Dealing with Missing Data

David Carlson

April 4, 2022

Intro to the Problem

- Studies often fail to describe, analyze, or acknowledge missing data

Intro to the Problem

- Studies often fail to describe, analyze, or acknowledge missing data
- This is a problem for scientific inquiry, but also can get your article rejected from (good) journals

Intro to the Problem

- Studies often fail to describe, analyze, or acknowledge missing data
- This is a problem for scientific inquiry, but also can get your article rejected from (good) journals
- Conclusions often change when accounting for missingness

Intro to the Problem

- Studies often fail to describe, analyze, or acknowledge missing data
- This is a problem for scientific inquiry, but also can get your article rejected from (good) journals
- Conclusions often change when accounting for missingness
- Generally, by default most (all?) regression software only utilize complete cases for standard modeling

Overview of Steps

- Ensure your data are coded correctly (e.g., NA often be coded as 99)

Overview of Steps

- Ensure your data are coded correctly (e.g., NA often be coded as 99)
- Identify missing values within each variable under study and describe / report

Overview of Steps

- Ensure your data are coded correctly (e.g., NA often be coded as 99)
- Identify missing values within each variable under study and describe / report
- Look for patterns in missingness

Overview of Steps

- Ensure your data are coded correctly (e.g., NA often be coded as 99)
- Identify missing values within each variable under study and describe / report
- Look for patterns in missingness
- Check for associations between missing and observed data

Overview of Steps

- Ensure your data are coded correctly (e.g., NA often be coded as 99)
- Identify missing values within each variable under study and describe / report
- Look for patterns in missingness
- Check for associations between missing and observed data
- Decide how to (transparently!) handle missingness

Types of Missingness

- Missing completely at random (MCAR)

Types of Missingness

- Missing completely at random (MCAR)
  - Missing data values do not relate to any other data in the dataset and there is no pattern to the actual values of the missing data themselves

Types of Missingness

- Missing completely at random (MCAR)
  - Missing data values do not relate to any other data in the dataset and there is no pattern to the actual values of the missing data themselves
  - This is easy to handle, but data are almost never missing completely at random (any social science examples? at all?)

Types of Missingness

- Missing completely at random (MCAR)
  - Missing data values do not relate to any other data in the dataset and there is no pattern to the actual values of the missing data themselves
  - This is easy to handle, but data are almost never missing completely at random (any social science examples? at all?)
  - We can ignore if MCAR (but how to convince?)

Types of Missingness (cont.)

- Missing at random (MAR)

Types of Missingness (cont.)

- Missing at random (MAR)
  - Missing data have relationship with other variables

Types of Missingness (cont.)

- Missing at random (MAR)
  - Missing data have relationship with other variables
  - Actual values that are missing are random

Types of Missingness (cont.)

- Missing at random (MAR)
  - Missing data have relationship with other variables
  - Actual values that are missing are random
  - Examples?

Types of Missingness (cont.)

- Missing not at random (MNAR)

Types of Missingness (cont.)

- Missing not at random (MNAR)
  - Missing data have relationship with other variables

Types of Missingness (cont.)

- Missing not at random (MNAR)
  - Missing data have relationship with other variables
  - Values of missing data are not random

Types of Missingness (cont.)

- Missing not at random (MNAR)
  - Missing data have relationship with other variables
  - Values of missing data are not random
  - MNAR most problematic; can alter your conclusions, and are the most difficult to diagnose and handle

Types of Missingness (cont.)

- Missing not at random (MNAR)
  - Missing data have relationship with other variables
  - Values of missing data are not random
  - MNAR most problematic; can alter your conclusions, and are the most difficult to diagnose and handle
  - They can only be detected by collecting and examining some of the missing data; this is often difficult or impossible to do

Types of Missingness (cont.)

- Missing not at random (MNAR)
  - Missing data have relationship with other variables
  - Values of missing data are not random
  - MNAR most problematic; can alter your conclusions, and are the most difficult to diagnose and handle
  - They can only be detected by collecting and examining some of the missing data; this is often difficult or impossible to do
  - Examples?

Correctly Coding Your Data

- The first step in any analysis is robust data cleaning and coding

Correctly Coding Your Data

- The first step in any analysis is robust data cleaning and coding
- Ensure all factors and numerics are correctly assigned; this is the most common reason to get an error (or incorrect results without an error!)

Correctly Coding Your Data

- The first step in any analysis is robust data cleaning and coding
- Ensure all factors and numerics are correctly assigned; this is the most common reason to get an error (or incorrect results without an error!)
- Ensure you know which variables have missing data; this presumes missing values are correctly assigned NA

Correctly Coding Your Data

- The first step in any analysis is robust data cleaning and coding
- Ensure all factors and numerics are correctly assigned; this is the most common reason to get an error (or incorrect results without an error!)
- Ensure you know which variables have missing data; this presumes missing values are correctly assigned NA
- Ensure factor levels and variable labels are assigned correctly

Describing and Analyzing Missingness

- Identify missing values in each variable: `missing_plot` (also include a table)

Describing and Analyzing Missingness

- Identify missing values in each variable: `missing_plot` (also include a table)
- Look for patterns of missingness: `missing_pattern` (table and plot of patterns)

Describing and Analyzing Missingness

- Identify missing values in each variable: `missing_plot` (also include a table)
- Look for patterns of missingness: `missing_pattern` (table and plot of patterns)
- Make sure you include missing data in demographics tables (do not silently drop NAs)

Describing and Analyzing Missingness

- Identify missing values in each variable: `missing_plot` (also include a table)
- Look for patterns of missingness: `missing_pattern` (table and plot of patterns)
- Make sure you include missing data in demographics tables (do not silently drop NAs)
- Check for associations between missing and observed data: `missing_pairs` | `missing_compare`

Describing and Analyzing Missingness

- Identify missing values in each variable: `missing_plot` (also include a table)
- Look for patterns of missingness: `missing_pattern` (table and plot of patterns)
- Make sure you include missing data in demographics tables (do not silently drop NAs)
- Check for associations between missing and observed data: `missing_pairs` | `missing_compare`
- If you work predominately with numeric rather than discrete data, you may find tests from the `MissMech` package useful

How to Handle?

