Regression and Mixed-effects Modelling in R : Session 2

Fang Jackson-Yang

2025-05-06

1 RECALL THE DATA & OUR ANALYSES FROM LAST WEEK

1.1 Preparation

We will continue using the simulated dataset called "vocabtrainingdata" from Session 1.

1.2 Descriptive Statistics

1.3 Think Point

Did you notice any problem with the Table 1 ?

What might have caused the issue? How can you fixed it?

2 Visulation

Recall that our goal is to check the effectiveness of the online course on the performance of the pupils in vocabulary tests. We assume that the effectiveness might differ among pupils. Before we fit a mixed-effect model to capture such individual difference, let's visualize the pattern for each pupil.

2.1 Think Point

What have you noticed from the Figure 1 ?

In what ways might this help you answer the research question?

3 Fit A Linear Mixed-effects Model

Now we have a good rough idea about the data, but it lacks precision. For that, we resort to mixed-effects modelling.

We are ready to fit a mixed-effects model to capture the main effect of time spent on the online course (number of weeks), main effect of proficiency level, as well as their interaction effect on pupil's vocab test scores. Importantly, we are going to account for individual difference as a random effect.

The random structure of our mixed-effects models can vary depending on what variances of random effects we include. For example, the simplest model would include only random intercepts but no random slops. But before we fit the model, let's check our understanding of the key concepts.

Table 1: Contingency Table of the Vocabulary Test Score Dataset

proficiency	week	n	Mean	SD	Min	Max
high	0	20	NA	NA	NA	NA
high	1	20	NA	NA	NA	NA
high	2	20	NA	NA	NA	NA
high	3	20	NA	NA	NA	NA
high	4	20	NA	NA	NA	NA
high	5	20	NA	NA	NA	NA
high	6	20	NA	NA	NA	NA
high	7	20	NA	NA	NA	NA
high	8	20	NA	NA	NA	NA
high	9	20	NA	NA	NA	NA
intermediate	0	20	51.55	6.15993	40	64
intermediate	1	20	NA	NA	NA	NA
intermediate	2	20	NA	NA	NA	NA
intermediate	3	20	NA	NA	NA	NA
intermediate	4	20	NA	NA	NA	NA
intermediate	5	20	NA	NA	NA	NA
intermediate	6	20	NA	NA	NA	NA
intermediate	7	20	NA	NA	NA	NA
intermediate	8	20	NA	NA	NA	NA
intermediate	9	20	NA	NA	NA	NA

Table 2: Contingency Table of the Vocabulary Test Score Dataset

proficiency	week	n	Mean	SD	Min	Max
high	0	20	75.16	6.69	62	87
high	1	20	78.32	4.28	69	85
high	2	20	77.50	4.62	64	83
high	3	20	77.42	5.53	66	86
high	4	20	77.74	6.18	66	89
high	5	20	77.00	6.24	66	91
high	6	20	77.67	6.70	68	89
high	7	20	77.79	6.43	66	90
high	8	20	80.33	5.42	73	92
high	9	20	82.76	4.32	75	92
intermediate	0	20	51.55	6.16	40	64
intermediate	1	20	53.11	6.51	39	63
intermediate	2	20	53.42	5.78	41	65
intermediate	3	20	56.56	7.66	41	69
intermediate	4	20	57.67	8.39	42	73
intermediate	5	20	61.78	10.37	43	83
intermediate	6	20	62.44	12.63	43	91
intermediate	7	20	63.83	10.01	44	83
intermediate	8	20	67.28	12.02	45	91
intermediate	9	20	70.11	13.41	47	93

