Linear Mixed-Effects Models

Andrew Zieffler

Random Coefficient Models (RCM)

RCMs are statistical models developed by focusing on individual trajectories in a subject-specific fashion.

The idea behind RCMs is that an individual's set of scores are determined by three components:

- **Population average effects** that influence the process of change for all subjects**,**
- **Subject-specific effects** unique to a particular subject, and
- **Lack of fit** or residual

The decomposition of scores into population effects (or fixed-effects) plus individual effects (random-effects) leads to the name **mixed-effects model**.

In other disciplines, for example in education research, the same framework is called a **multilevel model** or **hierarchical model**. These typically refer to an alternative way of specifying the mixed-effects model.

As an example, below is a mixed-effects model that includes population and individual effects for intercept and slope.

$$y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})t_{ij} + \epsilon_{ij}$$

In this equation:

- ***i*** indicates the *i*th subject or case; any term with an *i* subscript can vary across subjects
- ***j*** indicates the *j*th time point/wave**;** any term with a *j* subscript can vary across time points

Things to note:

- The beta terms (β_0 and β_1) have no *i* or *j* subscripts; they are the population average, or fixed-effects. These are the same for each individual and each time point.
- The *b* terms (b_{0i} and b_{1i}) have an *i* subscript; they are the individual-specific, or random-effects. These can be different for different subjects.
- The *t* term represents the time predictor; as such it has both an *i* and *j* subscript. This means the time predictor can vary across both subjects and time points.
- The residual term (ε_{ij}) has both an *i* and *j* subscript. This means the residuals can vary across both subjects and time points.

To further understand this, we will examine data from the *vocab.csv* file. These data include scores for 64 students measured on a scaled vocabulary test over four repeated measures (8th, 9th, 10th, and 11th grade).

```
# Read in the wide data
> vocab = read_csv(file = "~/Dropbox/epsy-8282/data/vocab.csv")
```

```
     id    t8    t9   t10   t11 female
  <int> <dbl> <dbl> <dbl> <dbl>  <int>
1     1  1.75  2.60  3.76  3.68      0
2     2  0.90  2.47  2.44  3.43      0
3     3  0.80  0.93  0.40  2.27      0
4     4  2.42  4.15  4.56  4.21      0
5     5 -1.31 -1.31 -0.66 -2.22      0
6     6 -1.56  1.67  0.18  2.33      0
```

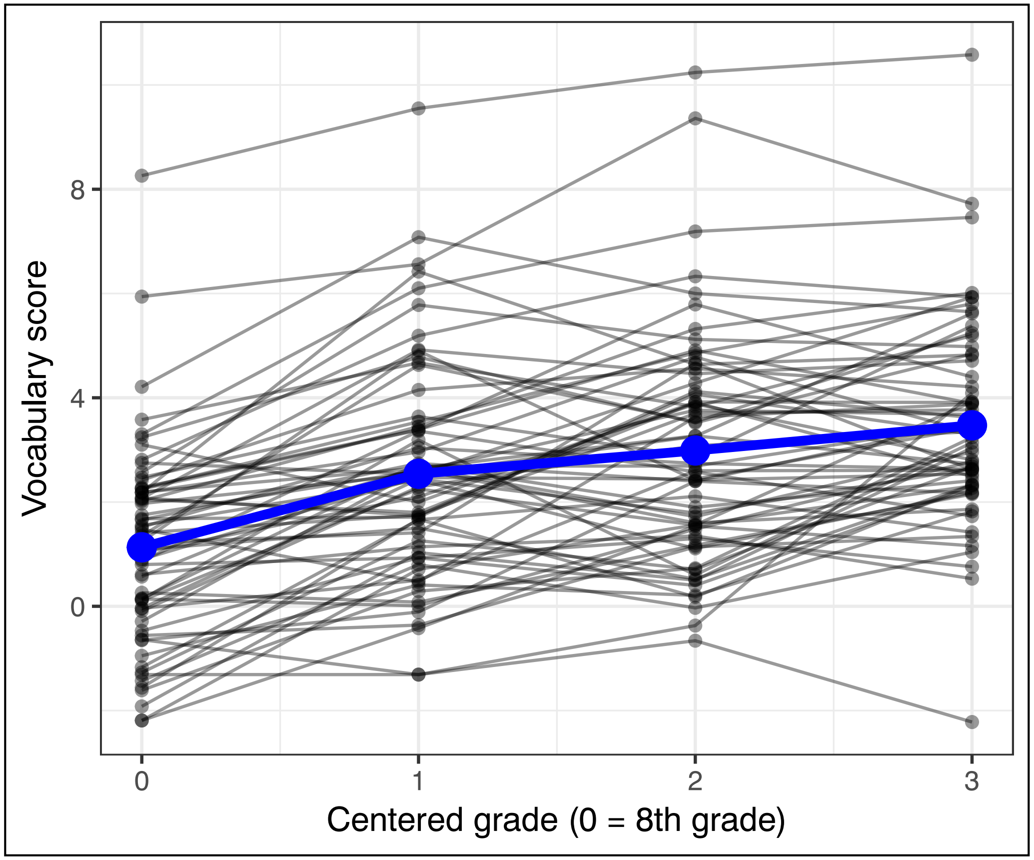
```
# Convert to long format
> vocab_long = vocab %>%
  gather(time, score, t8:t11) %>%
  mutate(
    grade = as.integer(str_replace(time, pattern = "t", replacement = "")) - 8
    ) %>%
  arrange(id, grade)
```

The mutate() function creates a numeric grade column we can use in analysis by replacing the "t" in the time column with nothing (""). Then this is converted from the character string to an integer. Finally, we subtract 8 to center the baseline time at 0.

