Linear Mixed Model Example: Test Score

Mila Sun

Winter 2022

1 Install and load required packages

```
#uncomment and run the following line if you did have these packages installed
#install.packages(c("nlme","lme4","mlmRev","ggplot2","lattice"))

# load packages
library(lme4)
```

```
## Warning: package 'lme4' was built under R version 3.6.2
```

```
library(nlme)
library(mlmRev)  #for Exam data
```

```
## Warning: package 'mlmRev' was built under R version 3.6.2
```

```
# for data visualization
library(ggplot2)
library(lattice)
```

2 Test score data

- 4,059 students in 65 schools
- Exam scores - one exam per student
- Correlation - If there are "high-performing" schools, maybe kids in these schools are more alike
- Outcome: normalized exam score
- Covariates:
  - Sex of the student
  - Pretest (LR test) score
  - School gender (mixed, boys, girls)
  - School average score
- All scores are standardized

3 Question of interest

- How does the pretest score affect the exam score?
  - In individuals
  - On the school level: are kids in high-performing schools also high-performing in exam scores?
- How much between-school variability is there?
- Is there between-school variability in the association between the scores?

4 Modelling

4.1 Fit a simple linear model

Fit model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

```r
# load built-in data
data(Exam)

# fit a simple linear model
m1 = lm(normexam ~ standLRT, data = Exam)
summary(m1)
```

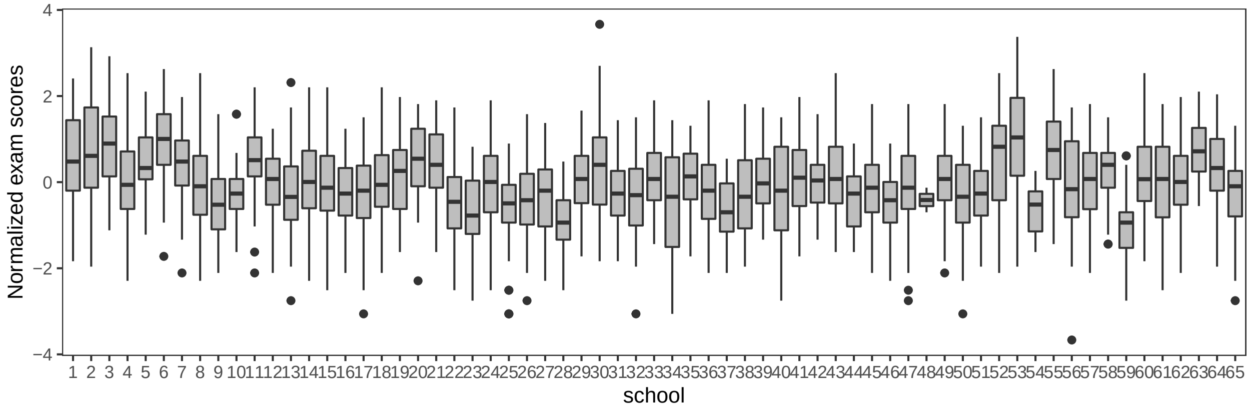
```
##
## Call:
## lm(formula = normexam ~ standLRT, data = Exam)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.65615 -0.51848  0.01265  0.54399  2.97399
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.001191   0.012642  -0.094    0.925
## standLRT     0.595057   0.012730  46.744   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8054 on 4057 degrees of freedom
## Multiple R-squared:   0.35,  Adjusted R-squared:  0.3499
## F-statistic:  2185 on 1 and 4057 DF,  p-value: < 2.2e-16
```

- **Fixed coefficients**: $\hat{\beta}_0 = -0.0011$ $(se = 0.012)$ and $\hat{\beta}_1 = 0.595$ $(se = 0.012)$
- **Variance of random term (error)**: $\hat{\sigma}^2 = 0.8054^2$

