UNSW SYDNEY | Australia's Global University

Dealing with Missing Data in Your Research

Nancy Briggs
Stats Central
Mark Wainwright Analytical Centre
July 11, 2019

Outline

- Why do we have missing data?
- What kind of missing data do we have?
- Approaches to dealing with it

Missing Data Happens

- Study Design
- Censoring
- Data collection or processing errors
- Non-response
  - Decline to answer
  - Attrition

- Or some other mechanism we can't see!

Types of Missing Data

- Missing Completely at Random (MCAR)
  - There is no relationship between the propensity to be a missing value and any observed or missing data.
  - Missing data (and the observed data) can be considered a simple random sample of the complete data.
- Missing at Random (MAR)
  - The propensity for a data point to be missing is not related to the missing data, but it is related to observed data.
- Missing Not at Random (MNAR)
  - There is a relationship between the propensity of a data point to be missing and its values.

What Happens with Missing Data

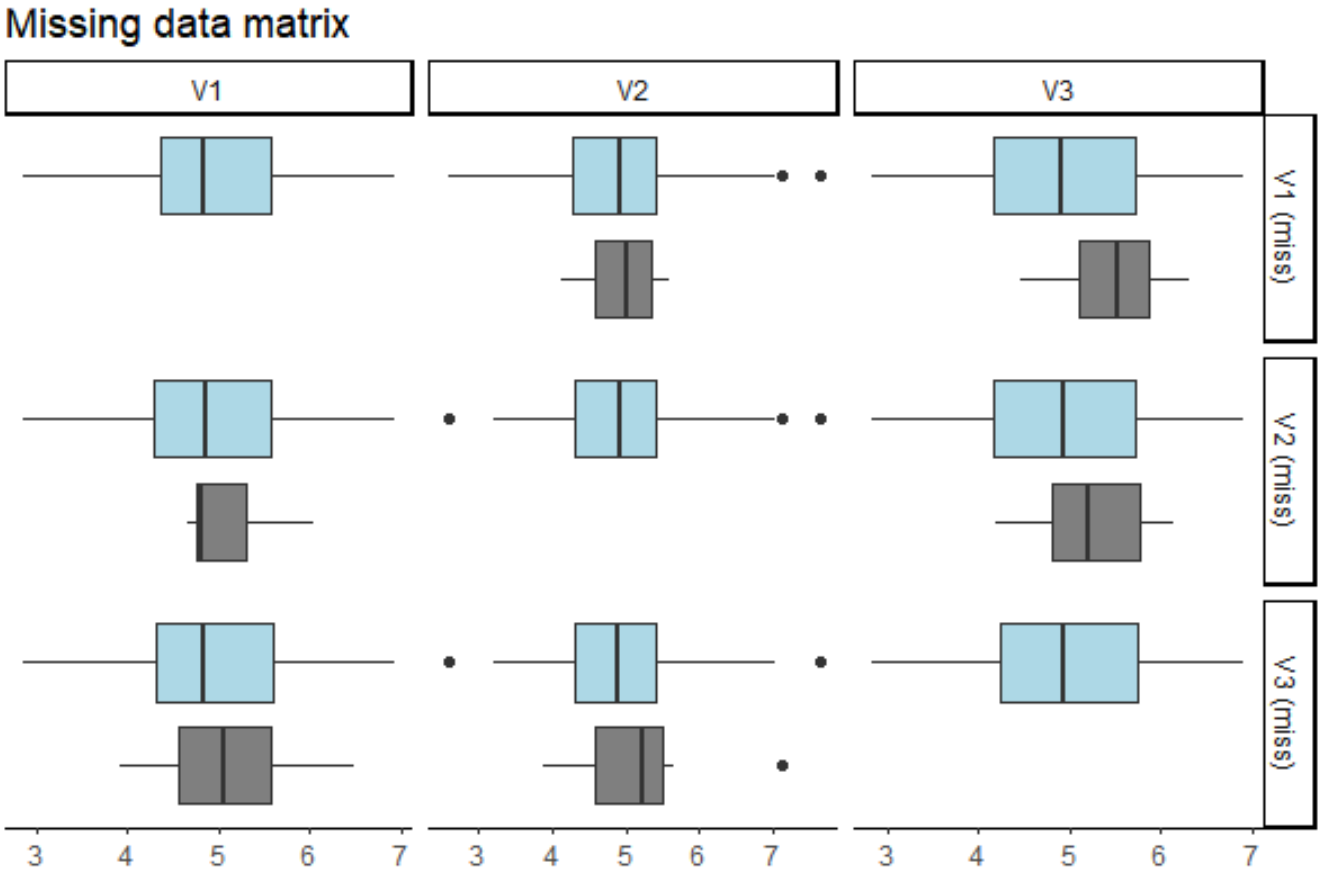
```
> describe(dat)
   vars   n mean   sd median trimmed  mad  min  max range
X1    1 100 4.81 0.97   4.73    4.80 0.93 2.57 7.50  4.94
X2    2 100 4.92 0.99   4.91    4.92 1.04 1.73 7.33  5.60
X3    3 100 4.86 0.96   4.86    4.85 1.01 2.94 7.39  4.46
```