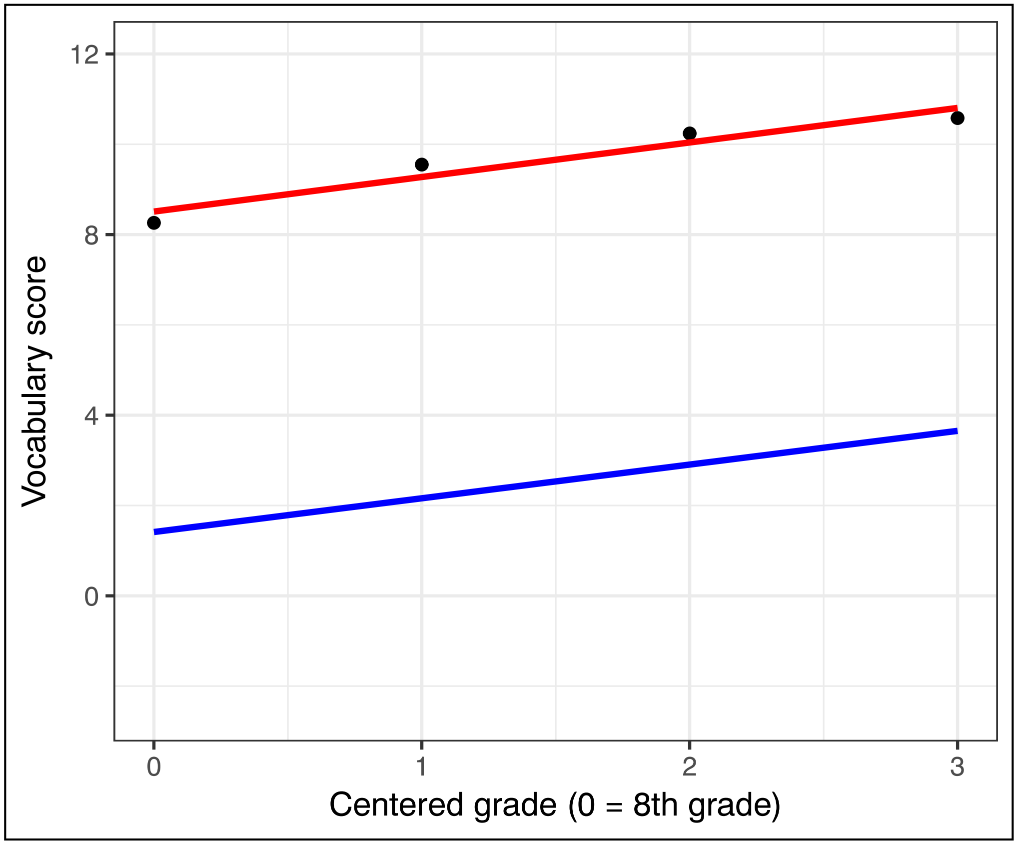
We will now create a spaghetti plot of the vocabulary trajectories and also indicate the average change profile. (For now we will ignore gender.)

```
> ggplot(data = vocab_long, aes(x = grade, y = score, group = id)) +
    geom_line(alpha = 0.4) +
    geom_point(alpha = 0.4) +
    stat_summary(group = 1, fun.y = mean, geom = "line", color = "blue", lwd = 1.5) +
    stat_summary(group = 1, fun.y = mean, geom = "point", color = "blue", size = 4) +
    theme_bw() +
    xlab("Grade") +
    ylab("Vocabulary score")
```

The vocabulary score show growth over time (on average). This growth looks relatively linear. There looks to be variation in students' vocabulary scores in 8th grade (intercepts). There also seems to be variation in students' change over time (slopes).



Let's focus on the the student with ID = 36.



