Multi-level Modeling

Gergana Todorova, Lancaster University

November 28, 2018

Abstract

The paper investigates the relationship between students' exam performance and their score on the London Reading Test score. Data consists of six other covariates, which significance for the model is also explored. Due to the fact that previous studies argued that the school facilities influence the performance of the students, multi-level modelling is applied to study this relationship. Results showed that school level is an important predictor for the models and explains 16% of the variance at a student level. The most important covariate for the prediction is the London Reading Test Score, which has both fixed and random effect. Other covariates for the prediction are the gender of the student, the type of the school and the result from the verbal reasoning score. Similar studies in the past proved that there is a correlation between the parental education, computer availability at home and the availability of a desk for homework at home to be also important predictors for the student performance. Therefore future studies should incorporate these factors in the analysis.

1 Introduction

This paper investigates the relationship between the exam performance of students and their score from the London Reading Test Score. Dataset consists of 4059 observation and 8 variables. The aim of the paper is to investigate the relationship between the students' performance of the exams and the London Reading Test score and test the significance of the other covariates. Students performance is a highly debated topic and many studies have investigated the relationship between the school and performance on the exam. Hanushek (1996) argues that the school and the additional resources some schools provide don't have to affect the performance of students [3]. Bernal(2016) criticized the methods used and proved that school plays a critical role for the students' accomplishments.[1] In addition, Eide (1997) performed regression and proved that differences in schools have an effect on student performance. In addition, factors like parental education, computer availability at home and availability of a desk for homework at home to be also important predictors for the student performance [2]. Multi-level modelling is applied in order to investigate if there is a difference in the students' performance in various schools. The relationship of the covariates is tested to have both fixed and random effect.

2 Exploratory Data Analysis

Dataset consists of 4059 observations and 8 variables. There is data for 65 schools, with a different number of observations - two of the schools have under 10 number of observations, whereas the biggest number of observations is 198. This difference is not a problem for applying multi-level modelling, however, the results for some school are going to be less accurate. Figure 1 represents a box plot of their student exam score. It can be seen that the mean and quantiles are quite different for the schools, which suggests that the school level might be important for the prediction. It should also be notified that the dataset consists

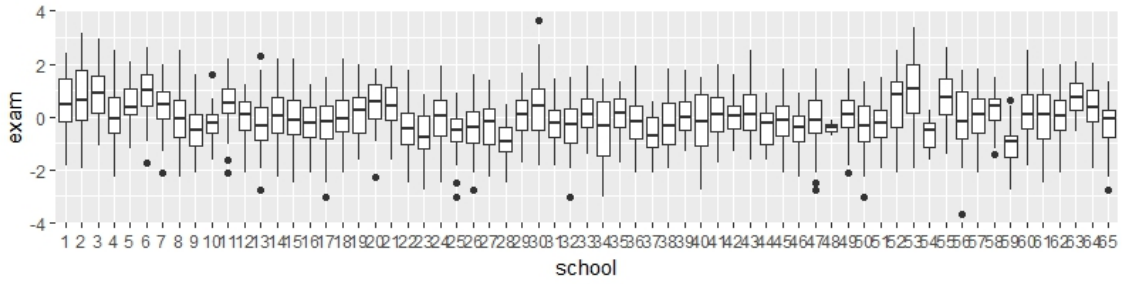


Figure 1: Box- plot with the exam scores for the different schools.

of 20% more girls than boys. Exam and score from the London test, verbal reasoning test and the intake ratio and are normally distributed.

3 Methods

Multi-level modelling is a very powerful tool, which allows us to analyse clustered data by introducing hierarchical levels to the analysis. They have three additional elements to normal regressions - a level, a fixed effect and a random effect. The level signalizes, which variable brings in the hierarchy. For example, in our dataset, there are students, which belong to different schools. This suggests that the variable school can be introduced as a level in order to see if the students are performing differently in different schools. The model will look like that if the level school is introduced:

$$y_i = \gamma_{00} + U_{0j} + \epsilon_{ij}, \qquad \epsilon - N(0, \sigma_\epsilon^2), \qquad U_{0j} - N(0, \sigma_{u0}^2),$$

where γ_{00} is the mean of the intercepts for all schools, U_{0j} is the deviation for each school mean from the grant mean, σ_ϵ^2 is the variance of children within schools and σ_{u0} is the variance between the schools' true means. In order to measure the effect of the dependency in the grouped observations is the intraclass correlation coefficient, which is calculated by the equation:

