# Linear mixed models on Students' scores

Timothy Lee

11/04/2020

## Context

The file `school.csv` contains data on 760 Grade 8 students (i.e., most are 11 years old) in 32 primary schools in the Netherlands. The data are adapted from Snijders and Boskers' Multilevel Analysis, 2nd Edition (Sage, 2012). our question of interest is "Which variables are associated with Grade 8 students' scores on an end-of-year language test?"

```
library(tidyverse)
library(lme4)
school_data =  read_csv("school.csv")
str(school_data)
```

```
## tibble [992 x 8] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ X1             : num [1:992] 1 2 3 4 5 6 7 8 9 10 ...
##  $ school         : num [1:992] 1 1 1 1 1 1 1 1 1 1 ...
##  $ ses            : num [1:992] -4.73 -17.73 -12.73 -4.73 -17.73 ...
##  $ test           : num [1:992] 46 45 33 46 20 30 30 57 36 36 ...
##  $ iq             : num [1:992] 3.13 2.63 -2.37 -0.87 -3.87 -2.37 -2.37 1.13 -2.37 -0.87 ...
##  $ sex            : num [1:992] 0 0 0 0 0 0 0 0 0 0 ...
##  $ minority_status: num [1:992] 0 1 0 0 0 1 1 0 1 1 ...
##  $ denomination   : num [1:992] 1 1 1 1 1 1 1 1 1 1 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   X1 = col_double(),
##   ..   school = col_double(),
##   ..   ses = col_double(),
##   ..   test = col_double(),
##   ..   iq = col_double(),
##   ..   sex = col_double(),
##   ..   minority_status = col_double(),
##   ..   denomination = col_double()
##   .. )
```

```
head(school_data)
```

```
## # A tibble: 6 x 8
##      X1 school    ses  test    iq   sex minority_status denomination
##   <dbl>  <dbl>  <dbl> <dbl> <dbl> <dbl>           <dbl>        <dbl>
## 1     1      1  -4.73    46  3.13     0               0            1
## 2     2      1 -17.7     45  2.63     0               1            1
```

```
## 3      3      1 -12.7     33 -2.37     0              0            1
## 4      4      1  -4.73    46 -0.87     0              0            1
## 5      5      1 -17.7     20 -3.87     0              0            1
## 6      6      1 -17.7     30 -2.37     0              1            1
```

Variables are defined as follows:

school: an ID number indicating which school the student attends
test: the student's score on an end-of-year language test
iq: the student's verbal IQ score
ses: the socioeconomic status of the student's family
sex: the student's sex
minority_status: 1 if the student is an ethnic minority, 0 otherwise

