

Part3-B

Bianca Brusco

4/28/2018

Final Project : Part 3

Create person-period file

```
#re-read data
classroom2 <- na.omit(classroom)
#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 $\zeta_{0k} \sim N(0, \sigma_{\zeta_0}^2)$ 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: 21794.2
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -5.0693 -0.6031  0.0030  0.6321  3.7529
##
## Random effects:
## Groups Name Variance Std.Dev.
## schoolid (Intercept) 337 18.36
## Residual 1288 35.89
```

```
## Number of obs: 2162, groups: schoolid, 105
##
## Fixed effects:
##           Estimate Std. Error      df t value Pr(>|t|)
## (Intercept)  465.053      2.136  133.580   217.76 <2e-16 ***
## year         57.844      1.544 2055.557    37.47 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##      (Intr)
## year -0.361
```

Add child-level random intercept

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

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

where $\delta_{0tijk} \sim N(0, \sigma_{\delta_0}^2)$, $\zeta_{0k} \sim N(0, \sigma_{\zeta_0}^2)$ and $\epsilon_{tijk} \sim N(0, \sigma_{\epsilon}^2)$ independently 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: 21425.2
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -4.7233 -0.4827  0.0125  0.4922  3.4892
##
## Random effects:
##      Groups              Name            Variance Std.Dev.
## childid:schoolid (Intercept)  722.0       26.87
## schoolid         (Intercept)  293.2       17.12
## Residual                        602.2       24.54
## Number of obs: 2162, groups:  childid:schoolid, 1081; schoolid, 105
##
## Fixed effects:
##           Estimate Std. Error      df t value Pr(>|t|)
## (Intercept)  465.288      2.057  116.653   226.2 <2e-16 ***
## year         57.844      1.056 1080.000    54.8 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##      (Intr)
## year -0.257
```

In model 0 the variance $\sigma_{\zeta_0}^2 = 377$ and in model 1 $\sigma_{\zeta_0}^2 = 293.2$.

In model 0, the variance for $\sigma_{\epsilon}^2 = 1288$ and in model 1 $\sigma_{\epsilon_0}^2 = 602.2$. 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 :

$$\frac{\sigma_{\zeta_0}^2(M_0) - \sigma_{\zeta_0}^2(M_1)}{\sigma_{\zeta_0}^2(M_0)} = \frac{377 - 293.2}{377} = 0.22$$

The between-school variance is reduced by 22% (or ‘explained’) with the introduction of student random effect.

Does the total variation stay about the same?

```
tot_m0 = 337 + 1288
tot_m1 = 722 + 293.2 + 602.2
paste("Tot variance for model 0 : ", tot_m0)
```

```
## [1] "Tot variance for model 0 : 1625"
```

```
paste("Tot variance for model 1: ", tot_m1)
```

```
## [1] "Tot variance for model 1: 1617.4"
```

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.

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

where $\delta_{0tijk} \sim N(0, \sigma_{\delta_0}^2)$, $\zeta_{0k} \sim N(0, \sigma_{\zeta_0}^2)$, $\zeta_{1k} \sim N(0, \sigma_{\zeta_1}^2)$ 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: 21403.1
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
```