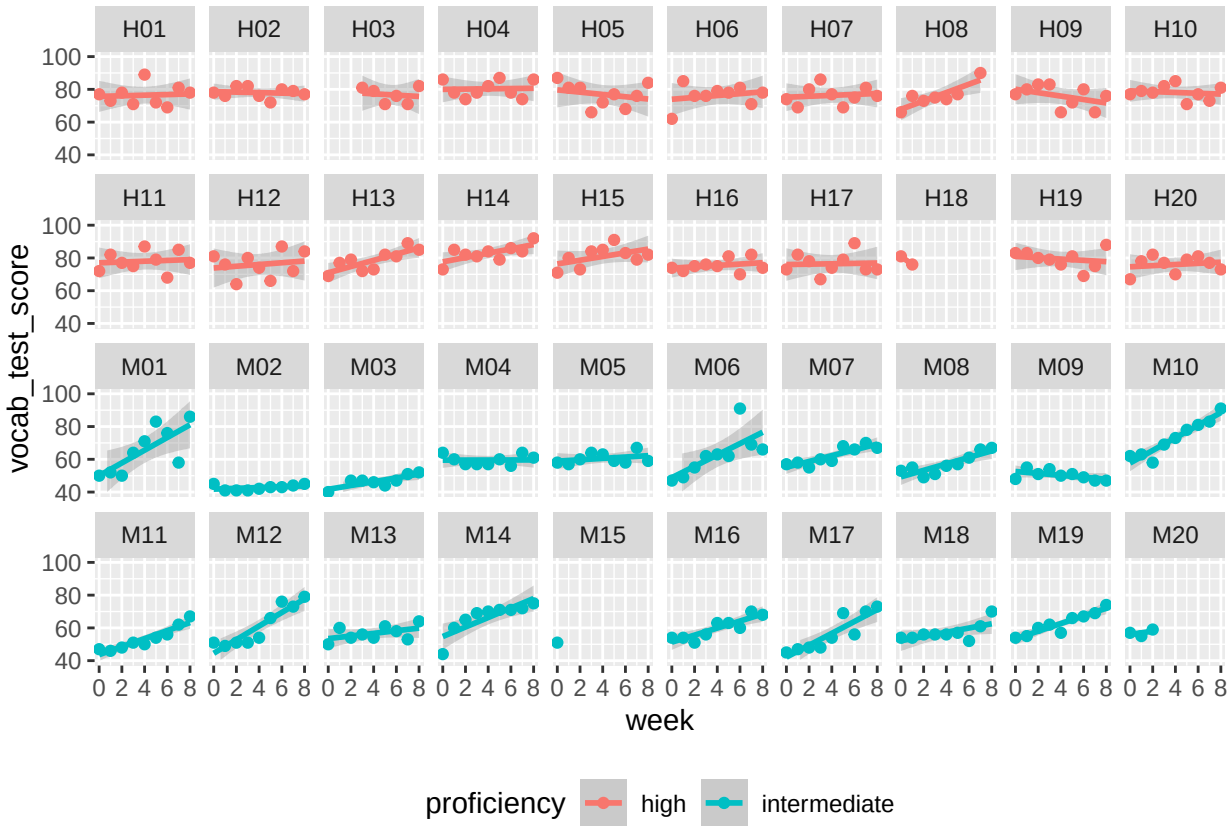


Figure 1: Individual Difference among Pupils Enrolled in the Online Course

3.1 Think Point

What is intercept?

What is slope?

How can we make sense of them in the context of random effects?

Recall the plots we had from Session 1 shown in Figure 2. Panel (a) with the Panel (b) both show the relationship between the number of weeks spent on the online course and the scores of vocabulary tests. How do the two plots differ? What does that tell you?

