Linear Mixed-Effects Models: Likelihood Ratio Tests

2020-03-11

In this set of notes, you will learn about the likelihood ratio test to analyze nested linear mixed-effects models.

Dataset and Research Question

In this set of notes, we will use data from the file *vocabulary.csv* (see the data codebook here). These data include repeated measurements of scaled vocabulary scores for $n = 64$ students.

```
# Load libraries
library(broom)
library(educate)
library(lme4) #for fitting mixed-effects models
library(MuMIn)
library(patchwork)
library(tidyverse)

# Read in data
vocabulary = read_csv(file = "~/Documents/github/epsy-8252/data/vocabulary.csv")
head(vocabulary)
```

```
# A tibble: 6 x 6
     id vocab_08 vocab_09 vocab_10 vocab_11 female
  <dbl>    <dbl>    <dbl>    <dbl>    <dbl>  <dbl>
1     1     1.75     2.6      3.76     3.68       1
2     2     0.9      2.47     2.44     3.43       0
3     3     0.8      0.93     0.4      2.27       0
4     4     2.42     4.15     4.56     4.21       1
5     5    -1.31    -1.31    -0.66    -2.22       0
6     6    -1.56     1.67     0.18     2.33       0
```

We will use these data to again explore the change in vocabulary over time (longitudinal variation in the vocabulary scores). We will focus on two primary research questions: (1) What is the growth pattern in the average vocabulary score over time? and (2) Is this growth pattern different for females and non-females?

Data Preparation

Prior to fitting any of the models, we will again create the tidy (long) dataset, create a quantitative (centered) predictor of grade-level that we will include in the models.

```
# Create lookup table
lookup_table = data.frame(
  grade = c("vocab_08", "vocab_09", "vocab_10", "vocab_11"),
  grade_quant_center = c(0, 1, 2, 3)
)

# Convert from wide to long structured data
vocabulary_long = vocabulary %>%
  pivot_longer(cols = vocab_08:vocab_11, names_to = "grade", values_to = "vocab_score") %>%
  left_join(lookup_table, by = "grade") %>%
  arrange(id, grade)

# View data
head(vocabulary_long)
```

```
# A tibble: 6 x 5
     id female grade    vocab_score grade_quant_center
  <dbl>  <dbl> <chr>          <dbl>              <dbl>
1     1      1 vocab_08        1.75                  0
2     1      1 vocab_09        2.6                   1
3     1      1 vocab_10        3.76                  2
4     1      1 vocab_11        3.68                  3
5     2      0 vocab_08        0.9                   0
6     2      0 vocab_09        2.47                  1
```

Fit a Set of Nested Models to Evaluate Effects of Interest

The likelihood ratio test allows us to use a p-value approach to evaluate models. This test does, however, require that we have nested models we are comparing. To this end, we need to fit a sequence of nested models that will allow us to evaluate different hypotheses of interest. Consider the following models:

$$\textbf{Model 0}: \text{Vocabulary Score}_{ij} = [\beta_0 + b_{0j}] + \epsilon_{ij}$$
$$\textbf{Model 1}: \text{Vocabulary Score}_{ij} = [\beta_0 + b_{0j}] + \beta_1(\text{Grade}_{ij}) + \epsilon_{ij}$$
$$\textbf{Model 2}: \text{Vocabulary Score}_{ij} = [\beta_0 + b_{0j}] + \beta_1(\text{Grade}_{ij}) + \beta_2(\text{Grade}^2_{ij}) + \epsilon_{ij}$$

Comparing Model 0 to Model 1 allows us to test whether there is an effect of grade-level on vocabulary scores. Then, comparing Model 1 to Model 2 would allow us to test the hypothesis that there is a nonlinear (quadratic) effect of grade-level on vocabulary scores.