- Delete the variable with the missing data

How to Handle?

- Delete the variable with the missing data
- Delete the cases with the missing data

How to Handle?

- Delete the variable with the missing data
- Delete the cases with the missing data
- Impute (fill in) the missing data

How to Handle?

- Delete the variable with the missing data
- Delete the cases with the missing data
- Impute (fill in) the missing data
- Model the missing data

MCAR

- If we know nothing about the missing values themselves, but we know of no plausible reason that the values of the missing data should be different to the values of the missing data

MCAR

- If we know nothing about the missing values themselves, but we know of no plausible reason that the values of the missing data should be different to the values of the missing data
- Depending on the number of data points that are missing, we may have sufficient power with complete cases to examine the relationships of interest (list-wise delete)

MCAR

- If we know nothing about the missing values themselves, but we know of no plausible reason that the values of the missing data should be different to the values of the missing data
- Depending on the number of data points that are missing, we may have sufficient power with complete cases to examine the relationships of interest (list-wise delete)
- If the variable in question is thought to be particularly important, you may wish to perform a sensitivity analysis; a sensitivity analysis in this context aims to capture the effect of uncertainty on the conclusions drawn from the model

MCAR

- If we know nothing about the missing values themselves, but we know of no plausible reason that the values of the missing data should be different to the values of the missing data
- Depending on the number of data points that are missing, we may have sufficient power with complete cases to examine the relationships of interest (list-wise delete)
- If the variable in question is thought to be particularly important, you may wish to perform a sensitivity analysis; a sensitivity analysis in this context aims to capture the effect of uncertainty on the conclusions drawn from the model
- If control is not associated with the explanatory variable of interest or the outcome, it may be considered not to be a confounder and so could be omitted (this neatly deals with the missing data issue, but of course may not be appropriate)

MAR

- The above is rarely the case

MAR

- The above is rarely the case
- If we simply drop all the cases in which variable is missing, then we disproportionately drop; this may have consequences for our conclusions if variable is associated with our explanatory variable of interest or outcome

MAR

- The above is rarely the case
- If we simply drop all the cases in which variable is missing, then we disproportionately drop; this may have consequences for our conclusions if variable is associated with our explanatory variable of interest or outcome
- Common solution: Multiple Imputation Through Chained Systems of Equations (package `mice`)

MAR

- The above is rarely the case
- If we simply drop all the cases in which variable is missing, then we disproportionately drop; this may have consequences for our conclusions if variable is associated with our explanatory variable of interest or outcome
- Common solution: Multiple Imputation Through Chained Systems of Equations (package `mice`)
- Fills in missing data using a best-estimate from all the other data that exists

MAR

- The above is rarely the case
- If we simply drop all the cases in which variable is missing, then we disproportionately drop; this may have consequences for our conclusions if variable is associated with our explanatory variable of interest or outcome
- Common solution: Multiple Imputation Through Chained Systems of Equations (package `mice`)
- Fills in missing data using a best-estimate from all the other data that exists
- If missingness in variable is predicted strongly by other observed variables, and the values of the missing data are random, then we can impute using other variables in the dataset

MAR

- The above is rarely the case
- If we simply drop all the cases in which variable is missing, then we disproportionately drop; this may have consequences for our conclusions if variable is associated with our explanatory variable of interest or outcome
- Common solution: Multiple Imputation Through Chained Systems of Equations (package `mice`)
- Fills in missing data using a best-estimate from all the other data that exists
- If missingness in variable is predicted strongly by other observed variables, and the values of the missing data are random, then we can impute using other variables in the dataset
- Imputing missing outcome is very controversial and no proof of (un-)biasedness exists (thoughts?)

MAR

- The above is rarely the case
- If we simply drop all the cases in which variable is missing, then we disproportionately drop; this may have consequences for our conclusions if variable is associated with our explanatory variable of interest or outcome
- Common solution: Multiple Imputation Through Chained Systems of Equations (package `mice`)
- Fills in missing data using a best-estimate from all the other data that exists
- If missingness in variable is predicted strongly by other observed variables, and the values of the missing data are random, then we can impute using other variables in the dataset
- Imputing missing outcome is very controversial and no proof of (un-)biasedness exists (thoughts?)
- You can also treat NA as a factor (I do not recommend)

MNAR

- Good luck!

MNAR

- Good luck!
- To determine if data are definitely MNAR, we need to know their value in a subset of observations

MNAR

- Good luck!
- To determine if data are definitely MNAR, we need to know their value in a subset of observations
- There is no easy way to handle this; if at all possible, try to get the missing data, otherwise, take care when drawing conclusions from analyses where data are thought to be missing not at random