Time Series and the Analysis of Longitudinal Data

Gemma Wallace & Neil Yetz
PSY 653 Module 6 Lab
Mar 4, 2020

Class Example

A research team is interested if student performance in coding skills increased over time during a coding class. During each week of the program, all participants completed a coding challenge. Each challenge had a set of coding skills that had to be employed to solve the challenge, but each challenge focused on solving some substantive problem (e.g., mapping social networks of users on an online forum, developing an algorithm to recommend new music based on a user's Spotify history, etc.). Each student's performance on the challenge was graded by the research team using a valid and reliable rubric able to detect growth in skills over time.

This dataset was provided by Kim Henry, PhD.

Variables

- kid_id: Subject ID
- week: Week in program (0-6)
- perform: Performance grade (scaled from 1-10)

Load & install packages

````
# Load Libraries
```{r}
library(psych)
library(tidyverse)

install.packages("lme4")
install.packages("lmerTest")
library(lme4)
library(lmerTest)

```

````

Read in Data

````
# Read in Data
```{r}
grow <- read_csv("grow.csv")
```
````

```
Parsed with column specification:
cols(
  kid_id = col_double(),
  week = col_double(),
  perform = col_double()
)
```

Describe the data

```{r}
describe(grow)
```

	vars <dbl>	n <dbl>	mean <dbl>	sd <dbl>	median <dbl>	trimmed <dbl>	mad <dbl>	min <dbl>	max <dbl>
kid_id	1	700	252.5	144.43	252.50	252.50	185.32	5.00	500
week	2	700	3.0	2.00	3.00	3.00	2.97	0.00	6
perform	3	700	4.9	1.33	4.88	4.87	1.32	1.27	10

3 rows | 1-10 of 13 columns

Describe the data

```{r}
describe(grow)
```

	vars	n	mean	sd	median	trimmed	mad	min	max
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
kid_id	1	700	252.5	144.43	252.50	252.50	185.32	5.00	500
week	2	700	3.0	2.00	3.00	3.00	2.97	0.00	6
perform	3	700	$\boxed{4.9}$	1.33	4.88	4.87	1.32	1.27	10

3 rows | 1-10 of 13 columns

Remember this number! This is called the "Mean of Means"

````
# Aggregate data
```{r}
agg_long <- aggregate(x=grow$perform,by=list(week = grow$week), FUN=mean)
agg_long
```
````

week <dbl>	x <dbl>
0	4.687369
1	4.491740
2	4.587813
3	4.952047
4	5.035082
5	5.108235
6	5.469560