Simulated data: MCAR

MCAR

```
> library(mice)
> result1 <- ampute(data = complete.data, mech="MCAR", prop=.25)

> describe(dat1)
   vars  n mean   sd median trimmed  mad  min  max range
V1    1 89 5.10 1.02   4.99    5.07 1.15 3.09 7.83  4.74
V2    2 96 5.15 0.89   5.28    5.18 0.78 2.93 7.17  4.24
V3    3 93 4.95 0.92   4.93    4.92 0.82 3.13 7.22  4.09


> library(finalfit)

> dat1 %>%
+   missing_compare(dependent, explanatory)


  Missing data analysis: V1                Missing Not missing     p
1                        V2 Mean (SD) 4.3 (1.4)   5.0 (0.9) 0.156
2                        V3 Mean (SD) 4.5 (0.9)   4.8 (0.9) 0.458
```

```
dat1 %>%
  missing_pairs(dependent, explanatory)
```

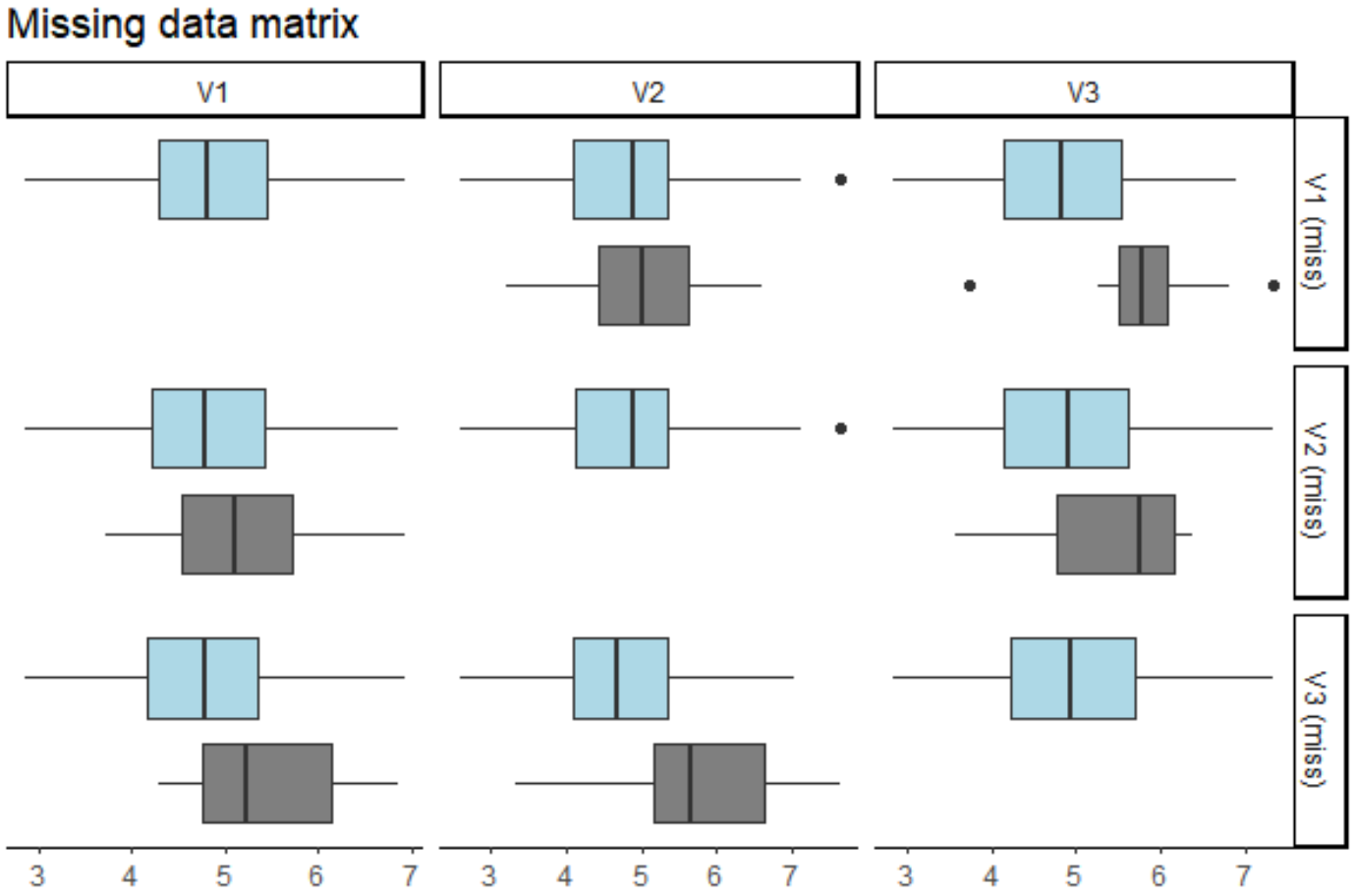


Simulated data: MAR

MAR

```
> library(mice)
> result2 <- ampute(data = complete.data, mech="MAR", prop=.25)

   vars  n mean   sd median trimmed  mad  min  max range
V1    1 94 4.83 0.96   4.76    4.82 0.92 2.65 7.50  4.85
V2    2 91 4.94 0.97   4.96    4.95 1.06 1.73 7.13  5.40
V3    3 94 4.87 0.96   4.86    4.84 1.02 2.94 7.39  4.46


> dat2 %>%
+   missing_compare(dependent, explanatory)


2  Missing data analysis: V1              Missing Not missing     p
3  1                        V2 Mean (SD) 5.0 (1.0)   4.9 (1.0) 0.423
4  2                        V3 Mean (SD) 5.8 (0.9)   4.8 (0.9) 0.002
```

```
dat2 %>%
  missing_pairs(dependent, explanatory)
```

