Missing Data Basics

Utrecht University Winter School: Missing Data in R

Kyle M. Lang

Department of Methodology & Statistics
Utrecht University

Outline

Missing Data Descriptives

- Missing Data Patterns
- Nonresponse Rates
- Coverage Measures

Missing Data Mechanisms

- Testing the Missing Data Mechanism

What are Missing Data?

Missing data are empty cells in a dataset where there should be observed values.

- The missing cells correspond to true population values, but we haven't observed those values.

What are Missing Data?

Missing data are empty cells in a dataset where there should be observed values.

- The missing cells correspond to true population values, but we haven't observed those values.

Not every empty cell is a missing datum.

- Quality-of-life ratings for dead patients in a mortality study
- Firm profitability after the company goes out of business
- Self-reported severity of menstrual cramping for men
- Empty blocks of data following "gateway" items

A Little Notation

$$
\begin{aligned}
Y &:= \text{An } N \times P \text{ Matrix of Arbitrary Data} \\
Y_{mis} &:= \text{The \textit{missing} part of } Y \\
Y_{obs} &:= \text{The \textit{observed} part of } Y \\
R &:= \text{An } N \times P \text{ response matrix} \\
M &:= \text{An } N \times P \text{ missingness matrix}
\end{aligned}
$$

The R and M matrices are complementary.

- $r_{np} = 1$ means y_{np} is observed; $m_{np} = 1$ means y_{np} is missing.
- $r_{np} = 0$ means y_{np} is missing; $m_{np} = 0$ means y_{np} is observed.
- M_p is the *missingness* of Y_p.

Missing Data Descriptives

Missing Data Pattern

Missing data (or response) patterns represent unique combinations of observed and missing items.

- P items $\Rightarrow 2^P$ possible patterns.

	X	Y
1	x	y
2	x	.
3	.	y
4	.	.

Patterns for $P = 2$

	X	Y	Z
1	x	y	z
2	x	y	.
3	x	.	z
4	.	y	z
5	x	.	.
6	.	.	z
7	.	y	.
8	.	.	.

Patterns for $P = 3$

Missing Data Pattern

The concept of a “missing data pattern” can also be used to classify the spatial arrangement of missing cells on a data set.

- Univariate
  - Missing data occur on only one variable

- Monotone
  - The proportion of complete elements, in both rows and columns, decreases when traversing the data set.
  - The observed cells can be arranged into a “staircase” pattern.

- Arbitrary
  - Missing values are “randomly” scattered throughout the data set.

Example Missing Data Patterns

	X	Y	Z
1	x	y	z
2	x	y	z
3	x	y	z
4	x	y	z
5	x	y	z
6	x	.	z
7	x	.	z
8	x	.	z
9	x	.	z
10	x	.	z

Univariate Pattern

	X	Y	Z
1	x	y	z
2	x	y	z
3	x	y	z
4	x	y	.
5	x	y	.
6	x	y	.
7	x	.	.
8	x	.	.
9	x	.	.
10	.	.	.

Monotone Pattern

	X	Y	Z
1	x	.	z
2	x	y	z
3	x	y	z
4	x	.	z
5	x	y	z
6	x	.	z
7	.	y	z
8	x	y	z
9	x	.	.
10	x	y	.

Arbitrary Pattern

Nonresponse Rates

Proportion Missing

- The proportion of cells containing missing data
- Good early screening measure
- Should be computed for each variable, not for the entire dataset

Attrition Rate

- The proportion of participants that drop-out of a study at each measurement occasion

Nonresponse Rates

Proportion of Complete Cases

- The proportion of observations with no missing data
- Often reported but nearly useless quantity

Fraction of Missing Information

- Associated with an estimated parameter, not with an incomplete variable
- Like an R^2 for the missing data
- Most important diagnostic value for missing data problems
- Can only be computed after treating the missing data

Coverage Measures

Covariance Coverage

$$CC_{jk} = N^{-1} \sum_{n=1}^{N} r_{nj} r_{nk}$$

- The proportion of cases available to estimate a given pairwise relationship (e.g., a covariance between two variables)
- Very important to have adequate coverage of the parameters you want to estimate

Coverage Measures

INBOUND STATISTIC

$$I_{jk} = \frac{\sum_{n=1}^{N}(1 - r_{nj})r_{nk}}{\sum_{n=1}^{N}(1 - r_{nj})}$$

- The proportion of missing cases in Y_j for which Y_k is observed

OUTBOUND STATISTIC

$$O_{jk} = \frac{\sum_{n=1}^{N} r_{nj}(1 - r_{nk})}{\sum_{n=1}^{N} r_{nj}}$$

- The proportion of observed cases in Y_j for which Y_k is missing

Coverage Measures

Influx Coefficient

$$I_j = \frac{\sum_{k=1}^{P} \sum_{n=1}^{N} (1 - r_{nj}) r_{nk}}{\sum_{k=1}^{P} \sum_{n=1}^{N} r_{nk}}$$