7 rows

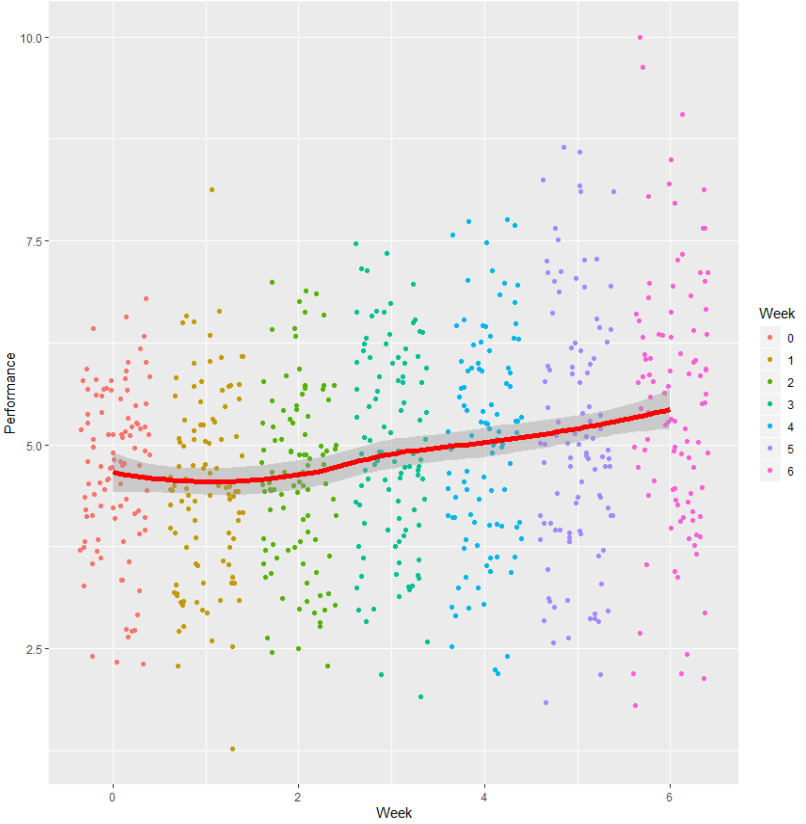
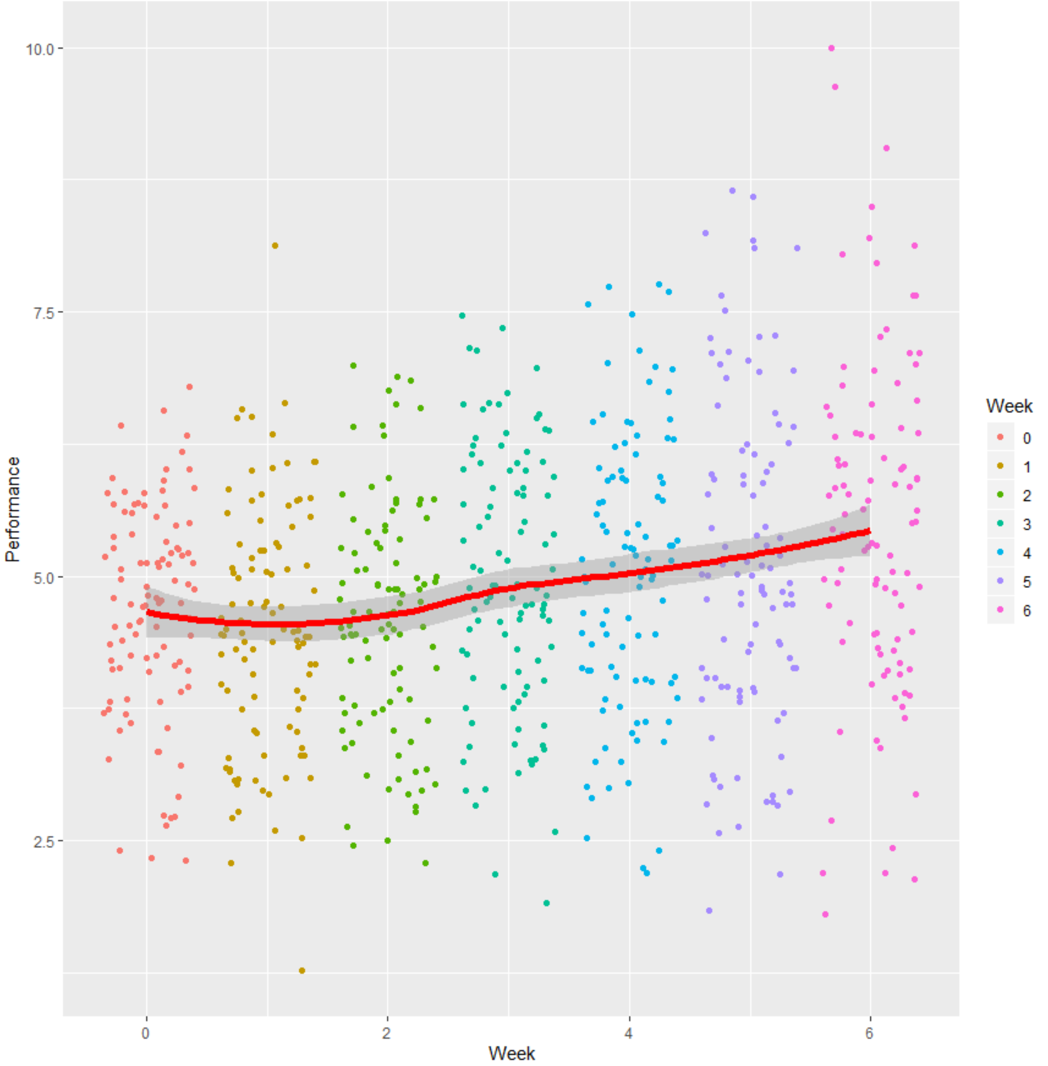
