

# 2042 MLM Mini Project (Spring 2020)

Group 1

May 14 2020

## Team members and division of work

Frank Jiang, Lisa Song, Yuyue Hua, Seeun Jang, Tong Jin

```
set.seed(2042001)
```

## Question 1

You will generate simulated data for a single school with 100 classrooms, each of which has 200 students.

- Outcome for student  $i$  in classroom  $j$ :  $Y_{ij}$ .
- There is a single predictor,  $X_{ij} \sim U(0, 1)$  (uniform on  $[0, 1]$ )
- There is a classroom random effect,  $\eta_j \sim N(0, \sigma_\eta^2)$ , where  $\sigma_\eta^2 = 2$ .
- Subject level error,  $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$ , where  $\sigma_\varepsilon^2 = 2$ .
- `set.seed(2042001)` once at the beginning of your code.
- Generate the random quantities in this order to ensure the same solution for everyone:  $X$ ,  $\eta_j$ ,  $\varepsilon_{ij}$
- The outcome has the following form (DGP, given the modeling parameters above):

$$Y_{ij} = 0 + 1X_{ij} + \eta_j + \varepsilon_{ij}, \eta_j \sim N(0, \sigma_\eta^2), \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2), \text{indep.}$$

- Generate a single simulated dataset (you will need a “classid” variable to track classrooms); you can optionally assign a “studentid”)
- Important:** construct classid such that classrooms appear consecutively within the dataframe. As per: `rep(1:J, each=n_j)`

```
# Compute variables related to the equation
N_j <- 100 # Total number of classrooms
n_i <- 200 # Number of students in each classroom
N_i <- N_j * 200 # Total number of students

X_ij <- runif( # Single predictor: Uniform on [0, 1]
  N_i,
  min = 0,
  max = 1
)
```

```

eta_j <- rnorm( # Classroom random effect: Normal on (0, 2)
  N_j,
  mean = 0,
  sd = sqrt(2)
)
epsilon_ij <- rnorm( # Subject level error: Normal on (0, 2)
  N_i,
  mean = 0,
  sd = sqrt(2)
)

# Create equation elements
eta_j <- rep( # Assign classroom random effect to students in each classroom
  eta_j,
  each = n_i
)
Y_ij <- 0 + 1 * X_ij + eta_j + epsilon_ij

classid <- rep(
  1:N_j,
  each = n_i
)
studentid <- seq(
  1:N_i
)

# Create a dataframe to store all elements
classroom_sim <- data.frame(
  outcome = Y_ij,
  predictor = X_ij,
  cls_raneff = eta_j,
  subject_error = epsilon_ij,
  classid = classid,
  studentid = studentid,
  row.names = studentid
)

```

## Question 2

Fit the model corresponding to the DGP on your simulated data.

```

# Fit the model in Q1
fit_q2 <- lmer(
  outcome ~ predictor + (1 | classid),
  data = classroom_sim
)

# Report the model fit
# summary(fit_q2)

# Calculate the coefficient estimate of slope on X
slope_X_q2 <- round(
  summary(fit_q2)$coefficients['predictor', 'Estimate'],

```

```
digits = 4
)
```

- a. Report coefficient estimate for slope on X.

**Response:** 0.9864.

- b. Does a 95% confidence band for this coefficient estimate cover the “truth” that you used to generate the data?

```
CI_Q2_b <- confint(
  fit_q2,
  parm = "predictor",
  level = 0.95
)
lower <- round(
  CI_Q2_b[1],
  digits = 4
)
upper <- round(
  CI_Q2_b[2],
  digits = 4
)
```

**Response:** Yes. The confidence band (95% of confidence) for this coefficient estimate is between 0.9179 and 1.0549. This confirms that a 95% confidence band covers the “truth” of the slope, which is 1.

### Question 3

3. Next, we simulate missing data in several ways. This is the first:

- a. Make a copy of the data, then modify the copy following these instructions:

```
classroom_miss <- classroom_sim
```

- b. Generate  $Z_{ij} \sim \text{Bernoulli}(p)$ , with  $p = 0.5$ .

```
Z_ij <- rbinom(
  N_i,
  size = c(0, 1),
  prob = 0.5
)
```

- c. Set  $Y_{ij}$  to NA when  $Z_{ij} == 1$ . This should look a lot like “MCAR” missingness.

```
classroom_miss$missing <- Z_ij
classroom_miss$outcome <- ifelse(
  classroom_miss$missing == 1,
  yes = NA,
  no = classroom_miss$outcome
)
```

- d. Refit the model on the new data and report the coefficient estimate for slope on X. Look at the other parameter estimates as well.

```
# Refit the model using missing data
fit_q3 <- lmer(
  outcome ~ predictor + (1 | classid),
  data = classroom_miss
)
# Report the model
# summary(fit_q3)

# Calculate the new coefficient estimate of slope on X
slope_X_q3 <- round(
  summary(fit_q3)$coefficients['predictor', 'Estimate'],
  digits = 4
)

# Calculate the confidence band of this model
CI_Q3_d <- confint(
  fit_q3,
  parm = "predictor",
  level = 0.95
)
```

```
## Computing profile confidence intervals ...
```

```
lower <- round(
  CI_Q3_d[1],
  digits = 4
)
upper <- round(
  CI_Q3_d[2],
  digits = 4
)
```

**Response:** 0.9612.

- e. Do you see any real change in the  $\beta_X$  estimate?

