Project 3: Generalized Linear Models

Group 17: Adrian Bruland, Mathias Opland

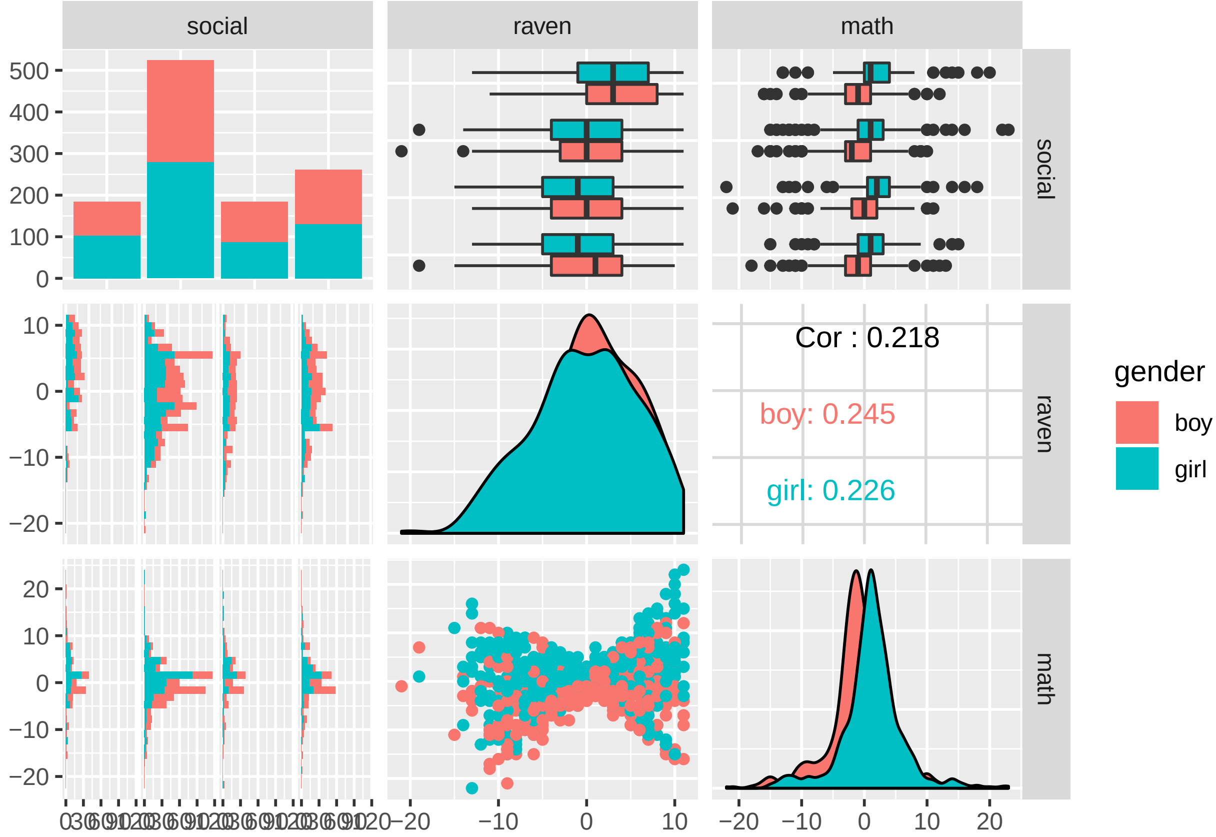
23.11.2018

Contents

a)

```
## Loading required package: ggplot2
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



- The plot in the bottom right corner shows that both the math score for boys and girls are approximately normally distributed, but that girls have a mean larger than zero and that boys have a mean smaller than zero. From this we can conclude that on a general basis, girls get better math score than boys.

In the the middle bottom window, math is plotted against raven. It seems that students with a high (10) or low (-10) value on the raven test score have a higher variance on the math variable than students with a middle (0) value for raven. The data set clusters together near raven value 0, giving a substantial correlation of 0.218. The correlation is even higher if we look only at boys og girls.

Judging by the univariate distributions, the subset of students that lie on the interval (-5,5) on the math variable makes up about half of the students, so it has a large impact on the math-raven correlation. On that interval, there seems to be a small, positive slope for math, which fits well with the estimated slope of .19.

When breaking down math across the 4 social classes, girls have a higher average math score than the boys in every class. Class number 2 seems to be the lowest scoring class on math, by a few points, while class number 3 looks to be the highest scoring class.