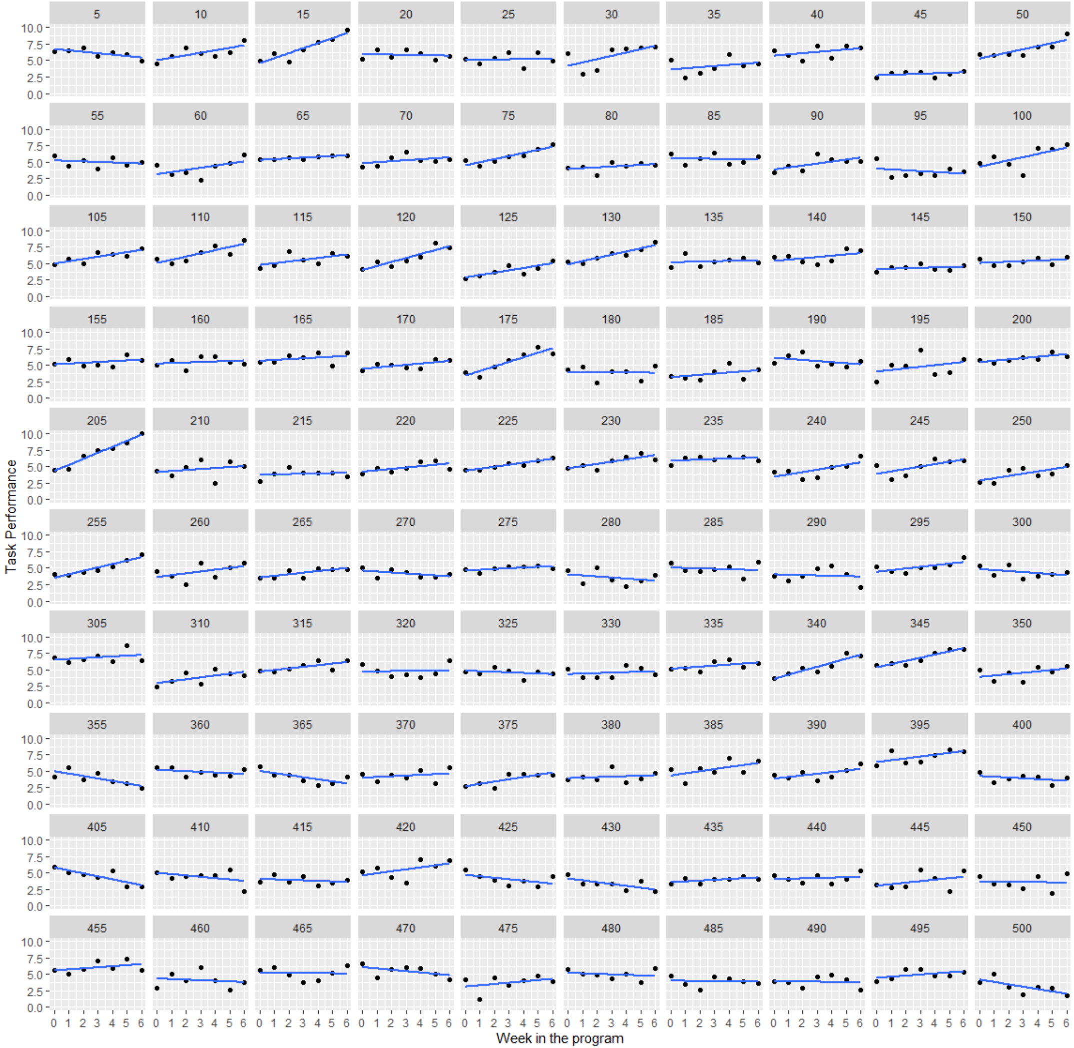
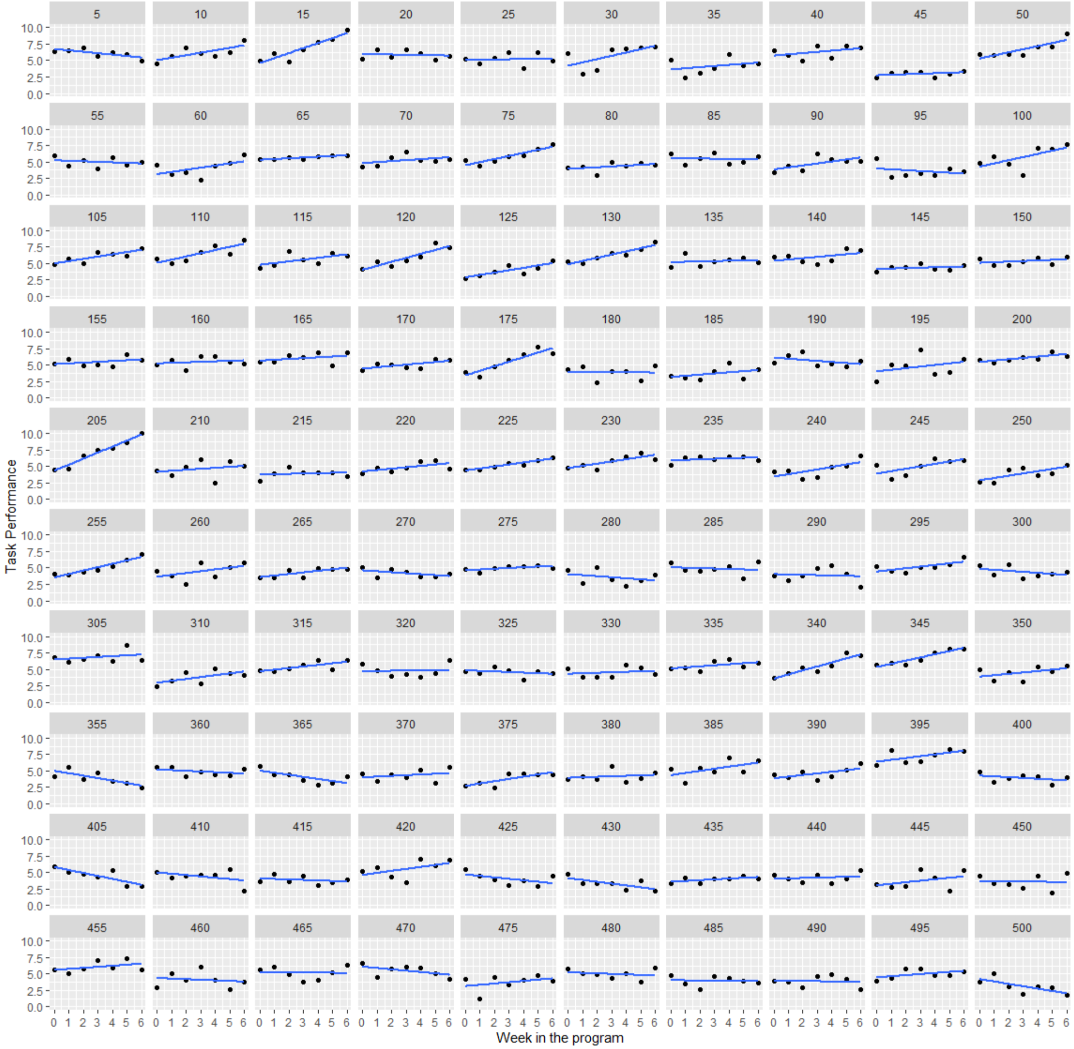
````
# Aggregate data
```{r}
agg_long <- aggregate(x=grow$perform,by=list(week = grow$week), FUN=mean)
agg_long
```
````