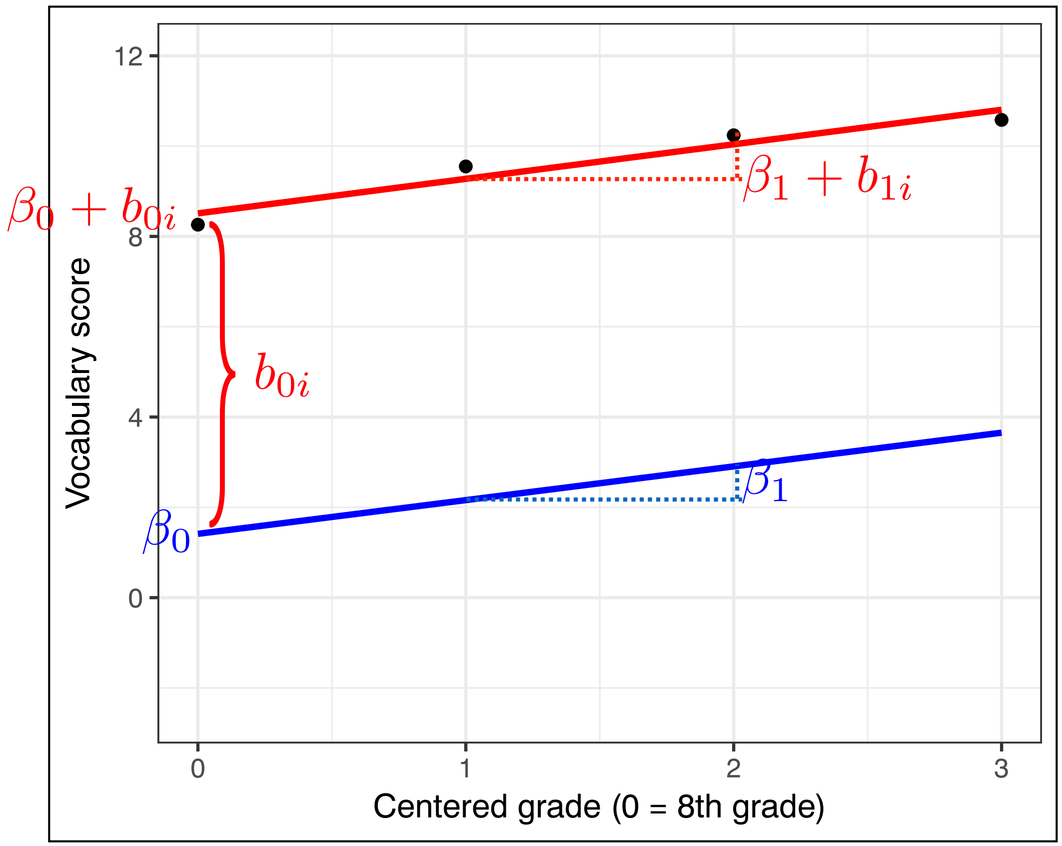
The blue line shows the population-average regression line.

The red line shows the individual-specific regression line.

The residuals are individual- and time-specific. They represent the vertical distance between the individual's data and their specific regression line.

The mixed-effects model specifies this student's individual-specific regression line as a function of the population-average regression line (blue) and deviation from this line.

$$y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})t_{ij} + \epsilon_{ij}$$



The blue line shows the population-average regression line.

- The intercept for this line is β_0.
- The slope for this line is β_1.

The red line shows the individual-specific regression line.

- The intercept for this line is $(\beta_0 + b_{0i})$.
- The slope for this line is denoted $(\beta_1 + b_{1i})$.

We can rearrange the terms in the mixed-effects model as follows:

$$y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})t_{ij} + \epsilon_{ij}$$

$$y_{ij} = \beta_0 + b_{0i} + \beta_1(t_{ij}) + b_{1i}(t_{ij}) + \epsilon_{ij}$$

$$y_{ij} = \beta_0 + \beta_1(t_{ij}) + \Big[b_{0i} + b_{1i}(t_{ij})\Big] + \epsilon_{ij}$$

The model is now rearranged so that the fixed-effects are written first and then the random-effects are written second.

This rearrangement mimics how we will write the syntax to fit the model in R.

$$y_{ij} = \beta_0 + \beta_1(t_{ij}) + \Big[b_{0i} + b_{1i}(t_{ij})\Big] + \epsilon_{ij}$$

```
lmer(y ~ 1 + x + (1 + x | id), data = …, REML = FALSE)
```

We specify the fixed-effects of intercept and an effect of x (similar to `lm()`)

We specify the random-effects of intercept and an effect of x. The `|id` says that we need to compute an intercept and slope for each ID.

We fit the model using Maximum Likelihood estimation.

The `lmer()` function is found in the **lme4** package.