- The proportion of observed cells in Y that exists in cases for which Y_j is missing
- How well the missing values in Y_j connect to the observed values in Y_{-j}

Coverage Measures

Outflux Coefficient

$$O_j = \frac{\sum_{k=1}^{P} \sum_{n=1}^{N} r_{nj}(1 - r_{nk})}{\sum_{k=1}^{P} \sum_{n=1}^{N} (1 - r_{nk})}$$

- The proportion of missing cells in Y that exists in cases for which Y_j is observed
- How well the observed values in Y_j connect to the missing values in Y_{-j}

Missing Data Mechanisms

Missing Data Mechanisms

Missing Completely at Random (MCAR)

- $P(R|Y_{mis}, Y_{obs}) = P(R)$
- Missingness is unrelated to any study variables.

Missing at Random (MAR)

- $P(R|Y_{mis}, Y_{obs}) = P(R|Y_{obs})$
- Missingness is related to only the *observed* parts of study variables.

Missing not at Random (MNAR)

- $P(R|Y_{mis}, Y_{obs}) \neq P(R|Y_{obs})$
- Missingness is related to the *unobserved* parts of study variables.

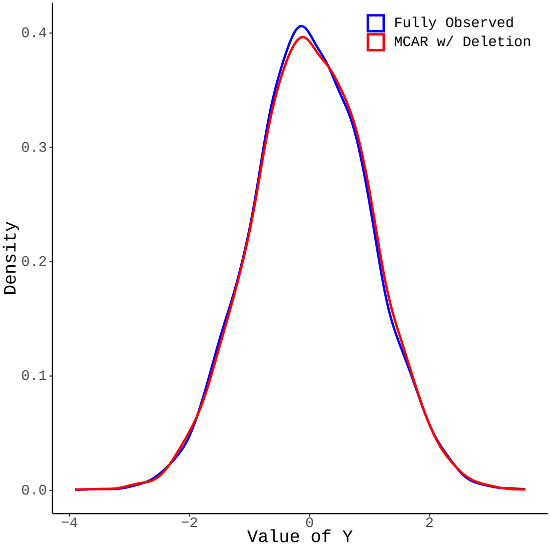
Simulate Some Toy Data

```
library(mvtnorm)
library(dplyr)
library(magrittr)
library(MASS)

nObs <- 5000 # Sample Size
pm   <- 0.3  # Proportion Missing

sigma <- matrix(c(1.0, 0.5, 0.3,
                  0.5, 1.0, 0.0,
                  0.3, 0.0, 1.0),
                ncol = 3)
dat0 <- rmvnorm(nObs, c(0, 0, 0), sigma) %>% data.frame()
colnames(dat0) <- c("x", "y", "z")

dat0 %$% cor(y, x)

[1] 0.5001822
```

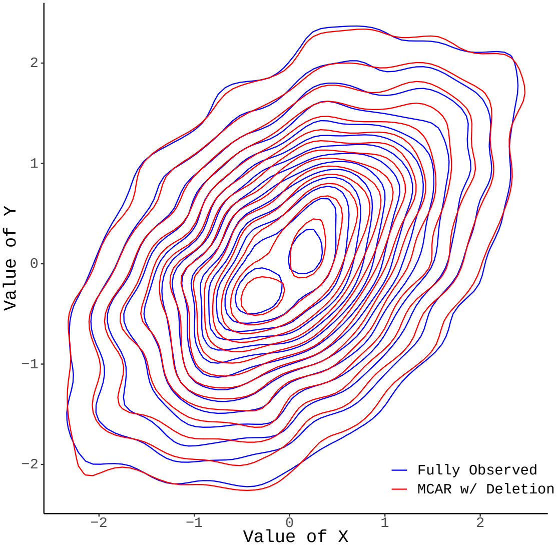
MCAR Example

```r
## Simulate MCAR Missingness:
m <- sample(1:nObs, size = pm * nObs)

## Impose MCAR missing on Y:
mcarData          <- dat0
mcarData[m, "y"] <- NA

## Check the correlation between X & Y:
mcarData %$% cor(y, x, use = "pairwise")

[1] 0.5197437
```

