

Homework 2

Leah Schultz

9/28/2017

Chapter 3: Growth Curves

```
library(dplyr)
library(tidyr)
library(lme4)
library(ggplot2)
library(broom)
library(merTools)
library(sjPlot)
oysup <- read.csv("~/1-descriptives-and-graphs-leahschultz/oysup_teacher_self.csv")
purpose <- read.csv("~/Dropbox/Lab & Research/OYSUP Project/oysup_self.csv")
oysup <- oysup %>%
  dplyr::select(FAMID, neuro_7s:neuro_10s)
dems <- purpose %>%
  dplyr::select(SEX2)
oysup <- cbind(oysup, dems)
```

First, restructuring data:

```
oysup_long <- tbl_df(oysup) %>%
  gather(c(neuro_7s:neuro_10s), key = "grade", value = "value") %>%
  separate(grade, into = c("variable", "grade"), sep = "_", convert = T) %>%
  separate(grade, into = c("grade", "delete"), sep = "s") %>%
  mutate(grade = as.numeric(grade)) %>%
  dplyr::select(-delete) %>%
  spread(variable, value)
oysup_long
```

```
## # A tibble: 4,296 x 4
##   FAMID SEX2 grade neuro
##   * <int> <int> <dbl> <dbl>
## 1  1001     2     7    NA
## 2  1001     2     8    NA
## 3  1001     2     9   3.5
## 4  1001     2    10   5.0
## 5  1002     2     7   3.5
## 6  1002     2     8   3.5
## 7  1002     2     9   2.0
## 8  1002     2    10   2.5
## 9  1003     1     7    NA
## 10 1003     1     8   4.0
## # ... with 4,286 more rows
```

1. Run linear models on all of your subjects (a basic regression). What is the average intercept, the average slope?

```
model1 <- lm(neuro ~ grade, data = oysup_long)
summary(model1)

##
## Call:
## lm(formula = neuro ~ grade, data = oysup_long)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.2537 -0.6753 -0.1361  0.8247  1.8639
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.52791    0.15216   23.18  <2e-16 ***
## grade       -0.03918    0.01741   -2.25  0.0245 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.05 on 3142 degrees of freedom
## (1152 observations deleted due to missingness)
## Multiple R-squared:  0.001609, Adjusted R-squared:  0.001291
## F-statistic: 5.063 on 1 and 3142 DF, p-value: 0.02451
```

Average intercept = 3.53

Average slope = -.04

2. Now run a mlm/lmer model with only a random intercept.

```
model2 <- lmer(neuro ~ (1 | FAMID), data = oysup_long)
summary(model2)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: neuro ~ (1 | FAMID)
## Data: oysup_long
##
## REML criterion at convergence: 8588.1
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -3.6193 -0.5530  0.0503  0.6092  3.0079
##
## Random effects:
## Groups   Name      Variance Std.Dev.
## FAMID    (Intercept) 0.5044   0.7102
## Residual                    0.6048   0.7777
## Number of obs: 3144, groups: FAMID, 960
##
## Fixed effects:
```

```
##           Estimate Std. Error t value
## (Intercept)  3.19688    0.02704   118.2
```

