MLM Mini Project

May 13 2020

Team Members and division of work:

- Bilal Waheed, Dennis Hilgendorf, Trey Dellucci, Joe Marlo, Yi-Hung Wang

```
# Insert code to set.seed
set.seed(2042001)
```

Question 1:

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

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

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

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

```
# Insert code to generate data and outcome variable, store variables in a
# dataframe

# set size assumptions
n.classrooms <- 100
n.stu.per.class <- 200

# generate data
X_ij <- runif(n.classrooms * n.stu.per.class, min = 0, max = 1)
eta_j <- rnorm(n.classrooms, mean = 0, sd = sqrt(2))
epsilon_ij <- rnorm(n.classrooms * n.stu.per.class, mean = 0, sd = sqrt(2))

# calculate outcome variable
Y_ij <- 0 + 1 * X_ij + rep(eta_j, each = n.stu.per.class) + epsilon_ij

# store variables in dataframe
dat <- data.frame(studentid = 1:(n.classrooms * n.stu.per.class), classid = rep(1:n.classrooms,
    each = n.stu.per.class), predictor = X_ij, outcome = Y_ij)
```

Question 2:

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

```r
# Insert code to fit model and print summary
lm1 <- lmerTest::lmer(outcome ~ predictor + (1 | classid), data = dat, REML = TRUE)
summary(lm1)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: outcome ~ predictor + (1 | classid)
##    Data: dat
##
## REML criterion at convergence: 71227.3
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.0143 -0.6761  0.0024  0.6711  3.7584
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  classid  (Intercept) 1.893    1.376
##  Residual             2.008    1.417
## Number of obs: 20000, groups:  classid, 100
##
## Fixed effects:
##                 Estimate   Std. Error           df t value            Pr(>|t|)
## (Intercept)    -0.007493     0.139072   102.234247  -0.054               0.957
## predictor       0.986417     0.034959 19900.411745  28.216 <0.0000000000000002
##
## (Intercept)
## predictor   ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##           (Intr)
## predictor -0.126
```

a. Report coefficient estimate for slope on X.

   Response: The coefficient estimate for the slope on X is 0.986.

b. Does a 95% confidence band for this coefficient estimate cover the "truth" that you used to generate the data? **comment**

```r
# Insert code to compute confidence interval
coefs <- summary(lm1)$coefficients
lower <- coefs[2, 1] - (coefs[2, 2] * 2)
upper <- coefs[2, 1] + (coefs[2, 2] * 2)
```

```
Response: Yes, the 95% confidence bound of [0.916, 1.056] covers the truth of 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:

```r
# Insert code to make a copy of the data
dat2 <- dat
```

b. Generate $Z_{ij} \sim \text{Bernoulli}(p)$, with $p = 0.5$
c. Set Y to NA when $Z_{ij} == 1$. This should look a lot like "MCAR" missingness.

```r
# Insert code the generate your data
Z_ij <- rbinom(n = n.classrooms * n.stu.per.class, size = 1, prob = 0.5)
dat2$outcome <- ifelse(Z_ij == 1, NA, dat2$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.

```r
# Insert code to fit model and compute confidence interval
lme2 <- lmerTest::lmer(outcome ~ predictor + (1 | classid), data = dat2, na.action = "na.omit")
summary(lme2)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: outcome ~ predictor + (1 | classid)
##    Data: dat2
##
## REML criterion at convergence: 35607.1
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.9102 -0.6698  0.0146  0.6663  3.8709
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  classid  (Intercept) 1.880    1.371
##  Residual             2.007    1.417
## Number of obs: 9945, groups:  classid, 100
##
## Fixed effects:
##               Estimate Std. Error         df t value            Pr(>|t|)
## (Intercept)   -0.02359    0.14005  105.47627  -0.168               0.867
## predictor      1.02485    0.04963 9846.41935  20.649 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##           (Intr)
## predictor -0.177
```

```r
# calculate confidence band
coefs <- summary(lme2)$coefficients
lower <- coefs[2, 1] - (coefs[2, 2] * 2)
upper <- coefs[2, 1] + (coefs[2, 2] * 2)
```

`Response: The coefficient estimate for the slope on X is 1.025.`

e. Do you see any real change in the β_X estimate? **comment**

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

   Response: The β_X estimate has increased slightly to 1.025, and the 95% confidence band as also slightly widened to [0.926, 1.124] but still covers the truth of 1. It does not appear to be a meaningful change from the original model.

f. What is the total sample size N used in the model fit? **comment**

   Response: The total sample size N is 9,945 which is expected as $p = 0.5$.

