Applied Missing data analysis with SPSS and R(Studio)

Martijn Heymans and Iris Eekhout

2018-09-03

Contents

Preface

The attention for missing data is growing and so will be the application of methods to solve the missing data problem. From our experience, researchers with missing data still find it difficult to reserve time to evaluate the missing data and from that to find a reasonable solution to handle their missing data for their main data analysis. This manual is developed for researchers that are looking for a solution of their missing data problem or want to learn more about missing data. The manual is developed as a result of a missing data course that we give. Further, we are also active in providing statistical advice in general and more specific about missing data. Because our time to give advice is mostly limited we wanted to give researchers a practical guide to help them get started with their missing data problem. Leading methodologists and statisticians and leading journals have published papers about the problems of missing data and warned researchers to take missing data seriously (Sterne et al., BMJ 2009, Little et al. NEJM 2012, Peng et al. 2015, JAMA). Hopefully this manual will help researchers to find the best solution for their missing data problem. We hope you will enjoy this manual and that you learn from it, at least to take missing data seriously and that you will use recommended methods to solve your missing data problem.

0.1 The goal of this Manual

In this manual the software packages SPSS and R play a central role. The combination of these two software packages may seem a coincidence, but it is not. For a long time, SPSS was the most popular software package worldwide to do statistical data analysis. Currently, R is growing in popularity fast and will probably become one of the most popular Software packages to do data analysis. Also for applied researchers. Both SPSS and R have their advantages and disadvantages. An advantage of SPSS is that it is a user-friendly software package compared to R and works with windows where you can for example drag your variables to. Subsequently, you can click the OK button and the statistical analysis procedure you prespecified gives you the output results. A disadvantage of SPSS may be that you are overloaded with statistical output that may not all needed to answer your research question. Compared to SPSS you could say that R is a more user-unfriendly software package where you need to use R code to activate statistical procedures and to get statistical results. R output will show more specific results, without extra information. Furthermore, R works much faster when it comes to running statistical procedures by using 1 or 2 lines of R code, compared to visiting a couple of windows in SPSS to activate the same statistical test. There is one other advantage of R and that is, that it is open source. This makes it possible for applied researchers to follow the calculations of complex procedures as the estimation of missing values closely along the line. You could say that R brings you to the heart of the matter. With R it is possible to turn complex data analysis functions and formula´s into computer code that can be used by everybody and vice versa. Because it is open source, you are able to read the code that is used for the analysis and to relate that code or pieces of code to the statistical output. This makes it possible to evaluate step by step the code and thus the statistical procedures and relate them to the subsequent results. You can copy specific parts of code from functions that others have written and evaluate what happens. This is one of the major advantages of R if you compare it to the closed source statistical package SPSS. R brings you a big learning environment when it comes to the understanding of all kind of statistical procedures as missing data analysis.

0.2 Multiple Imputation in SPSS and R

Multiple Imputation (MI) is a procedure that is developed in the 1970's by Donald Rubin. Later, around the 1990´s Multiple imputation was further developed and became more popular. For a long time, MI was only available for S-Plus and R software (S-plus is the commercial alternative of R), where it was further developed by Stef van Buuren, a statistician from TNO, Leiden, The Netherlands. For a long time, it was not possible to do MI analysis in SPSS because it was not available in SPSS. So, it was far out of reach for applied researchers for a long time. It became available from SPSS version 17. From that time MI is now used more by applied researchers. In this manual the handling of missing data is the main topic. We will also show how to apply these methods in both software packages SPSS and R. To apply the imputation methods that are discussed both software packages make use of random starting procedures. SPSS and R use for that intern random number generators. Because these are different, result might slightly differ. Our intention is not to compare the software packages SPSS and R and their output resultys. Both are trustful packages, it is more the estimation procedures that might lead to the differences. The imputation methods, will be applied in SPSS version 24 and with R software version 3.4.3. The R examples will be presented by using the output from RStudio version (version 1.1.383 – © 2009-2017 RStudio, Inc.). RStudio is an integrated development environment (IDE) for R. RStudio includes a wide range of productivity enhancing features and runs on all major platforms. As already stated, R allows you to program the statistical formula's yourself. We have therefore chosen to explain the formula's in more detail in combination with the application in R. The more applied researchers will be satisfied with the explanation and application of methods in SPSS.

0.3 Notation and annotation in this manual

The name of R packages, libraries and functions can be recognized by using Courier new lettertype, for example the package mice will be written as mice.