**Response:** No. The  $\beta_X$  estimate of the model with missing data values is 0.9612, which is very close to the original estimate, 0.9864.

- i. Does a 95% confidence band for this coefficient estimate cover the “truth” that you used to generate the data?

**Response:** Yes. The confidence band (95% of confidence) for this coefficient estimate is between 0.8822 and 1.0402. This confirms that a 95% confidence band covers the “truth” of the slope, which is 1.

- f. What is the total sample size  $N$  used in the model fit?

**Response:** The total sample size used in this model fit is 14963.

#### Question 4

Missing Data II: Make another copy of the original data, then modify the copy as follows:

```
classroom_miss2 <- classroom_sim
```

- a. Generate  $Z_{ij} \sim \text{Bernoulli}(X_{ij})$ , with  $X_{ij}$  your predictor generated previously.

```
Z_ij <- rbinom(  
  N_i,  
  size = c(0, 1),  
  prob = X_ij  
)
```

- b. Set Y to NA when  $Z_{ij} == 1$ . This should look a lot like “MAR” missingness.

```
classroom_miss2$missing <- Z_ij  
classroom_miss2$outcome <- ifelse(  
  classroom_miss2$missing == 1,  
  yes = NA,  
  no = classroom_miss2$outcome  
)
```

- c. Refit the model on the new data and report the coefficient estimate for slope on X. Look at the other parameter estimates as well.

```
# Refit the model using missing data  
fit_q4 <- lmer(  
  outcome ~ predictor + (1 | classid),  
  data = classroom_miss2  
)  
# Report the model  
# summary(fit_q4)  
  
# Calculate the new coefficient estimate of slope on X  
slope_X_q4 <- round(  
  summary(fit_q4)$coefficients['predictor', 'Estimate'],  
  digits = 4  
)  
  
# Calculate the confidence band of this model  
CI_Q4_d <- confint(  
  fit_q4,  
  parm = "predictor",  
  level = 0.95  
)  
lower <- round(  
  CI_Q4_d[1],  
  digits = 4  
)  
upper <- round(  
  CI_Q4_d[2],  
  digits = 4  
)
```

**Response:** 0.9842.

- d. Do you see any real change in the  $\beta_X$  estimate? Does a 95% confidence band for this coefficient estimate cover the “truth” that you used to generate the data?

**Response:** No. The  $\beta_X$  estimate of the model with missing data values is 0.9842, which is very close to the original estimate, 0.9864.

Yes. The confidence band (95% of confidence) for this coefficient estimate is between 0.9038 and 1.0645. This confirms that a 95% confidence band covers the “truth” of the slope, which is 1.

- e. What is the total sample size  $N$  used in the model fit?

**Response:** The total sample size used in this model fit is 14960.

### Question 5

Missing Data III: Make another copy of the original data, then modify the copy as follows:

```
classroom_miss3 <- classroom_sim
```

- a. First, define the `expit` function: `expit <- function(x) exp(x)/(1+exp(x))`

```
expit <- function(x) {  
  exp(x)/(1 + exp(x))  
}
```

- b. Generate  $Z_{ij} \sim \text{Bernoulli}(\text{expit}(Y_{ij}))$ , with  $Y_{ij}$  your *outcome* generated previously.

```
Z_ij <- rbinom(  
  N_i,  
  size = c(0, 1),  
  prob = expit(Y_ij)  
)
```

- c. Set  $Y$  to NA when  $Z_{ij} == 1$ . This should look like a violation of “MAR” missingness (missingness depends on outcome and cannot be *simply* predicted with the predictor set –  $Y$  should be correlated with  $X$ , though, so it might not be too bad a violation).

```
classroom_miss3$missing <- Z_ij  
classroom_miss3$outcome <- ifelse(  
  classroom_miss3$missing == 1,  
  yes = NA,  
  no = classroom_miss3$outcome  
)
```

- d. Refit the model on the new data and report the coefficient estimate for slope on  $X$ . Look at the other parameter estimates as well.

```

# Refit the model using missing data
fit_q5 <- lmer(
  outcome ~ predictor + (1 | classid),
  data = classroom_miss3
)
# Report the model
# summary(fit_q5)

# Calculate the new coefficient estimate of slope on X
slope_X_q5 <- round(
  summary(fit_q5)$coefficients['predictor', 'Estimate'],
  digits = 4
)

# Calculate the confidence band of this model
CI_Q5_d <- confint(
  fit_q5,
  parm = "predictor",
  level = 0.95
)
lower <- round(
  CI_Q5_d[1],
  digits = 4
)
upper <- round(
  CI_Q5_d[2],
  digits = 4
)

```

**Response:** 0.9429.

e. Do you see any real change in the  $\beta_X$  estimate?

**Response:** No. The  $\beta_X$  estimate of the model with missing data values is 0.9429, which is very close to the original estimate, 0.9864.

Yes. The confidence band (95% of confidence) for this coefficient estimate is between 0.8619 and 1.0239. This confirms that a 95% confidence band covers the “truth” of the slope, which is 1.

e. What is the total sample size  $N$  used in the model fit?

**Response:** The total sample size used in this model fit is 14253.