Agenda

1. Types of missing data
2. Modeling missing data
3. Estimation in R with brms

Types of missing data

Missing data

Example
Association between test scores

Variable	Mean	Standard deviation	Missing
Math score	530.5	43.1	86
Reading score	509.5	50.0	1409
Listening score	567.5	33.7	128

$n = 6684$

Missing data terminology

Variable	Mean	Standard deviation	Missing
Math score	530.5	43.1	86
Reading score	509.5	50.0	1409
Listening score	567.5	33.7	128

$n = 6684$

Missing completely at random (MCAR): The process that determines which reading scores are missing is independent of everything else.

Missing at random (MAR): The process that determines which reading scores are missing may depend on other covariates, but not on the outcome variable (e.g. students' reading ability).

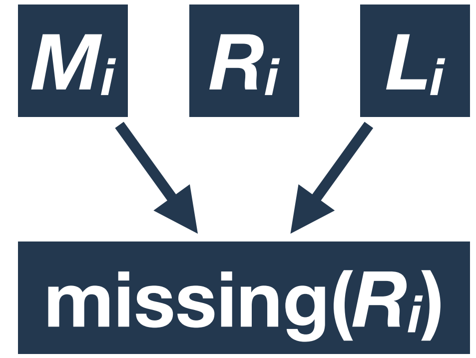
Missing not at random (MNAR): The process that determines which reading scores are missing *may* depend on the outcome variable (students' reading ability).

Missing data terminology

Missing completely at random (MCAR)

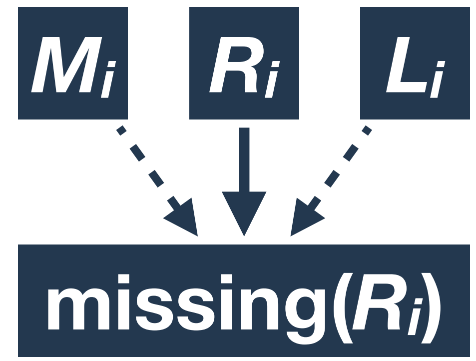
E.g. reading test administered to random subset of students.

Missing at random (MAR)



E.g. students with high listening scores could opt out of reading test.

Missing not at random (MNAR)



E.g. students with documented reading difficulties exempted from reading test.

Missing data in practice

Predicting math scores

$$MS_i \sim \text{Norm}(\mu_i, \sigma)$$
$$\mu_i = \beta_0 + \beta_1 RS_i + \beta_2 LS_i$$

$$\beta_0 \sim \text{Norm}(0, 5)$$
$$\beta_1 \sim \text{Norm}(0, 5)$$
$$\beta_2 \sim \text{Norm}(0, 5)$$
$$\sigma \sim \text{HalfCauchy}(0, 3)$$

MCAR

If reading scores are missing completely at random, we can simply drop incomplete cases with no risk of biasing our estimates.

MAR

If the missingness of reading scores depends on listening scores, we may be safe dropping incomplete cases *if* listening scores are included as a covariate, *and* the missingness does not lead to too-sparse of data.

MNAR

If the missingness of reading scores depends on student reading ability itself, dropping incomplete rows is almost certain to induce bias.

Testing type of missingness

Distinguishing MCAR, MAR, & MNAR

MCAR *can* be distinguished statistically

- A standard logistic regression can be used as a *partial* test for data missing completely at random.
- Create an indicator variable for the missing values; predict using ‘all’ relevant covariates.
- *Cannot* account for unobserved correlates.

MAR and MNAR *cannot* be distinguished statistically

- No quantitative way to tell whether a variable’s missingness depends on the value of the variable itself.
- Regardless of MAR or MNAR, imputation of missing values is often a good idea.