4.2 Preliminary graphical displays: the effect of schools

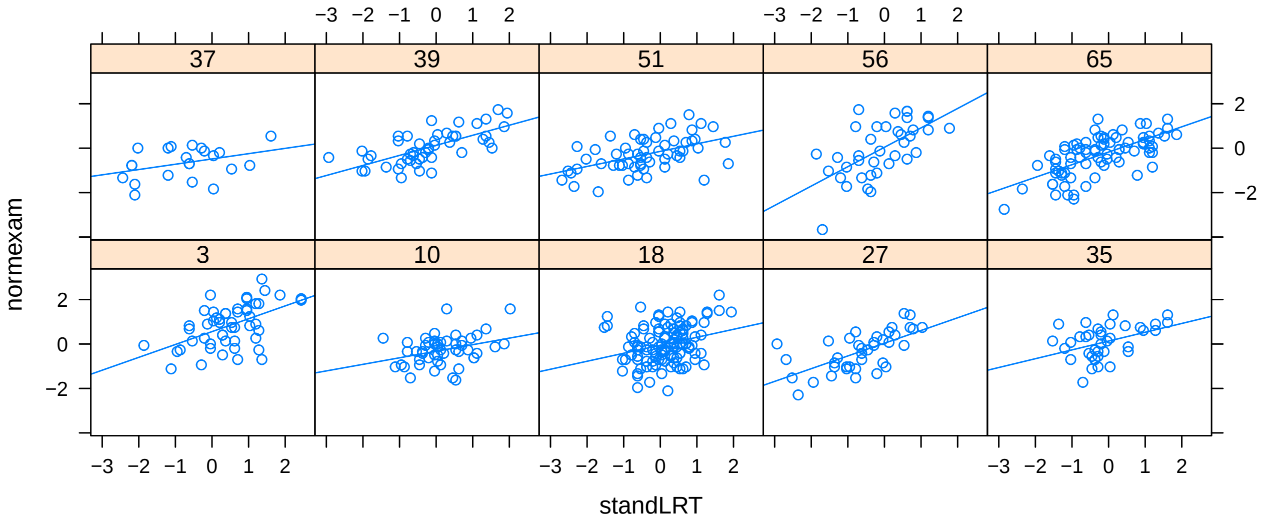
We plot the outcomes across school IDs, you see lots of individual variation, with some schools having relatively higher exam scores and others having relatively lower scores

```r
ggplot(Exam, aes(school, normexam)) +
  geom_boxplot(fill="grey") +
  ylab("Normalized exam scores") +
  theme_bw() + theme(panel.grid = element_blank())
```

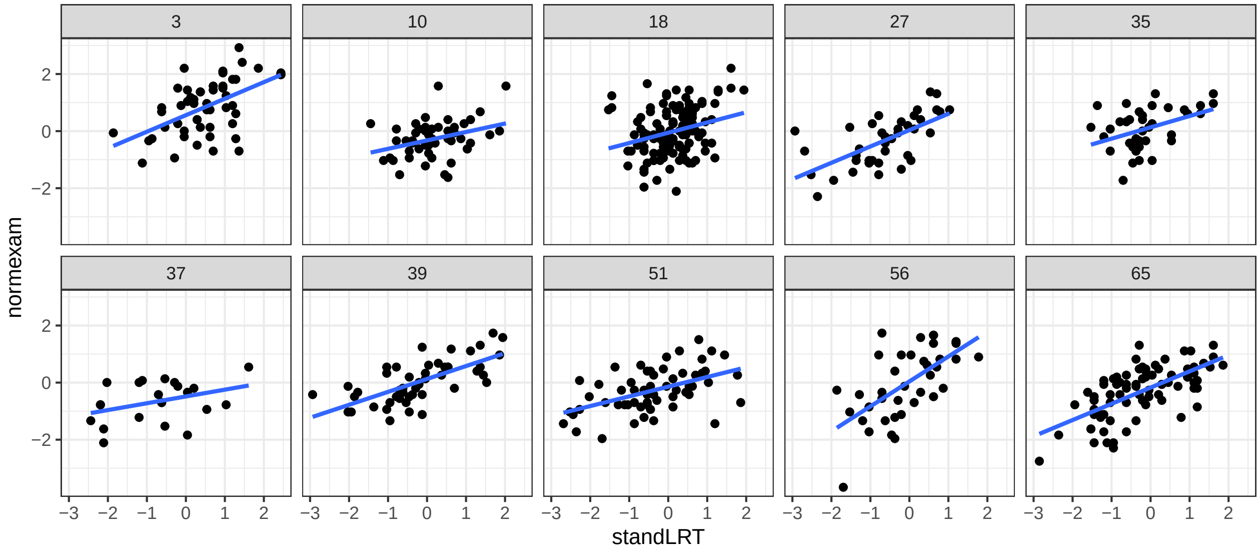


We can check for patterns within and between schools by plotting the response versus the pretest scores by school. The lines denotes simple linear regressions.

```
set.seed(202201)
school_id = sample(unique(Exam$school), size = 10, replace = F)
Exam_sub = Exam[Exam$school %in% school_id,]

# scatterplots group by school
xyplot(normexam ~ standLRT | school, data = Exam_sub, type = c("p", "r"))
```



```
# another way to display
ggplot(Exam_sub, aes(x=standLRT, y=normexam)) +
  geom_point() + geom_smooth(method="lm", se=FALSE) +
  facet_wrap(~school, nrow = 2) + theme_bw()
```