week <dbl>	x <dbl>
0	4.687369
1	4.491740
2	4.587813
3	4.952047
4	5.035082
5	5.108235
6	5.469560

7 rows

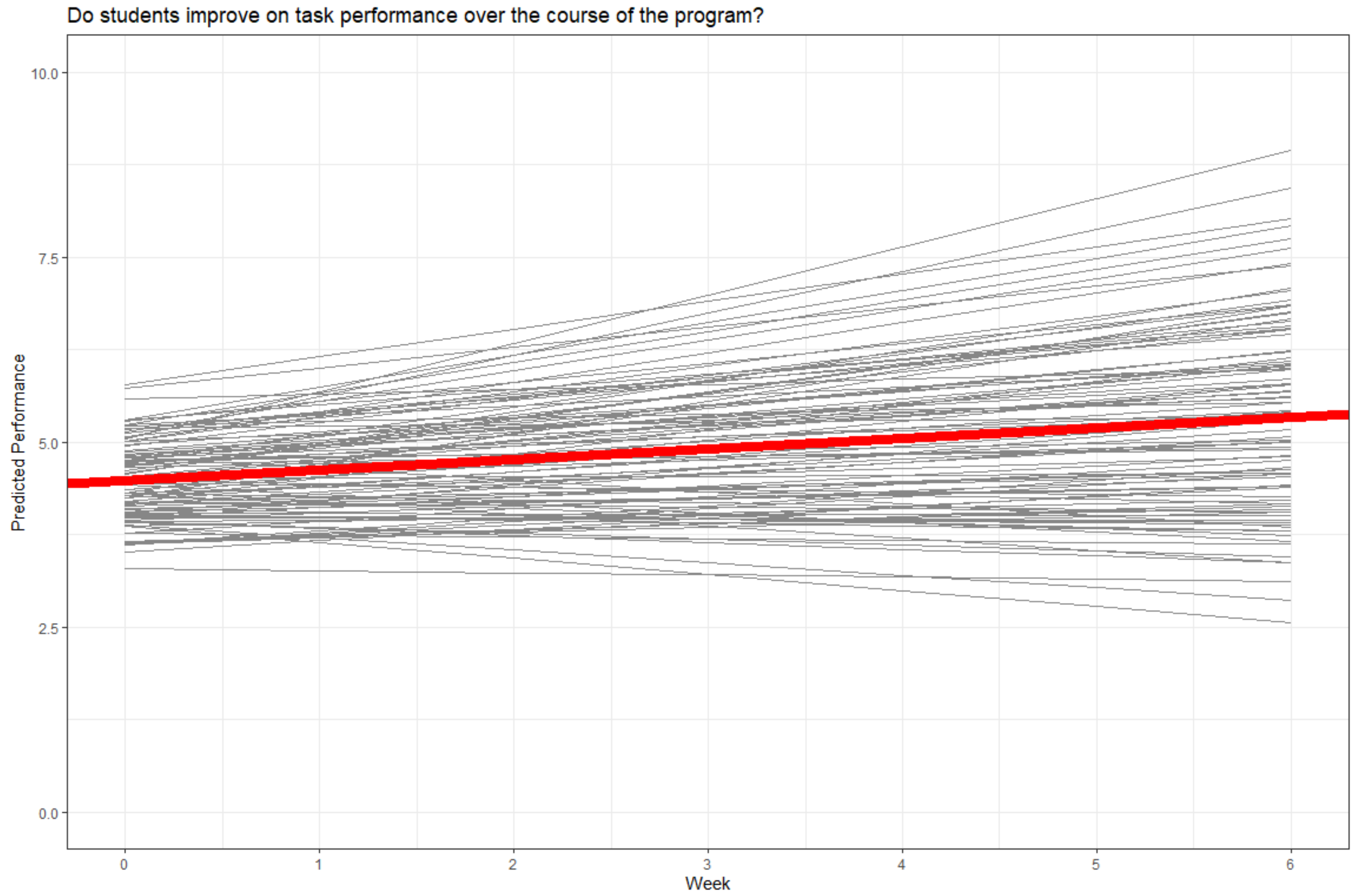
Mean of Means = 4.9

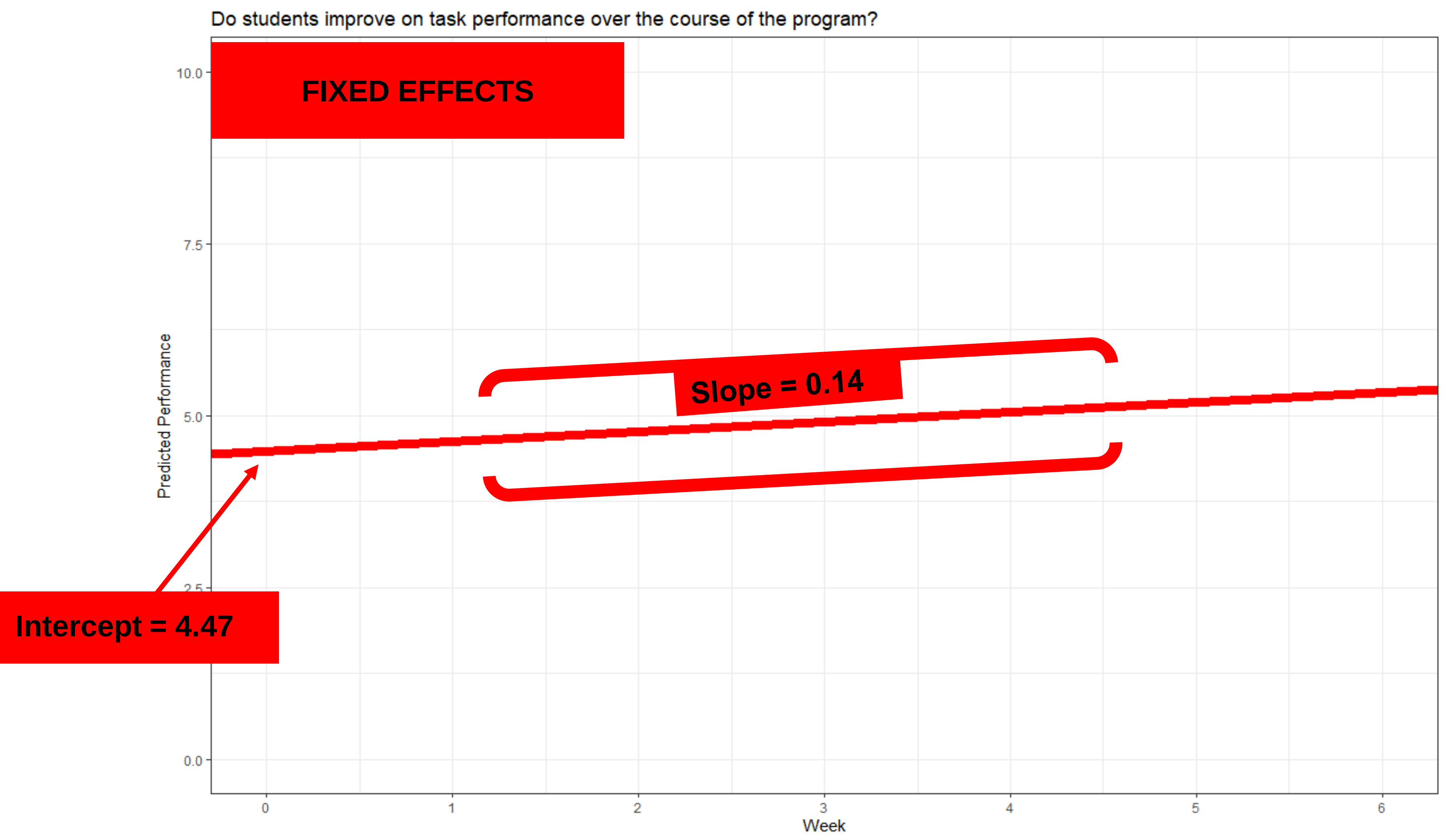
Visualize the data!

