An Investigation into the Impact of an Educational Intervention on Student Attainment

30th March 2023

Introduction

Throughout this report we aim to understand the effect of an educational intervention on student attainment, and how its impact evolves over time. We investigate the test performance of 210 pupils across 54 schools before and after the commencement of an intervention in a randomly controlled multisite trial. In such a trial, students across all involved schools are randomly allocated into either a control group (so receive no intervention), or the treatment (intervention) group. The randomized allocation increases the likelihood of evenly distributing the traits of the students across both groups, so there is a smaller chance of a systematic difference between the two groups, which might have confounded the true relationship between the variables we are investigating.

We first present the first couple rows of the data, which comprise 210 observations of 7 variables.

```
MST <- read.csv("https://andygolightly.github.io/teaching/MATH43515/mst1.csv",
                header=TRUE)
head(MST)
```

```
##   ID Posttest_Time1 Posttest_Time2 Posttest_Time3 Intervention Pretest School
## 1  1            101             87             91            1      99      1
## 2  2            119            109            115            1      95      1
## 3  3             94             79             83            0      82      1
## 4  4            100             83             89            0      94      1
## 5  5             69             73             78            1      69      2
## 6  6             78             69             72            1      74      2
```

```
dim(MST)
```

```
## [1] 210   7
```

Next we present a summary of the data collected, which describes the distributions of the test scores before the intervention (“Pretest), and at three points in time after the start of the programme (”Posttime_1”, “Posttime_2” and “Posttime_3”).

```
par(mfrow=c(1,4))
boxplot(MST$Pretest, xlab = "Pretest", ylab = "Score", ylim=c(60,140), col = 'red')
boxplot(MST$Posttest_Time1 ~ MST$Intervention, xlab = "Posttest 1", ylab = "Score",
        ylim=c(60,140), col = 'blue', )
boxplot(MST$Posttest_Time2 ~ MST$Intervention, xlab = "Posttest 2", ylab = "Score",
        ylim=c(60,140), col = 'yellow')
boxplot(MST$Posttest_Time3 ~ MST$Intervention, xlab = "Posttest 3", ylab = "Score",
        ylim=c(60,140), col = 'green')
```