Let's fit the following model to the vocabulary data.

$$\text{Vocabulary}_{ij} = \beta_0 + \beta_1(\text{Age}_{ij}) + \left[b_{0i} + b_{1i}(\text{Age}_{ij})\right] + \epsilon_{ij}$$

```
> library(lme4)

> lmer.1 = lmer(score ~ 1 + grade + (1 + grade | id),
    data = vocab_long, REML = FALSE)

> summary(lmer.1)

Random effects:
 Groups   Name        Variance  Std.Dev. Corr
 id       (Intercept) 3.138e+00 1.771334
          grade       7.514e-05 0.008668 1.00
 Residual             8.963e-01 0.946748
Number of obs: 256, groups:  id, 64

Fixed effects:
            Estimate Std. Error t value
(Intercept)  1.41330    0.24255   5.827
grade        0.74666    0.05294  14.105
```

The fixed-effects section of the output give the estimates for the fixed-effects.

```
Random effects:
 Groups   Name        Variance  Std.Dev. Corr
 id       (Intercept) 3.138e+00 1.771334
          grade       7.514e-05 0.008668 1.00
 Residual             8.963e-01 0.946748
Number of obs: 256, groups:  id, 64

Fixed effects:
            Estimate Std. Error t value
(Intercept)  1.41330    0.24255   5.827
grade        0.74666    0.05294  14.105
```

$$\hat{\beta}_0 = 1.41$$

$$\hat{\beta}_1 = 0.75$$

$$\hat{\text{Vocabulary}}_{ij} = 1.41 + 0.75(\text{Grade}_{ij})$$

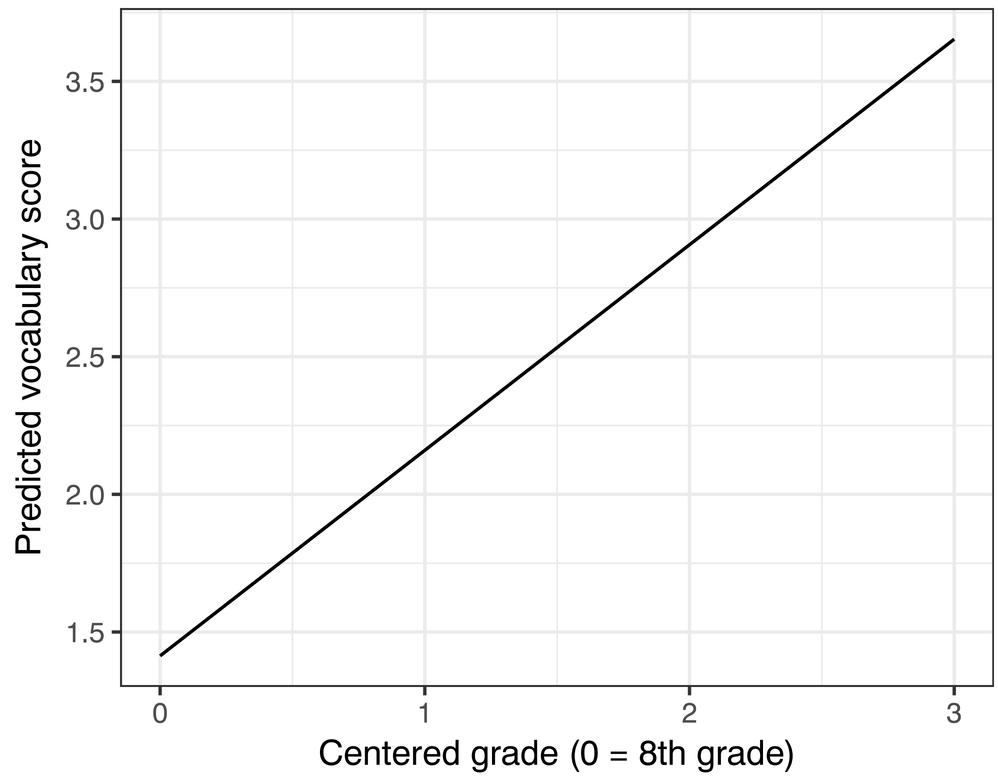
We interpret these effects similar to how we interpret fixed-effect from conventional regression models: (1) The average vocabulary score for students in Grade 8 (centered value of 0) is 1.41. (2) Each one-grade difference is associated with a 0.75-point difference in vocabulary scores, on average.

Let's plot the fitted fixed-effects model (the average model).

```
> plot_data = expand.grid(
    grade = seq(from = 0, to = 3, by = 0.1)
  )

> plot_data$yhat = predict(lmer.1, newdata = plot_data, re.form = NA)

> ggplot(data = plot_data, aes(x = grade, y = yhat)) +
    geom_line() +
    theme_bw() +
    xlab("Centered grade (0 = 8th grade)") +
    ylab("Predicted vocabulary score")
```



To obtain the estimates of the random-effects we use the `ranef()` function. These are the individual deviations.

```
> ranef(lmer.1)

$id
   (Intercept)         grade
1   0.38386117  0.0018784390
2  -0.20702144 -0.0010130671
3  -1.33126425 -0.0065145912
4   1.20628181  0.0059029850
5  -3.62996921 -0.0177633896
:       :              :
:       :              :
60 -1.00811515 -0.0049332490
61  1.00489036  0.0049174684
62  1.92779073  0.0094337158
63 -1.79139330 -0.0087662499
64 -0.76588384 -0.0037478811
```

$$\hat{b}_{01} = 0.384$$

$$\hat{b}_{11} = 0.002$$

$$\hat{b}_{02} = -0.207$$

$$\hat{b}_{12} = -0.001$$

Once we have the estimates of the fixed- and random-effects, we can create the individual-specific equations.