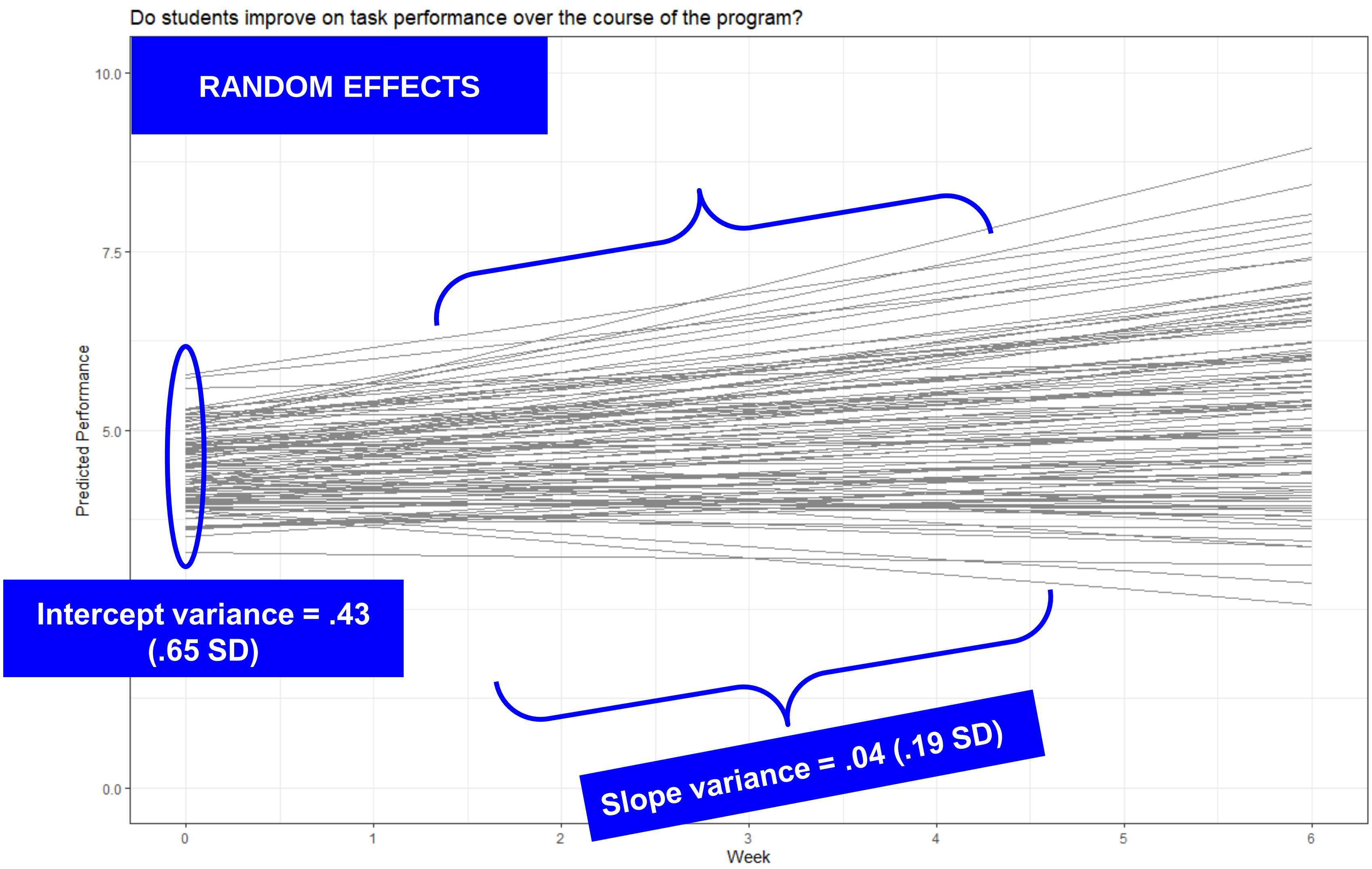
```
```{r, fig.width=9, fig.height=9}

ggplot(grow, aes(x = week, y = perform)) +
  geom_jitter(aes(color = factor(week))) +
  geom_smooth(method = "loess", color = "red", size = 2) +
  xlab("Week") +
  ylab("Performance") +
  labs(color = "Week")

```
```



Based on this plot, do you think you have justification to test for a linear effect of time on GPA?

What about a quadratic effect?



Visualize each individual subject!

````
```{r, fig.height = 12, fig.width=12}
ggplot(data = grow, aes(x = week, y = perform)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  scale_y_continuous(limits = c(0,10)) +
  scale_x_continuous(limits = c(0,6), breaks = c(0,1,2,3,4,5,6)) +
  facet_wrap(~kid_id) +
  labs(title = "Do students in the control condition improve their task performance over the course of the program?",
    x = "Week in the program", y = "Task Performance")
```
````

Do students in the control condition improve their task performance over the course of the program?



Average Intercept = 4.47
Average Slope = 0.14

Do students in the control condition improve their task performance over the course of the program?



Average Intercept = 4.47
Average Slope = 0.14

Remember these numbers!

Defining **fixed** and **random** effects in the lmer package

These definitions are a little different in longitudinal analyses than in ANOVAs.

Fixed effect = does not vary over subjects of groups – average value of slope or intercept

Random effect = might vary across subjects or groups – intercepts and slopes might be calculated for each group or each subject to see if they vary meaningfully

Build Baseline model

```
```{r}
mod1 <-  lmer(perform ~ 1 + (1|kid_id), REML = TRUE, data = grow)
summary(mod1)
```
```

Similar to a lm, the dependent variable is listed, then a tilde. Since, this is an unconditional model, there are no predictors, but we include a 1 to denote the intercept. This is called the fixed effects part of the model and will provide us with the mean of means across the groups.

lmer is the function used to specify a multilevel or growth model (it stands for linear mixed effects regression).

```
mod1 <- lmer(perform ~ 1 + (1 | kid_id), REML = TRUE, data = grow)
summary(mod1)
```

After the fixed effects, we provide the random effects. Here we list

the effects that we want to denote as random. In this case it is just the intercept (1), which will capture the between group variability. The bar (|) and then kid_id denotes the Level 2 grouping variable.

