Factors Affecting Test Scores

Junru Zhang

06/03/2020

The file school.csv (available on Quercus) contains data on 992 Grade 8 students (i.e., most are 11 years old) in 58 primary schools in the Netherlands. The data are adapted from Snijders and Boskers' Multilevel Analysis, 2nd Edition (Sage, 2012).

`Variables in the school.csv data set:`

Variable	Description
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

Question of interest: Which variables are associated with Grade 8 students' scores on an end-of-year language test?

```
library(tidyverse)
school_data <- read.csv("./school.csv")
# install.packages("Pmisc", repos = "http://R-Forge.R-project.org", type = "source")
```

What Model to Use?

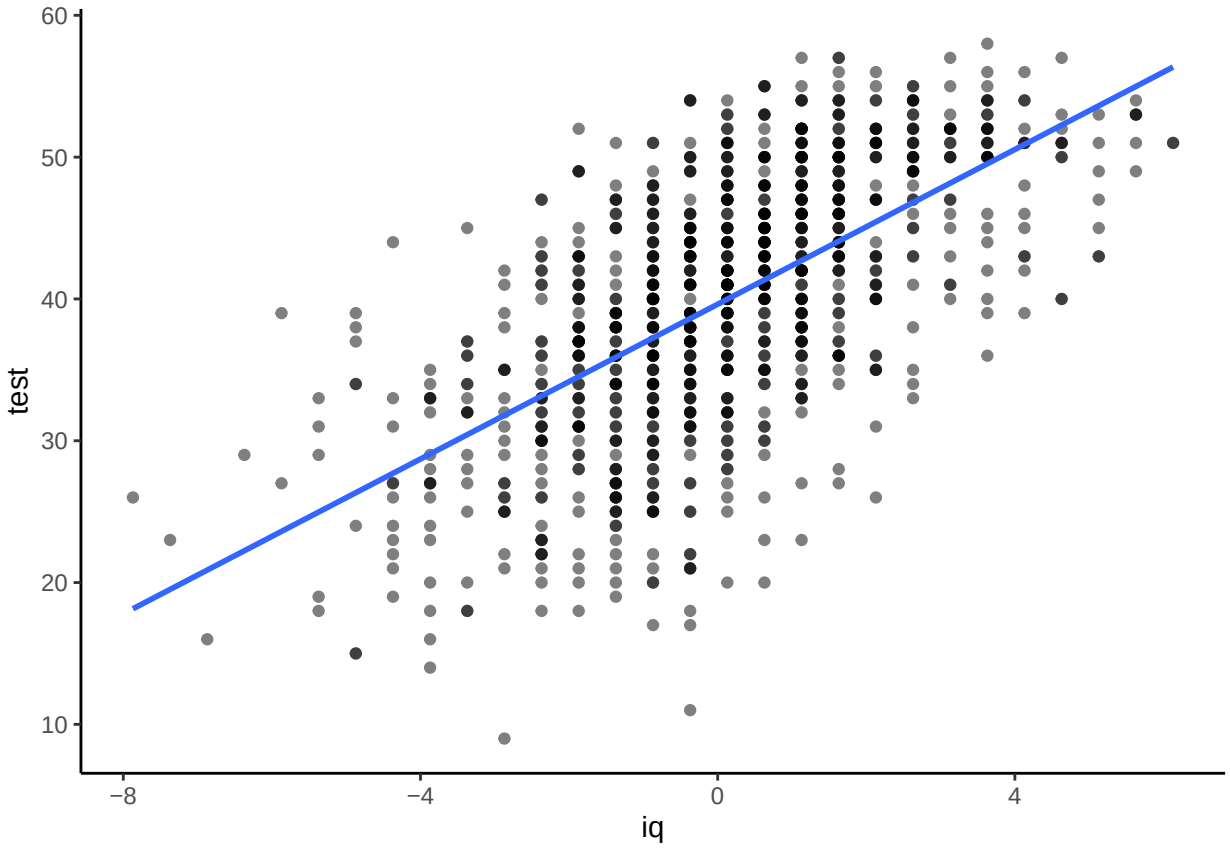
Linear regression is only a valid option if all obervations are independent. However, since the data is obtained from 58 primary schools, it is very likely that end-of-year language test scores of students in the same school are correlated. Thus, the independent observations assumption of the linear regression model will be violated. i.e. fitting a linear model is not a good choice.