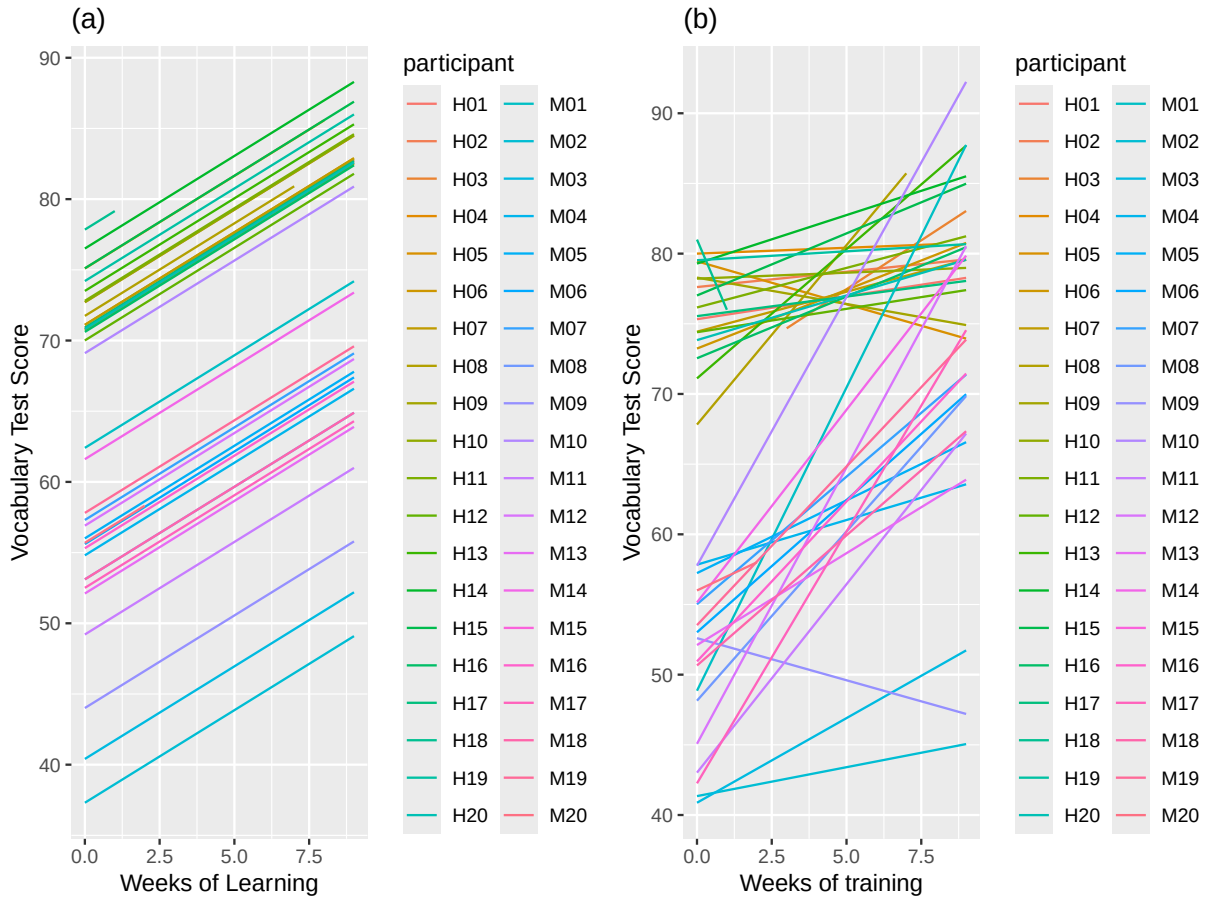


Figure 2: Random Intercepts (a) vs Random Intercepts and Slopes (b)

Looking at the model results, it does not look practical to include participants as a predictor. We will move to mixed-effect modelling by treating the individual differences as random effects.

4 MIXED-EFFECT MODELLING

4.1 Fit a Simplest Mixed-Model

A mixed-model with a simplest random structure means that it only includes random intercepts, but no random slopes. In our example, it means we assume that pupils differ in their test scores, but we assume

that the online course has the same effect on each pupil (the slope for each pupil is the same, i.e., no random slopes).