REML stands for Restricted Estimation Maximum Likelihood. This is one of the most common estimators for multilevel models, and for our intro, we will use this one exclusively.

````
```{r}
mod1 <-  lmer(perform ~ 1 + (1|kid_id), REML = TRUE, data = grow)
summary(mod1)
```
````

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: perform ~ 1 + (1 | kid_id)
   Data: grow

REML criterion at convergence: 2166.1

Scaled residuals: 
    Min      1Q  Median      3Q     Max 
-2.6469 -0.6424 -0.0085  0.5990  3.2746 

Random effects:
 Groups   Name        Variance Std.Dev.
 kid_id   (Intercept) 0.8013   0.8951  
 Residual             0.9821   0.9910  
Number of obs: 700, groups:  kid_id, 100

Fixed effects:
            Estimate Std. Error       df t value Pr(>|t|)    
(Intercept)  4.90455    0.09703 98.99999   50.55   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```{r}
mod1 <-  lmer(perform ~ 1 + (1|kid_id), REML = TRUE, data = grow)
summary(mod1)
```

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: perform ~ 1 + (1 | kid_id)
   Data: grow

REML criterion at convergence: 2166.1

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.6469 -0.6424 -0.0085  0.5990  3.2746

Random effects:
 Groups   Name        Variance Std.Dev.
 kid_id   (Intercept) 0.8013   0.8951
 Residual             0.9821   0.9910
Number of obs: 700, groups:  kid_id, 100

Fixed effects:
            Estimate Std. Error       df t value Pr(>|t|)
(Intercept)  4.90455    0.09703 98.99999   50.55   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- **Random intercept: On average, kids vary from the grand mean by .895 standard deviations**

- ***Fixed Intercept: In the absence of any fixed effects*, this intercept represents the "mean of means" of our outcome variable.**

Add week as a fixed *and* random effect

```{r}
mod2 <-  lmer(perform ~ 1 + week + (1 + week|kid_id), REML = TRUE, data = grow)
summary(mod2)
```

```{r}
mod2 <-  lmer(perform ~ 1 + week + (1 + week|kid_id), REML = TRUE, data = grow)
summary(mod2)
```

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: perform ~ 1 + week + (1 + week | kid_id)
   Data: grow

REML criterion at convergence: 2038.4

Scaled residuals: 
    Min      1Q  Median      3Q     Max 
-3.1997 -0.5740  0.0449  0.6342  3.0608 

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 kid_id   (Intercept) 0.42532  0.6522       
          week        0.03774  0.1943   0.10
 Residual             0.71247  0.8441       
Number of obs: 700, groups:  kid_id, 100

Fixed effects:
            Estimate Std. Error       df t value Pr(>|t|)    
(Intercept)  4.47310    0.08695 98.99841  51.442  < 2e-16 ***
week         0.14382    0.02514 99.00129   5.722 1.13e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
     (Intr)
week -0.292
```

```{r}
mod2 <-  lmer(perform ~ 1 + week + (1 + week|kid_id), REML = TRUE, data = grow)
summary(mod2)
```

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: perform ~ 1 + week + (1 + week | kid_id)
   Data: grow

REML criterion at convergence: 2038.4

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.1997 -0.5740  0.0449  0.6342  3.0608

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 kid_id   (Intercept) 0.42532  0.6522
          week        0.03774  0.1943   0.10
 Residual             0.71247  0.8441
Number of obs: 700, groups:  kid_id, 100

Fixed effects:
            Estimate Std. Error       df t value Pr(>|t|)
(Intercept)  4.47310    0.08695 98.99841  51.442  < 2e-16 ***
week         0.14382    0.02514 99.00129   5.722 1.13e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
     (Intr)
week -0.292
```

- **Random Intercept:** On average, subject intercepts vary by 0.652 standard deviations
- **Random Slope:** On average, subject slopes vary by 0.194 standard deviations

- **Fixed Intercept:** The average intercept, while incorporating week, is 4.473
- **Fixed Slope:** On average, subject scores increased at a rate of 0.144 units

```{r}
mod2 <-  lmer(perform ~ 1 + week + (1 + week|kid_id), REML = TRUE, data = grow)
summary(mod2)
```

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: perform ~ 1 + week + (1 + week | kid_id)
   Data: grow

REML criterion at convergence: 2038.4

Scaled residuals: 
    Min      1Q  Median      3Q     Max 
-3.1997 -0.5740  0.0449  0.6342  3.0608 

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 kid_id   (Intercept) 0.42532  0.6522       
          week        0.03774  0.1943   0.10
 Residual             0.71247  0.8441       
Number of obs: 700, groups:  kid_id, 100

Fixed effects:
            Estimate Std. Error       df t value Pr(>|t|)    
(Intercept)  4.47310    0.08695 98.99841  51.442  < 2e-16 ***
week         0.14382    0.02514 99.00129   5.722 1.13e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
     (Intr)
week -0.292
```

- **ICC Calculation**
  - **ICC = $\sigma^2_{RandomEffect}$ / $\sigma^2_{RandomTotal}$**
  - **ICC = .0377 / (.0377 + .7124)**
  - **ICC = .05028**
- ***There is only a small amount of variation in slopes across subjects (ICC = .050)***

One final plot (Optional)

````
```{r, fig.width=12, fig.height=8}
# add_predictions comes from the modelr package
install.packages(“modelr”)
library(modelr)