Visualizing the Data

Examining the relationship between verbal IQ scores and end-of-year language scores

Let's create a scatter plot!

```
ggplot(school_data, aes(x = iq, y = test)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE) +
  theme_classic()
```



From the graph, we see that as the higher verbal IQ score students have, the better their end-of-year test scores tend to be. In other words, we observe a positive association between students' verbal IQ scores and end-of-year test scores.

Creating New Variables

Before fitting the models, let's create 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. These are the variables we are interested in studying.

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

Modeling

Linear Model

Let's first fit a linear model that uses iq, sex, ses, minority_status, mean_ses and mean_iq as the covariates.

```
m1 <- lm(test ~ iq+sex+ses+minority_status+mean_ses+mean_iq, data= school_data)
summary(m1)
```

```
##
## 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 ***
## sex              2.34325    0.43385   5.401 8.30e-08 ***
## ses              0.19332    0.02641   7.319 5.19e-13 ***
## minority_status -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(m1)
```

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

The intercept is the average end-of-year test scores of the reference group, which are ethenic minority male students with a verbal IQ score of 0. Note that a score of 0 in our context means at the average level, rather than an actual score of 0. So students in the reference group have socioeconomic status and IQ scores that are at the average level. And all other coefficients represent the end-of-year test scores relative to the reference group, while holding all other covariates constant.

Among all these covariates, only the confidence interval for minority_status includes 0. This implies that we have strong evidence to claim that test scores has no association with minority status of students. The remaining covariates are all significantly associated with test scores and they all have confidence intervals which is entirely above 0 except mean socioeconomic status of schools. Having confidence intervals above 0 indicate increasing the unit of these covariates would cause an increase in test scores. Mean socioeconomic status has a confidence interal below 0, which indicates that the socioeconomic status has a negative impact on language test scores of Grade 8 studnets.

Linear Mixed Model

Let's now fit a linear mixed model with the same fixed effects as the linear model, and with a random intercept for school.

```
lmm1 <- lme4::lmer(test ~ iq + sex + ses + minority_status +
                     mean_ses + mean_iq +
                     (1|school),
                   data=school_data)
summary(lmm1)
```

```
## 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
## sex              2.29199    0.40260   5.693
## ses              0.19283    0.02396   8.047
## minority_status -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     sex    ses    mnrty_ men_ss
## iq          -0.035
## sex         -0.408  0.045
## ses          0.013 -0.284 -0.048
## minrty_stts -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(lmm1)
```

```
##                       2.5 %      97.5 %
## .sig01           2.1818595  3.51821014
## .sigma           5.9011373  6.46042873
## (Intercept)     37.4412106 39.31755070
## iq               2.0649432  2.49094360
## sex              1.5044771  3.08014874
## ses              0.1459275  0.23975452
## minority_status -2.5423935  1.24925972
## mean_ses        -0.3564217 -0.04606047
## mean_iq          0.6166461  2.63522563
```

We see that for the random effect U_i, which is the schools of the students, $Var(U_i)$=8.177 and Var(e_i) = 38.240. Therefore the random effect explains $\frac{Var(U_i)}{Var(U_i)+Var(e_i)} = \frac{8.177}{8.177+38.240} = 0.176 = 17.6\%$ of the variation of the data.

The confidence intervals of both fixed and random effects are significant except minority status, since their confidence intervals all exclude 0. Also, the first two rows of the confidence intervals shows that under a 5% significance level, random effect would capture at least $\frac{2.1818595^2}{2.1818595^2+5.9011373^2} = 0.136 = 13.6\%$ of the variation. So there is some evidednce that the test scores of students has some association with the schools that they

are in.

Fitting a linear mixed model is evidently better than fitting a linear model in this case. However, we still need to verify using data to see whether it the random effect term is worthy being included in the model.

