Compulsory exercise 3: Group 13

TMA4315 Generalized Linear Models

Benjamin Sigbjørnsen, Johan Vik Mathisen, and Martinius Theodor Singdahslen

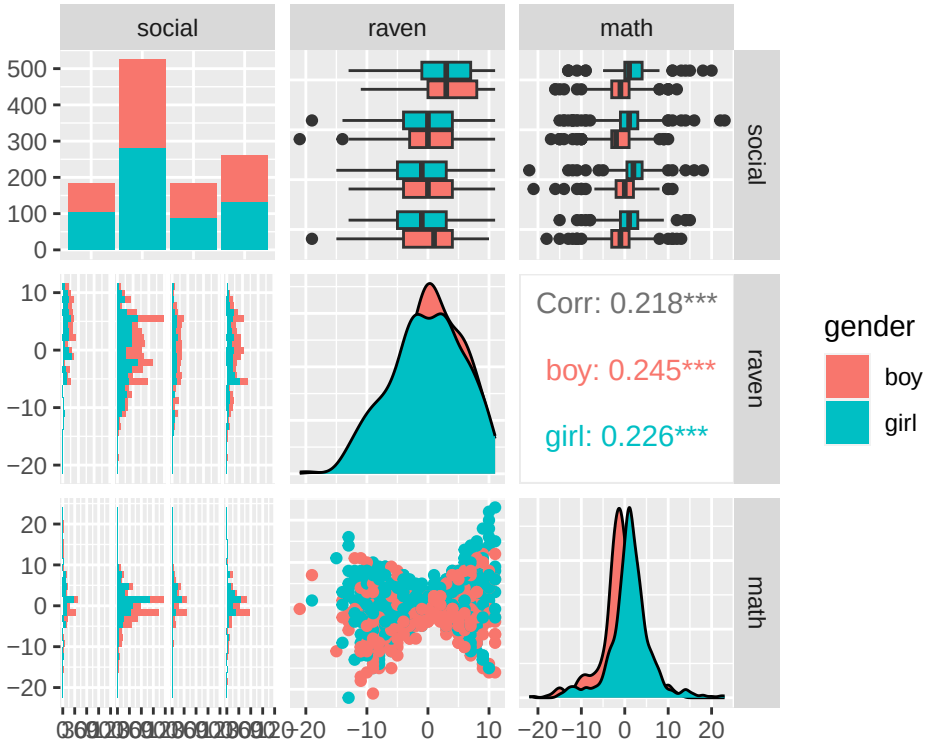
17 November, 2023

Problem 1

```
dataset <- read.table("https://www.math.ntnu.no/emner/TMA4315/2018h/jsp2.txt",
    header = TRUE)
```

a)

```
library(ggplot2)
library(GGally)
ggpairs(data = dataset, mapping = aes(col = gender), columns = c("social",
    "raven", "math"), legend = 1)
```



In the plot above, we see that the variables `social`, `raven` and `math` are plotted against each other distributed on `gender`. `Social` is a categorical variable with four levels. One of the categories in `social` have a higher frequency than the others, and one of the categories seems to have a higher mean value of `raven` than the others. `Math` seems to be evenly distributed between categories. `Math` and `raven` are continuous variables.

Raven seems to be similar distributed at each level of gender, while math seems to have a higher mean for girls. There is a significant correlation between raven and math.

```
# Fitting a linear model
mod = lm(math ~ raven + gender, data = dataset)
summary(mod)
```

```
##
## Call:
## lm(formula = math ~ raven + gender, data = dataset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.6704  -1.8791   0.1166   2.1166  19.6134
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.3131     0.2024  -6.488 1.29e-10 ***
## raven         0.1965     0.0240   8.188 6.98e-16 ***
## gendergirl    2.5381     0.2807   9.041  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.76 on 1151 degrees of freedom
## Multiple R-squared:  0.1105, Adjusted R-squared:  0.109
## F-statistic:  71.5 on 2 and 1151 DF,  p-value: < 2.2e-16
```

We have a linear model where Y_k is the response variable, X_k are the k-th row in the design matrix containing the covariates, β is a vector of coefficient that are found when fitting the model and ε_k are the error of the model

In the summary output above, we can see that both of the covariates are significant. Expected math score increases with raven and if the gender is girl. This is consistent with what we saw in the plot obtained from using ggpairs().