Modeling missing data

Modeling missing data

Data model

$$MS_i \sim \text{Norm}(\mu_i, \sigma)$$

$$\mu_i = \beta_0 + \beta_1 \boxed{RS_i} + \beta_2 LS_i$$

$$\beta_0 \sim \text{Norm}(0, 5)$$

$$\beta_1 \sim \text{Norm}(0, 5)$$

$$\beta_2 \sim \text{Norm}(0, 5)$$

$$\sigma \sim \text{HalfCauchy}(0, 3)$$

Missing data model

$$\boxed{RS_i} \sim \text{Norm}(m_i, s)$$

$$m_i = \alpha_0 + \alpha_1 LS_i$$

$$\alpha_0 \sim \text{Norm}(0, 5)$$

$$\alpha_1 \sim \text{Norm}(0, 5)$$

$$s \sim \text{HalfCauchy}(0, 3)$$

Modeling missing data

Data model	Missing data model
$MS_i \sim \text{Norm}(\mu_i, \sigma)$	$RS_i \sim \text{Norm}(m_i, s)$
$\mu_i = \beta_0 + \beta_1 RS_i + \beta_2 LS_i$	$m_i = a_0 + a_1 LS_i$
$\beta_0 \sim \text{Norm}(0, 5)$	$a_0 \sim \text{Norm}(0, 5)$
$\beta_1 \sim \text{Norm}(0, 5)$	$a_1 \sim \text{Norm}(0, 5)$
$\beta_2 \sim \text{Norm}(0, 5)$	$s \sim \text{HalfCauchy}(0, 3)$
$\sigma \sim \text{HalfCauchy}(0, 3)$	

Multiple imputation

Use missing data model to guess missing values of RS_i. Do this multiple times, creating multiple versions of the dataset.

Estimate the data model on *each* of these datasets.

Combine the results from all analyses to get (*hopefully*) unbiased estimates of β_1 and β_2.

Modeling missing data

Data model	Missing data model
$MS_i \sim \text{Norm}(\mu_i, \sigma)$	$RS_i \sim \text{Norm}(m_i, s)$
$\mu_i = \beta_0 + \beta_1 RS_i + \beta_2 LS_i$	$m_i = a_0 + a_1 LS_i$
$\beta_0 \sim \text{Norm}(0, 5)$	$a_0 \sim \text{Norm}(0, 5)$
$\beta_1 \sim \text{Norm}(0, 5)$	$a_1 \sim \text{Norm}(0, 5)$
$\beta_2 \sim \text{Norm}(0, 5)$	$s \sim \text{HalfCauchy}(0, 3)$
$\sigma \sim \text{HalfCauchy}(0, 3)$	

Model-based (Bayesian) imputation

Estimate the data model and the missing data model simultaneously.

Missing values of RS_i are treated as parameters, each with a ‘prior’ defined by the missing data model, and each with its own estimated posterior distribution.

(In essence, perform a new imputation for each step in the HMC chain)

Estimation in R with brms

Modelling missing data in brms

```
m <- bf(math_score ~ mi(reading_score) + listening_score) +
     bf(reading_score | mi() ~ listening_score)

fit_imputed <- brm(m,data=d)
```

Modelling missing data in brms

Data model

Combining models

```
m <- bf(math_score ~ mi(reading_score) + listening_score) +
     bf(reading_score | mi() ~ listening_score)

fit_imputed <- brm(m,data=d)
```

Missing data model

Modelling missing data in brms

`mi()` indicates imputed variable.

```
m <- bf(math_score ~ mi(reading_score) + listening_score) +
     bf(reading_score | mi() ~ listening_score)

fit_imputed <- brm(m,data=d)
```

`reading_score` contains missing and observed values.

Modelling missing data in brms

Priors are set for both models, using the `resp` argument to specify dependent variable.

```
pr <- c(
    prior(normal(0,5),class=b,        resp=mathscore),
    prior(cauchy(0,3),class=sigma,    resp=mathscore),
    prior(normal(0,5),class=b,        resp=readingscore),
    prior(cauchy(0,3),class=sigma,    resp=readingscore)
)
```

```
m <- bf(math_score ~ mi(reading_score) + listening_score) +
     bf(reading_score | mi() ~ listening_score)

fit_imputed <- brm(m, data=d, prior=pr)
```