$$\hat{\text{Vocabulary}}_{ij} = (\hat{\beta}_0 + \hat{b}_{0i}) + (\hat{\beta}_1 + \hat{b}_{1i})\text{Age}_{ij}$$

$$\hat{\beta}_0 = 1.41$$

$$\hat{\beta}_1 = 0.75$$

$$\hat{b}_{01} = 0.384$$

$$\hat{b}_{11} = 0.002$$

$$\hat{b}_{02} = -0.207$$

$$\hat{b}_{12} = -0.001$$

<u>Student ID = 1</u>

$$\hat{\text{Vocabulary}}_{1j} = (1.41 + 0.384) + (0.75 + 0.002)\text{Age}_{1j}$$

$$\hat{\text{Vocabulary}}_{1j} = (1.794) + (0.752)\text{Age}_{1j}$$

<u>Student ID = 2</u>

$$\hat{\text{Vocabulary}}_{2j} = (1.41 - 0.207) + (0.75 - 0.001)\text{Age}_{2j}$$

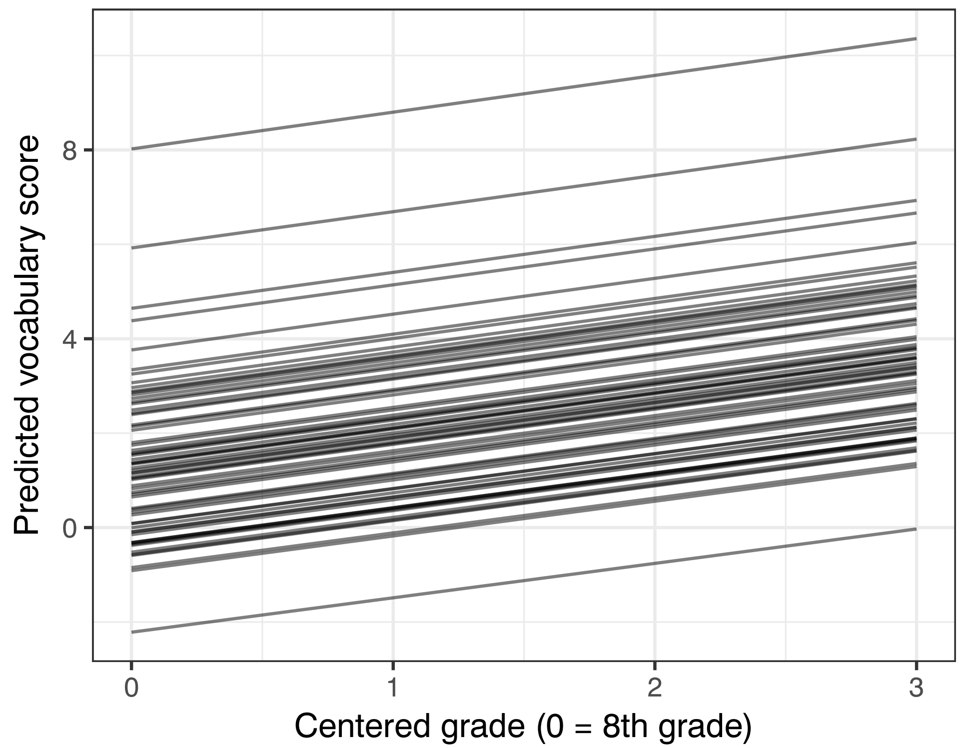
$$\hat{\text{Vocabulary}}_{2j} = (1.203) + (0.749)\text{Age}_{2j}$$

We can also plot the individual-specific lines.

```
> plot_data = expand.grid(
    grade = seq(from = 0, to = 3, by = 0.1),
    id = 1:64
  )

> plot_data$yhat = predict(lmer.1, newdata = plot_data)

> ggplot(data = plot_data, aes(x = grade, y = yhat, group = id)) +
    geom_line() +
    theme_bw() +
    xlab("Centered grade (0 = 8th grade)") +
    ylab("Predicted vocabulary score")
```



The random-effects section of the summary output give the variance estimates for the random-effects.

```
Random effects:
 Groups   Name        Variance  Std.Dev. Corr
 id       (Intercept) 3.138e+00 1.771334
          grade       7.514e-05 0.008668 1.00
 Residual             8.963e-01 0.946748
Number of obs: 256, groups:  id, 64

Fixed effects:
            Estimate Std. Error t value
(Intercept)  1.41330    0.24255   5.827
grade        0.74666    0.05294  14.105
```

$$\hat{\sigma}_0^2 = 3.14$$

$$\hat{\sigma}_1^2 = 0.00008$$

These estimates give us an indication of whether there is variation in the individual-specific intercepts and slopes.