In this model, we are investigating the linear relationship between the covariates, raven and gender, and the response, math.

b)

The components of the random intercept model are:

- $\mathbf{Y}_i$, a response vector that has dimensions $n_i \times 1$
- $\mathbf{X}_i$, a design matrix that has dimensions $n_i \times p$
- β, a vector of coefficients that has dimensions $n_i \times 1$
- $\mathbf{1}$, a vector of ones that has dimensions $n_i \times 1$
- γ_{0i}, the random effect of school i, dimensions 1×1 .
- ϵ_i, the random error vector that has dimensions $n_i \times 1$

We assume $\gamma_{0i} \sim N(0, \tau_0^2)$ and $\epsilon_i \sim N(\mathbf{0}, \sigma^2\mathbf{I})$

We also assume that the responses at different schools are independent of each other.

```
# Fitting a linear mixed model
library(lme4)
fitRI1 <- lmer(math ~ raven + gender + (1 | school), data = dataset)
summary(fitRI1)
```

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

```
## Formula: math ~ raven + gender + (1 | school)
##    Data: dataset
##
## REML criterion at convergence: 6772.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.4607 -0.4305 -0.0127  0.4083  4.2761
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  school   (Intercept)  3.879   1.969
##  Residual             19.220   4.384
## Number of obs: 1154, groups:  school, 49
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept) -1.26915    0.34375  -3.692
## raven        0.21442    0.02331   9.197
## gendergirl   2.51119    0.26684   9.411
##
## Correlation of Fixed Effects:
##            (Intr) raven
## raven      -0.017
## gendergirl -0.404  0.034
```

In the summary output above, we can see that the parameter estimates for `raven` and `gender` are similar to the estimates in the summary output in a). The deviation is less than one standard deviation. This means that adding `school` as random intercept has little effect on the parameter estimates.

As in **a)**, we can see that expected math score increases with raven test results and if the gender is girl.

There are no p-values, and this is also default for the `lmer` function. The reason is that the random effects complicates finding the p-value, it is not possible to find a analytical solution with a finite number of observations, we have to approximate.

The test $H_0 : \beta_{raven} = 0$ against $H_1 : \beta_{raven} \neq 0$ is performed using the `anova` function.

```
# Fitting a linear mixed model without raven as covariate
fitRI2 <- lmer(math ~ gender + (1 | school), data = dataset)
# Comparing the two mixed models
anova(fitRI2, fitRI1)
```

```
## Data: dataset
## Models:
## fitRI2: math ~ gender + (1 | school)
## fitRI1: math ~ raven + gender + (1 | school)
##        npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## fitRI2    4 6855.0 6875.2 -3423.5   6847.0
## fitRI1    5 6775.4 6800.7 -3382.7   6765.4 81.586  1  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value given in the output above is less than $2.2 * 10^{-22}$. This is a p-value smaller than any reasonable significant level. We therefore reject H_0 and conclude with $H_1 : \beta_{raven} \neq 0$

```
confint(fitRI1)
```

```
##                   2.5 %     97.5 %
## .sig01       1.5175722  2.5074552
## .sigma       4.2037481  4.5694076
## (Intercept) -1.9453109 -0.5910132
## raven        0.1686451  0.2600533
## gendergirl   1.9884877  3.0343260
```

We see in the summary output above that a 95% confidence interval for the effect of the female gender on the math score is: $[1.99, 3.03]$.

c)

We will now consider a model where we fit `math` as the response while `raven` is a fixed effect and `school` is a random intercept. Fitting the model we get the summary seen below.