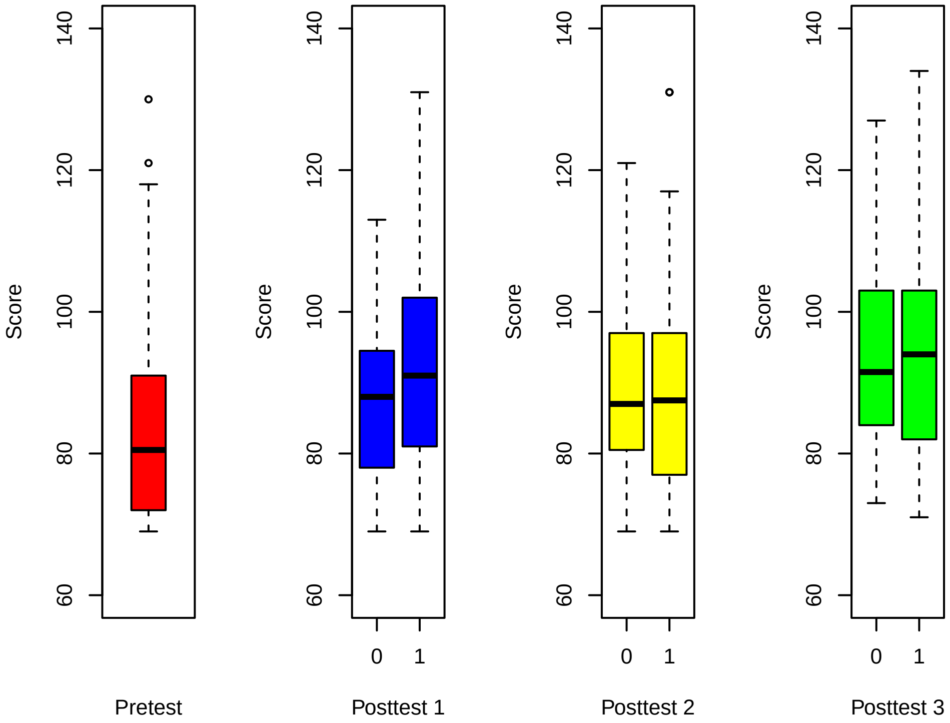


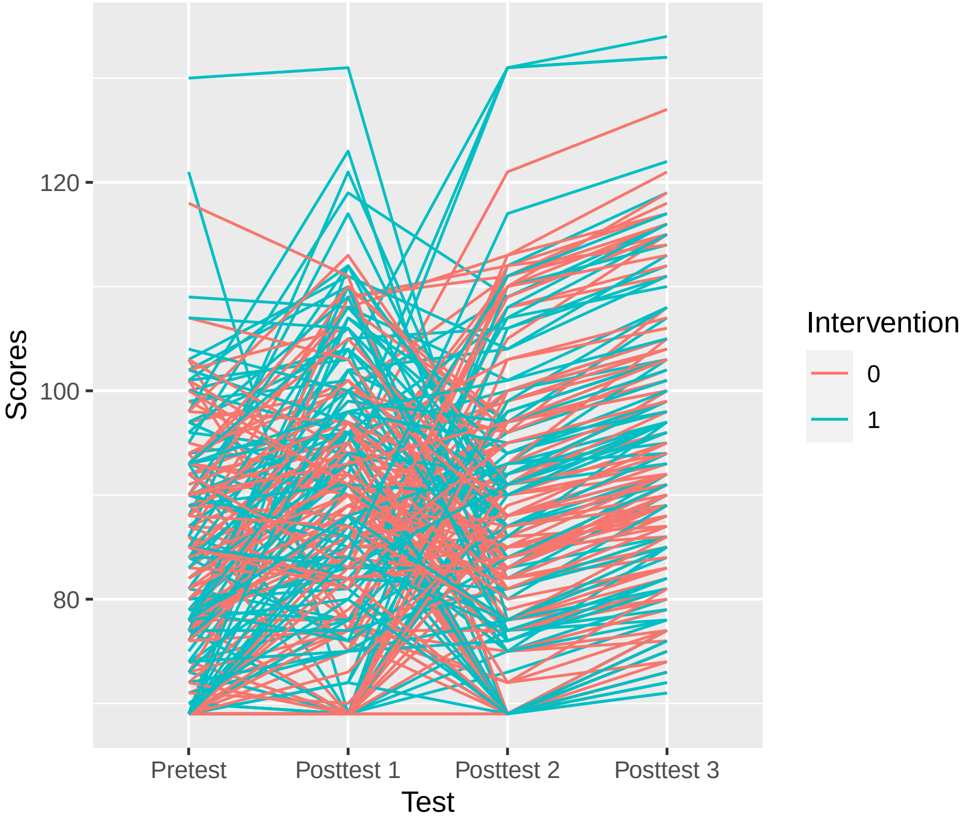
Figure 1: Boxplots for test scores

The value “0” refers to the control group and “1” refers to the intervention group. It appears that we can expect the intervention to have a slightly positive effect on student attainment, as shown by the generally higher medians and means of the scores from students in the intervention group, and the greater positive skew of the intervention boxplots compared to the control boxplots for Posttests 1 and 3. We see an increase of the mean score of the intervention groups from 91.27 for Posttest 1 to 93.95 for Posttest 3 and a similar increase is seen in the non-intervention group, suggesting that time is also an important factor in student attainment, as we would expect. There might also be an interaction between the intervention and time, that is, the time since start of the programme might influence the impact of the intervention on attainment.

We may wish to visualise the results grouped by intervention participation. We plot the results of every student below, split by colour into control or intervention group.

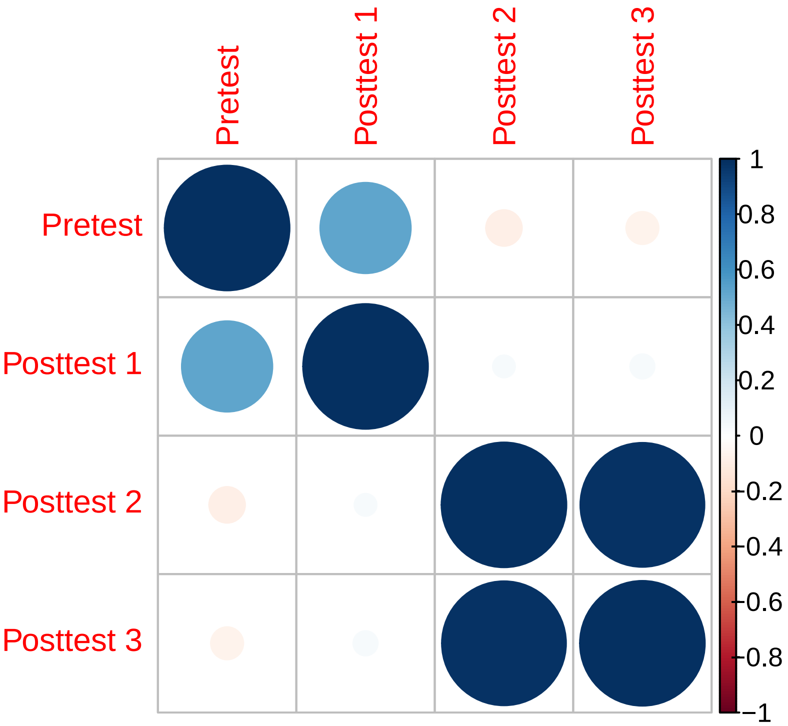
```
linedata = MST[,c(1,6,2,3,4,5)]
colnames(linedata) = c("ID", "Pretest", "Posttest 1",  "Posttest 2", "Posttest 3",
                       "Intervention")
longlinedata = pivot_longer(linedata,
                             cols =  c("Pretest", "Posttest 1",  "Posttest 2",
                                       "Posttest 3"),
                             names_to = "Test",
                             values_to = "Scores"
                             )[,c(1,3,4,2)]
longlinedata$Test = factor(x=longlinedata$Test,
                           levels = c("Pretest", "Posttest 1",
                                      "Posttest 2", "Posttest 3"),
                            ordered = TRUE)
longlinedata$Intervention = factor(x=longlinedata$Intervention)

ggplot(longlinedata, aes(x= Test, y = Scores, colour = Intervention)) +
  geom_line(aes(group = ID))
```