3. What is the ICC?

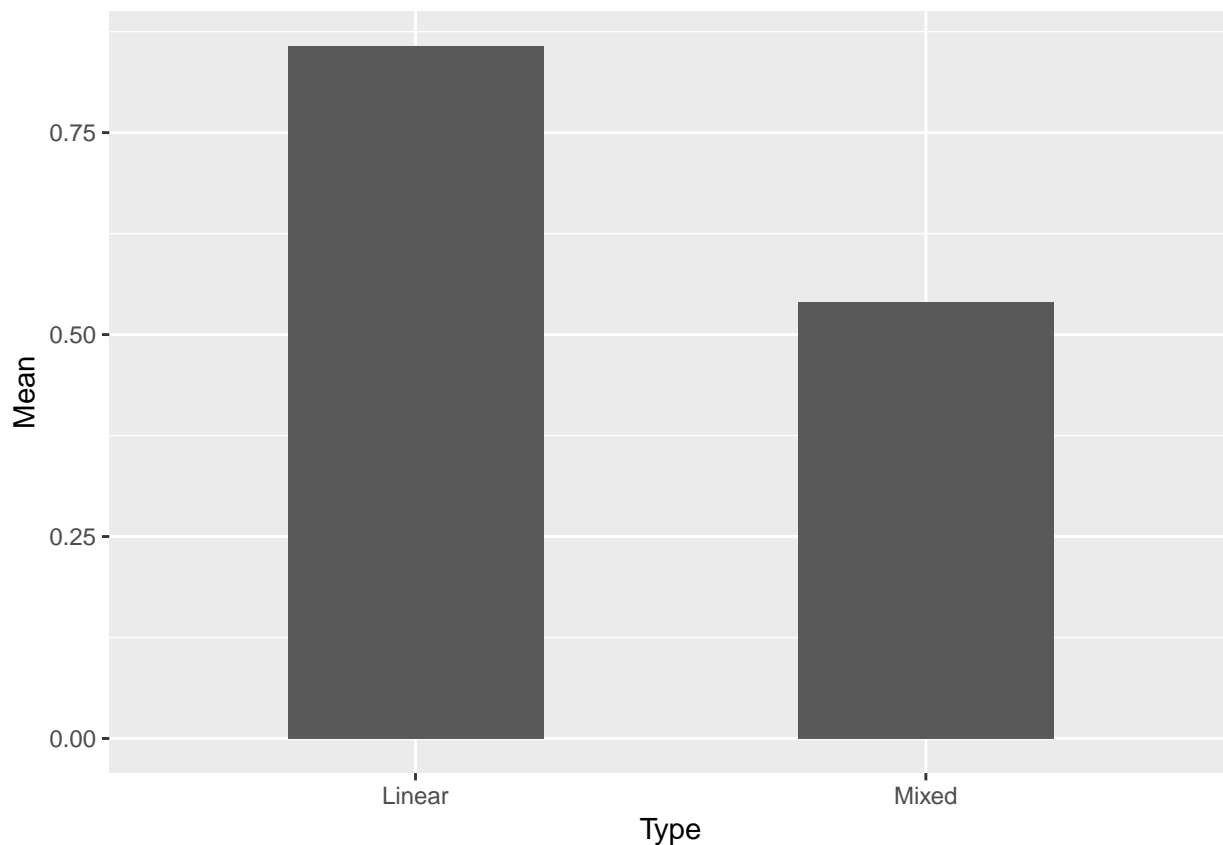
ICC = % between- vs. within-person variance

variance by ID / variance by ID + residual variance

$.50 / (.50 + .60) = .45$

What does residual variance look like compared to linear model? Create a graph to show this effect.

```
model1.aug <- augment(model1)
model2.aug <- augment(model2)
mod1_resid <- abs(model1.aug$.resid)
mod2_resid <- abs(model2.aug$.resid)
mean_mod1 <- mean(mod1_resid)
mean_mod2 <- mean(mod2_resid)
resid_df <- data.frame("Type" = c("Linear", "Mixed"),
                      "Mean" = c(mean_mod1, mean_mod2))
resid_plot <- ggplot(resid_df, aes(x = Type, y = Mean)) +
  geom_col(width = .5)
resid_plot
```



Average residual variance is lower in the mixed model, since we're accounting for individual-level, random effects.