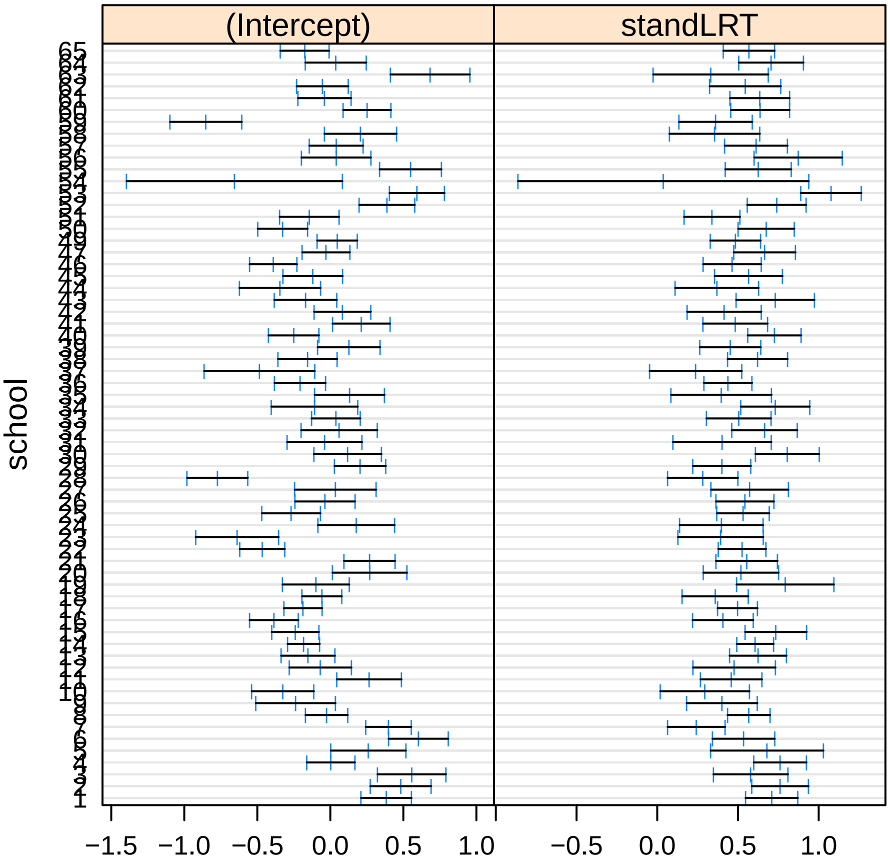
Although it is informative to plot the within-school regression lines, we need to assess the variability in the estimates of the coefficients before concluding if there is “significant” variability **between schools**. We can obtain the individual regression fits with the `lmList` function.

Note: school 48 only has two students taking the exam. Because within-school results based on only two students are unreliable, we will exclude this school from further plots (but we do include these data when fitting comprehensive models)

```
m1_lst = lmList(normexam ~ standLRT|school, Exam, subset = school!=48)
head(coef(m1_lst), n=10) # show the first 10 regression lines
```

```
##      (Intercept)  standLRT
## 1    0.383334972 0.7093415
## 2    0.482279914 0.7612884
## 3    0.557754447 0.5789886
## 4    0.003755186 0.7614459
## 5    0.260447309 0.6800207
## 6    0.603214394 0.5353444
## 7    0.398526891 0.2422785
## 8   -0.025194630 0.5674053
## 9   -0.237367229 0.4011999
## 10  -0.325234425 0.2950453
```

```
#and compare the confidence intervals on these coefficients.
ci = intervals(m1_lst)
plot(ci)
```



The confidence intervals for these separately fitted models indicate differences in the intercepts and the slopes.