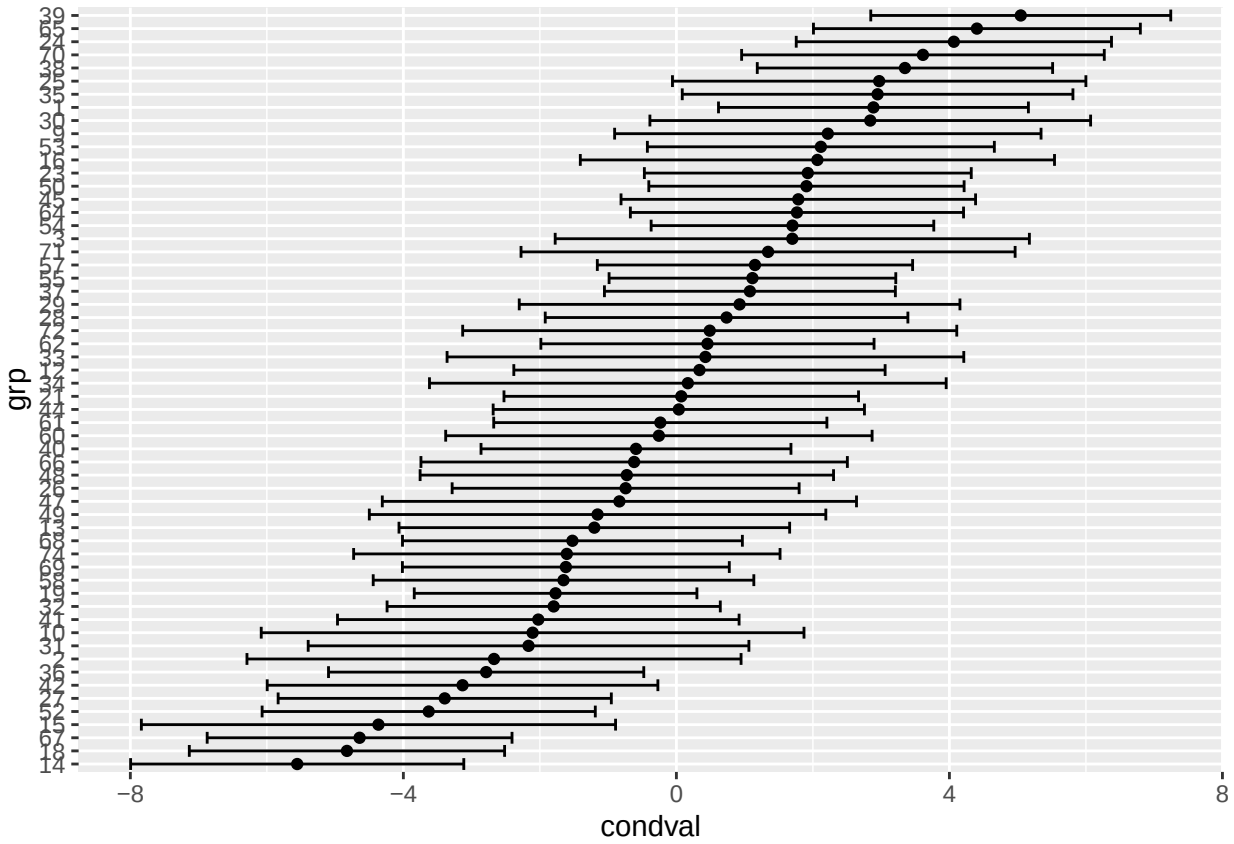
Fixed Effects

The coefficients, or the predicted test scores obtained from these two models are roughly the same. Since the sample size of linear model is larger than of the linear mixed model, the standard errors are small, therefore confidence intervals for the mixed model is wider than in the linear model in general. However, there are no random effects in the linear model to capture the differences caused by the schools, that is why some of the CIs for linear model is wider than for the linear mixed model (e.g. iq).

Ramdom Effects

Let's plot the random effects for the different schools to see if it is reasonable to have included these random effects in the linear mixed model.

```
random_effects <- lme4::ranef(lmm1, condVar=TRUE)
ranef_df <- as.data.frame(random_effects)
ranef_df %>%
  ggplot(aes(x = grp, y = condval, ymin = condval - 2*condsd, ymax = condval +
               2*condsd)) +
  geom_point() +
  geom_errorbar() +
  coord_flip()
```



It does seem reasonable to include these random effects since the intercepts of each group does vary, the estimates and confidence intervals do not line up vertically in a straight line.

Conclusion

We come to a conclusion that students' verbal IQ scores, gender, mean IQ score and mean socioeconomic status of their schools are all variables that are associated with Grade 8 students' scores on an end-of-year language test significantly. (Confidnce intervals excludes 0 for all these covariates.)

In particular, The average marks for female students are between 1.50 and 3.08 higher than for males. Also, higher IQ scores and higher mean IQ scores of the schools result in better end-of-year test performance. For a one unit increase in a student's IQ on this scale, the expected mark for a student increases by between 2.1 and 2.5 marks.

Students at schools with a unit higher average IQ (for the school) also have a higher expected mark of between about 0.6 to 2.6. Having personally higher SES is associated with a higher mark (by only about 0.1 to 0.2 marks for a one unit increase on the index). But a better mean socioeconomic status within each school results in worse end-of-year test performance (between 0.05 and 0.36 lower expected marks for a one unit increase in mean SES), after controlling for the other variables.

Also, we know that the differences between schools only explains 17.6% of the variation of the data.