3. Introduce a fixed slope term. What is the difference in terms of the fixed effects estimates between this estimate and the previous? Of the residual standard error? Create a graph to show both fixed effects estimates and the CIs around them.

```
model3 <- lmer(neuro ~ grade + (1 | FAMID), data = oysup_long)
summary(model3)

## Linear mixed model fit by REML ['lmerMod']
## Formula: neuro ~ grade + (1 | FAMID)
## Data: oysup_long
##
## REML criterion at convergence: 8584.6
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -3.6037 -0.5404  0.0331  0.6029  2.9295
##
## Random effects:
## Groups Name Variance Std.Dev.
## FAMID (Intercept) 0.5056  0.711
## Residual 0.6022  0.776
## Number of obs: 3144, groups: FAMID, 960
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept)  3.57268    0.12039  29.675
## grade       -0.04307    0.01345  -3.203
##
## Correlation of Fixed Effects:
##      (Intr)
## grade -0.974

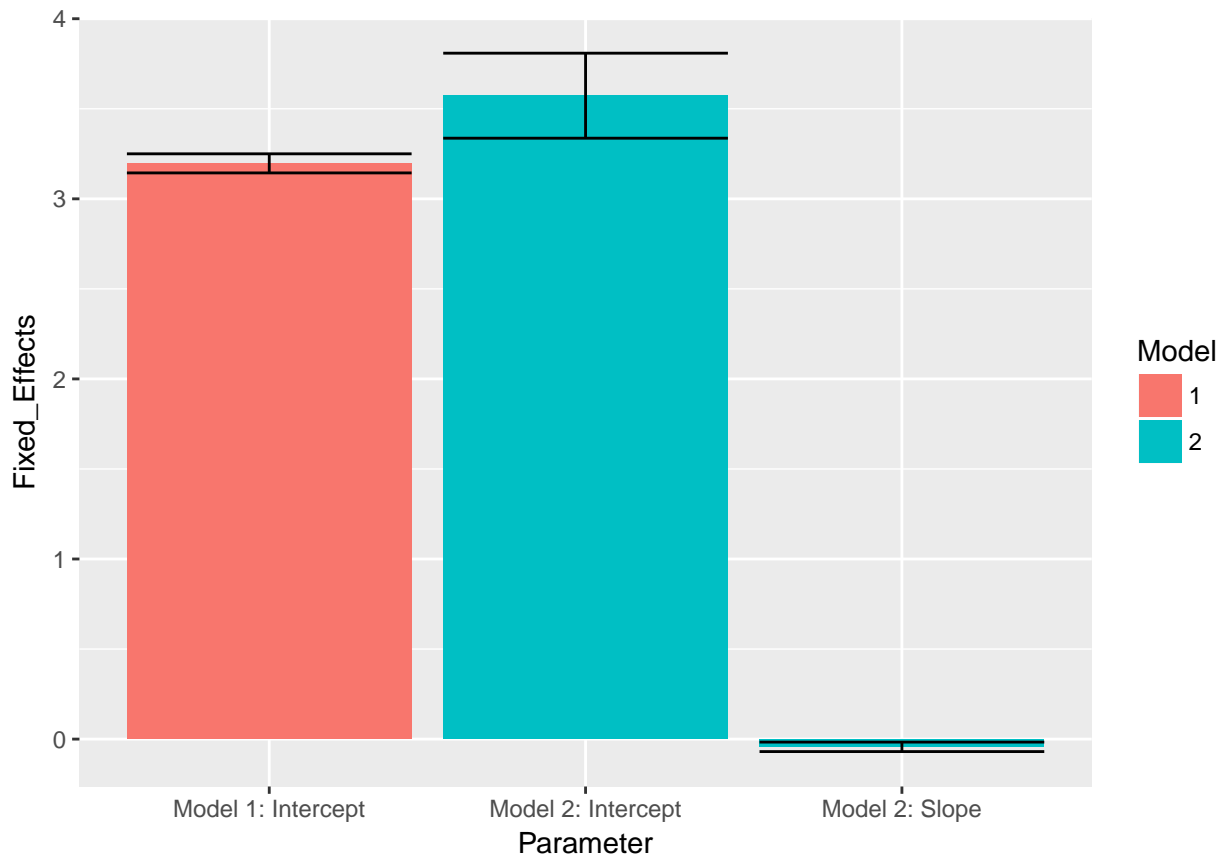
fixef(model2) - fixef(model3)[1]

## (Intercept)
## -0.3757999

fe_2 <- tidy(model2, effects = "fixed", conf.int = T, conf.level = 0.95)
fe_3 <- tidy(model3, effects = "fixed", conf.int = T, conf.level = 0.95)

fe_df <- data.frame("Parameter" = c("Model 1: Intercept", "Model 2: Intercept", "Model 2: Slope"),
                    "Fixed_Effects" = c(fe_2$estimate, fe_3[1,2], fe_3[2,2]),
                    "Lower_CI" = c(fe_2$conf.low, fe_3[1,5], fe_3[2,5]),
                    "Upper_CI" = c(fe_2$conf.high, fe_3[1,6], fe_3[2,6]),
                    "Model" = c("1", "2", "2"))

fe_plot <- ggplot(fe_df, aes(x = Parameter, y = Fixed_Effects)) +
  geom_col(aes(fill = Model)) +
  geom_errorbar(aes(ymin = Lower_CI, ymax = Upper_CI))
fe_plot
```



```
sigma(model2) - sigma(model3)[1]
```

```
## [1] 0.001725202
```