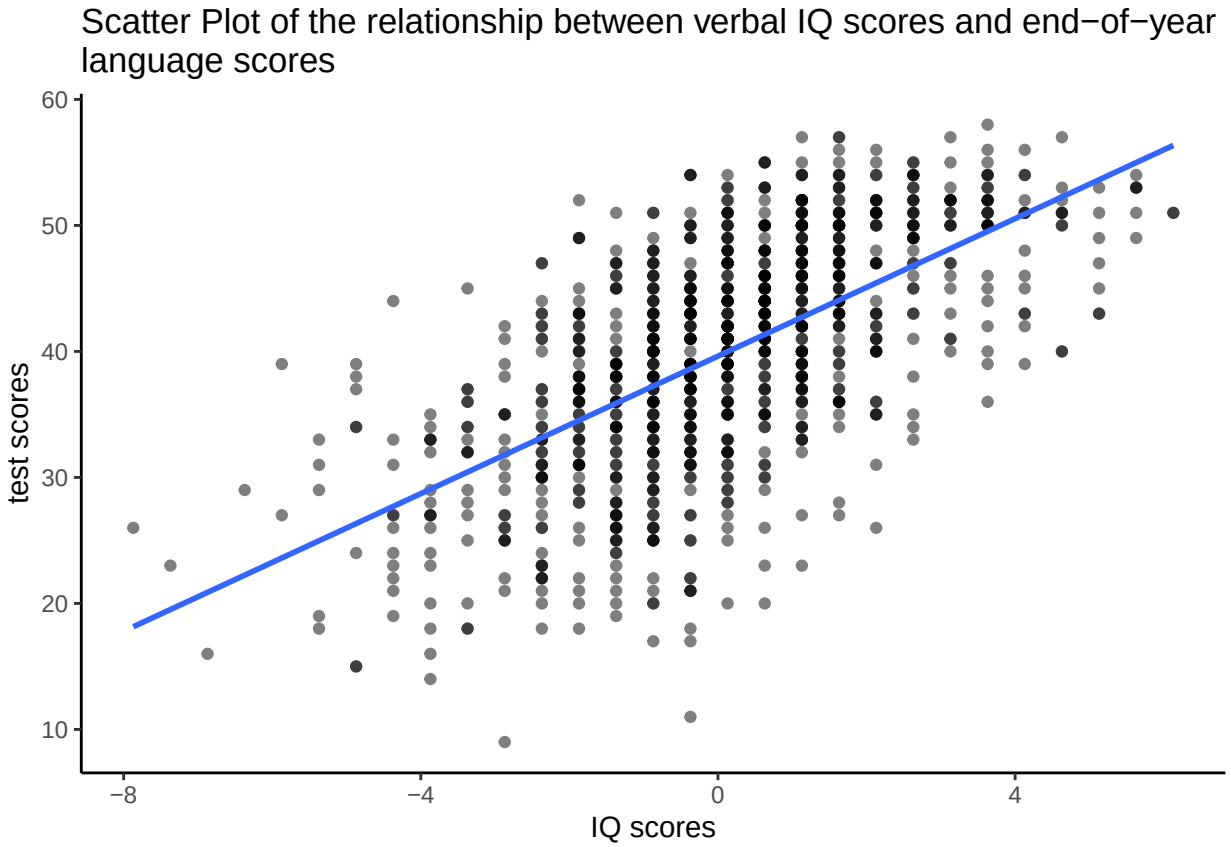
## Concerns if performing linear regression

The independence assumption (of individual observations/students) could be violated. For example, we have failed to consider the fact that perhaps students within the same school might be strongly correlated/dependent, i.e., perhaps the "quality of teaching" of a particular school leads to all the students of that school to perform better on the test in general, independent of all other variables.

## Scatter Plot of the relationship between verbal IQ scores and end-of-year language scores.

```
ggplot(school_data, aes(x = iq, y = test)) +
    geom_point(alpha = 0.5) +
    geom_smooth(method = "lm", se = FALSE) +
    theme_classic()+
    labs(y = "test scores", x = "IQ scores", title =
    "Scatter Plot of the relationship between verbal IQ scores and end-of-year
language scores")
```

```
## `geom_smooth()` using formula 'y ~ x'
```




Based on the scatter plot, we can see that there is somewhat a positive relationship between iq and test, but the points are very scattered (high variance) from the line of best fit. This indicates that a student's IQ might be associated with their test scores, but further analysis is required. This could also be caused by not having a random intercept for the different schools (i.e., some schools might have a higher intercept than others for reasons mentioned above).