```
fitRI2 <- lmer(math ~ raven + (1 | school), data = dataset)
summary(fitRI2)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: math ~ raven + (1 | school)
##    Data: dataset
##
## REML criterion at convergence: 6856.9
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.2705 -0.4725 -0.0045  0.4603  4.4890
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  school   (Intercept)  4.002   2.001
##  Residual             20.711   4.551
## Number of obs: 1154, groups:  school, 49
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  0.03840    0.32071   0.120
## raven        0.20682    0.02418   8.554
##
## Correlation of Fixed Effects:
##       (Intr)
## raven -0.004
```

The covariance between Y_{ij} and Y_{il} is given by,

$$\text{Cov}(Y_{ij}, Y_{il}) = \begin{cases} \tau_0^2, & \text{if } \ l \neq j \\ \tau_0^2 + \sigma^2, & \text{if } \ l = j \end{cases} \quad , \tag{1}$$

while the correlation between Y_{ij} and Y_{il} is given by,

$$\text{Corr}(Y_{ij}, Y_{il}) = \begin{cases} \dfrac{\tau_0^2}{\tau_0^2 + \sigma^2}, & \text{if } \ l \neq j \\ 1, & \text{if } \ l = j \end{cases} \quad . \tag{2}$$

Looking at the summary of the fit of the random intercept model we can see that the correlation of two different responses within the same school (Y_{ij} and Y_{il}) is

$$\frac{4.002}{20.711 + 4.002} = 0.162$$

For the random intercept model we can write the mathematical formula of the predicted random effects in each group as,

$$\hat{\gamma}_{0i} = \frac{n_i\hat{\tau}_0^2}{\hat{\sigma}_0^2 + n_i\hat{\tau}_0^2}e_i = \frac{n_i\hat{\tau}_0^2}{\hat{\sigma}_0^2 + n_i\hat{\tau}_0^2}\frac{1}{n_i}\sum_{j=1}^{n_i}(y_{ij} - (\hat{\beta}_0 + \hat{\beta}_1 x_{ij})),$$

where e_i is the raw mean residual of group i.

We can interpret the formula for $\hat{\gamma}_{0i}$ as the weighted sum of the mean raw residual of the group and the conditional expectation 0. We can view the mean raw resiuals as for each group as predictions ignoring the random error, and the conditional expectation as ignoring the groups. There would be no difference between any group if $\tau_{0i} = 0$ for all i.

The value $n_i\tau_0^2$ represents the intra class covariance (ICC) and σ_0^2 represents the variance in the entire population. Thus we can see that the weight,

$$\frac{n_i\hat{\tau}_0^2}{\hat{\sigma}_0^2 + n_i\hat{\tau}_0^2},$$

describes how much information we have gained about the random intercept. If σ_0^2 is large compared to the ICC it is difficult to distinguish between random noise and the predicted random intercept and we can see that it is closer to zero. In the opposite case it is easier to distinguish and we get that the weight goes towards one. Furthermore, we can see that if n_i is big that the weight approaches one, which does make sense as we obtain more information about the intercept.

```
library(ggplot2)
library(sjPlot)
gg1 <- plot_model(fitRI2, type = "re", sort.est = "(Intercept)",
    y.offset = 0.4, dot.size = 1.5) + theme(axis.text.y = element_blank(),
    axis.ticks.y = element_blank()) + labs(title = "Random intercept (RI)")
gg2 <- plot_model(fitRI2, type = "diag", prnt.plot = FALSE, title = "Quantile plot",
    geom.size = 1)
gg3 <- ggplot() + geom_density(aes(x = ranef(fitRI2)$school[[1]])) +
    labs(x = "x", y = "y", title = "Density of RI")
df <- data.frame(fitted = fitted(fitRI2), resid = residuals(fitRI2,
    scaled = TRUE))
gg4 <- ggplot(df, aes(fitted, resid)) + geom_point(pch = 21) +
    geom_hline(yintercept = 0, linetype = "dashed") + geom_smooth(se = FALSE,
    col = "red", size = 0.5, method = "loess") + labs(x = "Fitted values",
    y = "Residuals", title = "Residuals vs Fitted values")
gg5 <- ggplot(df, aes(sample = resid)) + stat_qq(pch = 19) +
    geom_abline(intercept = 0, slope = 1, linetype = "dotted") +
    labs(x = "Theoretical quantiles", y = "Standardized residuals",
        title = "Normal Q-Q")