Question 4:

Missing Data II: Make another copy of the original data, then modify the copy as follows: a. Generate $Z_{ij} \sim$ Bernoulli(X_{ij}), with X_{ij} your predictor generated previously. b. Set Y to NA when $Z_{ij} == 1$. This should look a lot like "MAR" missingness.

```r
# Insert code the generate your data
dat3 <- dat
Z_ij <- rbinom(n = n.classrooms * n.stu.per.class, size = 1, prob = dat3$predictor)
dat3$outcome <- ifelse(Z_ij == 1, NA, dat3$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. **comment**

```r
# Insert code to fit model and compute confidence interval
lme3 <- lmerTest::lmer(outcome ~ predictor + (1 | classid), data = dat3, na.action = "na.omit")
summary(lme3)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: outcome ~ predictor + (1 | classid)
##    Data: dat3
##
## REML criterion at convergence: 35850.3
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.8356 -0.6795  0.0052  0.6608  3.7058
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  classid  (Intercept) 1.874    1.369
##  Residual             2.015    1.420
## Number of obs: 10002, groups:  classid, 100
##
## Fixed effects:
##                Estimate  Std. Error          df t value            Pr(>|t|)
## (Intercept)    0.003442    0.139129  103.428326   0.025                0.98
## predictor      0.954720    0.060306 9903.323597  15.831 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##           (Intr)
## predictor -0.147
```

```r
# calculate confidence band
coefs <- summary(lme3)$coefficients
lower <- coefs[2, 1] - (coefs[2, 2] * 2)
upper <- coefs[2, 1] + (coefs[2, 2] * 2)
```

Response: The coefficient estimate for the slope on X is 0.955. Power has been reduced as our sample size is half of the original. The random effects are mostly unchanged.

d. Do you see any real change in the β_X estimate?

   i. Does a 95% confidence band for this coefficient estimate cover the "truth" that you used to generate the data? **comment**

   Response: The β_X estimate has decreased to 0.955, and the 95% confidence band as also widened to [0.834, 1.075] but still covers the truth of 1.

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

   Response: The total sample size N is 10,002 which is expected as the mean of our predictor (which is what the MAR process is dependent on) is 0.502.

Question 5:

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

```
# Insert code to make a copy of the original data
dat4 <- dat
```

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

```
# Insert code to define expit function
expit <- function(x) exp(x)/(1 + exp(x))
```

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

c. Set Y to NA when $Z_{ij} == 1$. This should look like a violation of "MAR" missingness (missingness depedents 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).

```
# Insert code the generate your data
Z_ij <- rbinom(n = n.classrooms * n.stu.per.class, size = 1, prob = expit(dat4$outcome))
dat4$outcome <- ifelse(Z_ij == 1, NA, dat4$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. **comment**

```
# Insert code to fit model and compute confidence interval
lme4 <- lmerTest::lmer(outcome ~ predictor + (1 | classid), data = dat4, na.action = "na.omit")
summary(lme4)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: outcome ~ predictor + (1 | classid)
##    Data: dat4
##
## REML criterion at convergence: 28257.5
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.0870 -0.6596  0.0090  0.6679  3.1897
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  classid  (Intercept) 1.078    1.038
##  Residual             1.539    1.240
## Number of obs: 8522, groups:  classid, 100
```

```
##
## Fixed effects:
##              Estimate Std. Error        df t value             Pr(>|t|)
## (Intercept)   -0.7488     0.1074  105.0594  -6.972       0.000000000286 ***
## predictor      0.7069     0.0475 8423.2269  14.881 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##           (Intr)
## predictor -0.208
```

```r
# calculate confidence band
coefs <- summary(lme4)$coefficients
lower <- coefs[2, 1] - (coefs[2, 2] * 2)
upper <- coefs[2, 1] + (coefs[2, 2] * 2)
```

Response: The coefficient estimate for the slope on X is 0.707. Power has been reduced as our sample size is less than half of the original. The random effects have decreased dramatically.

e. Do you see any real change in the β_X estimate? **comment**

   i. Does a 95% confidence band for this coefficient estimate cover the "truth" that you used to generate the data? **comment**

   Response: The β_X estimate has decreased signficantly to 0.707, and the 95% confidence band as also widened to [0.612, 0.802] and does not cover the truth of 1.

f. What is the total sample size N used in the model fit? **comment**

   Response: The total sample size N is 8,522 which corresponds to $= 1 - mean(expit(outcome))$.