As before, the results suggest that the intervention has a slight positive impact on attainment, with the highest scores for each of the tests after the start of the intervention held only by students in the intervention group. However, it looks like the distribution of scores changes over time - it appears that for Posttest 1, the intervention had a more noticeable positive effect on student attainment compared to no intervention, but for Posttests 2 and 3, the scores of the students in the control group caught up. It may be possible that time is a more important factor than intervention in the long run. Lastly, we would like to see if the pretest score is in any way correlated with the test scores, i.e. whether students with higher pretest scores end up getting higher Posttest scores. This is visualized in the correlation plot below.

```
test_cor = cor(linedata[,c(2,3,4,5)])
corrplot(test_cor)
```



The plot shows that "Pretest" is moderately positively correlated with "Posttest 1" but is not significantly correlated with neither "Posttest 2" nor "Posttest 3", both of which are correlated strongly amongst themselves. This might suggest that Pretest scores are not very useful in predicting the posttest scores after a period of time, and that instead other variables (such as time itself) are more influential in determining student attainment.

Methods

To determine whether the educational intervention, time, pretest scores, or combinations of them have any significant impact on the test scores, we will build a "multilevel model". A multilevel model takes into account that we would expect correlations between measurements of units that are classed in the same group. In this multisite trial, the selection bias was reduced by the random allocation of individuals from different schools into one of two groups. However, we still find the need to account for the hierarchical structure of the data in both groups, since accounting for the effects of the school and the individual may reveal a different trend in the results to the trend we find otherwise. We might expect that the results of students from the same schools are correlated, and we certainly expect that multiple results from the same student will be correlated. Ignoring these correlations can confound the true relationship between the variables and can lead to falsely claiming that certain effects are more significant than they really are. However, building a linear regression model for every school and student would be highly impractical. A multilevel model allows us to account for the group structures and the effects of their units on the measurements, allowing us to more accurately predict the impact of certain variables in one model.

The multilevel model will be a three level longitudinal model. We will have a hierarchy with schools at the top level, represented by j, students at the second level represented by i, and measurements at the bottom level, represented by t. We will have no predictor variables at the school level, two at the student level, $intervention_{ij}$ and $pretest_{ij}$, and one at the measurements level, $occas_{tij}$, the numeral of the test (e.g. Posttest 2 corresponds to $occas_{tij} = 2$) which is a proxy for time. At the measurements level we also have our response variable, $score_{tij}$, which represents the test score of student i in school j for measurement t. A longitudinal model allows us to evaluate the impact of intervention over time without having to build three separate models and comparing them. By looking at the parameter estimates, we will be able to determine the size of the impact of certain variables and cross-terms on the student attainment.

We will first start by building an empty, 'intercept-only' model. Then, by calculating the intraclass correlation coefficient (ICC), we can assess the need for the grouping structure in our model (values greater than 0.10 indicate significance), but we intend to we build a "full" multilevel model anyway. This is to allow us to manually determine whether grouping is needed at the individual and school level. Using a top-down modelling strategy, we remove irrelevant fixed effects by looking at p-values and, and random effects by performing likelihood-ratio tests. Lastly, model assumptions such as linearity and normality of residuals and random effects are checked by plotting the residuals against fitted values and by using Q-Q plots.

