

Part3-B

Bianca Brusco

4/28/2018

Part 3: Bianca

Create person-period file

In this part of the project, no variables of interested have missing observation. Therefore, the full dataset is used.

```
#new variables
classroom2 <- classroom2 %>% mutate(math0 = mathkind) %>% mutate(math1 = mathkind+mathgain)
#reshape the data
class_pp <- reshape(classroom2, varying = c("math0", "math1"), v.names = "math", timevar = "year",
times = c(0, 1), direction = "long")
```

Note: we ignore classroom in this analysis but keep it in the notation.

Initial longitudinal model

We fit a model with math as outcome, and fixed effect for time trend (year), as well as random intercept for school.

The equation for the model below:

$$Math_{tijk} = b_0 + \zeta_{0k} + b_1 * Time_{tijk} + \epsilon_{tijk}$$

where $\zeta_{0k} \sim N(0, \sigma_{\zeta_{eta0}}^2)$ and $\epsilon_{tijk} \sim N(0, \sigma_{\epsilon}^2)$

We refer to this as Model 0.

Below the model fit:

```
fit1 <- lmer(math ~ year + (1|schoolid), data = class_pp)
summary(fit1)

## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: math ~ year + (1 | schoolid)
## Data: class_pp
##
## REML criterion at convergence: 23951.7
##
## Scaled residuals:
## Min 1Q Median 3Q Max
## -5.2833 -0.6084 0.0037 0.6329 3.7761
##
## Random effects:
## Groups Name Variance Std.Dev.
## schoolid (Intercept) 348.7 18.67
## Residual 1268.4 35.62
```

```
## Number of obs: 2380, groups: schoolid, 107
##
## Fixed effects:
##           Estimate Std. Error      df t value Pr(>|t|)
## (Intercept)  464.932      2.116  132.154  219.73  <2e-16 ***
## year         57.566      1.460  2270.855   39.43  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##      (Intr)
## year -0.345
```

Add child-level random intercept

Q: BY “random intercept” do you mean just first obsrvation

To the previous model, we now add random intercepts for child:

$$Math_{tijk} = b_0 + \delta_{0ijk} + \zeta_{0k} + b_1 * Time_{tijk} + \epsilon_{tijk}$$

where $\delta_{0tijk} \sim N(0, \sigma_{\delta_0}^2)$, $\zeta_{0k} \sim N(0, \sigma_{\zeta_0}^2)$ and $\epsilon_{tijk} \sim N(0, \sigma_{\epsilon}^2)$ independetly of one another.

We refer to this as M1.

```
fit2 <- lmer(math ~ year + (1|schoolid/childid), data = class_pp)
summary(fit2)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: math ~ year + (1 | schoolid/childid)
##      Data: class_pp
##
## REML criterion at convergence: 23554.7
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -4.7492 -0.4811  0.0085  0.4881  3.4957
##
## Random effects:
##      Groups             Name             Variance Std.Dev.
## childid:schoolid (Intercept)  702.0       26.50
## schoolid         (Intercept)  307.5       17.54
## Residual                          599.1       24.48
## Number of obs: 2380, groups:  childid:schoolid, 1190; schoolid, 107
##
## Fixed effects:
##           Estimate Std. Error      df t value Pr(>|t|)
## (Intercept)  465.118      2.042  117.023  227.74  <2e-16 ***
## year         57.566      1.003  1189.000   57.37  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
```

```
##      (Intr)
## year -0.246
```

In model 0 the variance $\sigma_{\zeta_0}^2 = 348.7$ and in model 1 $\sigma_{\zeta_0}^2 = 307.5$.

In model 0, the variance for $\sigma_{\epsilon}^2 = 1268.4$ and in model 1 $\sigma_{\epsilon}^2 = 599.1$. We note that including child-level variation leads to a decrease in the variance of both the random effects.

Compute Pseudo-R²

Compute a pseudo R² relating the between school variation and ignoring between students in the same school.