### Attempt 1: Fitted Multiple Linear Regression

First, I created two new variables in the data set, `mean_ses` that is the mean of ses for each school, and `mean_iq` that is mean of iq for each school. Then, I fitted a linear model with `test` as the response and use `iq, sex, ses, minority_status, mean_ses` and `mean_iq` as the covariates.

```
school_data <- school_data %>%
    group_by(school) %>%
    mutate(mean_ses = mean(ses), mean_iq = mean(iq))
school_data$sex = factor(school_data$sex)
school_data$minority_status = factor(school_data$minority_status)
```

```
lin_model = lm(test ~ iq+ sex+ ses+ minority_status+ mean_ses+mean_iq, data = school_data)
summary(lin_model)
```

```
##
## Call:
## lm(formula = test ~ iq + sex + ses + minority_status + mean_ses +
##     mean_iq, data = school_data)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.4126  -4.5967   0.5543   4.9639  18.6042
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      38.45808    0.31251 123.061  < 2e-16 ***
## iq                2.28556    0.11979  19.079  < 2e-16 ***
## sex1              2.34325    0.43385   5.401 8.30e-08 ***
## ses               0.19332    0.02641   7.319 5.19e-13 ***
## minority_status1 -0.17083    0.97592  -0.175    0.861
## mean_ses         -0.21555    0.04641  -4.644 3.88e-06 ***
## mean_iq           1.42674    0.30264   4.714 2.77e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.818 on 985 degrees of freedom
## Multiple R-squared:  0.4511, Adjusted R-squared:  0.4477
## F-statistic: 134.9 on 6 and 985 DF,  p-value: < 2.2e-16
```

```
confint(lin_model) #1 is female and 1 is minority
```

```
##                      2.5 %     97.5 %
## (Intercept)      37.8448162 39.0713519
## iq                2.0504849  2.5206429
## sex1              1.4918849  3.1946222
## ses               0.1414857  0.2451566
## minority_status1 -2.0859568  1.7442963
## mean_ses         -0.3066319 -0.1244709
## mean_iq           0.8328516  2.0206247
```

Based on the summary table, the intercept represents the average test scores for male students who are not ethnic minority with the average IQ and social economic status, which is about 38.458. Also, test scores increases by about 2.286 per one unit increase of IQ score, and that female students will score better than male counterparts (holding all other covariates constant). Also, there is a very small but positive effect of social economic status (average test scores improve by around 0.193 for one unit increase of social economic status), and that being an ethnic minority could have no effect on test scores (since this co-variate has a high p-value, indicating for a potential of non-significant effect on the response). Furthermore, mean social economic status seems to have a slight negative effect on test scores for each school (i.e., test scores seem to slightly decrease if that particular school has a higher mean social economic status or test scores decrease by 0.307 per one unit increase of mean social economic status for a particular school), while students in schools with higher mean IQ scores tend to have better scores.

As for the confidence intervals, we see that the only covariate that contains a 0 in their respective confidence interval is minority status, indicating that we are 95% confident that this covariate might not be significant/has no effect on the average test scores (holding all other covariates constant). In other words, being an ethnic minority don't have a (statistically significant) effect on average test scores. All other covariates have intervals above 0 (indicating that we are 95% confident that there is a positive effect on test scores) with the only exception of mean social economic status to be below 0 (indicating that this covariate might slightly negatively effect a student's test scores).

**Attempt 2: Fitted a linear mixed model with the same fixed effects as before and with a random intercept for school.**

```
rand_model = lme4::lmer(test ~ iq+ sex+ ses+ minority_status+mean_ses + mean_iq + (1 |school),
data = school_data)
summary(rand_model)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: test ~ iq + sex + ses + minority_status + mean_ses + mean_iq +
##     (1 | school)
##    Data: school_data
##
## REML criterion at convergence: 6518.1
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.9926 -0.6304  0.0757  0.6945  2.6361
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  school   (Intercept)  8.177   2.859
##  Residual             38.240   6.184
## Number of obs: 992, groups:  school, 58
##
## Fixed effects:
##                  Estimate Std. Error t value
## (Intercept)      38.37951    0.48384  79.323
## iq                2.27784    0.10881  20.935
## sex1              2.29199    0.40260   5.693
## ses               0.19283    0.02396   8.047
## minority_status1 -0.65259    0.96943  -0.673
## mean_ses         -0.20131    0.08000  -2.517
## mean_iq           1.62512    0.52017   3.124
##
## Correlation of Fixed Effects:
##             (Intr) iq     sex1   ses    mnrt_1 men_ss
## iq          -0.035
## sex1        -0.408  0.045
## ses          0.013 -0.284 -0.048
## mnrty_stts1 -0.129  0.131  0.001  0.053
## mean_ses    -0.140  0.092  0.003 -0.296  0.039
## mean_iq      0.089 -0.199 -0.007  0.064  0.052 -0.494
```

```
confint(rand_model, oldNames=FALSE)
```

```
##                             2.5 %      97.5 %
## sd_(Intercept)|school  2.1818595  3.51821014
## sigma                  5.9011373  6.46042873
## (Intercept)           37.4412106 39.31755070
## iq                     2.0649432  2.49094360
## sex1                   1.5044771  3.08014874
## ses                    0.1459275  0.23975452
```

```
## minority_status1      -2.5423935  1.24925972
## mean_ses              -0.3564217 -0.04606047
## mean_iq                0.6166461  2.63522563
```

```
#summary(rand_model)$varcor
```