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: vocab_test_score ~ week * proficiency + (1 | participant)
##    Data: vocabdata
##
## REML criterion at convergence: 2438.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.3014 -0.5552  0.0434  0.6075  4.4641
##
## Random effects:
##  Groups      Name        Variance Std.Dev.
##  participant (Intercept) 25.47    5.046
##  Residual                35.45    5.954
## Number of obs: 369, groups:  participant, 40
##
## Fixed effects:
##                              Estimate Std. Error t value
## (Intercept)                   75.9310     1.3932  54.499
## week                           0.5023     0.1558   3.225
## proficiencyintermediate      -25.3754     1.9676 -12.896
## week:proficiencyintermediate   1.5755     0.2185   7.210
##
## Correlation of Fixed Effects:
##             (Intr) week   prfcnc
## week        -0.486
## prfcncyntrm -0.708  0.344
## wk:prfcncyn  0.346 -0.713 -0.477
```

Look at the results of the model, what did you notice? There is no p-value!

Do not panic. This can be calculated and the package "lmerTest" does this job for us. Install the package and run the library. Then fit your model again. Now you should get the p-values. If not, try to specify from which library you want to draw the lmer() function, e.g., lmerTest::lmer(). Now run your model again.

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: vocab_test_score ~ week * proficiency + (1 | participant)
##    Data: vocabdata
##
## REML criterion at convergence: 2438.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.3014 -0.5552  0.0434  0.6075  4.4641
##
## Random effects:
##  Groups      Name        Variance Std.Dev.
##  participant (Intercept) 25.47    5.046
##  Residual                35.45    5.954
## Number of obs: 369, groups:  participant, 40
```

```
##
## Fixed effects:
##                               Estimate Std. Error       df t value Pr(>|t|)
## (Intercept)                    75.9310     1.3932  65.6532  54.499  < 2e-16 ***
## week                            0.5023     0.1558 334.4194   3.225  0.00138 **
## proficiencyintermediate       -25.3754     1.9676  65.0011 -12.896  < 2e-16 ***
## week:proficiencyintermediate    1.5755     0.2185 334.4965   7.210 3.76e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) week   prfcnc
## week        -0.486
## prfcncyntrm -0.708  0.344
## wk:prfcncyn  0.346 -0.713 -0.477
```

Now you should get your p-values.You can see that the parameters are exactly the same, the only difference is that you now additionally get a column indicating significance.

4.1.1 Interpret the results

```
summary(mMixed1_pval)
```

What does the model results tell you?

Did you notice something that we did not have when we fitted simple regressions last week? (hint: individual difference among participants. which section of the model output gives us such information?)

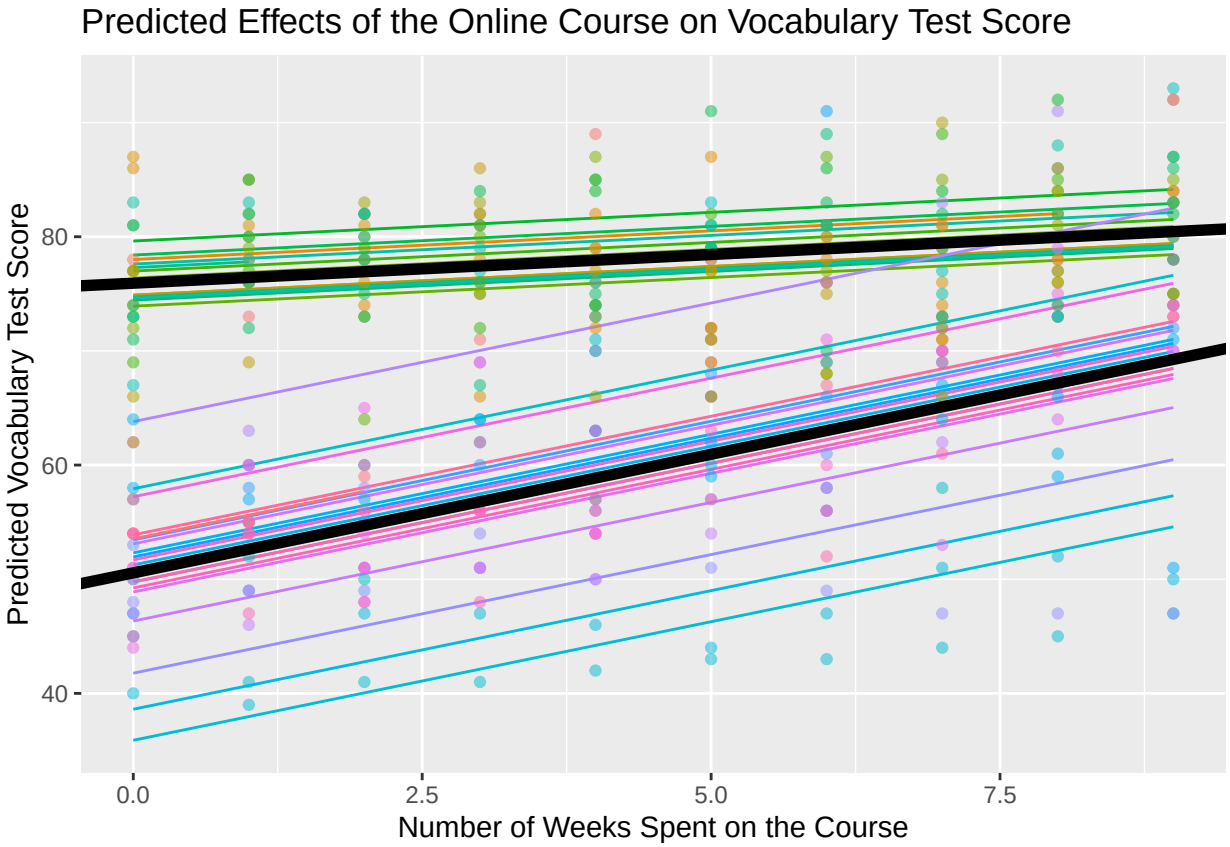
Did you notice something that was involved in the output of a simple regression model but disappeared here? (hint: recall overall model fit)