Fixed effect estimate for intercept increased by .37 points from 3.20 to 3.57.

Residual standard error decreased by .002. Seems like adding a fixed slope term did not do much to improve the model.

4. Run an additional model with a random slope. How does this change compare to the previous model? Should you keep the random slope or not?

```
model4 <- lmer(neuro ~ grade + (1 + grade | FAMID), data = oysup_long)
summary(model4)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: neuro ~ grade + (1 + grade | FAMID)
## Data: oysup_long
##
## REML criterion at convergence: 8574.7
##
## Scaled residuals:
##    Min      1Q  Median      3Q      Max
## -3.6964 -0.5298  0.0288  0.5854  2.6730
##
## Random effects:
```

```
## Groups      Name      Variance Std.Dev. Corr
## FAMID      (Intercept) 2.44479  1.5636
##           grade      0.02725  0.1651  -0.89
## Residual                    0.56256  0.7500
## Number of obs: 3144, groups: FAMID, 960
##
## Fixed effects:
##           Estimate Std. Error t value
## (Intercept)  3.56322    0.12723  28.007
## grade       -0.04210    0.01429  -2.946
##
## Correlation of Fixed Effects:
##      (Intr)
## grade -0.977
sigma(model3)[1] - sigma(model4)[1]

## [1] 0.02594421
anova(model3, model4)

## refitting model(s) with ML (instead of REML)
## Data: oysup_long
## Models:
## model3: neuro ~ grade + (1 | FAMID)
## model4: neuro ~ grade + (1 + grade | FAMID)
##      Df    AIC    BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## model3  4 8580.4 8604.7 -4286.2   8572.4
## model4  6 8574.7 8611.0 -4281.3   8562.7  9.7527     2  0.007625 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The residual error decreases .03 with the addition of a random slope. A likelihood ratio test indicates that the model fit is better with this new parameter, as well. I'll keep the random slope.

5. Interpret the correlation between the slope and the intercept.

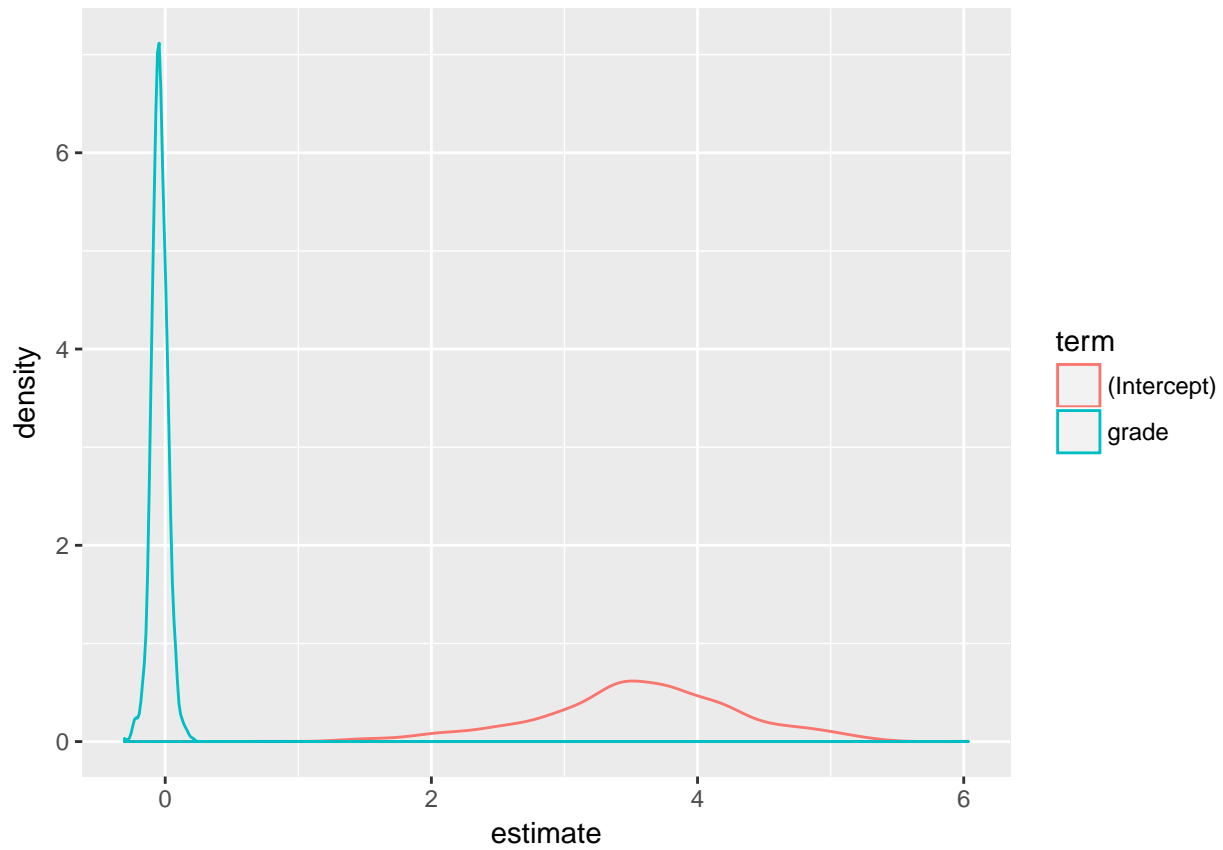
Correlation between slope and intercept = -.97. Adolescents who are initially higher on neuroticism will tend to decrease over time compared to children who are initially lower on neuroticism.

6. Create a density plot of the random effects from your final model.

```
random_params <- tidy(model4, effect = "ran_modes")