```
linear_model = lm(math ~ raven + gender, data=dataset)
summary(linear_model)
```

```
##
## 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
```

Formula:

$$Y_k = \mathbf{x_k}\beta + \epsilon_\mathbf{k}$$

- Y_k is the response, and tells us the expected math score for student k, given the covariates. The $\mathbf{x_k}$ is the covariate vector, and contains the given values for the student, as well as a 1 which is multiplied by the intercept estimate. In this case it is a vector with the raven test score and the gender, where girl gives 1 and boy zero. β is then the coefficients for the covariates, and explaines how much the will affect the response. In this case there is on coefficient for the raven test, which is multiplied by the test score and one coefficient for the gender which is multiplied by either 0 or 1, and the intercept score. ϵ is the variance in the math score (response) not explained by the model, and is centered around 0, with variance σ^2.

- The estimate of the intercept at -1.3131 says that given a student is a boy and has score zero at the raven test, the expected math score is -1.3131. Then the other covariates will affect the score either positively or negatively given a students values. The raven estimate says that every point will count positive on the math score with 0.1965 points. It also shows that girls will in average score 2.5381 better than boys, and with a variance of 0.2807. That means that there is a distinct difference between boys and girls in how well they do on the test, and that girls do a lot better. All the coefficients, as well as the regression is significant at 95% confidence interval, which makes us conclude that there is a correlation between the covariates and the response, as the model describes.

- This model shows the correlation between the score on the raven test and the gender, with the math score. It can be used to conclude if and how there is a general difference in how well boys and girls do math, and how a students raven test score will affect their math score (not the actual raven test score, but the knowledge and skills needed for achieving the given test score).

b)

$$\mathbf{Y_i} = \mathbf{X_i}\beta + \mathbf{1}\gamma_{\mathbf{0i}} + \epsilon_{\mathbf{k}}$$

- Here $\mathbf{Y_i}$ is the response for school i, and will be of dimension n_ix1, where n_i is number of student at school i. $\mathbf{X_i}$ is the covariate matrix, and contains the covariates of all the students in school i. The dimension of $\mathbf{X_i}$ is n_ixp, where p is the number of covariates (including intercept). β is the coefficients, and of dimension $px1$. $\mathbf{1}$ is of dimension n_ix1, and is a vector of ones. γ_{0i} is a scalar which describes the affect a given school has on the repsonse of the students at that school, i.e. the school intercept. Thus $\mathbf{1}\gamma\mathbf{i0}$ will be a n_ix1 matrix with all entries equal to γ_{0i}. ϵ_i will as before be the error, but as we have taken into account which school a student attends, the variance in the data not described by the model may be lower. ϵ_i is of dimension n_ix1.
- γ_{0i} is assumed to be distributed as γ_{0i} $N(0,\tau^2)$, and all the elements of ϵ_i is assumed to be i.i.d. $N(0,\sigma^2)$.
- The responses on school i, $\mathbf{Y_i}$ and school k, $\mathbf{Y_k}$ are independent, and will be equaly distributed with only a difference in the intercept of respectivly $\beta_0 + \gamma_{0i}$ and $\beta_0 + \gamma_{0k}$.

```r
library(lme4)
```

```
## Loading required package: Matrix
```

```r
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
```

- The parameter estimates of raven and gender are pretty equal to the linear model, as which school a student attends, even though there are differences between various schools, will in this case not affect the relationship between the covariates and the response. As the intercept could be looked at as a "mean" for what boys scores in math when their raven test score is zero, and since γ_0 is normal around zero we would not expect change in the intercept. As well the fixed effect does only affect the intercept and not the covariate with the raven score test (since it's not a random slope), and therefor can not affect the how the covariate raven affects the response.

- As with the linear model each points in the raven test score will affect the math score with 0.214 points, and girls will get 2.511 more points in general.
- β is only asymptotically normal with an unknown degrees of freedome, and we can only approximate the t-distribution for the parameters and the model. Therefor we assume normality, and find the p-value and confidence interval.