$$\rho = \frac{\sigma_{u0}^2}{\sigma_{u0}^2 + \sigma_\epsilon^2}$$

The second element of the multi-level models is the fixed effect of variables. This is similar to normal regression and represents how much a variable influences the prediction models and the assumption is that the level two is affected in the same way. For example, if the variable Normalized London Reading Score is introduced as a predictor to the analysis, the assumption for the fixed effects is that the coefficient for the prediction is the same for the different schools. The equation looks like that:

$$y_i = \gamma_{00} + \gamma_{01} lrt_j + U_{oj} + \epsilon_{ij}$$

In order to select the variables, which have a fixed effect on the model, the test on the likelihood deviance is performed and compared to a chi-square distribution. The task is repeated for each of the variables and the new model is compared to the previous new model via the ANOVA analysis.

The third element is the random effect. The random effect provides different slope for the second level and suggests that the variable affects the second level in different ways. The model with two levels and random intercept looks like:

$$y_i = (\gamma_{00} + \gamma_{01} X_{ij}) + (U_{oj} + U_{1j} X_{ij} + \epsilon_{ij}),$$

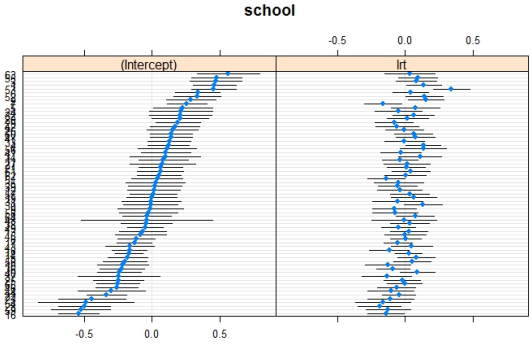


Figure 2: Random intercept U_{0j} values and U_{1j} is the predicted random slope

where the first part of the model represents the fixed part and the second part of the model represents the random part. In order to check whether there is a random effect model three tests will be taken into account. Firstly, the models will be compared with the ANOVA analysis which determined if there is a random effect or not. In addition, the values of the restricted maximum likelihood and Q-Q plots are considered.

4 Results

Firstly, a model is fitted introducing only school level, without any fixed and random effects. The result showed $\sigma^2_{u0} = 0.17$ and $\sigma^2_{\epsilon} = 0.84$, which makes the intraclass correlation coefficient 16%. This suggests that 16% of the variance is explained by the difference between schools. This is a significant result, therefore, school level is kept and considered important for the prediction model.

Next step is testing all variables for having fixed effect on the model. Results showed that the average intake score of the school is not important factor for the prediction models. Therefore the model looks like:

$$exam_i = \gamma_{00} + \gamma_{01} lrt_j + \gamma_{02} studgen_j + \gamma_{03} schogen_j + \gamma_{04} vr.score_j + U_{oj} + \epsilon_{ij},$$

where the parameter estimates γ are presented in Table 1.

	γ	2.5 %	97.5 %
(Intercept)	-0.25	-0.51	0.02
lrt	0.56	0.53	0.58
studgen	0.17	0.11	0.23
vr_score	0.14	0.04	0.25
schogen	-0.11	-0.21	-0.01

Table 1: Confidence intervals for the parameter estimates

Table 1 shows that the biggest impact for the prediction has the score from the London Reading Test - each unit increase in the London Reading Test score increases the students' performance on the exam by 0.56. Further interpretation is difficult to be given as both test scores are normalized in the dataset. It can be noticed that the values of the other parameters are close to 0 as they belong in the interval (0.1, 0.2). This suggests that their influence on the prediction is minimal. However, ANOVA tests showed that they are significant for the prediction, therefore, will be left in the model.