4.3 Fit a random intercept model

We begin with a model that has a random intercept by school plus additive fixed effects for the pretest score:

$$y_{ij} = \beta_0 + u_i + \beta_1 x_{ij} + \epsilon_{ij}.$$

Recall:

- Random errors $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_\epsilon^2)$
- u_i is the random intercept for cluster i, $u_i \sim \mathcal{N}(0, \sigma_u^2)$

```
# mixed model with random intercept
m2 = lmer(normexam ~ standLRT + (1|school), data = Exam)
summary(m2)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: normexam ~ standLRT + (1 | school)
##    Data: Exam
##
## REML criterion at convergence: 9368.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.7166 -0.6302  0.0294  0.6849  3.2673
##
```

```
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  school   (Intercept) 0.09384  0.3063
##  Residual             0.56587  0.7522
## Number of obs: 4059, groups:  school, 65
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept) 0.002323   0.040354   0.058
## standLRT    0.563307   0.012468  45.180
##
## Correlation of Fixed Effects:
##          (Intr)
## standLRT 0.008
```

- **Random intercept variance**: $\hat{\sigma}_u^2 = 0.0934$
- **Error variance**: $\hat{\sigma}_\epsilon^2 = 0.566$
- **Fixed effects**: $\hat{\beta}_0 = 0.002$ $(se = 0.040)$, $\hat{\beta}_1 = 0.566$ $(se = 0.012)$
- $Corr(\hat{\beta}_0, \hat{\beta}_1) = 0.008$ (not really of interest)

4.4 Fit a random coefficient model

Our data exploration indicated that the slope may vary by school. We can fit a model with random effects by school for both the slope and the intercept as

$$Y_{ij} = (\beta_0 + u_{0i}) + (\beta_1 + u_{1i})X_{ij} + \epsilon_{ij}$$

Recall:

- $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_\epsilon^2)$
- $u_{0i} \sim \mathcal{N}(0, \sigma_{u_0}^2)$
- $u_{1i} \sim \mathcal{N}(0, \sigma_{u_1}^2)$

```
m3 = lmer(normexam ~ standLRT + (standLRT|school), data = Exam, )
summary(m3)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: normexam ~ standLRT + (standLRT | school)
##    Data: Exam
##
## REML criterion at convergence: 9327.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.8323 -0.6317  0.0339  0.6834  3.4562
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  school   (Intercept) 0.09212  0.3035
##           standLRT    0.01497  0.1223   0.49
##  Residual             0.55364  0.7441
## Number of obs: 4059, groups:  school, 65
##
## Fixed effects:
```

```
##             Estimate Std. Error t value
## (Intercept) -0.01165    0.04011   -0.29
## standLRT     0.55653    0.02011   27.67
##
## Correlation of Fixed Effects:
##          (Intr)
## standLRT 0.365
```

- **Random intercept variance**: $\hat{\sigma}^2_{u_0} = 0.09212$
- **Random slope variance**: $\hat{\sigma}^2_{u_1} = 0.01497$
- **Error variance**: $\hat{\sigma}^2_{\epsilon} = 0.55364$

4.5 Model selection

Please be very, very careful when it comes to model selection. Focus on your question, don't just plug in and drop variables from a model haphazardly until you make something "significant". Always choose variables based on biology/ecology.

First think about your **experimental design, your system and data collected, as well as your questions**. Then you could use model selection to help you decide:

- AIC and/or BIC
- likelihood ratio test
- whether the variance of the random term ≈ 0
  - If $\sigma^2_{u_0} = 0$, this means that all "school-specific" intercepts are the same, which means that there is no variability due to school.

```
# compare m3 to the previous fit m2 with
anova(m2, m3)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: Exam
## Models:
## m2: normexam ~ standLRT + (1 | school)
## m3: normexam ~ standLRT + (standLRT | school)
##    npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## m2    4 9365.2 9390.5 -4678.6   9357.2
## m3    6 9328.9 9366.7 -4658.4   9316.9 40.372  2  1.711e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There is a strong evidence of a significant random effect for the slope by school, whether judged by AIC, BIC or the p-value for the likelihood ratio test.

Example with too small variance: add a fix and a random effect for the student's sex by school. Notice that the estimate of the variance of the `sex` term is essentially zero so there is no need to test the significance of this variance component.