Likelihood Ratio Test: p-Values for Mixed-Effects Models

So long as the assumptions of the linear mixed-effects model have been met (see Assumptions notes), we can obtain a p-value for testing whether the added effects in the more complex model explain additional variation in the outcome. Equivalently, this tests whether all of the added effects are zero. For example in comparing Model 0 to Model 1, we would be testing the hypothesis that:

$$H_0 : \beta_{\text{Grade-Level}} = 0$$

whereas, if we were comparing Model 0 to Model 2, we would be testing the hypothesis that:

$$H_0 : \beta_{\text{Grade-Level}} = \beta_{\text{Grade-Level}^2} = 0$$

To carry out a Likelihood Ratio Test, we use the `anova()` function and input the two mixed-effects models we want to compare. (Note: Both models need to be fitted with ML.)

```
# Fit unconditional random intercepts model
lmer.0 = lmer(vocab_score ~ 1 + (1|id), data = vocabulary_long, REML = FALSE)

# Fit unconditional linear growth model
lmer.1 = lmer(vocab_score ~ 1 + grade_quant_center + (1|id), data = vocabulary_long, REML = FALSE)

# LRT
anova(lmer.0, lmer.1)
```

```
Data: vocabulary_long
Models:
lmer.0: vocab_score ~ 1 + (1 | id)
lmer.1: vocab_score ~ 1 + grade_quant_center + (1 | id)
       Df  AIC  BIC logLik deviance Chisq Chi Df          Pr(>Chisq)
lmer.0  3 1015 1026   -505     1009
lmer.1  4  881  895   -436      873   137      1 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the null hypothesis being tested is that the reduced model (`lmer.0`) and the full model (`lmer.1`) fit the data equally well, the test needs to evaluates the misfit (error) of each model. Recall that the way we measure error when we have use likelihood to estimate models is via the deviance.

The deviance of the reduced model is 1009 and that for the full model is 873. If the two models fit equally well, we would expect the difference in deviance to be zero. The actual difference in deviance is 137. (This is often referred to as ΔG^2, for goodness-of-fit, or as χ^2.) This indicates that the more complex model fits the sample data better than the reduced model; the more complex model has a smaller deviance.

As with any difference, we wonder whether this is within what would be expected because of sampling (chance) variation. To test this, we evaluate ΔG^2 in a χ^2-distribution with df equal to the difference in K between the two models (K is the df for each model). This difference should be the difference in the complexity between the two models; the difference in the estimated number of parameters. Our reduced model has three parameters being estimated:

- $\hat{\beta}_0$,
- $\hat{\sigma}^2_\epsilon$, and
- $\hat{\sigma}^2_0$

And our complex model has four parameters being estimated:

- $\hat{\beta}_0$,
- $\hat{\beta}_{\text{Grade}}$,
- $\hat{\sigma}^2_\epsilon$, and
- $\hat{\sigma}^2_0$)

The difference in complexity between these models is $4 - 3 = 1$.

```
1 - pchisq(137, df = 1)
```

```
[1] 0
```

Note that all of these results are given in the `anova()` output. This is typically reported as something like:

> A likelihood ratio test indicated that the model that included the fixed-effects of grade-level fitted the data better that the unconditional random intercepts model, $\chi^2(1) = 137, p < .001$. This suggests that there is an effect of grade on average vocabulary scores.

Note that had we used the unconditional growth model that included the categorical predictor of grade-level, that this model would include six parameter estimates:

- $\hat{\beta}_0$,
- $\hat{\beta}_{\text{Vocab\_09}}$,
- $\hat{\beta}_{\text{Vocab\_10}}$,
- $\hat{\beta}_{\text{Vocab\_11}}$,
- $\hat{\sigma}^2_\epsilon$, and
- $\hat{\sigma}^2_0$)

In this case our LRT results would have been:

```
# Fit model with categorical grade predictor
lmer.1.cat = lmer(vocab_score ~ 1 + grade + (1|id), data = vocabulary_long, REML = FALSE)

# LRT
anova(lmer.0, lmer.1.cat)
```

```
Data: vocabulary_long
Models:
lmer.0: vocab_score ~ 1 + (1 | id)
lmer.1.cat: vocab_score ~ 1 + grade + (1 | id)
           Df  AIC  BIC logLik deviance Chisq Chi Df          Pr(>Chisq)
lmer.0      3 1015 1026   -505     1009
lmer.1.cat  6  865  886   -426      853   156      3 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Although the results are similar, there are some differences in them. The model complexity is greater for the unconditional growth model fitted with the categorical predictor ($k = 6$ rather than $k = 4$). The deviance is also slightly smaller (worse fit). These two values directly influence the size of χ^2 and of the p-value.