We calculate this as :

$$\frac{\sigma_{\zeta_0}^2(M_0) - \sigma_{\zeta_0}^2(M_1)}{\sigma_{\zeta_0}^2(M_0)} = \frac{348.7 - 307.5}{348.7} = 0.12$$

The between-school variance is reduced by 12% (or ‘explained’) with the introduction of student random effect.

Does the total variation stay about the same?

```
tot_m0 = 348.7 + 1268.4
tot_m1 = 702 + 307.5 + 599.1
paste("Tot variance for model 0 : ", tot_m0)

## [1] "Tot variance for model 0 : 1617.1"
paste("Tot variance for model 1: ", tot_m1)

## [1] "Tot variance for model 1: 1608.6"
```

There is only a slightly decrease in the total variance between Model 0 and Model1.

Add a random slope for time trend

We now add a random slope (ζ_1) for time trend within schools.

$$Math_{tijk} = b_0 + \delta_{0ijk} + \zeta_{0k} + (b_1 + \zeta_{1k}) * Time_{tijk} + \epsilon_{tijk}$$

where $\delta_{0tijk} \sim N(0, \sigma_{\delta_{0ijk}}^2)$, $\zeta_{0k} \sim N(0, \sigma_{\zeta_0}^2)$, $\zeta_{1k} \sim N(0, \sigma_{\zeta_1}^2)$ and $\epsilon_{tijk} \sim N(0, \sigma_{\epsilon}^2)$ – each independently of one another.

We refer to this as Model 2

We run the model and report the fit:

```
fit3 = lmer(math ~ year + (1 + year || schoolid) + (1 | childid), data = class_pp)
summary(fit3)

## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: math ~ year + (1 + year || schoolid) + (1 | childid)
## Data: class_pp
##
## REML criterion at convergence: 23529.1
##
```

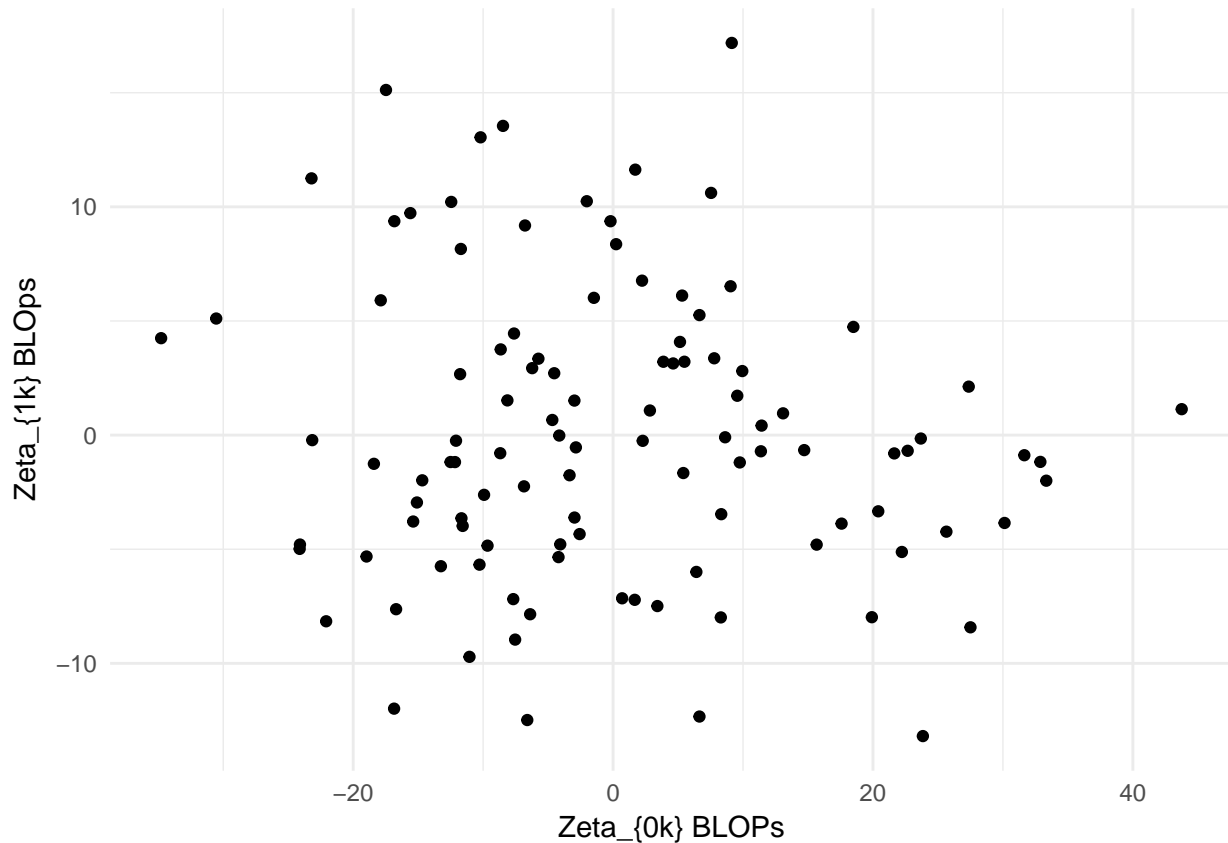
```
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -4.7665 -0.4721  0.0139  0.4686  3.6080
##
## Random effects:
##   Groups      Name      Variance Std.Dev.
##   childid    (Intercept) 725.13   26.928
##   schoolid   year        88.67    9.417
##   schoolid.1 (Intercept) 324.79   18.022
##   Residual                552.21   23.499
## Number of obs: 2380, groups:  childid, 1190; schoolid, 107
##
## Fixed effects:
##              Estimate Std. Error      df t value Pr(>|t|)
## (Intercept)  465.087      2.081 109.954  223.44  <2e-16 ***
## year         57.499      1.370  99.917   41.97  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##      (Intr)
## year -0.178
```