```
m4 = lmer(normexam ~ standLRT + sex+ (standLRT+sex|school), data = Exam)
summary(m4)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: normexam ~ standLRT + sex + (standLRT + sex | school)
##    Data: Exam
##
## REML criterion at convergence: 9302.6
```

```
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.8303 -0.6387  0.0216  0.6794  3.4447
##
## Random effects:
##  Groups   Name        Variance  Std.Dev. Corr
##  school   (Intercept) 0.0940989 0.30676
##           standLRT    0.0151222 0.12297   0.51
##           sexM        0.0008722 0.02953  -0.80  0.10
##  Residual             0.5501233 0.74170
## Number of obs: 4059, groups:  school, 65
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept)  0.06646    0.04292   1.548
## standLRT     0.55277    0.02015  27.433
## sexM        -0.18261    0.03229  -5.656
##
## Correlation of Fixed Effects:
##          (Intr) stnLRT
## standLRT  0.344
## sexM     -0.411  0.048
## optimizer (nloptwrap) convergence code: 0 (OK)
## boundary (singular) fit: see ?isSingular
```

5 Visualizing random effects

Examines the random effects (i.e things that are allowed to vary across schools, in this case each represents school-level effect of standLRT for our first 10 schools)

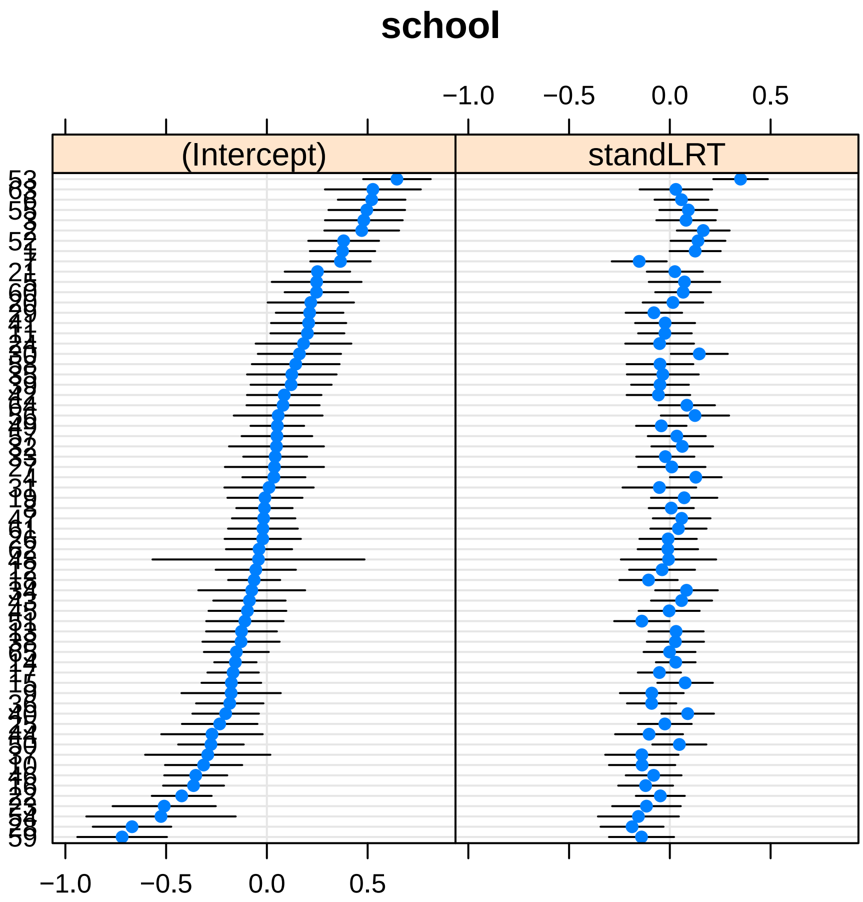
```
# The varying intercepts and slopes for each subject can be viewed by
ranef(m3)$school[1:10,]
```

```
##    (Intercept)     standLRT
## 1   0.37538382  0.125697850
## 2   0.47067577  0.165713931
## 3   0.48110095  0.080462260
## 4   0.03477527  0.128874470
## 5   0.24693327  0.072661033
## 6   0.51989850  0.057743604
## 7   0.36508132 -0.151940089
## 8  -0.01207589  0.007121824
## 9  -0.17738478 -0.089551620
## 10 -0.31392994 -0.137501408
```

Note: `ranef()` gives the conditional modes conditional on Y. You can think of these as school-level effects, i.e. how much does any school differ from the population?