MCAR Example



MCAR Example

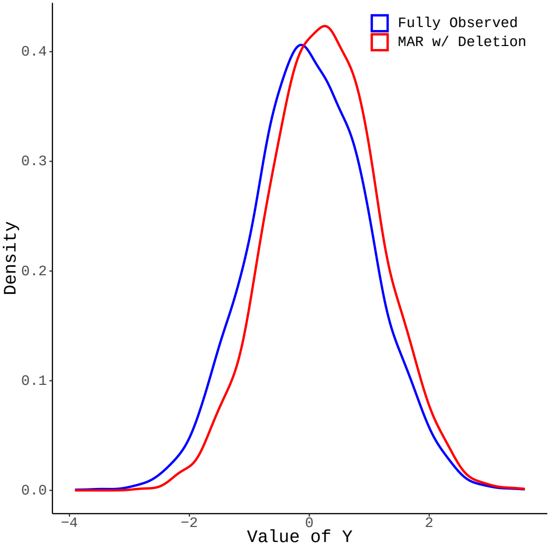


MAR Example

```r
## Simulate MAR Missingness:
m <- with(dat0, x < quantile(x, probs = pm))

## Impose MAR missing on Y:
marData             <- dat0
marData[m, "y"] <- NA

## Check the correlation between X & Y:
marData %$% cor(y, x, use = "pairwise")
```

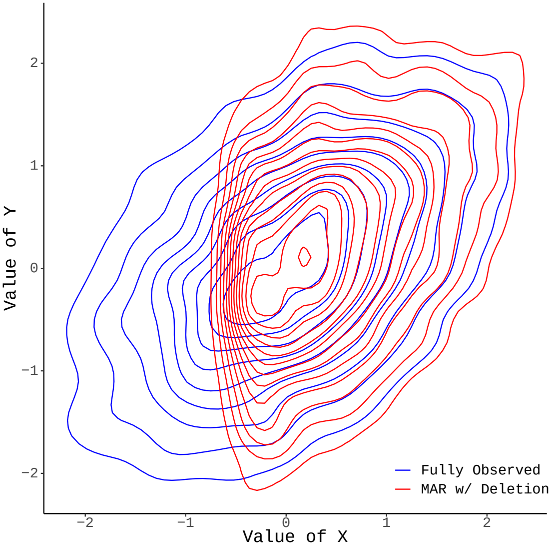
```
[1] 0.3825876
```

MAR Example



Fully Observed
MAR w/ Deletion
Density
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
Value of Y

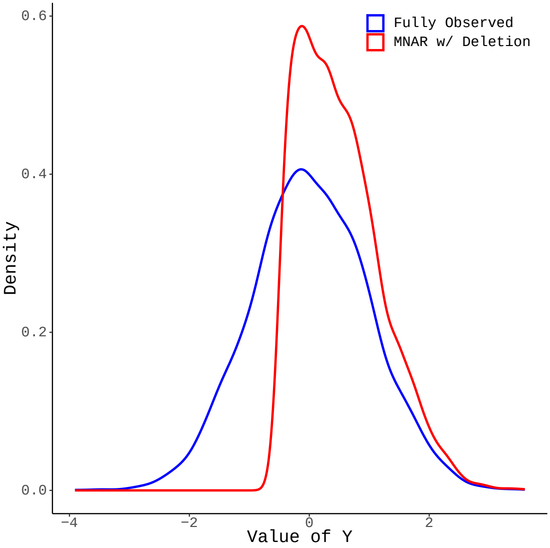
MAR Example


Value of Y
Value of X
2
1
0
−1
−2
−2
−1
0
1
2
Fully Observed
MAR w/ Deletion

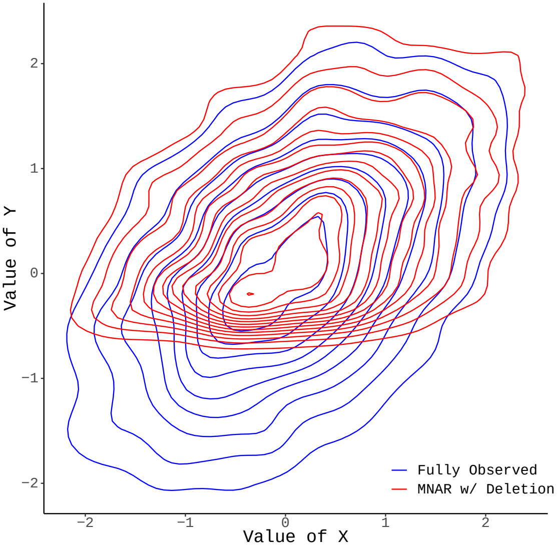
MNAR Example

```r
## Simulate MNAR Missingness:
m <- with(dat0, y < quantile(y, probs = pm))

## Impose MNAR missing on Y:
mnarData            <- dat0
mnarData[m, "y"] <- NA

## Check the correlation between X & Y:
mnarData %$% cor(y, x, use = "pairwise")

[1] 0.3901487
```

MNAR Example



MNAR Example



Crucial Nuance

In our previous MAR example, ignoring the predictor of missingness actually produces *Indirect MNAR*.

