan, second-level variables). Passive imputation can be used to maintain consistency between variables. Various diagnostic plots are available to inspect the quality of the imputations.

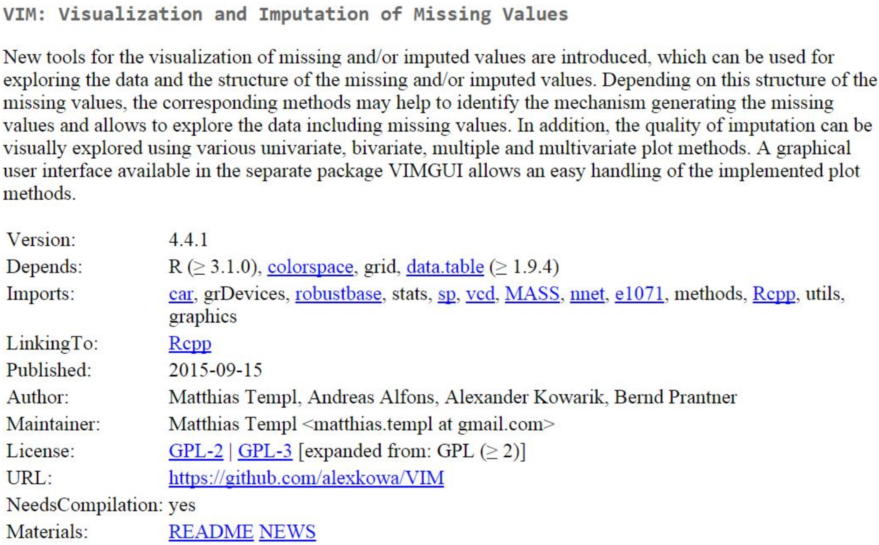
Version:	2.22
Depends:	R (≥ 2.10.0), methods, Rcpp (≥ 0.10.6), lattice
Imports:	MASS, nnet, randomForest, rpart
LinkingTo:	Rcpp
Suggests:	AGD, gamlss, lme4, mitools, nlme, pan, survival, Zelig
Published:	2014-06-11
Author:	Stef van Buuren [aut, cre], Karin Groothuis-Oudshoorn [aut], Alexander Robitzsch [ctb], Gerko Vink [ctb], Lisa Doove [ctb], Shahab Jolani [ctb]



MissingDataGUI: A GUI for Missing Data Exploration

Provides numeric and graphical summaries for the missing values from both categorical and quantitative variables. A variety of imputation methods are applied, including the univariate imputations like fixed or random values, multivariate imputations like the nearest neighbors and multiple imputations, and imputations conditioned on a categorical variable.

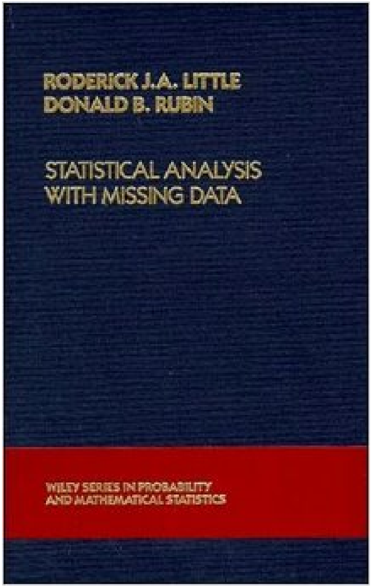
Version:	0.2-2
Depends:	gWidgetsRGtk2, ggplot2
Imports:	GGally, cairoDevice, grid, reshape
Suggests:	Hmisc, norm, mice, mi
Published:	2014-09-15
Author:	Xiaoyue Cheng, Dianne Cook, Heike Hofmann
Maintainer:	Xiaoyue Cheng <xycheng at iastate.edu>
License:	GPL-2 | GPL-3 [expanded from: GPL (≥ 2.0)]
NeedsCompilation:	no
Materials:	README NEWS
CRAN checks:	MissingDataGUI results



VIM: Visualization and Imputation of Missing Values

New tools for the visualization of missing and/or imputed values are introduced, which can be used for exploring the data and the structure of the missing and/or imputed values. Depending on this structure of the missing values, the corresponding methods may help to identify the mechanism generating the missing values and allows to explore the data including missing values. In addition, the quality of imputation can be visually explored using various univariate, bivariate, multiple and multivariate plot methods. A graphical user interface available in the separate package VIMGUI allows an easy handling of the implemented plot methods.