```
# The error bars represent 95% confidence intervals.
print(dotplot(ranef(m3,condVar=TRUE)))
```

```
## $school
```



6 Check model assumptions

- For a linear mixed model, the assumptions are
  - Linearity: the explanatory variables are related linearly to the response
  - $u_{0i} \sim \mathcal{N}(0, \sigma^2_{u_0})$, independent of each other
  - $u_{1i} \sim \mathcal{N}(0, \sigma^2_{u_1})$, independent of each other
  - $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2_{\epsilon})$, independent of each other and of u_i's
- Note: how to compute residuals ϵ_{ij}:

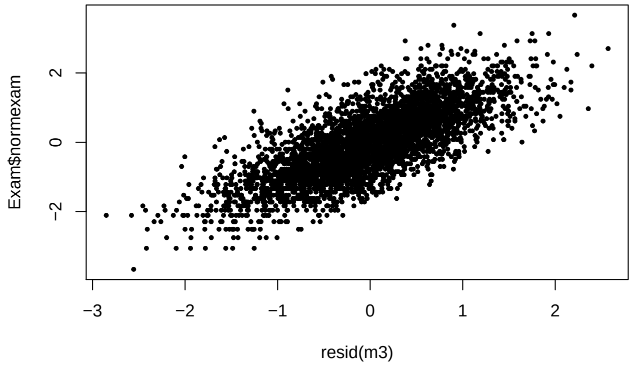
$$\hat{\epsilon}_{ij} = y_{ij} - \hat{\beta}_0 - \hat{u}_{0i} - (\hat{\beta}_1 + \hat{u}_{1i})X_{ij}$$

- In R:
  - ϵ's are called the "school residuals"
  - $y_{ij} - \hat{\beta}_0 - \hat{\beta}_1 X_{ij}$ are called the "fixed" residuals

6.1 Linearity

Graphically, plotting the model residuals against the response is one simple way to test and looking for any systematic shape. If an obvious pattern emerges, a higher order term may need to be included or you may need to mathematically transform a predictor/response.

```
plot(resid(m3), Exam$normexam, pch=19, cex=0.5)
```

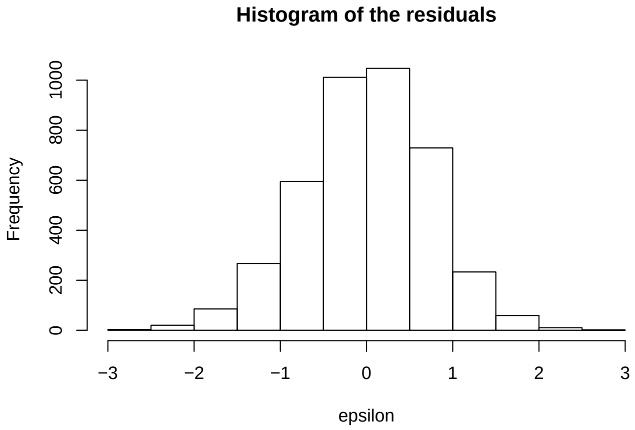


6.2 Check normality, independence and constant variance assumptions

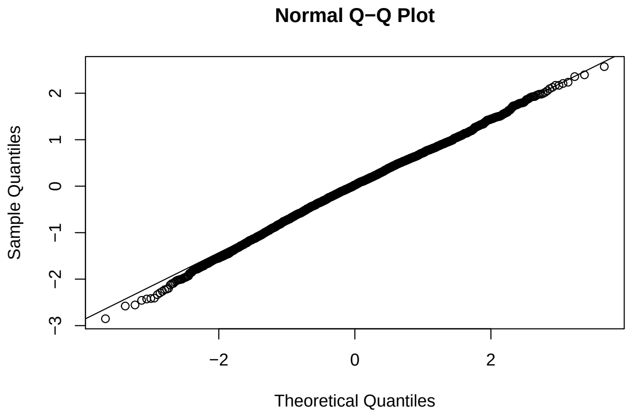
Note:

- `fitted()` computes $\beta_0 + u_{0i} + (\beta_1 + u_{1i})x_{ij}$
- A Q-Q plot or histogram of the residuals
- Plotting the residuals against the fitted values will indicate if there is non-constant error variance, i.e. if the variance increases with the mean, the residuals will fan out as the fitted value increases. Usually transforming the data, or using another distribution will help.

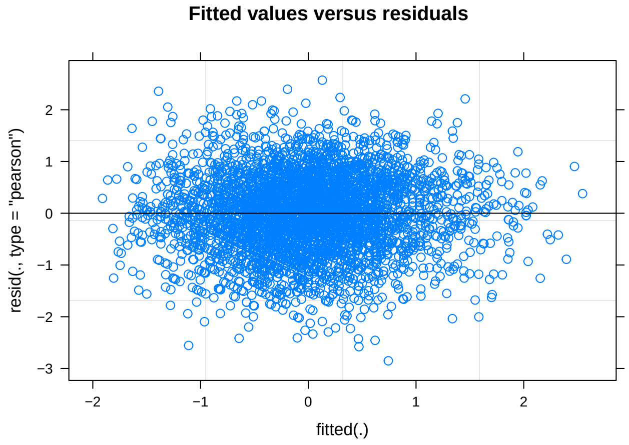
```
# check the normality assumption of the residuals
hist(residuals(m3), main = "Histogram of the residuals", xlab = "epsilon")
```



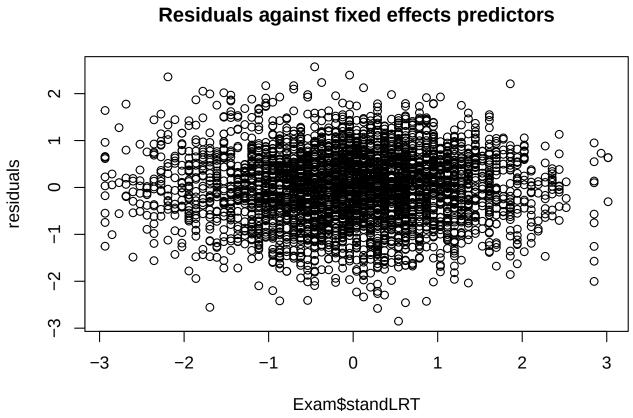
```
qqnorm(residuals(m3, type="pearson"))
qqline(residuals(m3, type="pearson"))
```



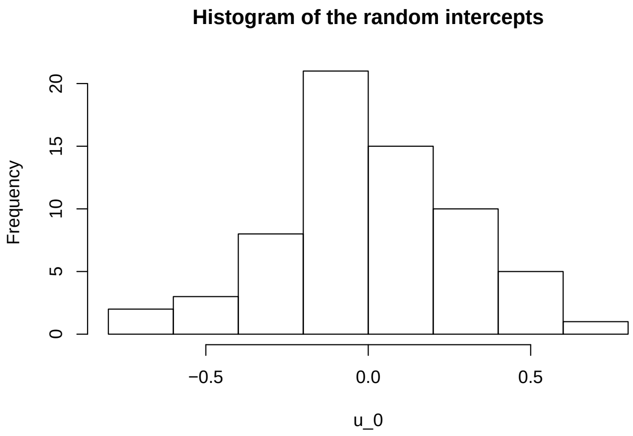
```
#or simply just run
#qqmath(m3)