Can you interpret and report the model results?

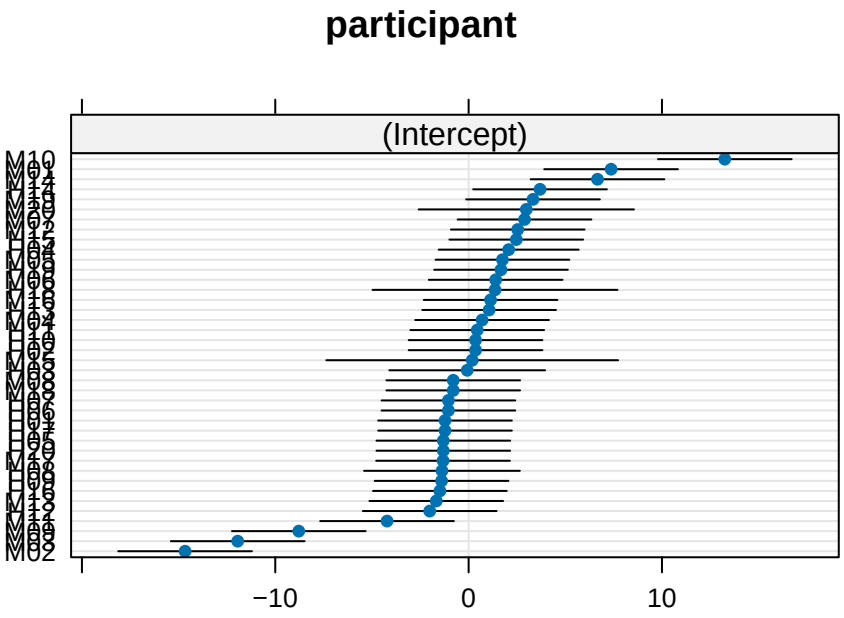
4.1.2 Visualise the model

4.1.2.1 Fixed effects

```
            (Intercept)                      week
             75.9310113                 0.5023305
proficiencyintermediate week:proficiencyintermediate
            -25.3753865                 1.5755482
```



4.1.2.2 Random effects The quick and easy way to visualise the the variance of random effects is to use the dotplot.ranef.mer() function in lme4.



