MLM Nested Project D

Xinming Dai, Chongjun Liao, Jeremy Lu, Yu Wang

Compiled on Thu May 05 11:26 EDT

Question 1: data generating process

```
set.seed(2042001)
# variance of the random effect
sigma_eta_2 <- 2
sigma_epsilon_2 <- 2
# generate data
dat <-
  tibble(classid = rep(c(1:100), each = 200),
         studentid = 1:(100*200),
         x = runif(100*200, min = 0, max = 1),
         eta_j = rep(rnorm(100, sd = sqrt(sigma_eta_2)), each = 200),
         epsilon = rnorm(100*200, sd = sqrt(sigma_epsilon_2)),
         y = x + eta_j + epsilon)
```

Question 2: fit the model

```
lmer_fit1 <- lmer(y ~ x + (1|classid), data = dat)
summary_lmer_fit1 <- summary(lmer_fit1)
summary_lmer_fit1
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: y ~ x + (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) -7.493e-03  1.391e-01  1.022e+02  -0.054    0.957
## x            9.864e-01  3.496e-02  1.990e+04  28.216   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##   (Intr)
## x -0.126
```

```
estimate_x <- summary_lmer_fit1$coefficients[2, 1]
se_x <- summary_lmer_fit1$coefficients[2, 2]
```

Question 2:

a. The estimated coefficient of X is 0.986.
b. The 95% confidence interval for this coefficient estimate is $[0.986 - 1.96 * 0.035, 0.986 + 1.96 * 0.035] = [0.9179, 1.0549]$. It covers the true coefficient, which is 1.

Question 3:

```
# 3a
dat_copy <- dat
# 3b
Z_Q3 <- rbinom(20000, 1, 0.5)
table(Z_Q3)
```

```
## Z_Q3
##     0     1
##  9945 10055
```

```
# 3c
dat_copy <- dat_copy %>% mutate(y = replace(y, 1:n(), ifelse(Z_Q3==1, NA, y)))
# 3d
lmer_fit_Q3 <- lmer(y ~ x +(1|classid), data = dat_copy)
summary(lmer_fit_Q3)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: y ~ x + (1 | classid)
##    Data: dat_copy
##
## 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 105.47622  -0.168    0.867
## x              1.02485    0.04963 9846.41936   20.649   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##   (Intr)
## x -0.177
```

```r
# 3f
N_Q3 <- nrow(dat)-sum(is.na(dat_copy$y))
N_Q3
```

```
## [1] 9945
```

e.

The estimate coefficient is 1.02, which does not change too much The 95% CI is $[1.02 - 1.96 * 0.05, 1.02 + 1.96 * 0.05]$, which is $[0.92, 1.12]$ almost converges to the true value

f.

The total sample size used in this Question is 9945

Question 4:

```r
# 4a
dat_copy_4 <- dat
z <- rbinom(100*200,1,dat_copy_4$x)
table(z)
```

```
## z
##     0     1
## 10002  9998
```

```r
# 4b
dat_copy_4$y <- ifelse(z==1,NA,dat_copy_4$y)
# 4c
lmer_fit_4 <- lmer(y ~ x + (1|classid), data = dat_copy_4)
summary(lmer_fit_4)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: y ~ x + (1 | classid)
##    Data: dat_copy_4
##
## 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) 3.442e-03  1.391e-01 1.034e+02   0.025     0.98
## x           9.547e-01  6.031e-02 9.903e+03  15.831   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##   (Intr)
## x -0.147
```

d.

i. The 95% confidence interval is [0.837,1.073], which covers the “truth”.

e.

```
N <- nrow(dat)-sum(is.na(dat_copy_4$y))
```

We use N = 10002 samples in the model fit.

Question 5:

```
dat_copy_5 <- dat
### a
expit <- function(x){exp(x)/(1+exp(x))}

### b
z <- rbinom(100*200,1, expit(dat_copy_5$y))
table(z)
```

```
## z
##     0     1
##  8522 11478
```

```
### c
dat_copy_5$y <- ifelse(z==1,NA,dat_copy_5$y)

### d
lmer_fit_5 <- lmer(y ~ x + (1|classid), data = dat_copy_5)
summary(lmer_fit_5)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: y ~ x + (1 | classid)
##    Data: dat_copy_5
##
## 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 2.86e-10 ***
## x              0.7069     0.0475 8423.2269  14.881  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##   (Intr)
## x -0.208
```

The new estimate for slope is 0.707.

e

The 95% confidence interval is [0.614,0.8], which does not cover the “truth”, besides the intercept also change.

f

The total sample size is 8522, based on number of observations.