In general, you should use a quantitative measure for time when you can!

Mathematics Behind the Likelihood Ratio Test

Why is this called a *likelihood ratio test*? Remember that the deviance is equal to $-2\ln(\mathcal{L})$. Thus the difference in deviance can be written as:

$$\Delta G^2 = -2\ln\left[\mathcal{L}(\text{Reduced Model})\right] - \left[-2\ln\left[\mathcal{L}(\text{Full Model})\right]\right]$$

Pulling out the -2 we get

$$\Delta G^2 = -2\left[\ln\left[\mathcal{L}(\text{Reduced Model})\right] - \ln\left[\mathcal{L}(\text{Full Model})\right]\right]$$

The difference between two logarithms, e.g., $\log(A) - \log(B)$ is the logarithm of the quotient ($\log(\frac{A}{B})$). Thus, we can re-write this as,

$$\Delta G^2 = -2\ln\left[\frac{\mathcal{L}(\text{Reduced Model})}{\mathcal{L}(\text{Full Model}}\right]$$

Now it should be a little more apparent why this test is called a likelihood RATIO test. Note that if both models fit the data equally well, their likelihood values would be equivalent and thus this equation would reduce to:

$$\begin{aligned}\Delta G^2 &= -2\ln\left[1\right] \\ &= -2(0) \\ &= 0\end{aligned}$$

Thus if the difference in the goodness-of-fit between the two models turns out to be zero (or are within chance variation of zero), both models fit the data equally and thus we should adopt the reduced model (Occam's Razor).

Back to the Example

The LRT allowed us to compare the baseline model to the model that included the linear effect of grade-level. Is there a nonlinear effect of grade-level?

```
# Fit unconditional quadratic growth model
lmer.quad = lmer(vocab_score ~ 1 + grade_quant_center + I(grade_quant_center ^ 2) +
                (1|id), data = vocabulary_long, REML = FALSE)

# LRT
anova(lmer.1, lmer.quad)
```

```
Data: vocabulary_long
Models:
lmer.1: vocab_score ~ 1 + grade_quant_center + (1 | id)
lmer.quad: vocab_score ~ 1 + grade_quant_center + I(grade_quant_center^2) +
lmer.quad:     (1 | id)
          Df AIC BIC logLik deviance Chisq Chi Df Pr(>Chisq)
lmer.1     4 881 895   -436      873
```

```
lmer.quad  5 867 884   -428      857    16      1   0.000064 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This test suggests that the model that included the linear and quadratic fixed-effects of grade-level fitted the data better that the model that only included the linear effect of grade-level, $\chi^2(1) = 16$, $p = .00006$. This suggests that there may be a nonlinear effect of grade-level on average vocabulary scores.

Would it be better to use the model with the log-transformed grade-level rather than the quadratic effect of grade-level? This is something we cannot test since the log model is not nested in the quadratic model. You will have to use residual plots and sound substantive judgment to make this determination.

As in the previous set of notes, we will adopt the log-linear model. We can compare the log-linear model to the baseline model, as the unconditional random intercepts model is nested in the log-linear model.

```
# Log-linear model
lmer.log = lmer(vocab_score ~ 1 + log(grade_quant_center + 1) + (1|id),
              data = vocabulary_long, REML = FALSE)

# LRT
anova(lmer.0, lmer.log)
```

```
Data: vocabulary_long
Models:
lmer.0: vocab_score ~ 1 + (1 | id)
lmer.log: vocab_score ~ 1 + log(grade_quant_center + 1) + (1 | id)
         Df  AIC  BIC logLik deviance Chisq Chi Df          Pr(>Chisq)
lmer.0    3 1015 1026   -505     1009
lmer.log  4  864  878   -428      856   153      1 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

These LRT results are similar to those we saw previously, and suggest there is likely an effect of grade-level.