Version:	4.4.1
Depends:	R (≥ 3.1.0), colorspace, grid, data.table (≥ 1.9.4)
Imports:	car, grDevices, robustbase, stats, sp, vcd, MASS, nnet, e1071, methods, Rcpp, utils, graphics
LinkingTo:	Rcpp
Published:	2015-09-15
Author:	Matthias Templ, Andreas Alfons, Alexander Kowarik, Bernd Prantner
Maintainer:	Matthias Templ <matthias.templ at gmail.com>
License:	GPL-2 | GPL-3 [expanded from: GPL (≥ 2)]
URL:	https://github.com/alexkowa/VIM
NeedsCompilation:	yes
Materials:	README NEWS



RODERICK J.A. LITTLE
DONALD B. RUBIN

STATISTICAL ANALYSIS
WITH MISSING DATA

WILEY SERIES IN PROBABILITY
AND MATHEMATICAL STATISTICS

What is multiple imputation?

Imputation

- Imputation, the practice of '*filling in*' missing data with plausible values, is an attractive approach to analyzing incomplete data.
- The intention is to solve the missing-data problem at the beginning of the analysis.
- However, a naive or unprincipled imputation method may create more problems than it solves, distorting estimates, standard errors and hypothesis tests, as documented by Little and Rubin (1987) and others.

Imputation

- For **MNAR**, imputation is not sufficient, because the missing data are totally different from the available data, i.e. your complete data has become a selective group of persons.
- For **MCAR** and **MAR**, there are roughly two kinds of techniques for imputation; Single and Multiple Imputation.

Single Imputation

Single Imputation

- Single imputation techniques are based on the idea that in a random sample every person can be replaced by a new person, given that this new person is randomly chosen from the same source population as the original person.
- In that case you can use the observed available data of the other persons to make an estimation of the distribution of the test result in the source population.
- It is called **single imputation**, because each missing is imputed once.

What is multiple imputation?

- There are many methods for single imputation, such as
  - replacement by the mean,
  - regression,
  - expected maximization (EM).
- **Expected maximization** is preferred, because in the other methods the variance and standard error are reduced and the chance for Type II errors increases.

Multiple Imputation

- Multiple imputation is a simulation-based approach to the statistical analysis of incomplete data.
- In multiple imputation, each missing datum is replaced by $m > 1$ simulated values.
- The resulting m versions of the complete data can then be analyzed by standard complete-data methods, and the results combined to produce inferential statements (e.g. interval estimates or p-values) that incorporate missing-data uncertainty.

What is multiple imputation?

- The difference with single imputation is that in MI the value is imputed for several times. There are more imputed datasets created.
- The different imputations are then based on random draws of different estimations of the underlying distribution in the source population.
- In this way, the imputed data comes from different distributions and therefore are less look alike.
- There is more uncertainty created in the dataset. Therefore the standard error increases.

What is multiple imputation?

- The amount of imputations is dependent on the amount of missing data, but mostly 5 to 10 imputations are enough.
- A drawback of this method it that several imputed datasets are created and that the statistical analysis has to be repeated in each dataset.
- Finally, results have to be pooled in a summary measure. Most statistical packages can do this automatically.

Multiple Imputation

- The question of how to obtain valid inferences from imputed data was addressed by Rubin's (1987) book on multiple imputation (MI).
- MI is a Monte Carlo technique in which the missing values are replaced by $m > 1$ simulated versions, where m is typically small (e.g. 3-10).

Multiple Imputation

- In Rubin's method for '**repeated imputation**' inference, each of the simulated complete datasets is analyzed by standard methods, and the results are combined to produce estimates and confidence intervals that incorporate missing-data uncertainty.
- Rubin (1987) addresses potential uses of MI primarily for large public-use data files from sample surveys and censuses.

Multiple Imputation

- With the advent of new computational methods and software for creating MI's, however, the technique has become increasingly attractive for researchers in the biomedical, behavioral, and social sciences whose investigations are hindered by missing data.
- These methods are documented in a recent book by Schafer (1997) on incomplete multivariate data.