MLM Mini Project

May 06 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, $\eta_j \sim N(0, \sigma^2_\eta)$, where $\sigma^2_\eta = 2$.
d. Subject level error, $\varepsilon_{ij} \sim N(0, \sigma^2_\varepsilon)$, where $\sigma^2_\varepsilon = 2$.
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):

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

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 = 2)
epsilon_ij <- rnorm(n.classrooms * n.stu.per.class, mean = 0, sd = 2)

# calculate outcome variable REVIEW THIS DGP !!!!!!!!!!!
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.stu.per.class,
    each = n.classrooms), predictor = X_ij, outcome = Y_ij)
```

Question 2: REVIEW THIS MODEL !!!!!!!!!!!

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

```
# 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: 85447.3
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.9119 -0.6757  0.0004  0.6679  3.9138
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  classid  (Intercept) 3.773    1.942
##  Residual             4.010    2.002
## Number of obs: 20000, groups:  classid, 200
##
## Fixed effects:
##                 Estimate   Std. Error           df t value            Pr(>|t|)
## (Intercept)    -0.009762     0.140309   212.153529   -0.07               0.945
## predictor       0.979130     0.049576 19804.740987   19.75 <0.0000000000000002
##
## (Intercept)
## predictor   ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##           (Intr)
## predictor -0.177
```

a. Report coefficient estimate for slope on X.

   Response: The coefficient estimate for the slope on X is 0.979.

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

```
# 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)
# print(paste0('The 95% confidence band is [', round(lower, 3), ', ',
# round(upper, 3), ']'))
```

```
Response: Yes, the 95% confidence bound of [0.88, 1.078] 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)
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: 42804.3
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.9127 -0.6611  0.0144  0.6574  3.8739
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  classid  (Intercept) 3.750    1.936
##  Residual             4.008    2.002
## Number of obs: 9945, groups:  classid, 200
##
## Fixed effects:
##                Estimate Std. Error         df t value            Pr(>|t|)
## (Intercept)    -0.03563    0.14280  225.49369   -0.25               0.803
## predictor       1.04113    0.07051 9753.28880   14.77 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##           (Intr)
## predictor -0.246
```

```r
# calculate confidence band
coefs <- summary(lme2)$coefficients
lower <- coefs[2, 1] - (coefs[2, 2] * 2)
upper <- coefs[2, 1] + (coefs[2, 2] * 2)
# print(paste0('The 95% confidence band is [', round(lower, 3), ', ',
# round(upper, 3), ']'))
```

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 1.041, and the 95% confidence band as tightened to [0.9, 1.182].

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

```
# number of observations used sum(!is.na(dat2$outcome))
# length(summary(lme2)$residuals)

# number of groups used summary(lme2)$ngrps
```

```
Response: The total sample size $N$ is 9,945.
```

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.

```
# 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**

```
# Insert code to fit model and compute confidence interval
lme3 <- lmerTest::lmer(outcome ~ predictor + (1 | classid), data = dat3)
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: 43056.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.7593 -0.6668  0.0089  0.6592  3.8904
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  classid  (Intercept) 3.726    1.930
##  Residual             4.013    2.003
## Number of obs: 10002, groups:  classid, 200
##
## Fixed effects:
##              Estimate Std. Error         df t value            Pr(>|t|)
## (Intercept)    0.00925    0.14098  217.06476   0.066               0.948
## predictor      0.92519    0.08551 9809.75042  10.820 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##           (Intr)
## predictor -0.206
```

```
# calculate confidence band
coefs <- summary(lme3)$coefficients
lower <- coefs[2, 1] - (coefs[2, 2] * 2)
upper <- coefs[2, 1] + (coefs[2, 2] * 2)
# print(paste0('The 95% confidence band is [', round(lower, 3), ', ',
# round(upper, 3), ']'))
```

```
Response: The coefficient estimate for the slope on X is 0.925.
```

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 slope estimate has decreased along with the confidence band: [0.754, 1.096].

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

```
# number of observations used sum(!is.na(dat3$outcome))
# length(summary(lme3)$residuals)

# number of groups used summary(lme3)$ngrps
```

```
Response: The total sample size $N$ is 10,002.
```

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 $Z_{ij} \sim \text{Bernoulli}(expit(Y_{ij}))$, 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)
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: 34273
##
```

```
## Scaled residuals: 
##     Min      1Q  Median      3Q     Max 
## -4.4580 -0.6522  0.0258  0.6684  3.3833 
## 
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  classid  (Intercept) 1.487    1.219   
##  Residual             2.755    1.660   
## Number of obs: 8741, groups:  classid, 199
## 
## Fixed effects:
##                Estimate Std. Error         df t value             Pr(>|t|)    
## (Intercept)   -1.33795    0.09412  236.61886 -14.215 <0.0000000000000002 ***
## predictor      0.58527    0.06285 8557.52816   9.312 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Correlation of Fixed Effects:
##           (Intr)
## predictor -0.318
```

```
# calculate confidence band
coefs <- summary(lme4)$coefficients
lower <- coefs[2, 1] - (coefs[2, 2] * 2)
upper <- coefs[2, 1] + (coefs[2, 2] * 2)
# print(paste0('The 95% confidence band is [', round(lower, 3), ', ',
# round(upper, 3), ']'))
```

```
Response: The coefficient estimate for the slope on X is 0.585.
```

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 slope estimate has decreased signficantly along with the confidence band: [0.46, 0.711].

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

```
# number of observations used sum(!is.na(dat4$outcome))
# length(summary(lme4)$residuals)

# number of groups used summary(lme4)$ngrps
```

```
Response: The total sample size $N$ is 8,741.
```