It also gives the variance estimate for the errors.

$$\hat{\sigma}_\epsilon^2 = 0.896$$

What if you thought the trajectory of growth was quadratic rather than linear.

$$\text{Vocabulary}_{ij} = \boxed{\beta_0 + \beta_1(\text{Grade}_{ij}) + \beta_2(\text{Grade}_{ij}^2)} + \left[\boxed{b_{0i} + b_{1i}(\text{Age}_{ij}) + b_{2i}(\text{Age}_{ij}^2)}\right] + \epsilon_{ij}$$

Here we include a fixed-effect of intercept, linear effect of grade, and a quadratic effect of grade.

We also include random-effects of intercept, linear effect of grade, and quadratic effect of grade.

```
> lmer.2 = lmer(score ~ 1 + grade + I(grade^2) + (1 + grade + I(grade^2) | id),
    data = vocab_long, REML = FALSE)
```

```
> summary(lmer.2)

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 id       (Intercept) 2.81670  1.6783
          grade       0.14753  0.3841    0.90
          I(grade^2)  0.02495  0.1579   -0.73 -0.96
 Residual             0.75215  0.8673
Number of obs: 256, groups:  id, 64

Fixed effects:
            Estimate Std. Error t value
(Intercept)  1.18158    0.23489   5.030
grade        1.44181    0.17635   8.176
I(grade^2)  -0.23172    0.05769  -4.017
```

Fitted Fixed-Effects Model

$$\hat{\text{Vocabulary}}_{ij} = 1.18 + 1.44(\text{Grade}_{ij}) + -0.23(\text{Grade}_{ij}^2)$$

Interpretations: (1) The average vocabulary score for students in Grade 8 (centered value of 0) is 1.18. (2) We do not interpret the linear effect since it is a constituent main-effect of an interaction term in the model. (3) The effect of grade on vocabulary score varies depending on grade-level. (To understand this further, plot it.)

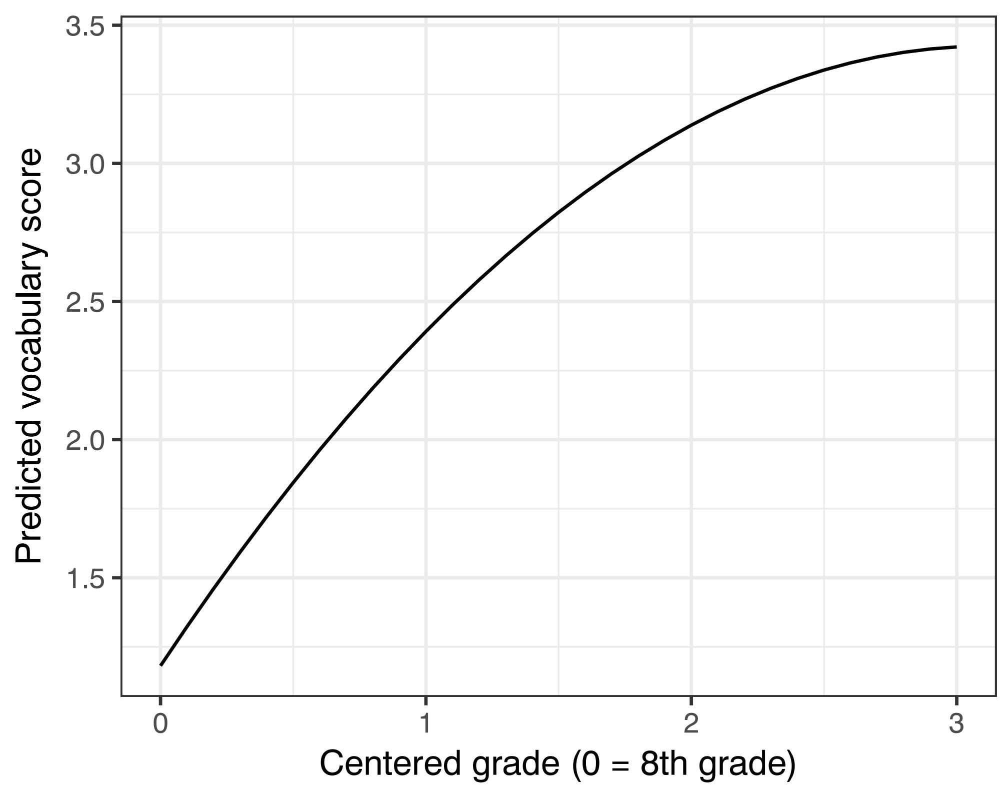
Let's plot the fitted fixed-effects model (the average model).

```
> plot_data = expand.grid(
    grade = seq(from = 0, to = 3, by = 0.1)
  )

> plot_data$yhat = predict(lmer.2, newdata = plot_data, re.form = NA)

> ggplot(data = plot_data, aes(x = grade, y = yhat)) +
    geom_line() +
    theme_bw() +
    xlab("Centered grade (0 = 8th grade)") +
    ylab("Predicted vocabulary score")
```

