Agenda

1. Presentations overview
2. Frequentist estimation of multilevel models using ‘lme4’
3. Using data with oversampled populations
4. Handling missing data

Presentations

Format | 20 slides, automatically advancing every 20 seconds.
(Practice!)

Slot	Tue, April 9	Thu, April 11
1	Yildirim, Irem	Moloney, Kate
2	McCormack, Andrew	Hequet, Céline
3	Traves, Samantha	Nossek, Sean
4	Jutras, Kevin	Yang, Winnie
5	Carter-Rau, Rohan	Lee, Martha
6	Song, Sumin	Gounden Rock, Alyson
7	Amsden, Ryan	Zhao, Qiao
8	Jeong, Tay	Ng, Ka U
9	Isaac, Maike	Zhou, Lingyu
10	Moody, Alayne	

lme4

Overview

lme4 is the ‘standard’ R package for estimating mixed-effects models. It uses a frequentist approach, finding maximum-likelihood estimates for model parameters and approximating standard errors.

Benefits

When lme4 can estimate a model, it tend to do so *much* faster than brms (minutes instead of hours).

Drawbacks

brms can estimate a *much* broader set of models (e.g. zero-inflated Poisson models) and can handle *many* more scenarios (e.g. missing data imputation).

Furthermore, lme4 fails to ‘converge’ on many models, and troubleshooting is difficult.

Finally, lme4 cannot incorporate prior distributions on parameters.

lme4

Define a random-intercept model

```
m <- listening_score ~
     female + (1 | teacher_id)
```

Fit using brms

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

Fit using lmer

For generalized linear models (logistic, poisson, etc.), use the ‘glmer’ or ‘glmer.nb’ functions.

```
library(lme4)
fit_lmer <- lmer(m,data=d)
```

Oversampling

The problem

A truly uniform sample from a population may not include enough cases from smaller groups for meaningful analysis. This is especially true for intersecting categories (e.g. Asian students with Black teachers).

Full sample

White	4440
Black	2191
Asian	20
Hispanic	9
Native American	9
Other	11

~5% subsample

White	225
Black	101
Asian	1
Hispanic	1
Native American	0
Other	0

Oversampling

The solution

Deliberately sample populations you know to be small with higher probability. In this case, we could sample 3% of white students, 6% of Black students, and 100% of remaining students.

Full sample

White	4440
Black	2191
Asian	20
Hispanic	9
Native American	9
Other	11

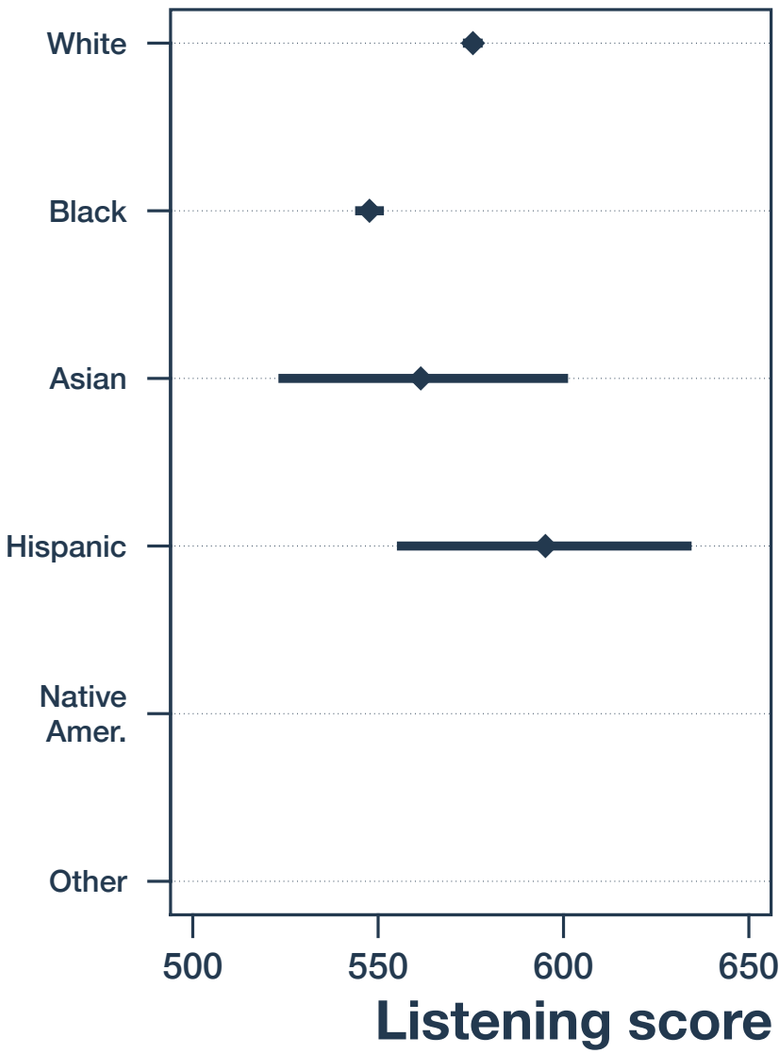
~5% subsample (with oversampling)