Regression Table of the Simplest Model

	Vocabulary Test Score			
Predictors	*Estimates*	*CI*	*Statistic*	*p*
(Intercept)	75.93	73.19 – 78.67	54.50	**<0.001**
Weeks of Learning	0.50	0.20 – 0.81	3.22	**0.001**
proficiencyintermediate	-25.38	-29.24 – -21.51	-12.90	**<0.001**
week:proficiencyintermediate	1.58	1.15 – 2.01	7.21	**<0.001**
Random Effects				
σ^2	35.45			
$\tau_{00\ participant}$	25.47			
ICC	0.42			
$N_{\ participant}$	40			
Observations	369			
Marginal R^2 / Conditional R^2	0.631 / 0.785			

4.1.2.3 Report the results "We fitted a mixed model including week of learning, proficiency level as well as their interaction as fixed effects, and by-participant intercept as a random effect (random effects for participant had variance of 25.47 and SD of 5.05). Proficiency level was dummy coded using"high" proficiency level as the reference level. The model showed a significant intercept, indicating that students' test scores differed. Number of weeks into the online learning course was a significant predictor for the test score (β_1=0.50, SE=0.16, 95%CI = [.20, .81], t= 3.22, p< .01). Students' proficiency level was also a significant predictor for the test score; those with low proficiency performed significantly worse than the high proficiency group (β_2=-25.38, SE=1.97, 95%CI = [-29.24, -21.51], t= -12.9, p< .001). Moreover, the model also revealed a significant interaction effect between week and proficiency. The effect of the online learning course on improving pupils' vocab test scores was larger for the low proficient group than high proficient group (β_3=1.58, SE=.22, 95%CI = [1.15, 2.01], t = 7.21, p< .001)."

Also note the model fit. Note here we get Marginal R-squared and conditional R-squared. The former represents how much variance in the date that your fixed effects (i.e., predictors) can explain, whereas the latter represents how much variance in the data that the model overall can explain.

4.1.2.4 Other aspects of the results Often not needed when reporting model results in your paper.

To view the results of a mixed model, instead of using the built-in function summary(), we can use the augment() function in the broom.mixed package. This is a handy function that will give you a summary table including the fitted values, residuals, hat values, and so forth. If you are unfamiliar with these concepts, do not worry, you often do not need to report these in your paper. If you are curious, hat values and Cook's D are parts of model diagnostics, used to identify influential data or outliers. More on this next week.

5 TIME FOR A BREAK

In the first hour of today's session, we built a mixed model accounting for by-subject random intercepts. What about by-subject random slopes?

Let's build models with different structures of random effects.

5.1 Account for random slopes

We start with a mixed-model with a full/max random structure. This means that the model includes all sources of random variances, including both random intercepts and random slopes for all predictors.

In our example, it means we assume that pupils differ in their test scores. We also assume that the online course has different effects on each pupil (the slope for each pupil is different) and that proficiency level shows difference influence on each pupil. Moreover, we assume proficiency level also has a different effect on the influence of the online course on test score for each pupil.

We get a warning message telling us the model failed to converge. One way to deal with convergence issue is to adjust optimizer.

Still failed to converge. This means our model cannot explain the data with all our hypothesised effects. We need to simplify the random structure.

5.2 Fit a Reduced Mixed-Model

Now we simply our random structure. Fit a model that captures the random effect slope of the online course (but not proficiency level) on each pupil.

Think point:

How can we reduce the random-effect structure while still accounting for by-subject random slope?

Which element can we drop from the random-effect structure?

Below is an example. Can you explain the rationale of the random-effect strauctur in this model?

5.2.1 Think Point

What results did you get from this reduced model?

Can you create plots to visualise the fixed effects?