Analysis

To build our model, we start by converting the data frame "MST" into a long format, where each row corresponds to a measurement.

```
extract_occasion = function(colname){
  occasion_text = substr(colname, nchar(colname),nchar(colname))
  occasion_no = as.integer(occasion_text)
  return(occasion_no)
}

trialdata = pivot_longer(MST,
                         cols = c('Posttest_Time1', 'Posttest_Time2', 'Posttest_Time3'),
                         names_to = 'occas',
                         names_transform = extract_occasion,
                         values_to = 'score'
                         )[,c(1,5,6,2,3,4)]
trialdata$ID = factor(trialdata$ID)
trialdata$School = factor(trialdata$School)
head(trialdata)
```

```
## # A tibble: 6 x 6
##   ID    occas score Intervention Pretest School
##   <fct> <int> <int>        <int>   <int> <fct>
## 1 1         1   101            1      99 1
## 2 1         2    87            1      99 1
## 3 1         3    91            1      99 1
## 4 2         1   119            1      95 1
## 5 2         2   109            1      95 1
## 6 2         3   115            1      95 1
```

We present the empty model as

$$score_{tij} = \gamma + \alpha_{ij} + \beta_j + \epsilon_{tij},$$

where γ is the overall intercept, $\alpha_{ij} \sim N(0, \sigma_\alpha^2)$ is the random effect for the student, $\beta_j \sim N(0, \sigma_\beta^2)$ is the random effect for the school, and $\epsilon_{tij} \sim N(0, \sigma_e^2)$ is the residual error term for the test score. Now we fit and summarize the empty model.

```
empty_model = lmer(formula = score ~ 1+(1|School) + (1|School:ID), data = trialdata)

#summary(empty_model)
# Output:
```

```
# Random effects:
#  Groups    Name        Variance Std.Dev.
#  School:ID (Intercept)  60.062   7.750
#  School    (Intercept)   3.494   1.869
#  Residual              123.492  11.113
# Number of obs: 630, groups:  School:ID, 210; School, 54
#
# Fixed effects:
#             Estimate Std. Error      df t value Pr(>|t|)
# (Intercept)  90.5820     0.7409 48.7779   122.3   <2e-16 ***

icc(empty_model, by_group = TRUE)
```

```
## # ICC by Group
##
## Group     |   ICC
## -----------------
## School:ID | 0.321
## School    | 0.019
```

We obtain values of $0.321 + 0.019 = 0.34$ for the student ICC and 0.019 for the school ICC. 0.34 is greater than 0.10, which we stated was the value past which we accept the significance of the grouping structure. However, the value of the school ICC is ≈ 0.02, which suggests that the school grouping structure is not needed, but as mentioned before we build the full model anyway. We start with

$$score_{tij} = a_{ij} + b_{ij} occas_{tij} + \epsilon_{tij}$$

where for student i in school j,

$$a_{ij} = \gamma + \alpha_{ij} + \beta_j + \delta_j intervention_{ij} + \zeta_j pretest_{ij}$$

and

$$b_{ij} = \eta + \kappa_{ij} + \lambda_j + \mu_j intervention_{ij} + \nu_j pretest_{ij}.$$

Let ω represent any of (β, ζ), and let ψ represent any of (α, κ). Then $\omega_j \sim N(0, \sigma^2_\omega)$ and $\psi_{ij} \sim N(0, \sigma^2_\psi)$. Furthermore, we have

$$\delta_j = \xi + \pi_j, \zeta_j = \theta + \phi_j, \mu_j = \rho + \tau_j, \text{ and } \nu_j = \psi + \Delta_j,$$

where if Λ represents any of $(\pi, \phi, \tau, \Delta)$ then $\Lambda_j \sim N(0, \sigma^2_\Lambda)$.

By substituting the expressions for $\delta_j, \zeta_j, \mu_j,$ and ν_j into a_{ij} and b_j, and then substituting the resulting expressions for a_{ij} and b_j into the expression for $score_{tij}$, we get:

$$\begin{aligned} score_{tij} =& \gamma + \alpha_{ij} + \beta_j + \xi \cdot intervention_{ij} + \theta \cdot pretest_{ij} + \eta \cdot occas_{tij} \\ &+ \rho \cdot intervention_{ij} occas_{tij} + \psi \cdot pretest_{ij} occas_{tij} \\ &+ \pi_j intervention_{ij} + \phi_j pretest_{ij} + \kappa_{ij} occas_{tij} \\ &+ \lambda_j occas_{tij} + \tau_j intervention_{ij} occas_{tij} \\ &+ \Delta_j pretest_{ij} occas_{tij} + \epsilon_{tij}, \end{aligned}$$

which is our ‘full’ model. We now fit and summarise the model, aiming to simplify it using statistical tests.

```r
full_model = lmer(score ~ 1 + Intervention + Pretest + occas + Intervention:occas
                  +Pretest:occas + (1+occas|School:ID)
                  +(1+Intervention+Pretest+occas+Intervention:occas+Pretest:occas|School),
                  data = trialdata)

coef(summary(full_model))
```

```
##                      Estimate Std. Error        df   t value     Pr(>|t|)
## (Intercept)        17.9760429 7.93005640 135.74139  2.266824 2.498200e-02
## Intervention        4.6866745 1.84459157  97.98889  2.540765 1.263081e-02
## Pretest             0.7977992 0.09681035  74.18492  8.240847 4.462850e-12
## occas              30.6291328 4.36081782 130.52160  7.023713 1.069281e-10
## Intervention:occas -1.6509631 1.13126079 116.23267 -1.459401 1.471521e-01
## Pretest:occas      -0.3340925 0.05210350 110.77319 -6.412093 3.628889e-09
```