Taxonomy of Sequential Models to Test Research Question 2

The second research question we had was, whether the growth pattern is different for females and non-females. To evaluate this with statistical tests we might initially compare the unconditional growth model (`lmer.5`) to a model that also includes the effect of `female`. Then we might compare that model to a model that also includes an interaction between `female` and the log-transformed grade-level.

```
# Main effects model
lmer.main = lmer(vocab_score ~ 1 + log(grade_quant_center + 1) + female +
               (1|id), data = vocabulary_long, REML = FALSE)

# Interaction model
lmer.int = lmer(vocab_score ~ 1 + log(grade_quant_center + 1) + female + log(grade_quant_center + 1):female +
               (1|id), data = vocabulary_long, REML = FALSE)

# LRT
anova(lmer.log, lmer.main, lmer.int)
```

```
Data: vocabulary_long
Models:
lmer.log: vocab_score ~ 1 + log(grade_quant_center + 1) + (1 | id)
lmer.main: vocab_score ~ 1 + log(grade_quant_center + 1) + female + (1 |
lmer.main:     id)
lmer.int: vocab_score ~ 1 + log(grade_quant_center + 1) + female + log(grade_quant_center +
lmer.int:     1):female + (1 | id)
          Df AIC BIC logLik deviance Chisq Chi Df    Pr(>Chisq)
lmer.log   4 864 878   -428      856
lmer.main  5 823 840   -406      813 43.62      1 0.00000000004 ***
lmer.int   6 823 844   -406      811  1.49      1          0.22
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This sequence of tests suggest that there is likely a main-effect of sex, after controlling for differences in grade-level ($\chi^2(1) = 43.62$, $p < .001$). When comparing the interaction model to the main effects model, however, the evidence suggests that there does not seem to be an interaction between sex and grade-level ($\chi^2(1) = 1.49$, $p = .22$).

Magnitude of the t-Values

Recall that another informal method of determining whether predictors were important is to examine the *t*-value associated with that predictor. In general, predictors that have *t*-values with an absolute value greater than or equal to two should be retained in the model. For example, if we look at the coefficient-level output of the interaction model:

```
tidy(lmer.int)
```

```
# A tibble: 6 x 5
  term                              estimate std.error statistic group
  <chr>                                <dbl>     <dbl>     <dbl> <chr>
1 (Intercept)                         -0.109     0.254    -0.427 fixed
2 log(grade_quant_center + 1)          1.79      0.149    12.0   fixed
3 female                               2.81      0.371     7.57  fixed
4 log(grade_quant_center + 1):female  -0.266     0.217    -1.22  fixed
5 sd_(Intercept).id                    1.23     NA        NA     id
6 sd_Observation.Residual              0.903    NA        NA     Residual
```

The coefficient for the interaction term in this model has a *t*-value with an absolute value of 1.223. Since this is less than two, we would likely drop the interaction term from the model. Before evaluating the other predictors, we would re-fit the model without the interaction (main effects model).

This informal method of evaluating predictor importance is related to the *p*-value approach to model selection. Namely, if a predictor has a *t*-value greater than or equal to two, it typically has a *p*-value that is near (or less than) .05. For example,

```
2 * pt(q = -2, df = 100)
```

```
[1] 0.0482
```

Even for a small sample size this is a reasonable rule-of-thumb, although not perfect,

```
2 * pt(q = -2, df = 10)
```

```
[1] 0.0734
```

Summary

Although in this example both frameworks for evaluating evidence (information criteria and statistical tests) resulted in the same adopted model, the *p*-value approach should not be seen as confirming evidence! The results from evaluating models using a *p*-value approach and those employing a model evidence approach are not always consistent with each other. This is because they represent two very different philosophical approaches to model selection. As such, the decision about how you will make decisions about model adoption (*p*-value vs. model evidence) should be decided prior to carrying out any data analysis.

Neither framework should be used in isolation from exploratory work (e.g., density plots of the outcome and predictors, scatterplots of the relationships) and residual checking. Both types of evidence are only worthwhile if the model assumptions are valid. As in life, there are no shortcuts to rigorous scientific inquiry.