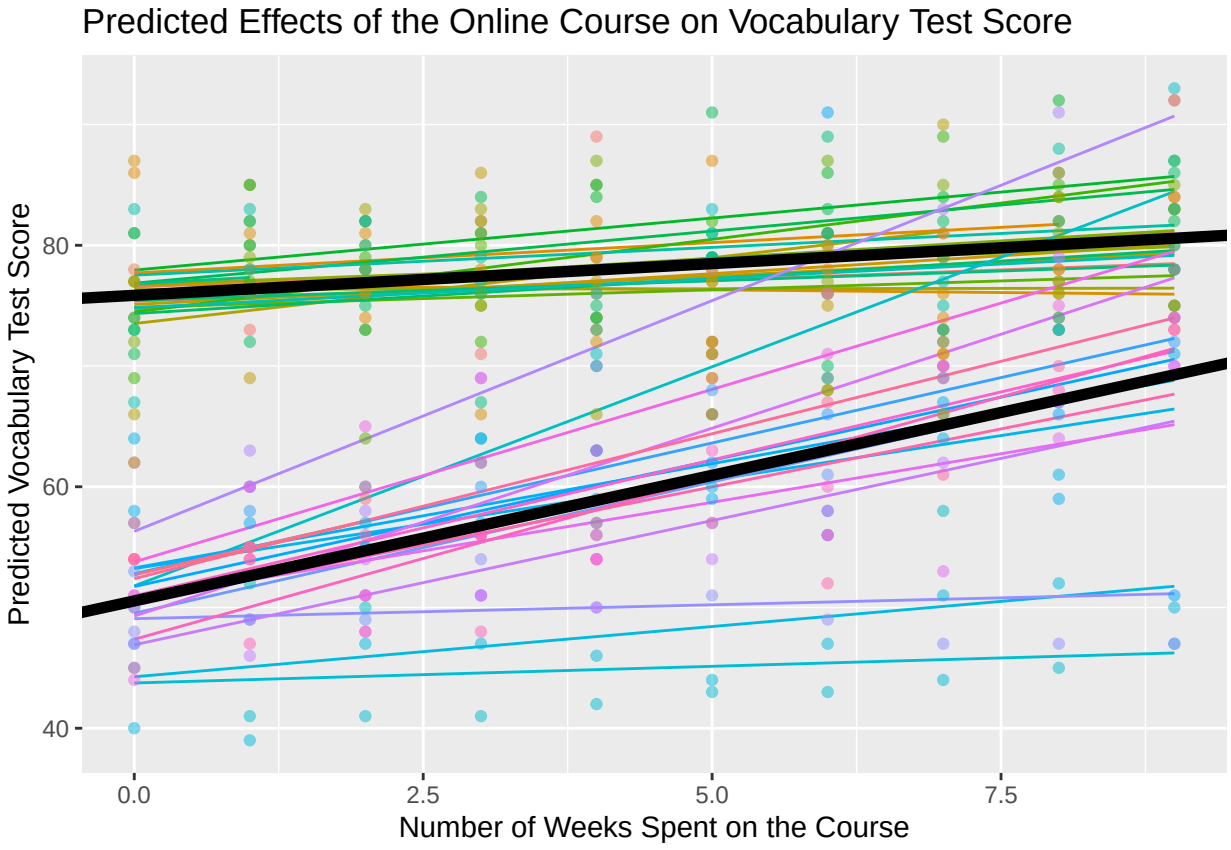
Can you plot the random effects?