The p-value for the fixed effect of $intervention_{ij}occas_{tij}$ is greater than 0.05, and so we do not need the $\rho \cdot intervention_{ij}occas_{tij}$ term in our model. This suggests that the time does not impact how effective the intervention is. Now we perform a likelihood-ratio test using the `ranova` function to determine the necessity of the random effects.

```r
# ranova(full_model)
```

The output is summarised below.

```r
# Output:
#                                                                                Pr(>Chisq)
# occas in (1 + occas | School:ID)                                               <2e-16 ***
# Intervention:occas in (1 + Intervention + Pretest + occas
#                          + Intervention:occas + Pretest:occas | School)        0.7496
# Pretest:occas in (1 + Intervention + Pretest + occas
#                     + Intervention:occas + Pretest:occas | School)             0.9404
```

We see that we do not need random slopes for school for neither $intervention_{ij}occas_{tij}$ nor $pretest_{ij}occas_{tij}$, so we remove the terms $\tau_j intervention_{ij}occas_{tij}$ and $\Delta_j pretest_{ij}occas_{tij}$. Now we check the significance of fixed and random effects of our updated model.

```r
second_model = lmer(score ~ 1+ Intervention + Pretest + occas + Pretest:occas +
                      (1+occas|School:ID) +
                      (1+Intervention+Pretest+occas|School), data = trialdata)
summary(second_model)$coef
```

```
##                 Estimate Std. Error        df   t value     Pr(>|t|)
## (Intercept)   19.8154411 7.83301044 185.17899  2.529735 1.224845e-02
## Intervention   2.5792637 1.14819924  52.39505  2.246355 2.892160e-02
## Pretest        0.7873277 0.09479364 155.02792  8.305701 4.615395e-14
## occas         29.2136142 4.25793749 200.27360  6.860978 8.344147e-11
## Pretest:occas -0.3264181 0.05080988 203.87132 -6.424304 9.171399e-10
```

```r
# ranova(second_model)
```

```
#Output:
#                                                                 Pr(>Chisq)
# occas in (1 + occas | School:ID)                                < 2.2e-16 ***
# Intervention in (1 + Intervention + Pretest + occas | School)   0.503018
# Pretest in (1 + Intervention + Pretest + occas | School)        0.789613
# occas in (1 + Intervention + Pretest + occas | School)          0.004705 **
```

Here we see that all the fixed effects are significant but we do not need random slopes for school for $intervention_{ij}$ or $pretest_{ij}$. Thus we remove the terms $\pi_j intervention_{ij}$ and $\phi_j pretest_{ij}$ from our model, and fit the next model.

```
third_model = lmer(score ~ 1+ Intervention + Pretest + occas + Pretest:occas +
                     (1+occas|School:ID) + (1+occas|School), data = trialdata)
#summary(third_model)
```

```
#Output:
# Random effects:
#  Groups    Name        Variance Std.Dev. Corr
#  School:ID (Intercept)  8.162   2.857
#            occas       32.431   5.695    -1.00
#  School    (Intercept) 74.561   8.635
#            occas        7.932   2.816    -1.00
#  Residual              64.139   8.009
# Number of obs: 630, groups:  School:ID, 210; School, 54
#
# Fixed effects:
#                 Estimate Std. Error        df t value Pr(>|t|)
# (Intercept)    20.51690    7.66220 295.48217   2.678  0.00783 **
# Intervention    2.47763    1.05389 208.73614   2.351  0.01966 *
# Pretest         0.77954    0.09069 312.21646   8.596 4.04e-16 ***
# occas          29.07661    4.20648 216.68289   6.912 5.25e-11 ***
# Pretest:occas  -0.32488    0.05019 220.52413  -6.473 6.15e-10 ***
```

```
#ranova(third_model)
```

```
#Output:
#                                                  Pr(>Chisq)
# occas in (1 + occas | School:ID)              < 2.2e-16 ***
# occas in (1 + occas | School)                  0.0009285 ***
```

Finally, we see that all the fixed and random effects are significant! We check to see that assumptions made in building the model are satisfied, namely the assumptions of linearity and normality of the residuals and random effects, before taking this as our final model. First is the plot of residuals against fitted values.

```
plot(third_model)
```