R code of the procedures used in the manual is marked grey and the explanation in these grey parts can be found in the grey parts itself annotated by the # symbol. The lines that start with the symbol > are R Code lines that have been running in the R Console in RStudio. Example:

R code XX

```r
# Activate the foreign package and read in the SPSS dataset

library(foreign)
dataset <- read.spss(file="Backpain 50 missing.sav", to.data.frame=T)
```

```
## re-encoding from UTF-8
```

Chapter 1

Introduction

You can label chapter and section titles using `{#label}` after them, e.g., we can reference Chapter 1. If you do not manually label them, there will be automatic labels anyway, e.g., Chapter 3.

Figures and tables with captions will be placed in `figure` and `table` environments, respectively.

```r
par(mar = c(4, 4, .1, .1))
plot(pressure, type = 'b', pch = 19)
```

Reference a figure by its code chunk label with the `fig:` prefix, e.g., see Figure 1.1. Similarly, you can reference tables generated from `knitr::kable()`, e.g., see Table 1.1.

```r
knitr::kable(
  head(iris, 20), caption = 'Here is a nice table!',
  booktabs = TRUE
)
```

You can write citations, too. For example, we are using the **bookdown** package (Xie, 2018) in this sample book, which was built on top of R Markdown and **knitr** (Xie, 2015).

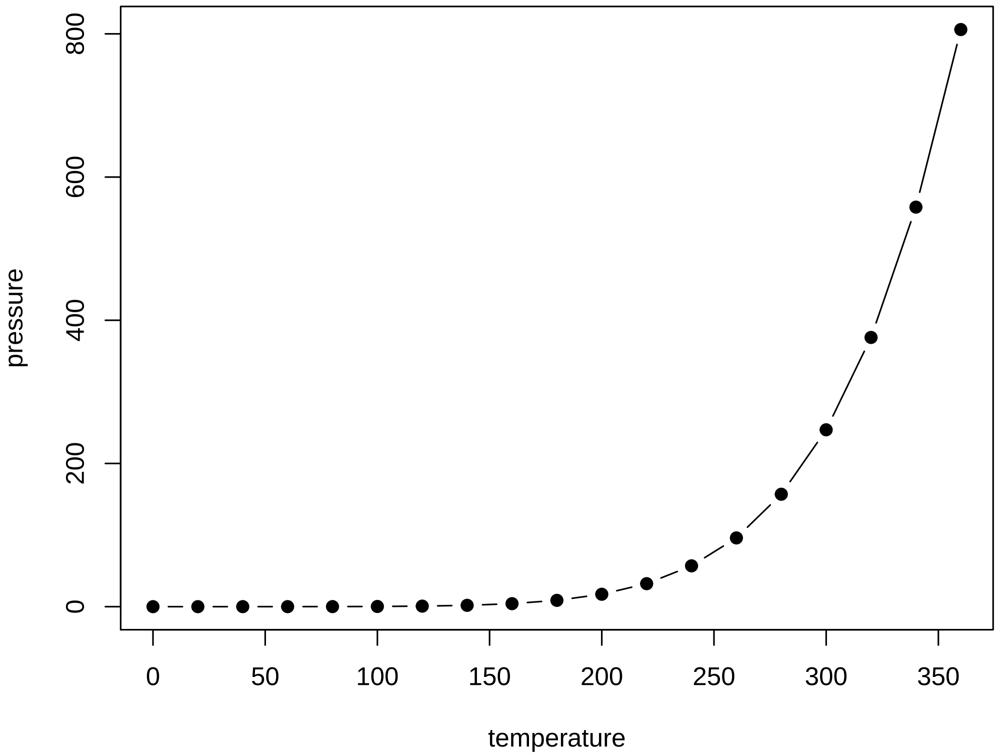

Figure 1.1: Here is a nice figure!

Table 1.1: Here is a nice table!

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5.0	3.4	1.5	0.2	setosa
4.4	2.9	1.4	0.2	setosa
4.9	3.1	1.5	0.1	setosa
5.4	3.7	1.5	0.2	setosa
4.8	3.4	1.6	0.2	setosa
4.8	3.0	1.4	0.1	setosa
4.3	3.0	1.1	0.1	setosa
5.8	4.0	1.2	0.2	setosa
5.7	4.4	1.5	0.4	setosa
5.4	3.9	1.3	0.4	setosa
5.1	3.5	1.4	0.3	setosa
5.7	3.8	1.7	0.3	setosa
5.1	3.8	1.5	0.3	setosa

Chapter 2

Literature

Here is a review of existing methods.

Chapter 3

Methods

We describe our methods in this chapter.

Chapter 4

Applications

Some *significant* applications are demonstrated in this chapter.

4.1 Example one

4.2 Example two

Chapter 5

Final Words

We have finished a nice book.

Bibliography

Xie, Y. (2015). *Dynamic Documents with R and knitr.* Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition. ISBN 978-1498716963.

Xie, Y. (2018). *bookdown: Authoring Books and Technical Documents with R Markdown.* R package version 0.7.