White	139
Black	140
Asian	20
Hispanic	9
Native American	9
Other	11

Oversampling

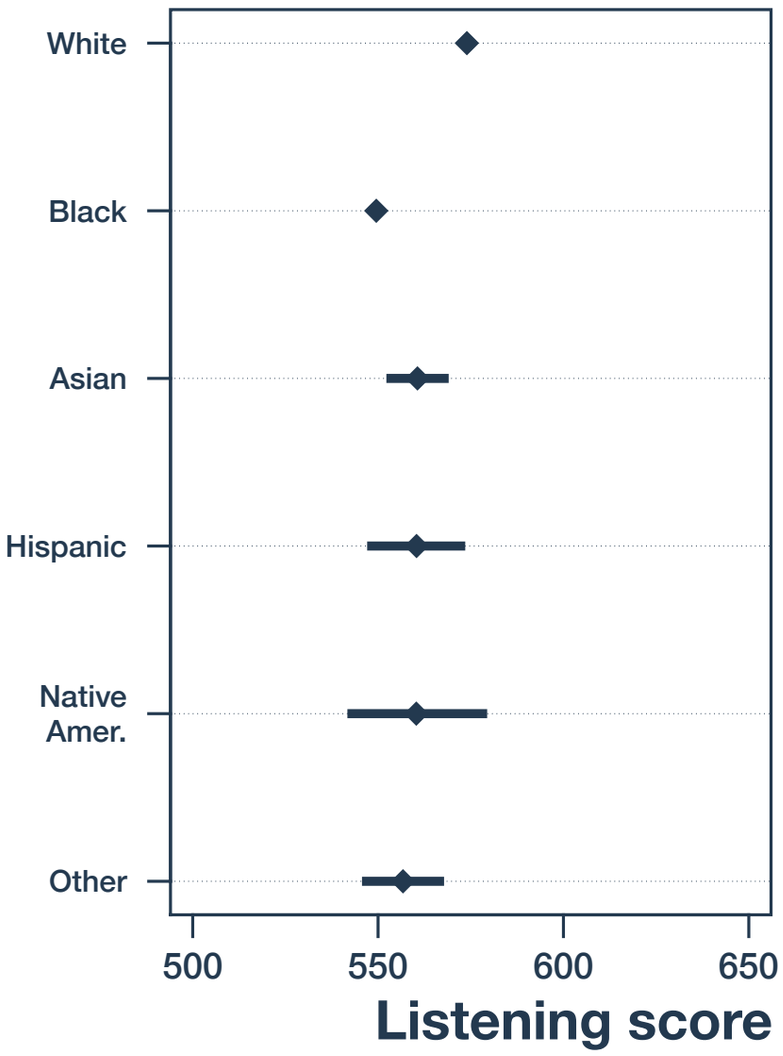
~5% subsample

White	225
Black	101
Asian	1
Hispanic	1
Native American	0
Other	0



~5% subsample
(with oversampling)

White	139
Black	140
Asian	20
Hispanic	9
Native American	9
Other	11



Using oversampled data

Sampling weights tell us how many cases this data point represents in the population.

ID	listening_score	race_ethnicity	s_w
4	556	Black	16.66667
20	—	Hispanic	1.00000
43	568	Other	1.00000
60	531	White	33.33333
86	592	White	33.33333
122	611	Asian	1.00000
⋮	⋮	⋮	⋮

Using oversampled data

```
listening_score | weights(s_w)~
    re_black + re_asian + re_hispanic +
    re_native_american + re_other
```

Sampling weights are indicated in brms with a pipe (‘|’) after your outcome variable and the special “weights” function that indicates the variable containing case weights (in our case, ‘s_w’).