Firstly, we can see from the first 2 rows of the confint table that we are 95% confident that the estimated standarad deviation of the variance of the random effects when added the random intercept is between 2.18 to 3.52 (i.e., this is the standard deviation of the variation explained by adding the random effect for each schools), and that we are also 95% confident that the standard deviation of the remaining variation that is not accounted for by the model (i.e., the variation of the residuals) is between 5.90 to 6.46.

As for the summary table, it is also similar to that of the linear model (previous question). In other words, the intercept represents the average test scores for male students who are not ethnic minority with the average IQ and social economic status, which is about 38.3795. Also, test scores increases by about 2.278 per one unit increase of IQ score, and that female students will score better than male counterparts (holding all other covariates constant). Also, there is a very small but positive effect of social economic status (average test scores improve by around 0.193 for one unit increase of social economic status), and that being an ethnic minority could have no effect on test scores (since this co-variate has a 0 in its confidence interval, indicating for a potential of non-significant effect on the response). Furthermore, the mean social economic status seems to have a slight negative effect on test scores for each school (i.e., test scores seem to slightly decrease if that particular school has a higher mean social economic status or test scores decrease by 0.201 per one unit increase of mean social economic status for a particular school), while students in schools with higher mean IQ scores tend to have better scores.

For the results in the confidence interval table, it is also somewhat similar to that of the linear model from previous question (in terms of the numerical values for each C.I.), where the only covariate that contain a 0 in the confidence interval is minority status, indicating that we are 95% confident that this covariate might not be significant/has no effect on test scores. All other co-variates have intervals above 0 (indicating that we are 95% confident that there is a positive effect of these covariates on test scores) with the only exception of mean social economic status of a school to be below 0 (indicating that we are 95% confident this covariate might negatively effect a student's test scores).

## Similarities and differences between the coefficients of the fixed effects

We can see that in general, the coefficients of most of the covariates are mostly similar, with the only difference being the width of the CI's for some covariates, especially school level covariates. In other words, since the LMM assumes that there might be some dependency between the observations of each student (i.e., school-level factors), we "lose" some "unique" information for each observation, resulting in a smaller sample size and hence a wider confidence interval). For example, the covariates of mean social economic status and mean IQ scores are school-related, grouping level factors, and hence give us no "extra" information in our LMM as all the students within that same school will have the same mean ses and mean IQ scores. This is reflected in the wider CI for these two factors in the LMM compared to the linear model (as we account for more uncertainty due to smaller sample size). The rest of the other student-level covariates are somewhat similar.