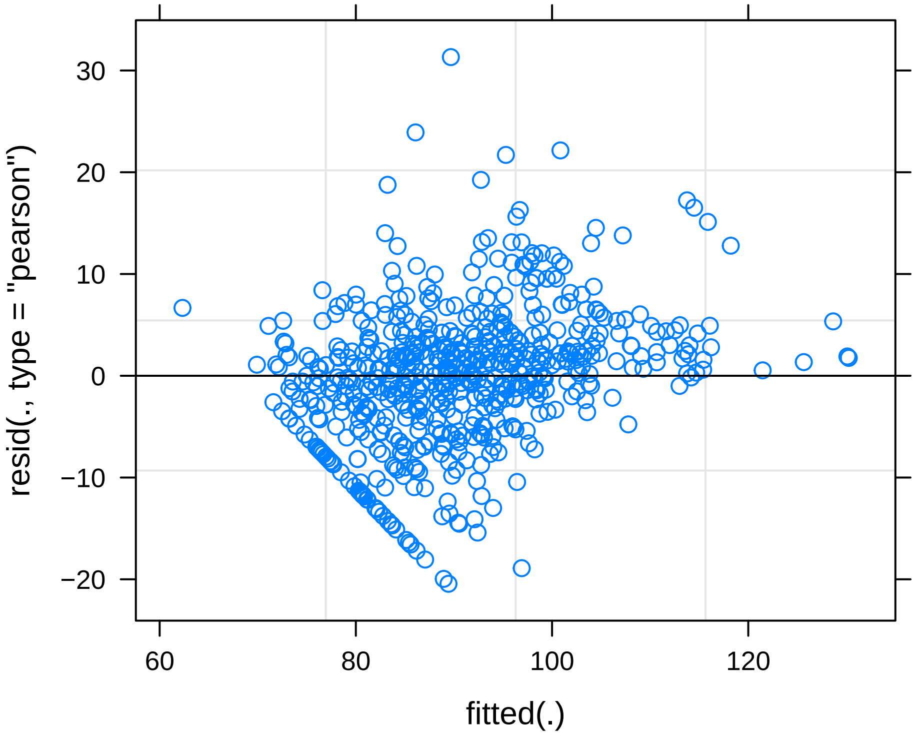


Figure 2: Residuals against fitted values

The diagonal line in the bottom left comes from the minimum recorded values of the pretest and tests, 69, which a high number of students obtained. Otherwise, there is generally random scatter in the chart, which agrees with the assumption of linearity for the model. Now we test for normality of residuals.

```
qqnorm(resid(third_model))
qqline(resid(third_model))
```

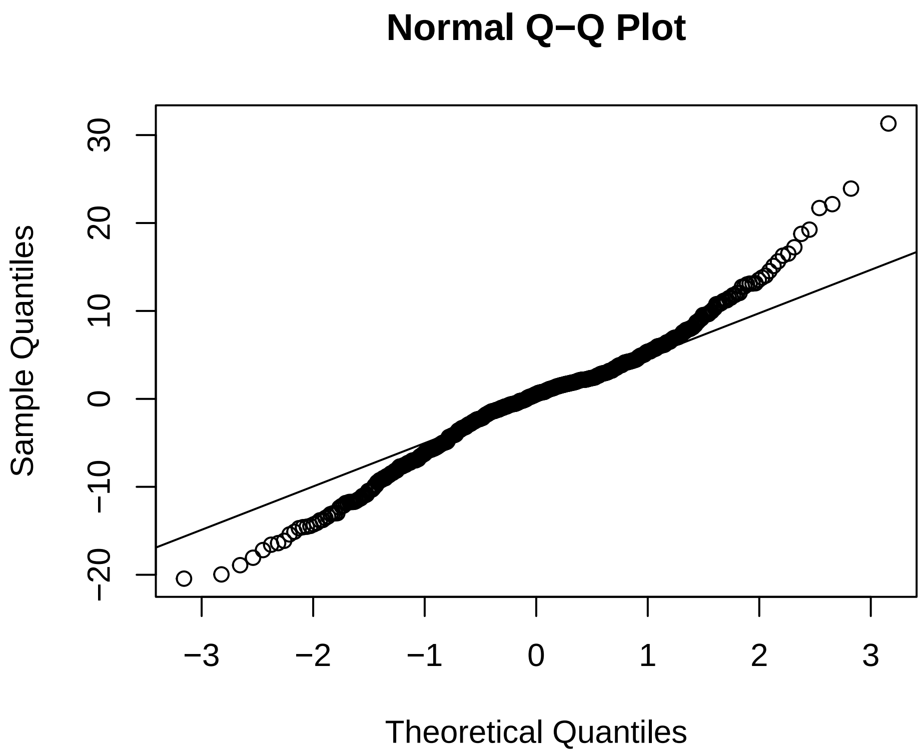


Figure 3: Q-Q plot for normality of residuals

The plotted points lie approximately on the $y = x$ lines, but we notice deviations at either end. To ensure we have normality, we use the Shapiro-Wilk test.

```
shapiro.test(resid(third_model))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(third_model)
## W = 0.97642, p-value = 1.542e-08
```