# check the constant variance and independence assumption
plot(m3, main="Fitted values versus residuals")
```



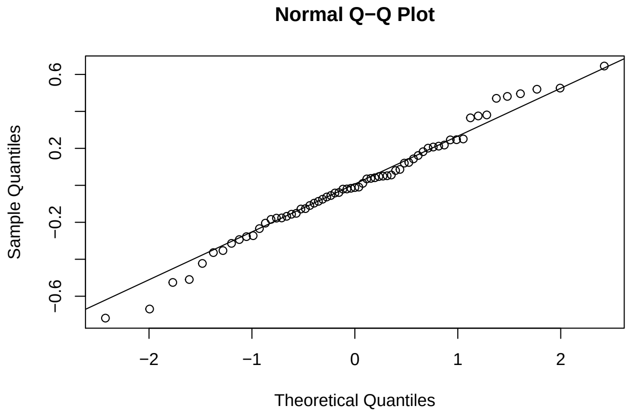
```
plot(Exam$standLRT, residuals(m3, type="pearson"), ylab="residuals", main="Residuals against fixed effe
```



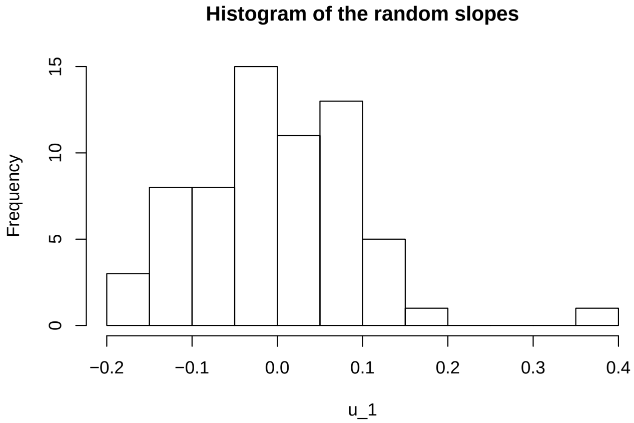
```
#The even spread of the residuals suggest that the model is a good fit for the data.