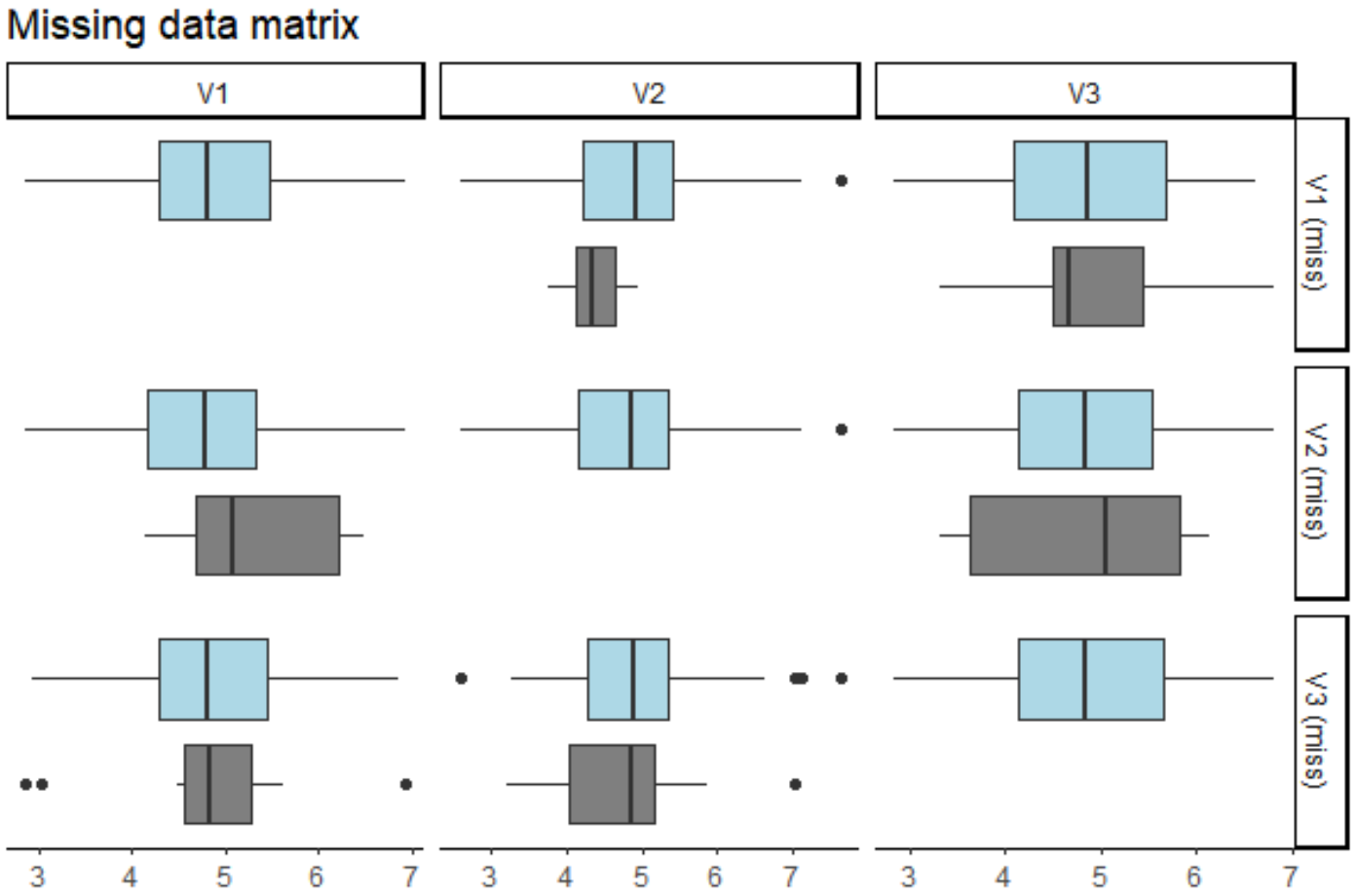


Simulated data: MNAR

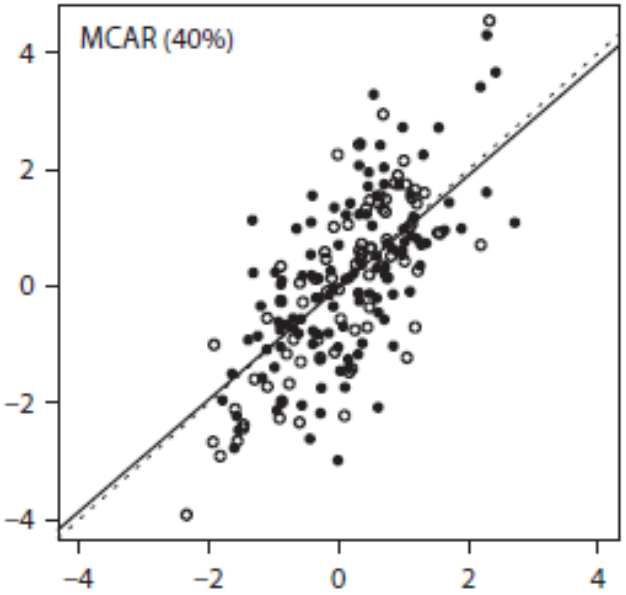
MNAR

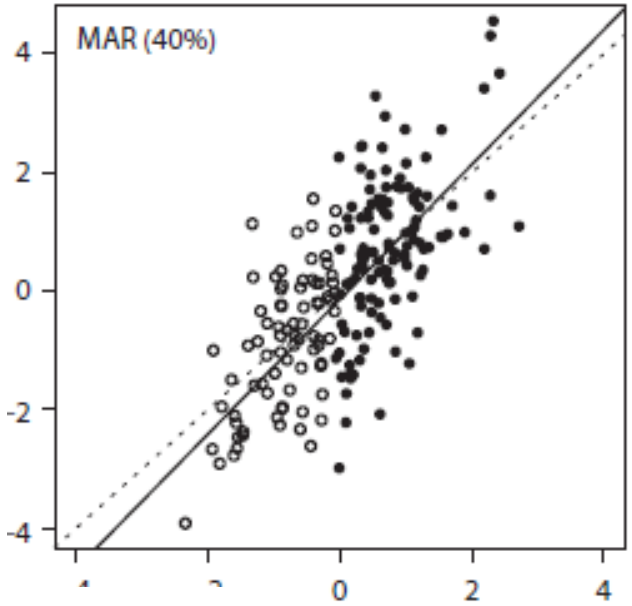
```
> library(mice)
> result1 <- ampute(data = complete.data, mech="MNAR", prop=.25)

   vars  n mean   sd median trimmed  mad  min  max range
V1    1 89 4.77 1.01   4.58    4.75 1.01 2.57 7.50  4.94
V2    2 93 4.86 0.96   4.88    4.88 1.05 1.73 6.98  5.25
V3    3 81 4.71 0.94   4.66    4.69 0.96 2.94 6.94  4.00


> dat3 %>%
+   missing_compare(dependent, explanatory)
  Missing data analysis: V1                Missing Not missing     p
1                        V2 Mean (SD) 4.4 (0.4)   4.9 (1.0) 0.082
2                        V3 Mean (SD) 4.9 (1.0)   4.8 (0.9) 0.989
```

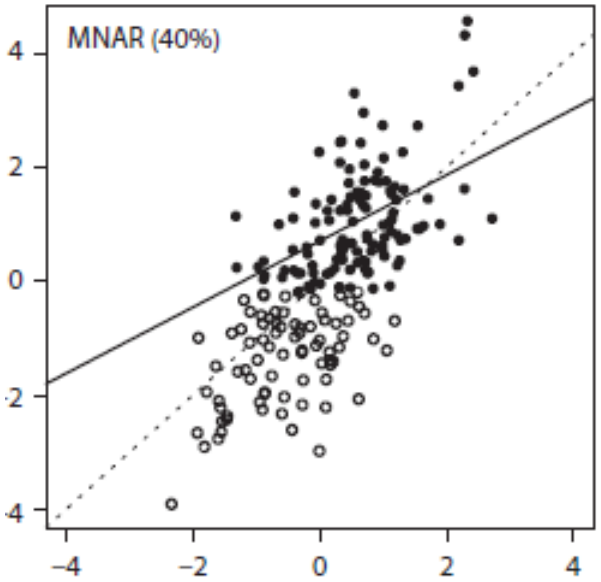
```
dat2 %>%
  missing_pairs(dependent, explanatory)
```



(a)







Nakagawa, S. (2015).

Results from a regression of X1 on X2 & X3

	Full		MCAR		MAR		MNAR	
Int	3.5 (0.65)	<0.001	3.21 (0.78)	<0.001	3.46 (0.75)	<0.001	3.12 (0.78)	<0.001
X2	0.04 (0.1)	0.702	0.03 (0.12)	0.817	0.07 (0.11)	0.51	-0.03 (0.12)	0.825
X3	0.23 (0.1)	0.023	0.29 (0.12)	0.015	0.19 (0.12)	0.11	0.37 (0.13)	0.005

Good News!
(about some types of missing data)

Some missing data are ***ignorable***.

Good News!
(about some types of missing data)

Some missing data are ***ignorable***.

Ironically, this does not imply that you can actually just ignore the missing data!

Good News!
(about some types of missing data)

Some missing data are ***ignorable***.

Ironically, this does not imply that you can actually just ignore the missing data!

Ignorability refers to whether we can ignore the way in which data are missing when we impute or augment missing data.

What to do?