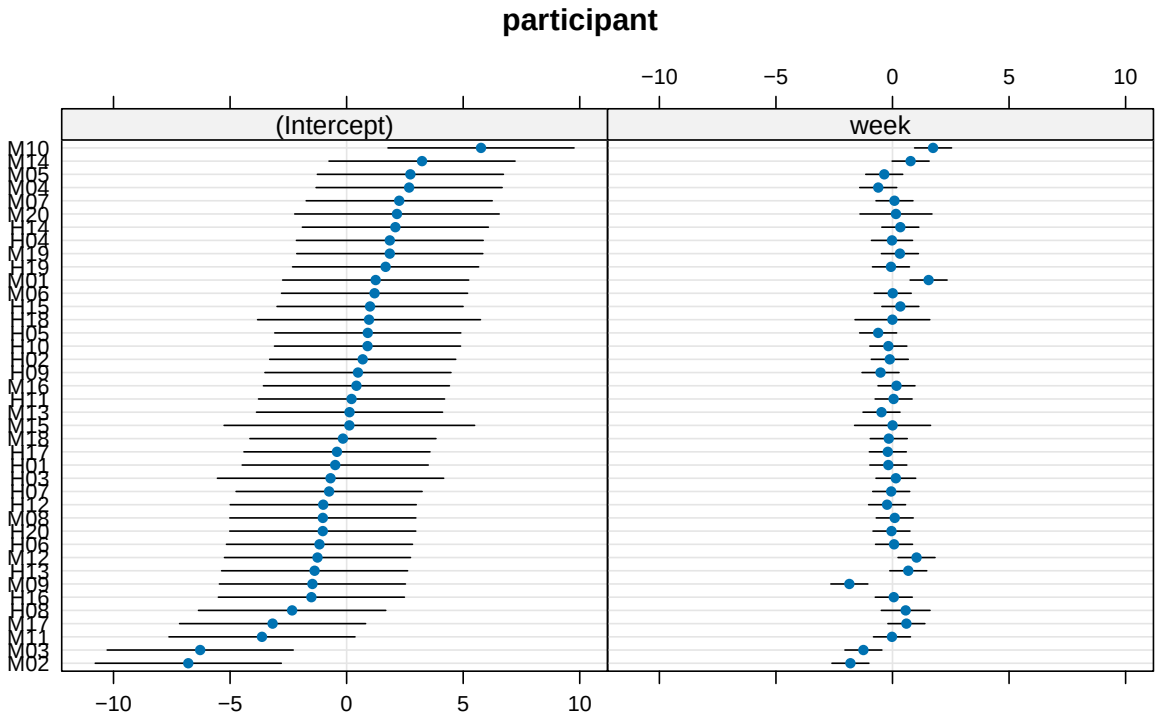
How would you report the results?

It's your turn. Try to reuse the code we use for the simplest model to address above questions based on the results of the reduced model.

Try to write your own code in the following empty chunck.

Check the sample code below if you struggle.


Predicted Effects of the Online Course on Vocabulary Test Score
Predicted Vocabulary Test Score
80
60
40
0.0
2.5
5.0
7.5
Number of Weeks Spent on the Course


participant
(Intercept)
week
−10
−5
0
5
10
M10
M14
M05
M04
M07
M20
H14
H04
M19
H19
M01
M06
H15
H13
H05
H10
H02
H09
M16
H11
M13
M15
M18
H17
H01
H03
H07
H12
M08
H20
H06
M12
H18
M09
H16
H08
M17
M11
M03
M02

Regression Table of the Simplest Model

	Vocabulary Test Score			
Predictors	*Estimates*	*CI*	*Statistic*	*p*
(Intercept)	75.85	73.83 – 77.88	73.53	**<0.001**
Weeks of Learning	0.53	0.06 – 1.00	2.20	**0.029**
proficiencyintermediate	-25.31	-28.17 – -22.46	-17.42	**<0.001**
week:proficiencyintermediate	1.56	0.89 – 2.22	4.58	**<0.001**
Random Effects				
σ^2	29.51			
$\tau_{00\ participant}$	10.09			
$\tau_{11\ participant.week}$	0.69			
$\rho_{01\ participant}$	0.11			
ICC	0.52			
$N_{participant}$	40			
Observations	369			
Marginal R^2 / Conditional R^2	0.630 / 0.822			

6 Model Comparison & Selection

Now we have two models, one with the simplest structure (intercept only) and the other includes both random intercept and random slope. Which model has a better goodness of fit?

```
anova(mMixed1_pval, mMixed_reduced)
```

```
## Data: vocabdata
## Models:
## mMixed1_pval: vocab_test_score ~ week * proficiency + (1 | participant)
## mMixed_reduced: vocab_test_score ~ week * proficiency + (1 + week | participant)
##                npar    AIC    BIC  logLik -2*log(L)  Chisq Df Pr(>Chisq)
## mMixed1_pval      6 2451.4 2474.8 -1219.7    2439.4
## mMixed_reduced    8 2416.2 2447.5 -1200.1    2400.2 39.151  2  3.151e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

What can you conclude?

7 What's next?

7.1 Exercise (rest of today and Friday)

Can you fit a model with a different structure of random effects and interpret the results?

Hint: consider including some or all of the following: - random intercept?
- random slope of one predictor? - random slopes of both predictors? - random slope of the interaction between the two predictors?

7.2 Also for Friday: BYOD (Bring your own data)