Next step is to check whether there is random effect or not. As mentioned each variable is added to the model, its significance is compared to the previous best models and left if

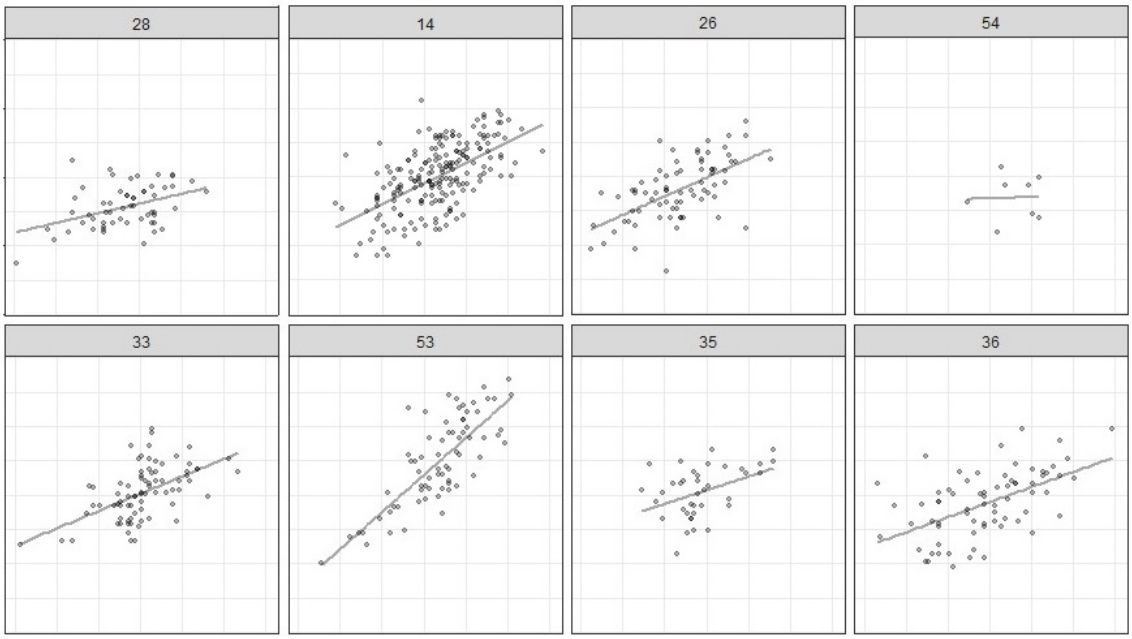


Figure 3: Fitted values for the model with random effect

the p value is below 0.05. Results showed that there is only one significant random effect for this models, which is the London Reading Test variable. As an equation the models looks like:

$$exam_i = (\gamma_{00} + \gamma_{01} lrt_j + \gamma_{02} studgen_j + \gamma_{03} schogen_j + \gamma_{04} vr.score_j) + (U_{oj} + U_{1j} lrt_j + \epsilon_{ij}),$$

where the first brackets represent the fixed effect and the second bracket the random effect. The estimated values for the random effect can be seen in Figure 2. The left part of the picture shows the random intercept U_{0j} for a school j and the right part shows the random slope variation U_{1j}. It can be noticed that both observations vary as the confidence interval for some are not even interacting with the overall mean. This shows clearly that the score of the London Reading test has a different effect in the schools.

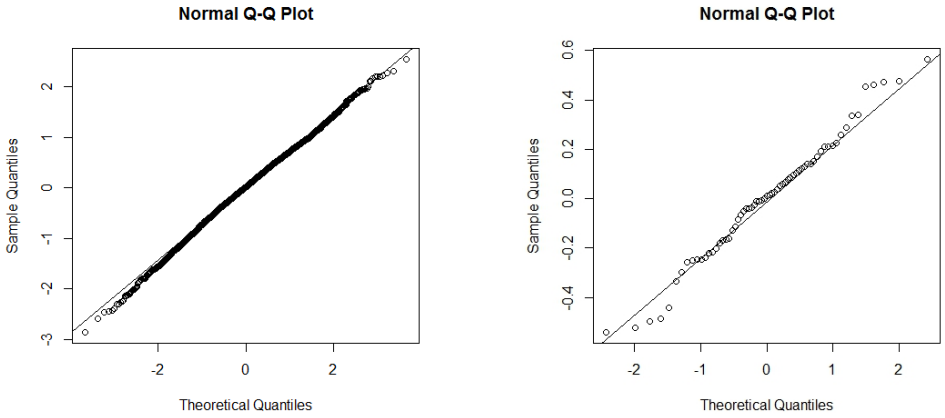
Having introduced the random effect, now the model represents 65 different models, which intercepts and coefficient for lrt vary but the rest variables are the same. For example for the model for the students belonging to school number 53 looks like:

$$exam_i = 0.30 + 0.89 lrt_j + 0.17 studgen_j + 0.1 schogen_j - 0.1 vr.score_j + \epsilon_{ij},$$

which means that if we hold other variables fixed, a unit increase in the London Test Score leads to 0.89 rises in the students' performance in the exam.