Generate the BLUPs for this model (Model 2)

Examine then whether the independence between zeta_0 and zeta_1 is reflected in a scatterplot of these two sets of effects.

```
pp_ranefs <- ranef(fit3)

if (vanillaR) {
  plot(pp_ranefs$schoolid[,2],pp_ranefs$schoolid[,1])
}else{
  ggplot(pp_ranefs$schoolid, aes(x = pp_ranefs$schoolid[,2], y = pp_ranefs$schoolid[,1] )) +
  geom_point() + labs(x = "Zeta_{0k} BLOPs", y = "Zeta_{1k} BLOPs") + theme_minimal()
}
```



From the plot, the BLOPs for ζ_{0k} and for ζ_{1k} appear uncorrelated, reflecting the way in which the model was built. In the BLUPS, we have a correlation of:

```
cor(pp_ranefs$schoolid[,2],pp_ranefs$schoolid[,1])
```

```
## [1] -0.1118599
```

That is, a very low negative correlation as we see from the plot. (test for sig)

Heteroscedasticity in the random effects

Question: What are: $V_S(\text{year} = 0)$, $V_S(\text{year} = 1)$?

The model we are considering is :

$$\text{Math}_{tijk} = b_0 + \delta_{0tijk} + \zeta_{0k} + (b_1 + \zeta_{1k})\text{Time}_{tijk} + \epsilon_{tijk}$$

So we have that (in this model, in which we are forcing correlation of 0 between slope and intercept):

- $V_S(\text{year} = 0) = \sigma_{\zeta_{0k}}^2 = 324.79$
- $V_S(\text{year} = 1) = \sigma_{\zeta_{0k}}^2 + \sigma_{\zeta_{1k}}^2 = 88.67 + 324.79 = 413.46$