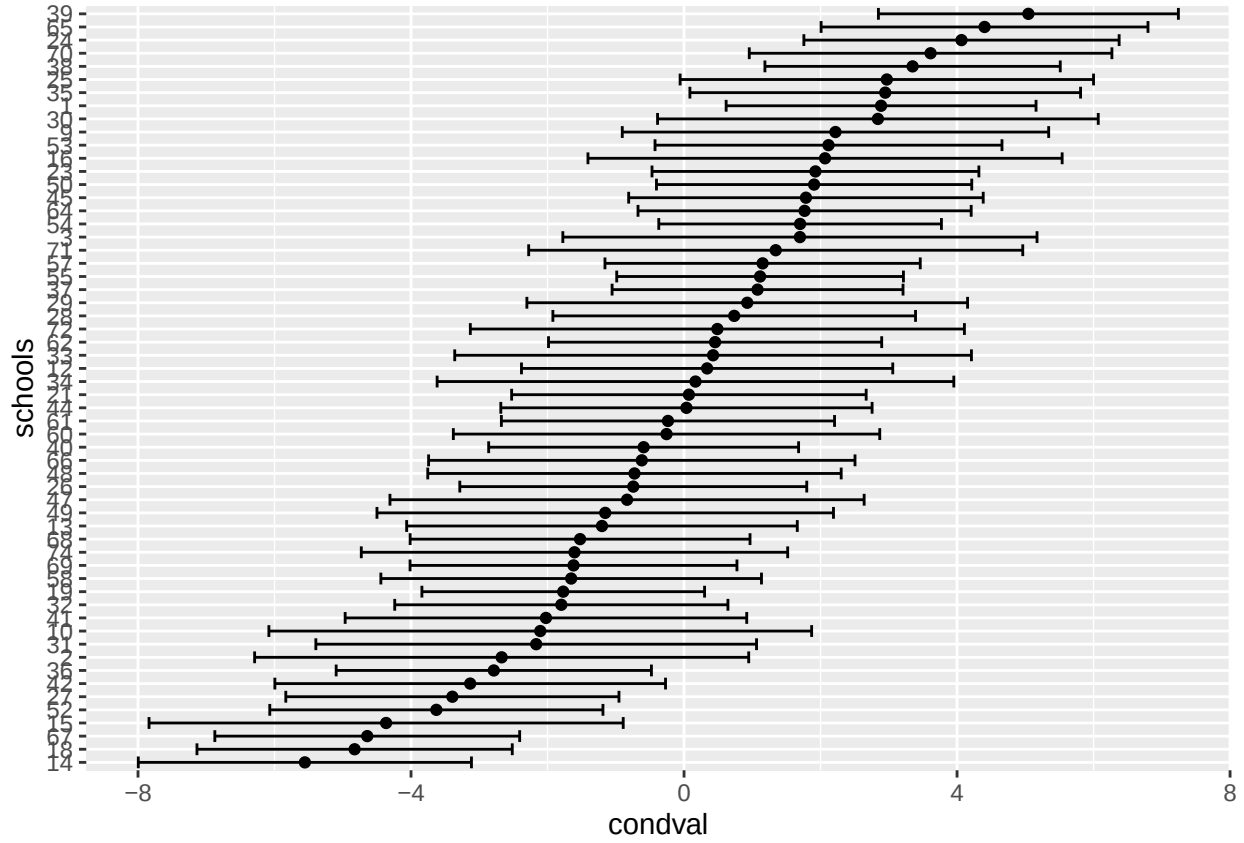
## Plot of random effects for the different schools

```
rand_model
```

```
## Linear mixed model fit by REML ['lmerMod']
```

```
## Formula: test ~ iq + sex + ses + minority_status + mean_ses + mean_iq +
##     (1 | school)
##    Data: school_data
## REML criterion at convergence: 6518.126
## Random effects:
##  Groups   Name        Std.Dev.
##  school   (Intercept) 2.859
##  Residual             6.184
## Number of obs: 992, groups:  school, 58
## Fixed Effects:
##      (Intercept)               iq             sex1              ses
##          38.3795           2.2778           2.2920           0.1928
## minority_status1         mean_ses          mean_iq
##          -0.6526           -0.2013           1.6251
```

```
my_random_effects <- ranef(rand_model, condVar=TRUE)
ranef_df <- as.data.frame(my_random_effects)
ranef_df %>%
 ggplot(aes(x = grp, y = condval, ymin = condval - 2*condsd, ymax = condval + 2*condsd)) +
 geom_point() +
 geom_errorbar() +
 coord_flip()+
 labs(x = "schools")
```