Crucial Nuance

In our previous MAR example, ignoring the predictor of missingness actually produces *Indirect MNAR*.

QUESTION: What happens if we ignore the predictor of missingness, but that predictor is independent of our study variables?

Crucial Nuance

In our previous MAR example, ignoring the predictor of missingness actually produces *Indirect MNAR*.

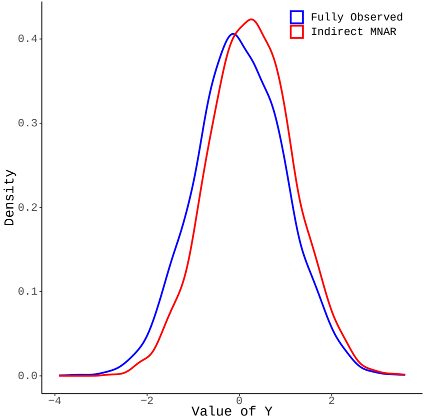
QUESTION: What happens if we ignore the predictor of missingness, but that predictor is independent of our study variables?

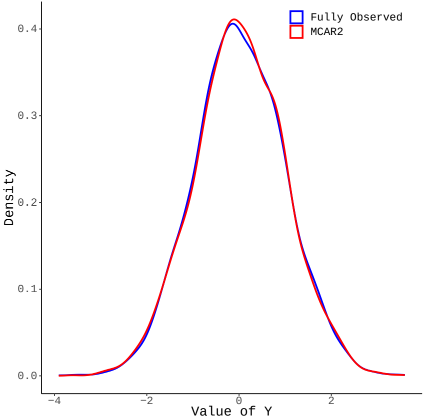
```
m <- with(dat0, z < quantile(z, probs = pm))

mcarData2          <- dat0
mcarData2[m, "y"] <- NA

mcarData2 %$% cor(y, x, use = "pairwise")

[1] 0.5119953
```

ANSWER: We get back to MCAR :)

Crucial Nuance

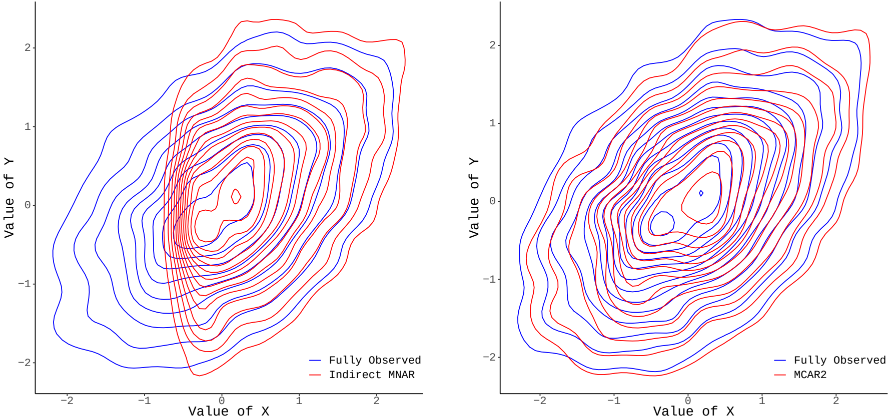
The missing data mechanisms are not simply characteristics of an incomplete dataset; we also need to account for the analysis.





Crucial Nuance

The missing data mechanisms are not simply characteristics of an incomplete dataset; we also need to account for the analysis.



Testing the Missing Data Mechanism

We cannot fully test the MAR or MNAR assumptions.

- To do so would require knowing the values of the missing data.
- We can find observed predictors of missingness.
    - Use classification algorithms to predict missingness from Y_{obs}.
    - We can never know that we have discovered all MAR predictors.
- In practice, MAR and MNAR live on the ends of a continuum.
    - Our missing data problem exists at some unknown point along this continuum.
    - We can do a lot to nudge our problem towards the MAR side.

Testing the Missing Data Mechanism

We can (partially) test the MCAR assumption.

- With MCAR, the missing data and the observed data should have the same distribution.
- We can test for MCAR by testing the distributions of *auxiliary variables*, $\mathbf{Z}$.
    - Use a t-test to compare the subset of Z_p that corresponds to Y_{mis} to the subset corresponding to Y_{obs}.
    - The Little (1988) MCAR test is a multivariate version of this.

These procedures actually test if the data are *observed* completely at random.

References

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, *83*(404), 1198–1202. doi: 10.1080/01621459.1988.10478722