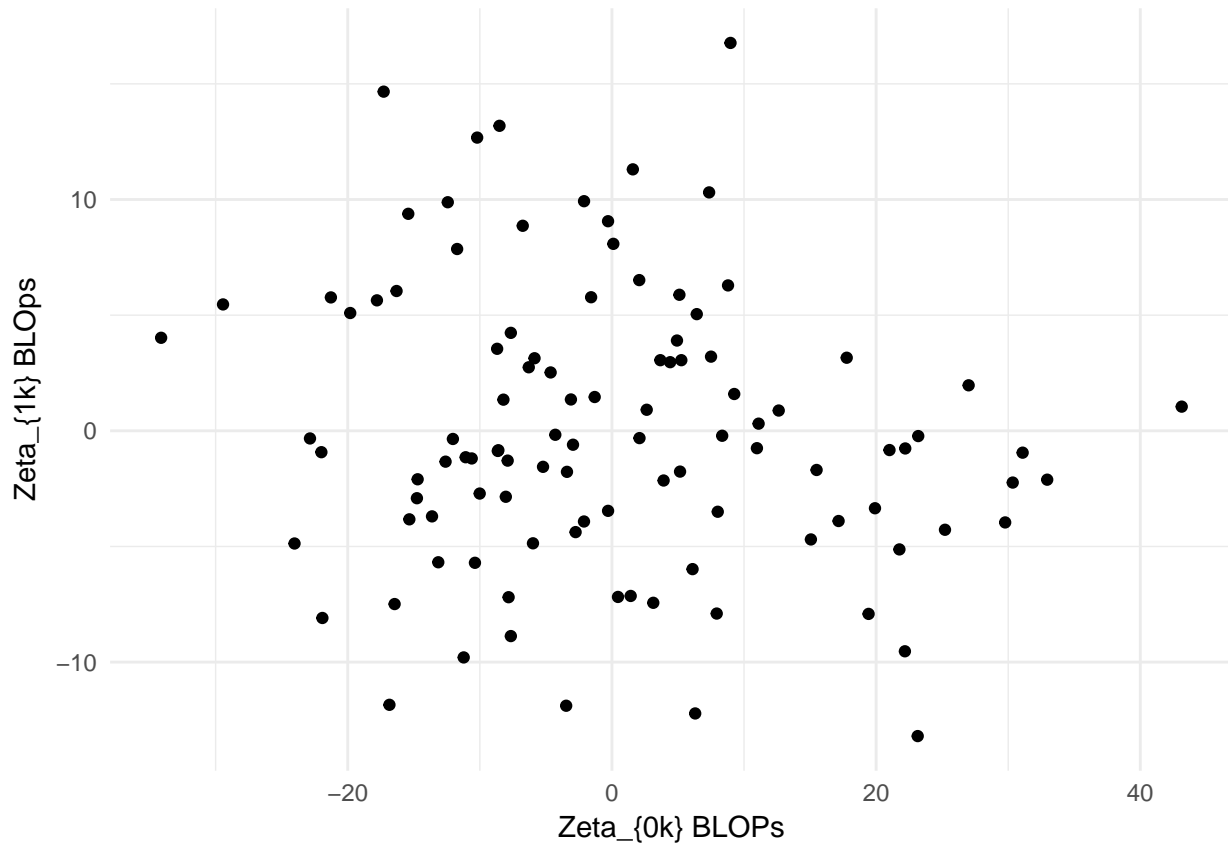
```
## -4.7461 -0.4788 0.0119 0.4719 3.5976
##
## Random effects:
## Groups      Name      Variance Std.Dev.
## childid     (Intercept) 745.36   27.301
## schoolid    year        85.96    9.272
## schoolid.1  (Intercept) 315.69   17.768
## Residual                    554.92   23.557
## Number of obs: 2162, groups:  childid, 1081; schoolid, 105
##
## Fixed effects:
##              Estimate Std. Error      df t value Pr(>|t|)
## (Intercept)  465.257      2.108 108.631  220.73   <2e-16 ***
## year          57.751      1.402 101.342   41.18   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##      (Intr)
## year -0.188
```

Generate the BLUPs for this model (Model 2)

Examine then whether the independence between `zeta0` and `zeta1` 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.

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):

- $V_S(\text{year} = 0) = \sigma_{\zeta_{0k}}^2 = 85.96$
- $V_S(\text{year} = 1) = \sigma_{\zeta_{0k}}^2 + \sigma_{\zeta_{1k}}^2 = 85.96 + 315.69 = 401.65$

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: 11017.5
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -4.7416 -0.5655  0.0086  0.6286  3.5648
##
## Random effects:
## Groups Name Variance Std.Dev.
## schoolid (Intercept) 364.1 19.08
## Residual 1391.4 37.30
## Number of obs: 1081, groups: schoolid, 105
##
## Fixed effects:
## Estimate Std. Error df t value Pr(>|t|)
## (Intercept) 465.382 2.244 101.588 207.4 <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: 10851.5
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -3.5824 -0.6158 -0.0063  0.6191  3.8063
##
## Random effects:
## Groups Name Variance Std.Dev.
## schoolid (Intercept) 279.1 16.71
## Residual 1202.6 34.68
## Number of obs: 1081, groups: schoolid, 105
##
## Fixed effects:
## Estimate Std. Error df t value Pr(>|t|)
## (Intercept) 523.2 2.0 101.9 261.6 <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 $\hat{\sigma}_{\zeta_{0k}}^2 = 364.1$, while for Year 1 Model, we have $\hat{\sigma}_{\zeta_{0k}}^2 = 279.1$. 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: 21391.3
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -4.6737 -0.4699  0.0038  0.4683  3.4882
##
## Random effects:
## Groups Name Variance Std.Dev. Corr
## childid (Intercept) 749.0 27.37
## schoolid (Intercept) 373.5 19.33
## year 112.4 10.60 -0.53
## Residual 547.8 23.41
## Number of obs: 2162, groups: childid, 1081; schoolid, 105
##
## Fixed effects:
## Estimate Std. Error df t value Pr(>|t|)
## (Intercept) 465.257 2.241 101.265 207.6 <2e-16 ***
## year 58.006 1.491 95.409 38.9 <2e-16 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
## (Intr)
## year -0.486
```

Correlation between ζ_0 and $\zeta_1 = -0.53$.

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 21415 21449 -10702 21403
## fit6 7 21405 21445 -10696 21391 11.879 1 0.0005678 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With a p-value $p = 0.0005678$, 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):

- $V_S(year = 0) = \sigma_{\zeta_{0k}}^2 = 373.5$
- $V_S(year = 1) = \sigma_{\zeta_{0k}}^2 + \sigma_{\zeta_{1k}}^2 + 2\rho_{01}\sigma_{\zeta_{0k}}\sigma_{\zeta_{1k}} = 373.5 + 112.4 - 2 * 0.53 * \sqrt{373.5}\sqrt{112.4} = 268.7$

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 for year 0 and 279 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.