The p-value obtained, $1.542*10^{-8}$, is much lower than 0.05, so we can now confidently say we have normality of residuals. Next we check for normality of random effects.

```
par(mfrow=c(2,2))
qqnorm(ranef(third_model)$School[,1])
qqline(ranef(third_model)$School[,1], col ='red')

qqnorm(ranef(third_model)$School[,2])
qqline(ranef(third_model)$School[,2],col = 'red')

qqnorm(ranef(third_model)$`School:ID`[,1])
qqline(ranef(third_model)$`School:ID`[,1], col='red')
```

```r
qqnorm(ranef(third_model)$`School:ID`[,2])
qqline(ranef(third_model)$`School:ID`[,2], col='red')
```

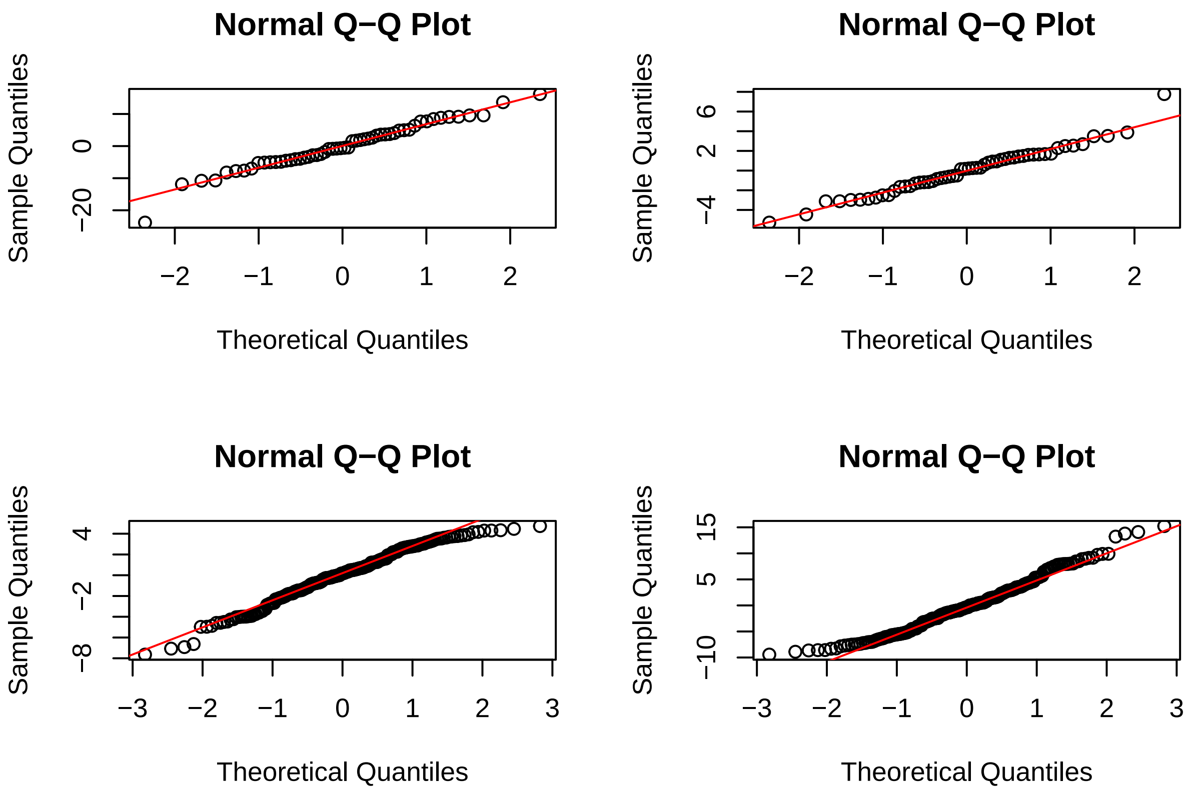


Figure 4: Q-Q plots for normality of random effects

We have that the plotted points generally lie close to the $y = x$ lines, so normality of random effects also holds. So the third model is our final model:

$$\begin{aligned}
score_{tij} =& 20.51690 + \alpha_{ij} + \beta_j + 2.47763 intervention_{ij} + 0.77954 pretest_{ij} \\
& + 29.07661 occas_{tij} - 0.32488 pretest_{ij} occas_{tij} + \kappa_{ij} occas_{tij} + \lambda_j occas_{tij} \\
& + \epsilon_{tij}
\end{aligned}$$

where

$$\alpha_{ij} \sim N(0, 8.162), \beta_j \sim N(0, 74.561), \kappa_{ij} \sim N(0, 32.431), \lambda_j \sim N(0, 7.932) \text{ and } \epsilon_{tij} \sim N(0, 64.139).$$

```r
final_model=third_model
(summary(final_model)$coef)
```

```
##                  Estimate Std. Error       df  t value     Pr(>|t|)
## (Intercept)    20.5168979 7.66220364 295.4822 2.677676 7.828093e-03
## Intervention    2.4776335 1.05388976 208.7361 2.350942 1.965815e-02
## Pretest         0.7795411 0.09068769 312.2165 8.595888 4.043085e-16
## occas         29.0766115 4.20647501 216.6829  6.912346 5.254399e-11
## Pretest:occas -0.3248784 0.05019074 220.5241 -6.472875 6.147464e-10
```