Run model separately by year

We now examine what happens if we run the model separately by year. Do we get the same estimates for the variance between schools?

```

class_year0 = class_pp[class_pp$year == 0,]

# Run model for year 0
fit4 = lmer(math ~ (1 | schoolid), data = class_year0)
summary(fit4)

## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: math ~ (1 | schoolid)
## Data: class_year0
##
## REML criterion at convergence: 12085.7
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -4.8223 -0.5749  0.0005  0.6454  3.6237
##
## Random effects:
## Groups Name Variance Std.Dev.
## schoolid (Intercept) 364.3 19.09
## Residual 1344.5 36.67
## Number of obs: 1190, groups: schoolid, 107
##
## Fixed effects:
## Estimate Std. Error df t value Pr(>|t|)
## (Intercept) 465.23 2.19 103.20 212.4 <2e-16 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# Run model for year 1
class_year1 = class_pp[class_pp$year == 1,]
fit5 = lmer(math ~ (1 | schoolid), data = class_year1)
summary(fit5)

## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: math ~ (1 | schoolid)
## Data: class_year1
##
## REML criterion at convergence: 11950.8
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -5.291 -0.612 -0.005  0.613  3.793
##
## Random effects:
## Groups Name Variance Std.Dev.
## schoolid (Intercept) 306.8 17.52
## Residual 1205.0 34.71
## Number of obs: 1190, groups: schoolid, 107
##
## Fixed effects:
## Estimate Std. Error df t value Pr(>|t|)

```

```
## (Intercept) 522.698      2.027 103.069   257.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In this case, for the Year 0 Model, we get an estimated $\hat{\sigma}_{\zeta_{0k}}^2 = 364.3$, while for Year 1 Model, we have $\hat{\sigma}_{\zeta_{0k}}^2 = 306.8$. We note that these estimates are different from the ones computed above.

Allow for correlation

We now allow for correlation between the random effects for the intercept and the slope. We call this Model 3.

```
fit6 = lmer(math ~ year + (1 + year | schoolid) + (1 | childid), data = class_pp)
summary(fit6)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: math ~ year + (1 + year | schoolid) + (1 | childid)
## Data: class_pp
##
## REML criterion at convergence: 23520.3
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -4.7030 -0.4686  0.0066  0.4669  3.5142
##
## Random effects:
## Groups Name Variance Std.Dev. Corr
## childid (Intercept) 728.0 26.98
## schoolid (Intercept) 370.6 19.25
## year 109.1 10.44 -0.45
## Residual 547.0 23.39
## Number of obs: 2380, groups: childid, 1190; schoolid, 107
##
## Fixed effects:
## Estimate Std. Error df t value Pr(>|t|)
## (Intercept) 465.099 2.188 102.919 212.60 <2e-16 ***
## year 57.668 1.440 94.575 40.04 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
## (Intr)
## year -0.439
```

Correlation between ζ_0 and $\zeta_1 = -0.45$.

To test whether the correlation is statistically significant, we can compare Model 2 with Model 3 using an anova test.

```
anova(fit3, fit6, refit = F)
```

```
## Data: class_pp
## Models:
## fit3: math ~ year + (1 + year || schoolid) + (1 | childid)
## fit6: math ~ year + (1 + year | schoolid) + (1 | childid)
## Df AIC BIC logLik deviance Chisq Chi Df Pr(>Chisq)
```

```
## fit3  6 23541 23576 -11764    23529
## fit6  7 23534 23575 -11760    23520 8.8241      1    0.002973 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With a p-value $p = 0.003$, we reject the null hypothesis at $\alpha = 0.05$ significance level, and conclude that there correlation term is statistically significant.

So we have that (in this model where we are allowing for correlation between slope and intercept):

- $V_S(\text{year} = 0) = \sigma_{\zeta_{0k}}^2 = 370.6$
- $V_S(\text{year} = 1) = \sigma_{\zeta_{0k}}^2 + \sigma_{\zeta_{1k}}^2 + 2\rho_{01}\sigma_{\zeta_{0k}}\sigma_{\zeta_{1k}} = 370.6 + 109.1 - 2 * 0.45 * \sqrt{370.6}\sqrt{109.1} = 298.72$

These estimates are a lot closer to the school variances that result from fitting the models for the two years separately (in which we have σ_{ζ}^2 respectively be 364.3 for year 0 and 306.8 for year 1.

Therefore, it seemt that the model that allows for correlation between the two random effects has a better fit than the one forcing that correlation to be 0.