```
2*pnorm(-abs(summary(fitRI1)$coef[2,3]))
```

```
## [1] 3.68749e-20
```

```
p_value = c(summary(fitRI1)$coef[3,1] - 1.96 * summary(fitRI1)$coef[3,2], summary(fitRI1)$coef[3,1] + 1
p_value
```

```
## [1] 1.988179 3.034194
```

c)

- On our model, the covariance and correlation of the math score for two students attending the same school is

$$\mathrm{Cov}(Y_{ij}, Y_{il}) = \tau_0^2 \quad , \quad \mathrm{Corr} = \frac{\tau_0^2}{\tau_0^2 + \sigma^2}$$

```
library(lme4)
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
```

```
#summary(fitRI2)$coef
```

The correlation for model fitRI2 is 4.002/(4.002+20.711)=0.162.

- The random intercept parameter

  $$\gamma_{0i}$$

  is predicted by

  $$\gamma_{0i} = \frac{n_i \hat{\tau}_0^2}{\hat{\sigma}^2 + n_i \hat{\tau}_0^2} e_i = \frac{n_i \hat{\tau}_0^2}{\hat{\sigma}^2 + n_i \hat{\tau}_0^2} \cdot \frac{1}{n_i} \sum_{j=1}^{n_i} (Y_{ij} - x_{ij}^T \hat{\beta})$$

  .

Here, n_i is the number of residuals (i.e. number of datapoints) in cluster i. The variances $\hat{\sigma}^2, \hat{\tau}_0^2$ are for random error $\varepsilon_(ij)$ and random intercept γ_{0i}. The factor e_i is the average of the random errors ε_{ij} for all datapoints j in cluster i.

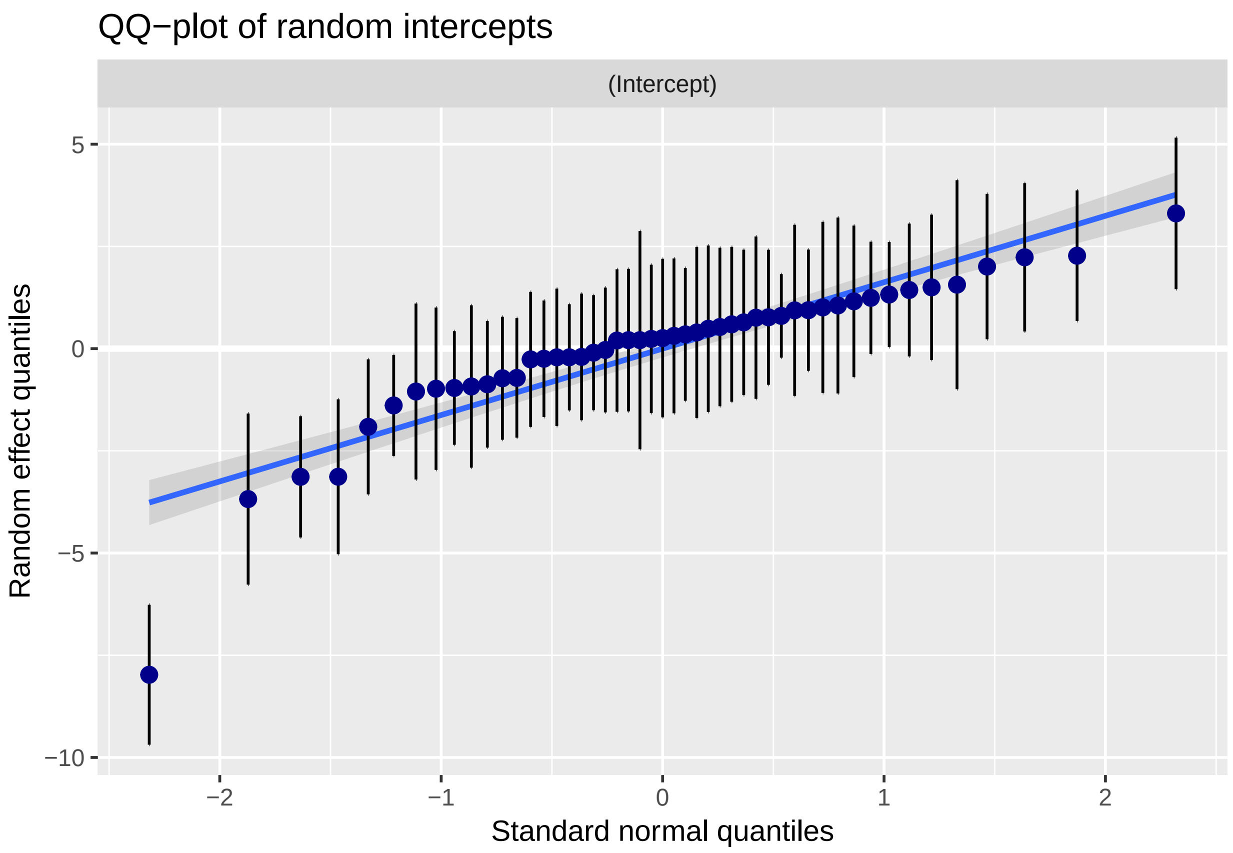
```
library(ggplot2)
library(sjPlot)
```