Figure 3 shows the fitted values and the observations for eight random schools. It can be noticed that there is a significant difference in the slope for the schools. On the x-axis is plotted the score from the London Reading Test and on the Y-axis the result from the exam. For example, for the school number 53, a high influence of the London Test Score on the exams prediction can be noticed whereas the slope for the school number 54 is almost flat. It should be taken into account that school number 54 has fewer observations. However comparing the slopes with school number 28, which seem to provide a reasonable number of observations, a difference in the slope can be noticed.

Finally, model checking is performed. Firstly, the value of the Restricted Maximum Likelihood has decreased by 40, from the model, which has only a fixed effect to the model with a random effect. This shows that the random effect should be introduced into the model. Secondly, the Q-Q plots are considered, which can be seen in Figure 4. Student



(a) student level (b) school level

Figure 4: Q-Q Plot for level 1 and 2

level Q-Q plot seems to follow the theoretical quantiles, which shows that the model is a good fit. The picture on the right shows that the school level deviates from the Theoretical Quantiles at the edges, which signalizes for a bad models fit. However, it should be taken into account that the number of observations at the school level is significantly lower than those at the student level. In addition, looking at the fitted values for the school in Figure 3 it can be seen that the model fits well with the observations. Following the last two observations, the model with the random effect can be accepted as valid.

5 Conclusion

The paper investigated the effect of the London Reading Test Score on the exams performance for students in 65 schools. The data has a hierarchical structure, therefore, multi-level modelling was introduced. Including the second level of school showed to be important as it reduced the variance by 16%. In addition, the gender of the student, score from the verbal reasoning test and the type of school proved to be also important for the prediction. Tests also showed that the variable score form the London Reading Test Score has fixed and a random effect for the prediction, which means that there is the different slope and intercept for the various schools. There are few limitations in regards to the study. Firstly, the interactions between variables are not tested. Secondly, previous research in the performance of students in Italy showed that factors like a computer at home and parental education are relevant for the performance of students. These covariates are not observed in our dataset therefore not investigated. Future studies should investigate their relationship on the students in London.

References

[1] Pedro Bernal, Nikolas Mittag, and Javaeria A. Qureshi. Estimating effects of school quality using multiple proxies. *Labour Economics*, 39(C):1–10, 2016.

[2] Eric Eide and Mark H. Showalter. The effect of school quality on student performance: A quantile regression approach. *Economics Letters. March*, 1998.