# Get predicted values
mod2.plot <- add_predictions(data = grow, model = mod2)

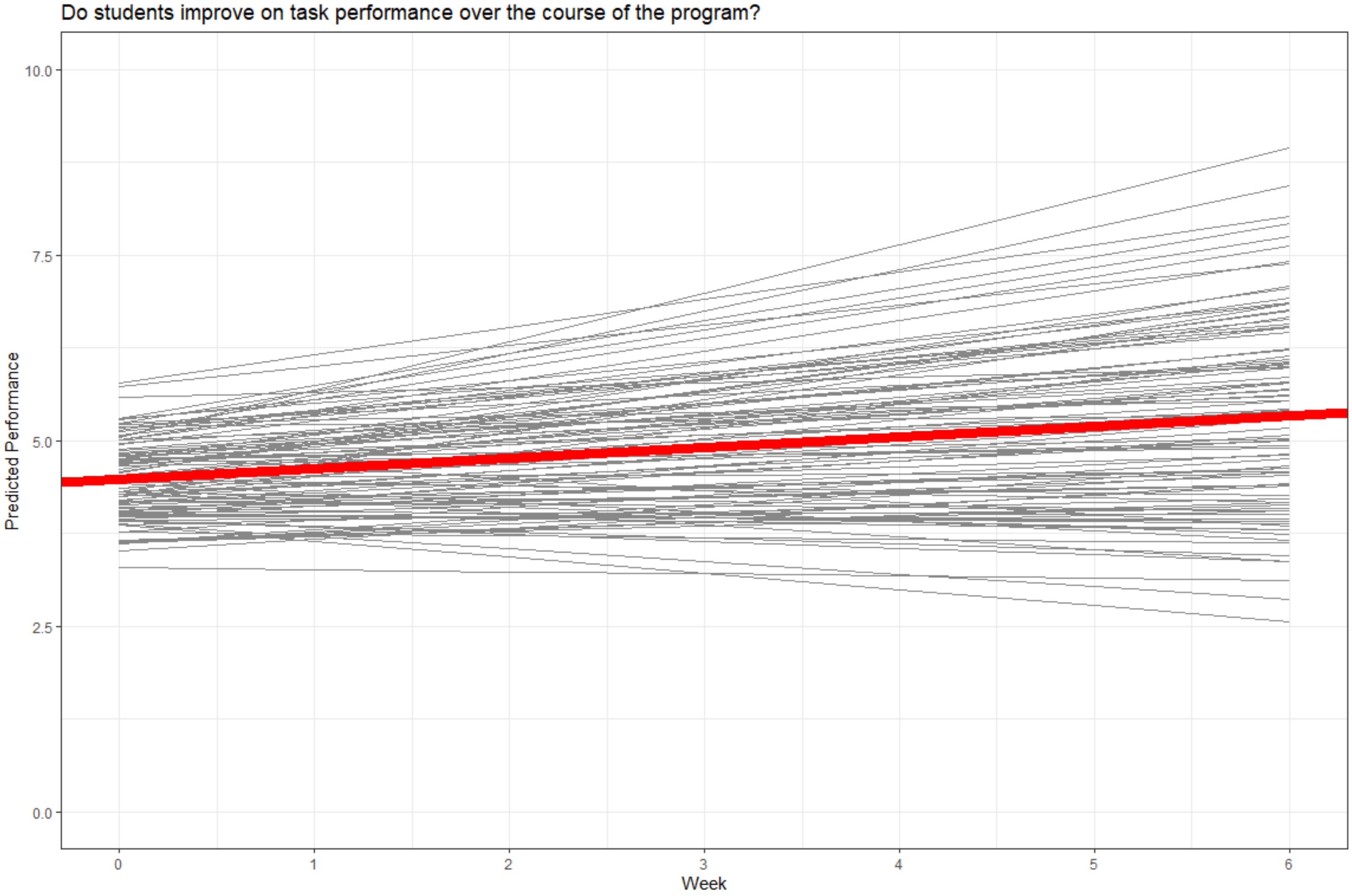
# Make plot
ggplot(data = mod2.plot, aes(x = week, y = pred, group = kid_id)) +
  geom_line(color = "grey53") +
  geom_abline(intercept = 4.4731, slope = .1438, color="red", size=3) +
  scale_y_continuous(limits = c(0,10)) +
  scale_x_continuous(limits = c(0,6), breaks = c(0,1,2,3,4,5,6)) +
  labs(title = "Do students improve on task performance over the course of the program?",
    x = "Week", y = "Predicted Performance") +
  theme_bw()
```
````


Do students improve on task performance over the course of the program?
Predicted Performance
10.0
7.5
5.0
2.5
0.0
0
1
2
3
4
5
6
Week


Do students improve on task performance over the course of the program?
FIXED EFFECTS
Slope = 0.14
Intercept = 4.47
Predicted Performance
10.0
7.5
5.0
2.5
0.0
0
1
2
3
4
5
6
Week



Do students improve on task performance over the course of the program?
RANDOM EFFECTS
Intercept variance = .43
(.65 SD)
Slope variance = .04 (.19 SD)
Predicted Performance
10.0
7.5
5.0
0.0
0
1
2
3
4
5
6
Week


Do students improve on task performance over the course of the program?
Predicted Performance
10.0
7.5
5.0
2.5
0.0
0
1
2
3
4
5
6
Week