Although the confidence intervals overlap for most schools, it seems that the point estimate of average test scores for each school is somewhat different (especially on the 2 extremes), indicating a need for a random intercept. This could also be shown from the large deviation of both tails from the average test scores

(condval = 0 line), suggesting some schools have very high test scores (i.e., school number 39) while other schools (i.e., school number 14) have very low average test scores. The large confidence interval could be mainly due to a small sample size (of each individual school).

## Conclusion

```
summary(rand_model)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: test ~ iq + sex + ses + minority_status + mean_ses + mean_iq +
##     (1 | school)
##    Data: school_data
##
## REML criterion at convergence: 6518.1
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.9926 -0.6304  0.0757  0.6945  2.6361
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  school   (Intercept)  8.177   2.859
##  Residual             38.240   6.184
## Number of obs: 992, groups:  school, 58
##
## Fixed effects:
##                  Estimate Std. Error t value
## (Intercept)      38.37951    0.48384  79.323
## iq                2.27784    0.10881  20.935
## sex1              2.29199    0.40260   5.693
## ses               0.19283    0.02396   8.047
## minority_status1 -0.65259    0.96943  -0.673
## mean_ses         -0.20131    0.08000  -2.517
## mean_iq           1.62512    0.52017   3.124
##
## Correlation of Fixed Effects:
##             (Intr) iq     sex1   ses    mnrt_1 men_ss
## iq          -0.035
## sex1        -0.408  0.045
## ses          0.013 -0.284 -0.048
## mnrty_stts1 -0.129  0.131  0.001  0.053
## mean_ses    -0.140  0.092  0.003 -0.296  0.039
## mean_iq      0.089 -0.199 -0.007  0.064  0.052 -0.494
```

```
summary(lin_model)
```

```
##
## Call:
## lm(formula = test ~ iq + sex + ses + minority_status + mean_ses +
##     mean_iq, data = school_data)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.4126  -4.5967   0.5543   4.9639  18.6042
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      38.45808    0.31251 123.061  < 2e-16 ***
## iq                2.28556    0.11979  19.079  < 2e-16 ***
## sex1              2.34325    0.43385   5.401 8.30e-08 ***
## ses               0.19332    0.02641   7.319 5.19e-13 ***
## minority_status1 -0.17083    0.97592  -0.175    0.861
## mean_ses         -0.21555    0.04641  -4.644 3.88e-06 ***
## mean_iq           1.42674    0.30264   4.714 2.77e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.818 on 985 degrees of freedom
## Multiple R-squared:  0.4511, Adjusted R-squared:  0.4477
## F-statistic: 134.9 on 6 and 985 DF,  p-value: < 2.2e-16
```

In summary, we can see that although the coefficients or covariates of both models are roughly similar (except for school-level covariates like mean social economic status and mean IQ scores), the LMM is preferred as it accounts for the dependency for each students within the same school. This is further reinforced by the plots of 1h, indicating potential school-level factors we need to account for by adding a random effect for each school. This is further emphasised by the proportion of the (residual) variation explained by adding in the random effects the difference between schools accounts for, which is is 8.177/(8.177 + 38.240) or about 17%.

As for the confidence intervals (as concluded before), the only the covariate of that has 0 in between its interval is minority status, suggesting that we are 95% confident that that this factor have no effect on the average test scores.

Hence, in conclusion, we can say that the variables of IQ, social economic status, sex will all have an effect on the students' test scores, but also, we need to account for the effect of the difference between each individual schools (which also have an effect for all the students within that particular school).