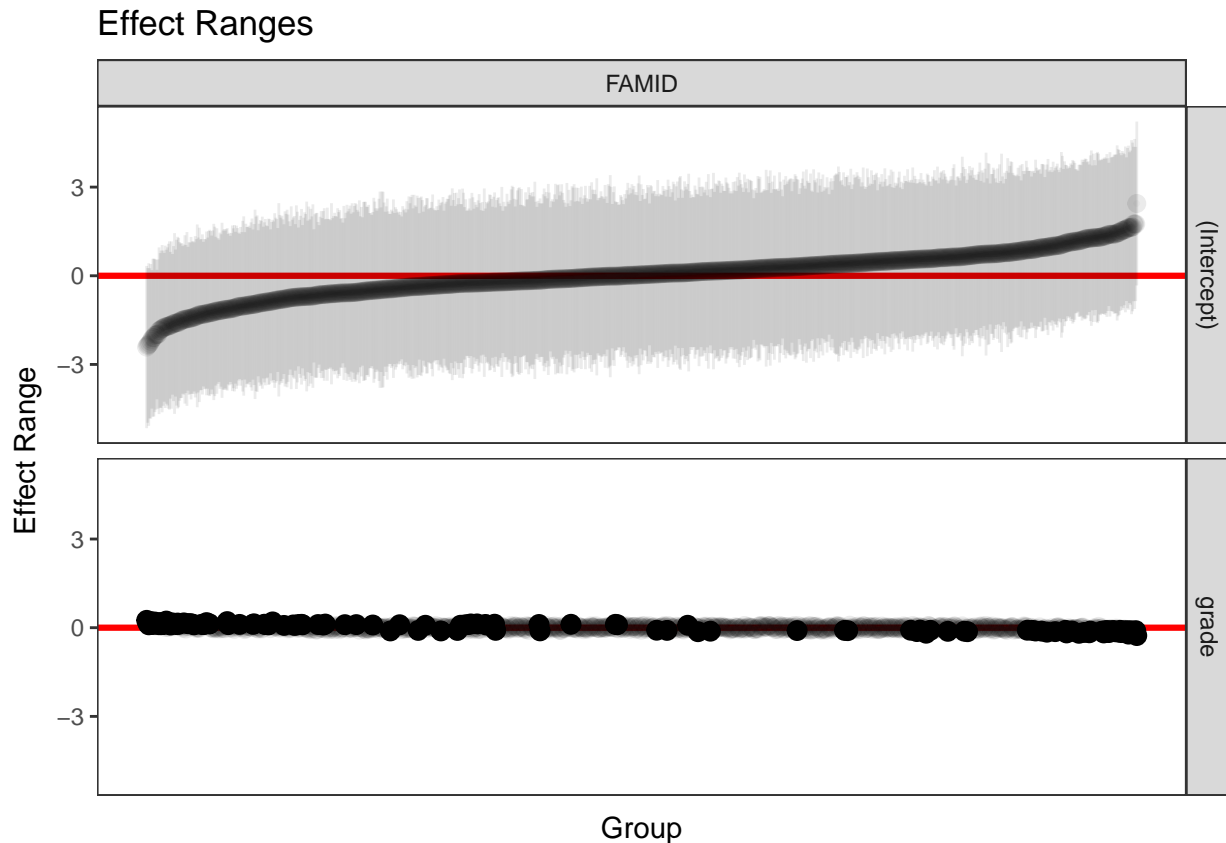
raneff_plot <- ggplot(random_params, aes(x = estimate, color = term)) +
  geom_density()

raneff_plot
```



7. Create a caterpillar plot of the random effects. Is there any person that seems odd in terms of large standard errors around intercept and slope estimates?

```
re.sim <- RESim(model4)
plotRESim(re.sim)
```



It looks as if there are about many individuals who deviate from the fixed slope, and are thus driving the model to be better represented by a random slope. There is only one student who seems to deviate notably from the fixed intercept.

8. Create a plot of the trajectory, along with a spaghetti plot of each person's individual slope. Set the alpha level (transparency) on the individual slopes to make them easier to see.

```
predict <- predictInterval(merMod = model4, newdata = oysup_long,
                           level = 0.9, n.sims = 100,
                           stat = "median", include.resid.var = TRUE)
growth_df <- cbind(oysup_long, predict$fit)

growth_plot <- ggplot(oysup_long, aes(x = grade, y = predict$fit)) +
  geom_line(aes(group = FAMID), alpha = .2) +
  stat_smooth(method = lm) +
  theme_bw() +
  xlab("Grade") +
  ylab("Self-Reported Neuroticism") +
  ylim(1,5)
growth_plot
```