Hadley Wickham
@hadleywickham

Following

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

RETWEETS	FAVORITES
42	62

9:01 PM - 17 Aug 2015

Kieran Healy @kjhealy · Aug 17

@hadleywickham

4

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.