- Listwise / Pairwise Deletion
- Single Imputation
- Model Based Methods: Full Information Maximum Likelihood
- Multiple Imputation
- Model for the missing data

What to do? Bad Ideas

- Listwise / Pairwise Deletion
  - Bad Idea!
- Single Imputation Methods
  - Mean Imputation
    - Use the mean of existing values to replace
      - Affects variability estimates
      - Sometimes overestimate coefficients, sometimes underestimate coefficients
  - Regression
    - Use predicted values from a regression using complete cases
      - Affects variability estimates
      - Affects covariances among variables
  - “Last one Carried Forward” (longitudinal data)
    - assumes that the value of the outcome remains unchanged by the missing data

What to do? Good Ideas

- Maximum Likelihood
  - The algorithm uses all the available information
  - Not based on artificial (imputed) data
  - But!
    - Model-Dependent
    - Not necessarily always an option
- Expectation Maximization Algorithm
- Multiple Imputation

What to do? Good Ideas

- Maximum Likelihood
- Expectation Maximization Algorithm
  - Uses a two-step method to estimate data that preserves means, variances and covariances of the variables
  - Fast and Easy!
  - Can use in a wide range of data analysis scenarios
  - You actually end up with a complete dataset.
  - But!
    - Standard errors for inferential analyses will be biased downward
    - The analyses of the EM-based data do not properly account for the uncertainty inherent in imputing missing data.
- Multiple Imputation

What to do? Good Ideas

- Maximum Likelihood
- Expectation Maximization Algorithm
- Multiple Imputation
  - Multiple possible datasets are created, and each one is analysed. Results are aggregated.
  - Incorporates uncertainty into predictions of unknown values unlike single imputation methods
  - Can improve both the accuracy and often the statistical power of results
  - But
    - Results are dependent on your imputation model (model for the missing data).
    - It can be hard, especially with multilevel data
    - Sample size & proportion missing may be an issue
    - Need to be careful with interaction terms in your model

“For most of our scientific history, we have approached missing data much like a doctor from the ancient world might use bloodletting to cure disease or amputation to stem infection (e.g. removing the infected parts of one’s data by using list-wise or pair-wise deletion)”

Todd Little (cited in Enders, 2010)

How to do Multiple Imputation?

```
library (mice) # loading the mice library

# the imputation step with 100 copies

imputation <- mice (dat2, m = 100, seed = 34634)

analysis <- with (imputation, lm (V1 ~ V2 + V3 ))

pooling <- pool (analysis) # the pooling step

summary (pooling)
```

	Full		MAR		MI with 100 datasets	
Int	3.5 (0.65)	<0.001	3.46 (0.75)	<0.001	3.66 (0.73)	<0.001
X2	0.04 (0.1)	0.702	0.07 (0.11)	0.51	0.003 (0.1)	0.97
X3	0.23 (0.1)	0.023	0.19 (0.12)	0.11	0.23 (0.11)	0.04

What About Non-ignorable Missing Data?

- All of the above suggestions are only recommended if you have MCAR or MAR data.
- There are no tests to detect MNAR (non-ignorable) missingness.
- Suggested approaches:
  - Incorporate a model for the missing data into your data analysis
    - This can be really complicated & hard!
    - You have to assume a lot about the relationship between the missing values and how they came to be missing. This is hard because you can't actually know.
    - If these assumptions are incorrect, these non-ignorable models may perform *worse* than the models for ignorable missingness.
  - Sensitivity Analysis?

References

Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford Press.

Laird, N.M. (1988). Missing data in longitudinal studies. Statistics in Medicine, 7(1-2).

Nakagawa, S. (2015). Missing Data : Mechanisms, Methods and Messages. In G.A. Fox, et al. Editor (Ed.), Ecological statistics : contemporary theory and application.

Nguyen, C. D., Carlin, J. B., & Lee, K. J. (2017). Model checking in multiple imputation: an overview and case study. *Emerging themes in epidemiology*, *14*, 8. doi:10.1186/s12982-017-0062-6

Shylaja B., & Kumar, S. (2018). Traditional versus modern missing data handling techniques: An overview. International Journal of Pure and Applied Mathematics, 118(14).

Useful R packages:

finalfit: Harrison, E. https://cran.r-project.org/web/packages/finalfit/finalfit.pdf

mice: Noghrehchi, F. Missing Data Analysis with mice. https://web.maths.unsw.edu.au/~dwarton/missingDataLab.html

Hacky Hour!

- From 3-4 pm at Penny Lane, Thursdays
- Research Technology Services has lots of people available for informal chats, advice, input
- Join us!