library(ggpubr)
gg2[[2]]$school
```

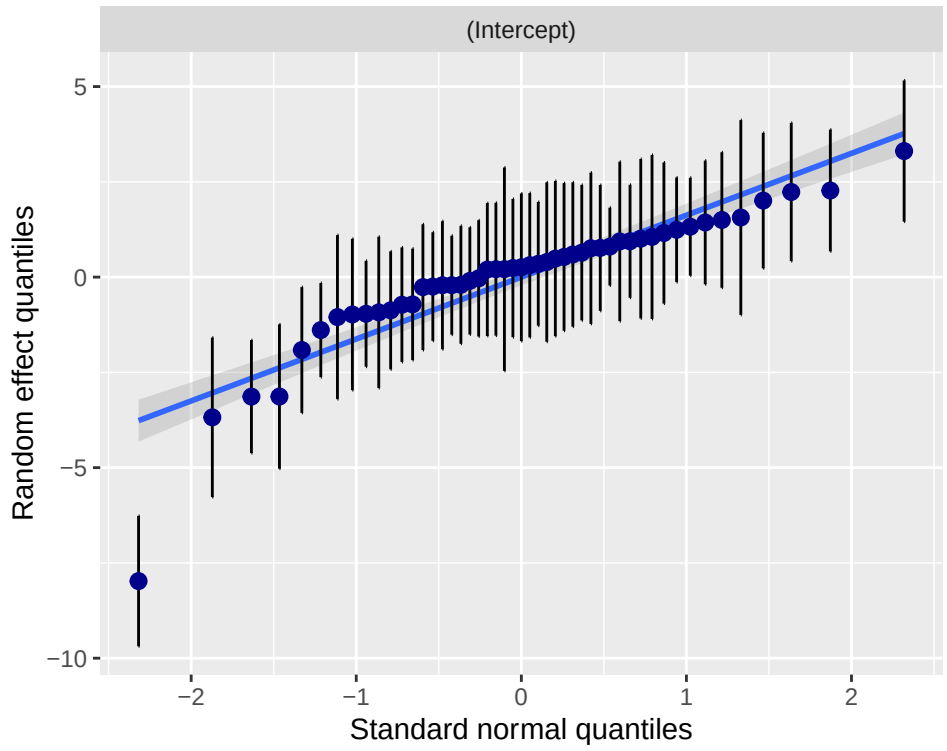


Figure 1:

This first plot, Figure 1 is a Q-Q plot of the random intercept effect. We can also see that there is a confidence interval drawn around each. Since it is assumed that the random effects is normally distributed it makes sense to plot them in a Q-Q plot to see if the assumption of normality is violated. We can see in this Q-Q plot that the assumption of normality of the random effect seems to hold up quite well. I would like to point out that we have one outlier that has a value of approximately -7.5.

```
ggarrange(gg1, gg3, gg4, gg5)
```

Now we will look at the four different plots in figure 2.

In the "Random intercept (RI)" plot we have the predicted random intercepts plotted on the x-axis and sorted from lowest to highest. It can be wived as the empirical CDF of the random intercept. We would expect that about half would be grater than zero and this is the case of our CDF plot. Again we see the outlier at -7.5.

The plot with titel "Density of RI" is a histogram of the random effects that has been smoothed out by the `geom_density` function to make it represent the probability density function of the random intercept. We can see that it is approximately centered around zero, which we would expect if the model assumptions are not violated. However, compared to the density of the Gaussian, it seems to have a lot of mass centered around zero making it seem pointy. This is however something that can happen when we only have 49 samples.

The next plot with the title "Residuals vs Fitted values" is simply the residuals plotted against the predicted math scores. Here as in a linear model (not a LMM) we expect the residuals to have a constant variance (homoscedasticity). This plot seems to indicate that the assumption is not violated. Note that we have some outliers around 5 and one at -10.

This last plot with the titel "Normal Q-Q" in the group of four plots, is the standardized residuals plotted in a Q-Q plot. This plot shows if the assumption of normality of the residuals is violated. The Q-Q plot for the residuals of our model seem to hold the normality assumption quite well.

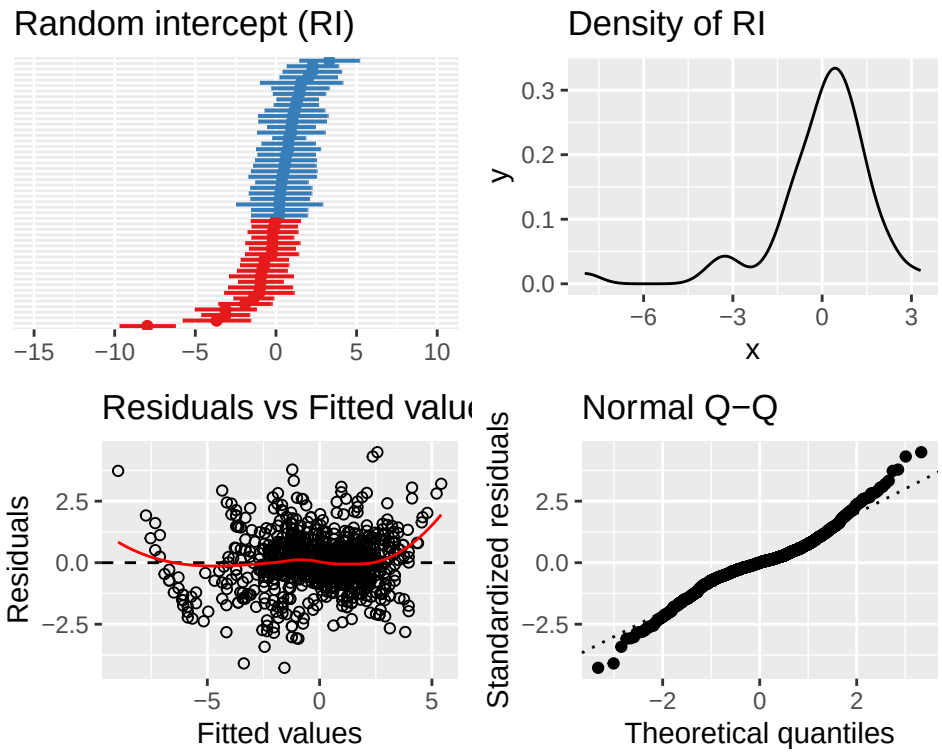


Figure 2:

o