# check the normality of the random effects
ranef_m3 = ranef(m3)$school

hist(ranef_m3[,1], main = "Histogram of the random intercepts", xlab = "u_0")
```



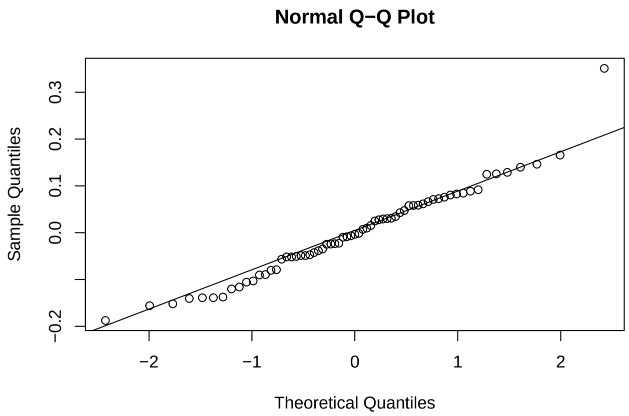
```
qqnorm(ranef_m3$`(Intercept)`)
qqline(ranef_m3$`(Intercept)`)
```



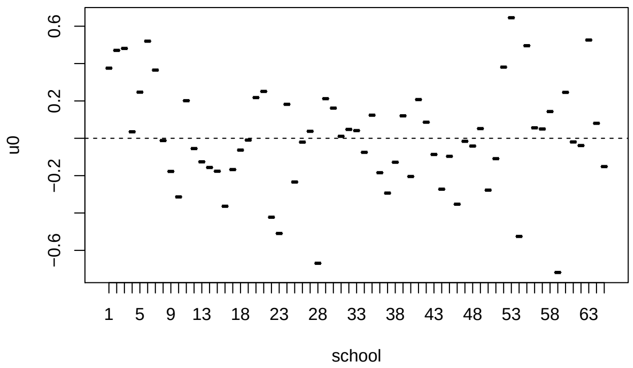
```
hist(ranef_m3[,2], main = "Histogram of the random slopes", xlab = "u_1")
```



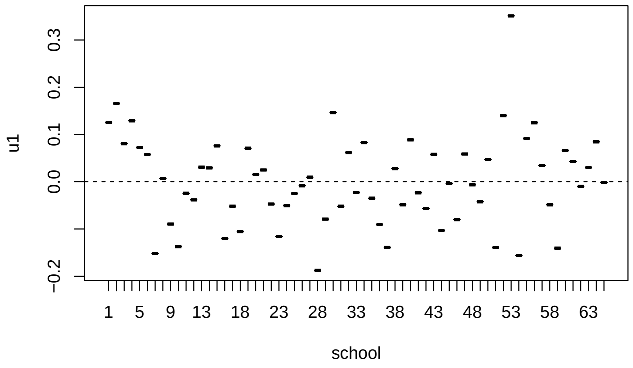
```
qqnorm(ranef_m3$standLRT)
qqline(ranef_m3$standLRT)
```



```r
# check the constant variance of the random effects
ranef_m3$school_id = unique(Exam$school)
plot(as.factor(ranef_m3$school_id), ranef_m3$`(Intercept)`, xlab="school", ylab="u0")
abline(h=0, lty=2)
```



```r
plot(as.factor(ranef_m3$school_id), ranef_m3$standLRT, xlab="school", ylab="u1")
abline(h=0, lty=2)
```



7 Parameter interpretation for linear mixed models

Consider model m3 with both random intercepts and random slopes:

$$y_{ij} = (\beta_0 + u_{0i}) + (\beta_1 + u_{1i})x_{ij} + \epsilon_{ij}.$$

Then

$$E[Y|u_0, u_1, X] = (\beta_0 + u_0) + (\beta_1 + u_1)X,$$

and the marginal model is

$$E[Y|X] = \beta_0 + \beta_1 X.$$

So:

- β_0 is the expected response at $X = 0$ (which is the mean pretest score)
- β_1 is the expected change in response for a unit increase in X
- These expectations are with respect to the distribution of random effects and are averages across the population of individuals

For a generic individual (i.e., nested within school in our example)

- $\beta_0 + u_0$ is the expected response at $X = 0$
- $\beta_1 + u_1$ is the expected change in response for a unit increase in X
- In a linear model, an alternative interpretation is that β_1 is the change in response for a unit change in X for a "typical" school with $u_1 = 0$.