[3] Eric A. Hanushek. Measuring investment in education. *Journal of Economic Perspectives*, 10(4):9–30, December 1996.

```
library(lme4)
library(foreign)
library(lattice)
library(dplyr)
library(ggplot2)
library(boot)
student_data <- read.dta("part1.dta")
attach(student_data)
####~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
####~~~~~~~~~~~~~~~~~~~Exploratory data analysis~~~~~~~~~~~~~~~~~~~~~~~~~~~
####~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
corr_stud<-cor(student_data)
student_data
head(student_data)
hist(student_data$vr_score)
table(student_data$studgen,student_data$school)
student_data$school =as.factor(student_data$school)
student_data$student<-as.factor(student_data$student)
student_data$studgen<-as.factor(student_data$studgen)
student_data$vr_score<-as.factor(student_data$vr_score)
student_data$schogen<-as.factor(student_data$schogen)
#student_data$uniqueID<-as.factor(student_data$uniqueID)

p10 <- ggplot(student_data, aes(x = school, y = exam)) +
  geom_boxplot()
#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
#~~~~~~~~~~~~~~~~fitting the model~~~~~~~~~~~~~~~~~~~~#
#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
m1 <- lmer(exam ~ 1 + (1 | school))
summary(m1)
#school as a level; ltr as a fixed effect
m2 <- lmer(exam ~ lrt + (1 | school))
summary(m2)
anova(m2,m1) #  we preffer m2

v1 <- lme4::VarCorr(m1)
v1
intvar <- v1$school[1]
intvar
resvar <- attr(v1, "sc")^2
resvar
icc <- intvar / (resvar + intvar)
icc
##########Result : we want lrt as fixed effect

#~~~~~~~~~~~~~~~~~~~~~
#effect of gender
#~~~~~~~~~~~~~~~~~~~~~
m5 <- lmer(exam ~ lrt + studgen + (1 | school))
anova(m5,m2) # we preffer m5
#~~~~~~~~~~~~~~~~~~~~~
#effect of vr_score
#~~~~~~~~~~~~~~~~~~~~~
m6 <- lmer(exam ~ lrt + studgen + vr_score + (1 | school))
anova(m5,m6) #  we preffer m6
#~~~~~~~~~~~~~~~~~~~~~
#effect of schogen
```

```
#~~~~~~~~~~~~~~~~~~~~
m9 <- lmer(exam ~ lrt + studgen + vr_score + schogen +(1 | school))
anova(m6,m9) #  we preffer m9
#~~~~~~~~~~~~~~~~~~~~
#effect of intake
#~~~~~~~~~~~~~~~~~~~~
m10 <- lmer(exam ~ lrt + studgen +vr_score +schogen + intake + (1 | school))
anova(m10,m9) #  we preffer m9
# this one doesn't have an effect
#################################################################
##############Latex tables#######################################
#################################################################
m9$coefficient
tabl<-cbind(fixef(m9), confint(m9, "beta_", method="Wald"))
confint(m9, "theta_")
confint(m9, "theta_")^2
xtable(tabl)
summary(student_data)
#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
#~~~~~~~~~~~~~~~~~~~~~~~~~Random effect tests~~~~~~~~~~~~~~~~~~~~~~~~~
#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
#################################################################
#~~~~~~~~~~~~~~~~~~~~~~Likelihood test~~~~~~~~~~~~~~~~~~~~~~~~~~~
m9 <- lmer(exam ~ lrt + studgen + vr_score + schogen +(1 | school))
mr1 <- lmer(exam ~ lrt + studgen + vr_score + schogen +(lrt | school))
anova(m9,mr1)
summary(mr1)
#ltr is random effect
m9 <- lmer(exam ~ lrt + studgen + vr_score + schogen +(1 | school))
mr2 <- lmer(exam ~ lrt + studgen + vr_score + schogen +(studgen | school))
anova(m9,mr2)
# studgen is no random effect
m9 <- lmer(exam ~ lrt + studgen + vr_score + schogen +(1 | school))
mr3 <- lmer(exam ~ lrt + studgen + vr_score + schogen +(vr_score | school))
anova(m9,mr3)
# vr_score is no random effect
m9 <- lmer(exam ~ lrt + studgen + vr_score + schogen +(1 | school))
mr4 <- lmer(exam ~ lrt + studgen + vr_score + schogen +(schogen | school))
anova(m9,mr4)
# schogen is no random effect
#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
allcoef <- coef(mr1)
summary(mr1)
xtable(allcoef)
# (d) Obtain level 1 residuals
res <- residuals(mr1)
#' Posterior means of random effects
re <- ranef(mr1, condVar=T)
names(re)
plot(re)
#' Remember due to shrinkage - variance of these less than
#' estimated variance component from fitted model
var(re$school)
summary(mr1)
#' Investigate normality of level 1 residuals
qqnorm(res)
qqline(res)
#' Investigate normality of level 2 predicted random effects
```

```
qqnorm(re$school[,1])
qqline(re$school[,1])
#' Always/usually a more difficult judgement
#' as level 2 sample size small
##########################################################
##### Graphs
##########################################################
n<-nrow(student_data)
vr_score_1 <- rep(2, each = n)
studgen_1 <- rep(1, each = n)
schogen_1 <- rep(1, each = n)
df_stundet = data.frame(school, student, exam, lrt, studgen_1, schogen_1, vr_score
head(student_data)
student_data$pred <-fitted(mr1)
mr1 <- lmer(exam ~ lrt + studgen + vr_score + schogen +(lrt | school))
student_data <- resid(mr1)
student_data <- mutate(student_data,
                       prediction_score = fitted(mr1))
v1 <- lme4::VarCorr(mr1)
v1
intvar <- v1$school[1]
intvar
resvar <- attr(v1, "sc")^2
resvar
icc <- intvar / (resvar + intvar)
icc
gg <- ggplot(student_data, aes(x = lrt, y = exam, group = school)) +
  geom_smooth(method = "lm", se = FALSE, color = "darkgrey") +
  geom_point(alpha = 0.2, size = 1) +
  facet_wrap(~school) +
  theme_bw()
print(gg)
coef(mr1)
```