```
df <- data.frame(x = rep(range(dataset$raven), each = 49), y = coef(fitRI2)$school[,
    1] + coef(fitRI2)$school[, 2] * rep(range(dataset$raven),
    each = 49), School = factor(rep(c(1:42, 44:50), times = 2)))
ggplot(df, aes(x = x, y = y, col = School)) + geom_line()
```

In this last plot, figure 3 we have plotted the response against the models only covariate **raven** for each school. Since each school has its own intercept and we have 49 schools we get 49 parallel lines in the random intercept model. Here we can clearly see one outlier as we did in the plot "Random intercept (RI)". However, this plot is more to visualize the relationship between **raven** and **math** scores in the different schools when fitting a linear random intercept model.

d)

Model with social status of father as fixed effect and ANOVA table of models with and without social status of father as fixed effect.

```
fitRI2 <- lmer(math ~ raven + (1 | school), data = dataset)
fitRI3 <- lmer(math ~ raven + social + (1 | school), data = dataset)
anova(fitRI2, fitRI3)
```

```
## Data: dataset
## Models:
## fitRI2: math ~ raven + (1 | school)
## fitRI3: math ~ raven + social + (1 | school)
##        npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## fitRI2    4 6858.9 6879.1 -3425.4   6850.9
## fitRI3    7 6856.8 6892.1 -3421.4   6842.8 8.1175  3    0.04364 *
```

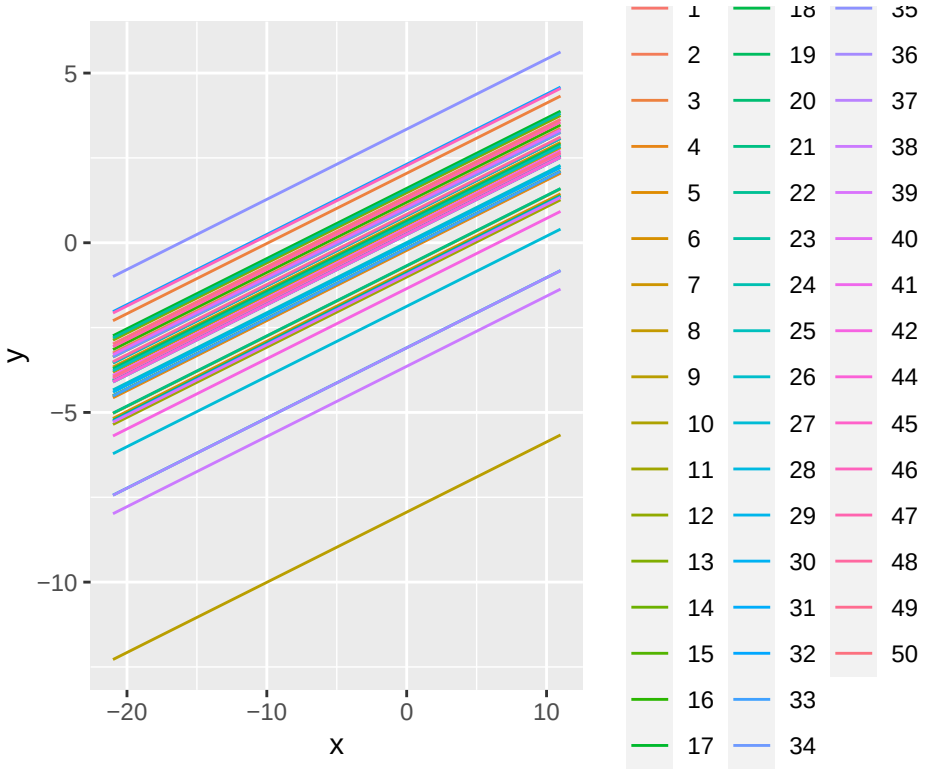


Figure 3:

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
fitRI2t <- lmer(math ~ raven + (1 | school), data = dataset,
    REML = FALSE)
fitRI3t <- lmer(math ~ raven + social + (1 | school), data = dataset,
    REML = FALSE)
anova(fitRI2t, fitRI3t)
```

```
## Data: dataset
## Models:
## fitRI2t: math ~ raven + (1 | school)
## fitRI3t: math ~ raven + social + (1 | school)
##         npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## fitRI2t    4 6858.9 6879.1 -3425.4   6850.9
## fitRI3t    7 6856.8 6892.1 -3421.4   6842.8 8.1175  3    0.04364 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- *Compare the model with and without the social status of the father using hypothesis test from the anova above. Which of the two models do you prefer?*

From the anova table above we see that the p-value is 0.043. This is significant at significance level 0.05. We would thus prefer the larger model including `social` as a fixed effect.