The effect of grade on vocabulary score varies depending on grade-level. While the effect is generally positive, it tapers off at higher grade levels.



```
> ranef(lmer.2)

$id
   (Intercept)        grade   I(grade^2)
1   0.36731761  0.0786089736 -0.027322586
2  -0.20568090 -0.0485376266  0.018180223
3  -1.29602290 -0.2573244342  0.083634718
4   1.18559996  0.2780796539 -0.103709527
5  -3.32356108 -0.4975656725  0.111022343
:       :             :             :
:       :             :             :
60 -0.88706437 -0.1072434306  0.013343589
61  0.99153792  0.2334634710 -0.087308062
62  1.89386180  0.4311924288 -0.157373741
63 -1.72188315 -0.4051010670  0.151409186
64 -0.70367231 -0.1206750841  0.033275589
```

Fitted Individual-Specific Model for Student 1

$$\hat{\text{Vocabulary}}_{1j} = (1.18 + 0.37) + (1.44 + 0.08)(\text{Grade}_{1j}) + (-0.23 - 0.03)(\text{Grade}^2_{1j})$$

$$\hat{\text{Vocabulary}}_{1j} = 1.55 + 1.52(\text{Grade}_{1j}) - 0.26(\text{Grade}^2_{1j})$$

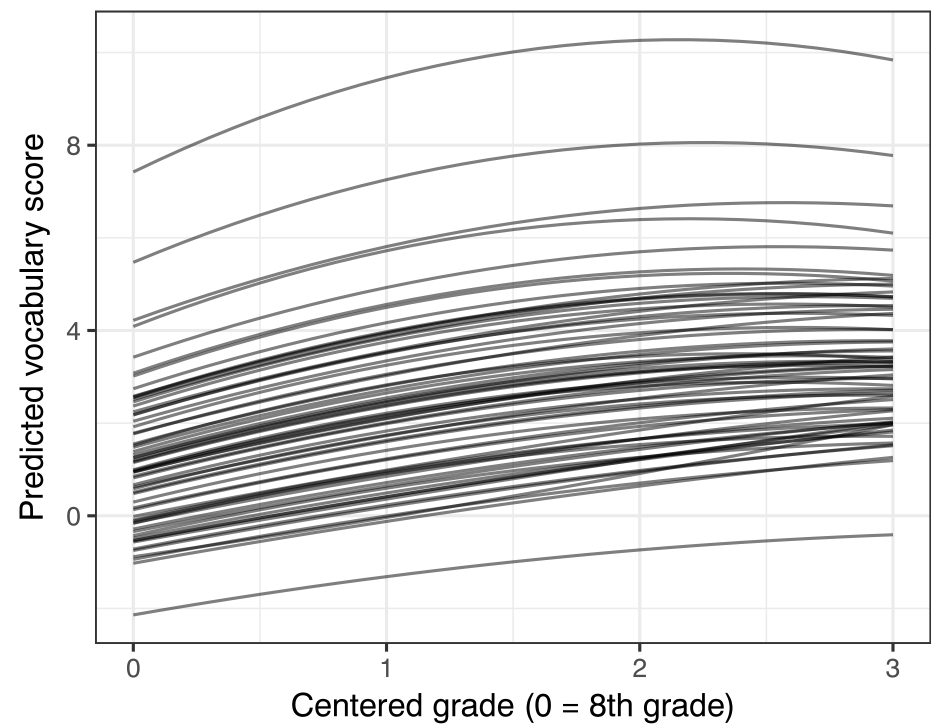
We can also plot the individual-specific lines.

```
> plot_data = expand.grid(
    grade = seq(from = 0, to = 3, by = 0.1),
    id = 1:64
  )

> plot_data$yhat = predict(lmer.2, newdata = plot_data)

> ggplot(data = plot_data, aes(x = grade, y = yhat, group = id)) +
    geom_line() +
    theme_bw() +
    xlab("Centered grade (0 = 8th grade)") +
    ylab("Predicted vocabulary score")
```

We can see the growth-and-taper pattern varies quite a lot between students.



Let's go back to our liner effect of grade model. But this time, let's include a fixed-effect of sex.

$$\text{Vocabulary}_{ij} = \beta_0 + \beta_1(\text{Grade}_{ij}) + \beta_2(\text{Female}_{ij}) + \left[b_{0i} + b_{1i}(\text{Age}_{ij})\right] + \epsilon_{ij}$$

Note that we are not including a random-effect for sex. This is because we want to only model the group-effect (average) of sex. We do not think the effect varies by individuals.

```
> lmer.3 = lmer(score ~ 1 + grade + female + (1 + grade | id),
    data = vocab_long, REML = FALSE)

> summary(lmer.3)

Fixed effects:
            Estimate Std. Error t value
(Intercept)  1.36185    0.31633   4.305
grade        0.74666    0.05294  14.104
female       0.11760    0.46484   0.253
```

Interpretations: (1) Males at grade 8 are expected to have a vocabulary score of 1.36, on average. (2) Females at grade 8 are expected to have a vocabulary score of 1.48 (0.12-points higher than males), on average.
(3) Each grade-level difference is associated with an increase in vocabulary score of 0.75, on average, for both males and females.