Our model suggests that the significant predictors are the intervention, pretest score, time (*occas*), and a cross-level interaction between pretest and time. Therefore we have an answer to the original question behind our investigation - we expect the educational intervention to increase the score of a student i in school j by 2.47763353 ± 2.10778 marks at the 95% confidence level. Additionally, at the 95% confidence level we also expect that each increase in the pretest score affects the final score by adding 0.7795411 ± 0.1813754, then subtracting $(0.3248784 \pm 0.1003815) * occas_{tij}$, and each increase in the time *occas* (that is, each subsequent posttest) adds on 29.0766115 ± 8.41295 marks.

There does not seem to be any significant random effect for the intervention between schools, since the p-value obtained for the likelihood-ratio test was $0.503018 > 0.05$. However, there is in fact evidence of heterogeneity of the effect of time at the school and student level - the likelihood-ratio tests produced p-values of $2.2 * 10^{-16}$ for ‘School:ID’ and 0.0009285 for ‘School’, and both are magnitudes smaller than 0.05. We note that we also have a statistically significant interaction between $pretest_{ij}$ and $occas_{tij}$, which is negative in size, and suggests that as time grows, the pretest score becomes less relevant. There does not seem to be any evidence of any other cross-level effects.

Discussion of Results and Conclusion

We conclude by discussing the importance of our results, providing possible interpretations, and any limitations of our model. Firstly, we conclude that the educational intervention does have an impact on student attainment at the 95% significance level, contributing 2.47763353 ± 2.10778 marks. However, we note that this is a relatively small impact considering that the marks can theoretically reach 200, meaning we only expect a 1.2% increase in marks. We also have a large positive fixed effect for *occas*, making time an important factor in predicting test scores, which we expect as with more time students can practice and improve. A noticeable impact is seen from the pretest score, which we can view as a proxy for student ability (although not a perfect representation), and so unsurprisingly higher pretest scores predict higher test scores. Note that since pretest scores are generally large in size compared to the other variables, the pretest score provides a meaningful contribution to the predicted score despite the size of its coefficient. The random effects suggest that factors such as the particular student and school impact both the predicted baseline score and also the impact of time on the predicted test score, which could be explained by factors such as varying student motivation and school expectations. Lastly, we can interpret the cross term to mean that with increasing time, the pretest score is less useful at predicting future scores, probably because a gap in time between tests allows for other factors such as revision time to influence posttest scores. With every future test, the contribution to the predicted score by $pretest_{ij}$ decreases by $0.3248784 * pretest_{ij}$. This result agrees with the high correlation found between the pretest and posttest 1 scores in our initial exploratory analysis, while there was virtually no correlation between the pretest and subsequent posttest scores.

We highlight some limitations of the model. An interpretation of the intercept relies on the other predictors being 0, which in our model is not possible. This is because *occas* has been defined to take values $1, 2, 3 \ldots$ which correspond to the numeral of the posttest. Thus, we find that our intercept only serves to help fit the model to appropriate values and cannot be interpreted. Furthermore, we have assumed in building the model that *occas* represents times that are evenly split apart, that is, the tests were taken at constant intervals. Otherwise, the meaning of *occas* falls apart, because the difference in *occas* (numeral of the test) between any two consecutive tests is always 1.

We have found in our investigation that the educational intervention does have a statistically significant impact on the student attainment, but its impact over time does not seem to change. Furthermore, although units such as the school and the student have an effect on the impact of time on the predicted test score, they do not seem to have an impact on the effect of the intervention at all. Overall, while the intervention does have a measurable impact on the predicted test score, the size of the impact is not particularly large.

```
require(wordcountaddin)
```

```
## Loading required package: wordcountaddin
```

```
word_count()
```

```
## For information on available language packages for ’koRpus’, run
##
##   available.koRpus.lang()
##
## and see ?install.koRpus.lang()
```

```
##
## Attaching package: ’koRpus’
```

```
## The following object is masked from ’package:wordcountaddin’:
##
##     readability
```

```
## [1] 2750
```

```
text_stats()
```

Method	koRpus	stringi
Word count	2750	2437
Character count	15268	15371
Sentence count	133	Not available
Reading time	13.8 minutes	12.2 minutes