This tells brms to multiply the likelihood for each case by that case’s value of s_w.

Missing data

Example
Test score association

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 students' reading ability.

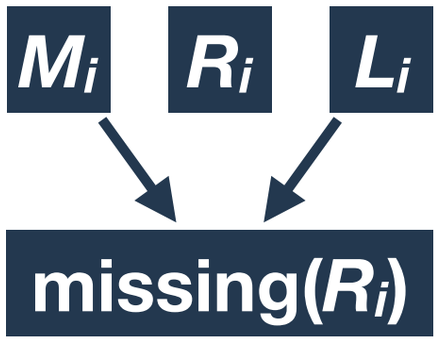
Missing not at random (MNAR) | The process that determines which reading scores are missing may depend on students' reading ability.

Missing data terminology

Missing completely at random (MCAR)

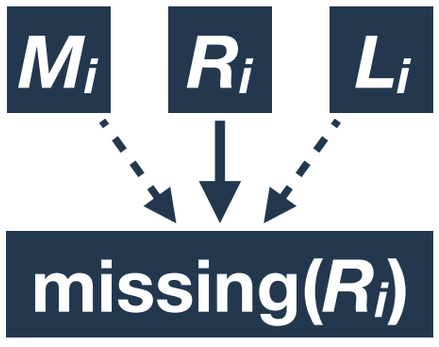
E.g. reading test administered to random subset of students.

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

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

$$\mu_i = \beta_0 + \beta_1 RS_i + \beta_2 LS_i$$

$$\beta_0 \sim \text{Norm}(500, 100)$$

$$\beta_1 \sim \text{Norm}(0, 50)$$

$$\beta_2 \sim \text{Norm}(0, 50)$$

$$\sigma \sim \text{HalfCauchy}(0, 50)$$

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 math or listening scores, we may be safe dropping incomplete cases *unless* the missingness limits leads to sparse data.

MNAR

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

Modelling missing data

Data model

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

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

$$\beta_0 \sim \text{Norm}(500, 100)$$

$$\beta_1 \sim \text{Norm}(0, 50)$$

$$\beta_2 \sim \text{Norm}(0, 50)$$

$$\sigma \sim \text{HalfCauchy}(0, 50)$$

Missing data model

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

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

$$\alpha_0 \sim \text{Norm}(500, 100)$$

$$\alpha_1 \sim \text{Norm}(0, 50)$$

$$s \sim \text{HalfCauchy}(0, 50)$$

Modelling missing data

Data model

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

$$\mu_i = \beta_0 + \beta_1 RS_i + \beta_2 LS_i$$

$$\beta_0 \sim \text{Norm}(500, 100)$$

$$\beta_1 \sim \text{Norm}(0, 50)$$

$$\beta_2 \sim \text{Norm}(0, 50)$$

$$\sigma \sim \text{HalfCauchy}(0, 50)$$

Missing data model

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

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

$$\alpha_0 \sim \text{Norm}(500, 100)$$

$$\alpha_1 \sim \text{Norm}(0, 50)$$

$$s \sim \text{HalfCauchy}(0, 50)$$

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 unbiased estimates of β_1 and β_2.