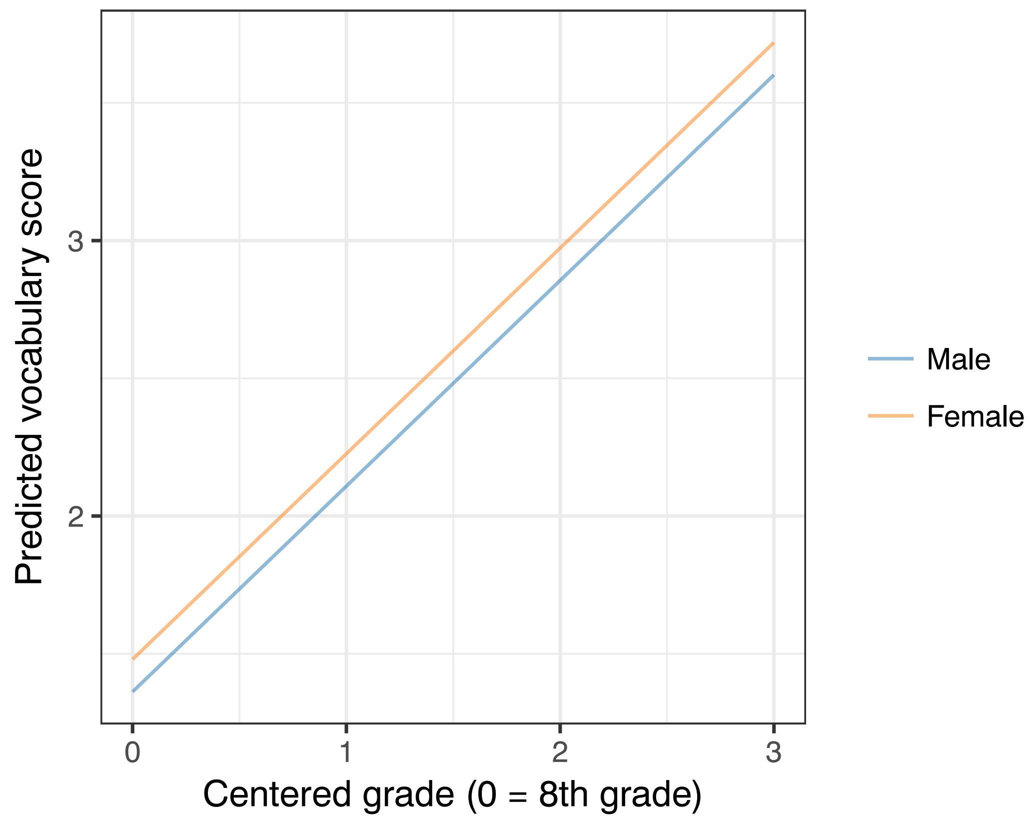
$$\hat{\text{Vocabulary}}_{ij} = 1.36 + 0.75(\text{Grade}_{ij}) + 0.12(\text{Female}_{ij})$$

```
> plot_data = expand.grid(
    grade = seq(from = 0, to = 3, by = 0.1),
    female = 0:1
  )

> plot_data$yhat = predict(lmer.2, newdata = plot_data, re.form = NA)

> plot_data$sex = factor(plot_data$female, levels = c(0, 1), labels = c("Male", "Female"))


> ggplot(data = plot_data, aes(x = grade, y = yhat, color = sex)) +
    geom_line() +
    theme_bw() +
    xlab("Centered grade (0 = 8th grade)") +
    ylab("Predicted vocabulary score") +
    ggsci::scale_color_d3(name = "")
```



To get the individual-specific fitted equations we need the random-effects, and whether the individual was male or female.

```
> ranef(lmer.3)

$id
   (Intercept)         grade
1   0.431194356  2.327778e-03
2  -0.159169084 -8.592653e-04
3  -1.282702876 -6.924599e-03
:          :              :
```

```
> vocab_long %>% filter(id == 1)

# A tibble: 4 x 5
     id female  time score grade
  <int>  <int> <chr> <dbl> <dbl>
1     1      0    t8  1.75     0
2     1      0    t9  2.60     1
3     1      0   t10  3.76     2
4     1      0   t11  3.68     3
```

Fitted Individual-Specific Model for Student 1

$$\hat{\text{Vocabulary}}_{1j} = (1.36 + 0.43) + (0.75 + 0.002)(\text{Grade}_{1j}) + 0.12(0)$$

$$\hat{\text{Vocabulary}}_{1j} = 1.79 + 0.752(\text{Grade}_{1j})$$

References and Source Material

- Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2011). *Applied longitudinal analysis*. New York: Wiley.
- Kohli, N. (2016). *Linear mixed-effects models (course notes)*. Minneapolis, MN: Author.
- Long, J. D. (2012). *Longitudinal data analysis for the behavioral sciences using R*. Thousand Oaks, CA: Sage. (Minneapolis reading data)

- Graphics used from Open Clip Art Library (https://openclipart.org/)