- *Comment on the AIC and BIC*

By looking at the ANOVA table we see that the AIC is lower and the BIC is higher for the larger model than for the smaller. As the AIC and BIC penalizes the number of parameters differently, and BIC scales the

penalty by a factor depending on sample size. We know that selecting model based on AIC gives gives a possibility of overfitting, and by BIC a possibility of underfitting.

- *Why does the print out say "refitting model(s) with ML (instead of REML)".*
  - *why do we not want REML when comparing models with the same random terms but with different fixed terms?*

The print out does not say "refitting model(s) with ML (instead of REML)". The default for `lmer()` is REML, but as we see above the ANOVA tables when fitted with REML and ML are the same. This is likely because ANOVA refits the models with ML, but a sortware update has removed the info from the print out.

The reason for not using REML when comparing models with the same random terms is that the REML estimates the random effects by considering linear combinations of the data that remove the fixed effects (the transformation method). If the fixed effects are changed, the likelihoods of the two models will not be directly comparable.

Random intercept and random slope model.

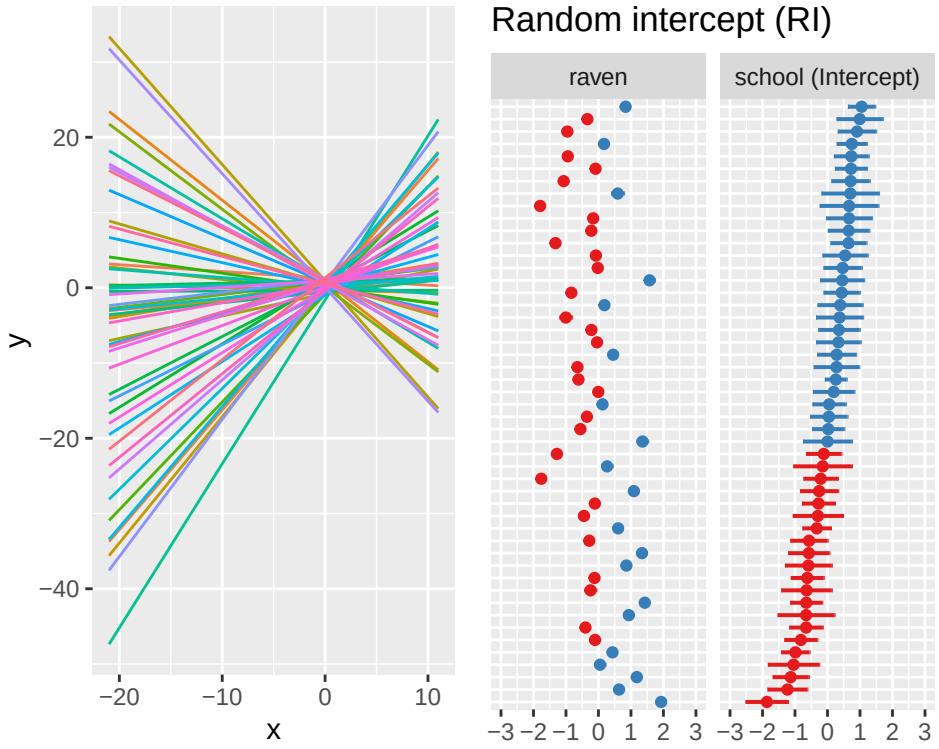
```
fitRIS <- lmer(math ~ raven + (1 + raven | school), data = dataset)
summary(fitRIS)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: math ~ raven + (1 + raven | school)
##    Data: dataset
##
## REML criterion at convergence: 4537.6
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.87463 -0.66206 -0.03913  0.65818  3.09716
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  school   (Intercept) 0.5519   0.7429
##           raven       0.7293   0.8540   -0.40
##  Residual             2.2094   1.4864
## Number of obs: 1154, groups:  school, 49
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   0.2603     0.1183   2.200
## raven         0.2498     0.1223   2.042
##
## Correlation of Fixed Effects:
##       (Intr)
## raven -0.356
```