Modelling missing data

Data model

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

$$\mu_i = \beta_0 + \beta_1 RS_i + \beta_2 LS_i$$

$$\beta_0 \sim \text{Norm}(500, 100)$$

$$\beta_1 \sim \text{Norm}(0, 50)$$

$$\beta_2 \sim \text{Norm}(0, 50)$$

$$\sigma \sim \text{HalfCauchy}(0, 50)$$

Missing data model

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

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

$$\alpha_0 \sim \text{Norm}(500, 100)$$

$$\alpha_1 \sim \text{Norm}(0, 50)$$

$$s \sim \text{HalfCauchy}(0, 50)$$

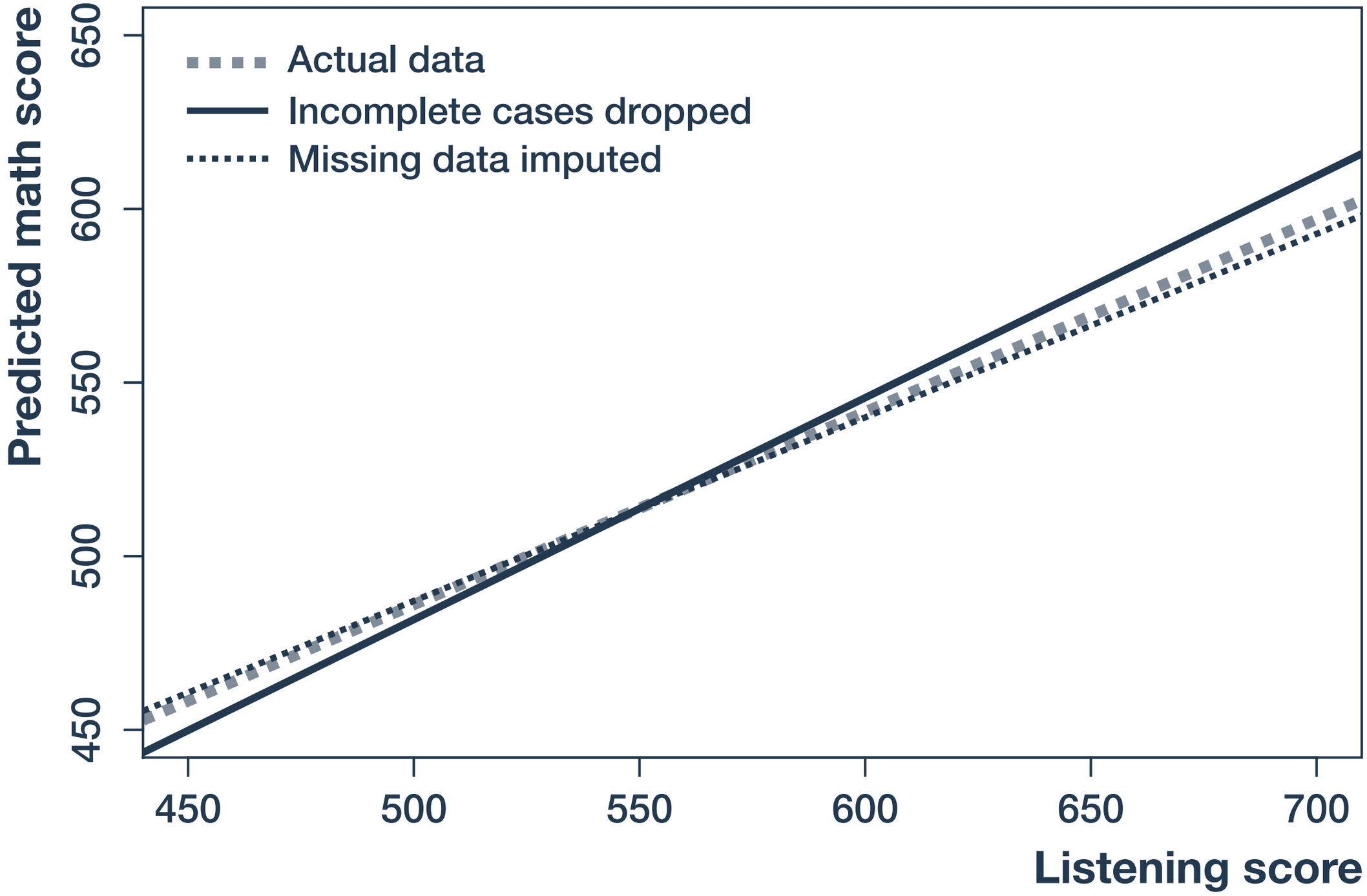
Model-based (Bayesian) imputation

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)

Modelling missing data



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

`bf` is short for `brmsformula`.
Used when combining multiple formulas.

```
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

Data model

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

$$\mu_i = \beta_0 + \beta_1 RS_i + \beta_2 LS_i$$

$$\beta_0 \sim \text{Norm}(500, 100)$$

$$\beta_1 \sim \text{Norm}(0, 50)$$

$$\beta_2 \sim \text{Norm}(0, 50)$$

$$\sigma \sim \text{HalfCauchy}(0, 50)$$

Missing data model

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

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

$$\alpha_0 \sim \text{Norm}(500, 100)$$

$$\alpha_1 \sim \text{Norm}(0, 50)$$

$$s \sim \text{HalfCauchy}(0, 50)$$

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

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