```
df <- data.frame(x = rep(range(dataset$raven), each = 49), y = coef(fitRIS)$school[,
    1] + coef(fitRIS)$school[, 2] * rep(range(dataset$raven),
    each = 49), School = factor(rep(c(1:42, 44:50), times = 2)))
gg1 <- ggplot(df, aes(x = x, y = y, col = School)) + geom_line()

gg2 <- plot_model(fitRIS, type = "re", sort.est = "(Intercept)",
    y.offset = 0.4, dot.size = 1.5) + theme(axis.text.y = element_blank(),
    axis.ticks.y = element_blank()) + labs(title = "Random intercept (RI)")

ggarrange(gg1, gg2, ncol = 2, legend = FALSE)
```



Write the mathematical formula for the random intercept and slope model and comment on what you see from fitting the model.

For each school $i = 1, ..., 50$ but not 43:

$$Y_i = X_i\beta + U_i\gamma_i + \varepsilon_i$$

The global model is

$$Y = X\beta + U\gamma + \varepsilon$$

Where U is block diagonal and Y, X, γ and ε are just all school specific versions concatenated.

e)

Now imagine that we want to model the probability for a student to fail maths instead of the the individual grades given in maths. A student fails if they score less than -10.

- *Why is it not suitable to use a linear mixed effects model?*

The range of the predictions of an LMM can be, as for a LM, greater than the interval $[0, 1]$, making it unsuitable for modelling probabilities.

- *What type of model would be more suitable? (hint: IL module 7)*

A GLMM would be more suitable to this problem. In particular a Binomial (with $n = 1$) random intercept model.

- *How would we add a random school intercept into this model (in which part of the model)?*

We add the random intercept to the linear predictor η_{ij}, similarly to how we add the random intercept to a linear model.

- *What is the main challenge with this type of models? (hint: marginal model)*

In order to to parameter estimation we need the marginals $f(y_{ij}) = \int_{\gamma_i} f(y_{ij}|\gamma_i)f(\gamma_i)d\gamma_i$. As general exponential family is not closed under conditioning, the conditionals $f(y_{ij}|\gamma_i)$ are often times intractible. In order to do parameter estimation we must use numerics and computer intensive methods.

- *Fit the equivalent model to fitRI1, and comment on the differences in the model inferences.*

First transform the `math` covariate to binary pass = 1 and fail = 0, then fit a binomial random intercept model.

```
pass_fail <- function(x) {
    if (x < -10) {
        return(0)
    } else {
        return(1)
    }
}
dataset$math <- lapply(dataset$math, pass_fail)
dataset$math <- as.integer(dataset$math)
dataset$math <- factor(dataset$math)


fitRI1GLMM <- glmer(math ~ raven + gender + (1 | school), data = dataset,
    family = binomial(link = "logit"))
```

```
summary(fitRI1GLMM)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: math ~ raven + gender + (1 | school)
##    Data: dataset
##
##      AIC      BIC   logLik deviance df.resid
##    268.3    288.5   -130.2    260.3     1150
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -7.8247  0.0072  0.0139  0.0493  3.5581
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  school (Intercept) 28.22    5.312
## Number of obs: 1154, groups:  school, 49
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   8.6331     1.8553   4.653 3.27e-06 ***
## raven         0.2409     0.0444   5.427 5.74e-08 ***
## gendergirl    0.2374     0.3885   0.611    0.541
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##            (Intr) raven
## raven       0.405
## gendergirl -0.130  0.021
```