```
## Warning in checkMatrixPackageVersion(): Package version inconsistency detected.
## TMB was built with Matrix version 1.2.15
## Current Matrix version is 1.2.14
## Please re-install 'TMB' from source using install.packages('TMB', type = 'source') or ask CRAN for a
```

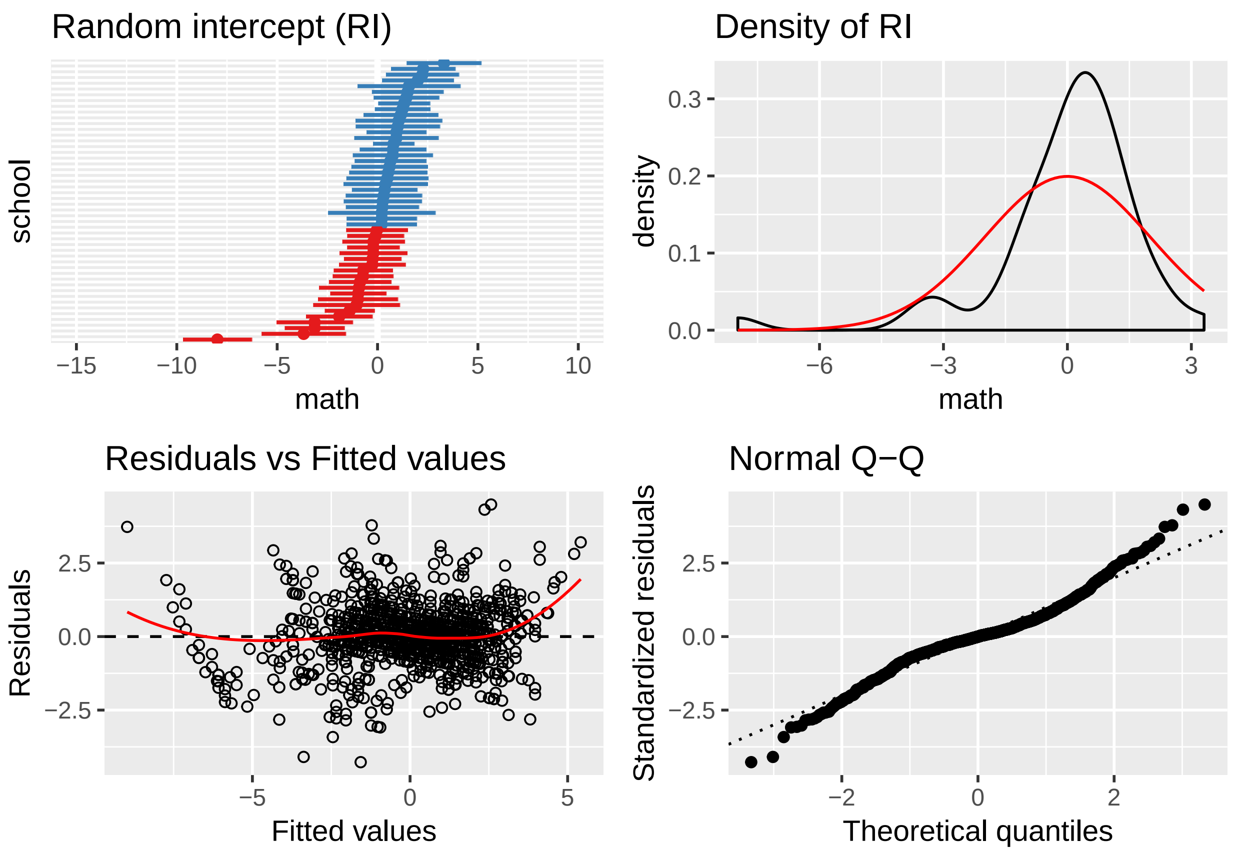
```
library(ggpubr)
```

```
## Loading required package: magrittr
```

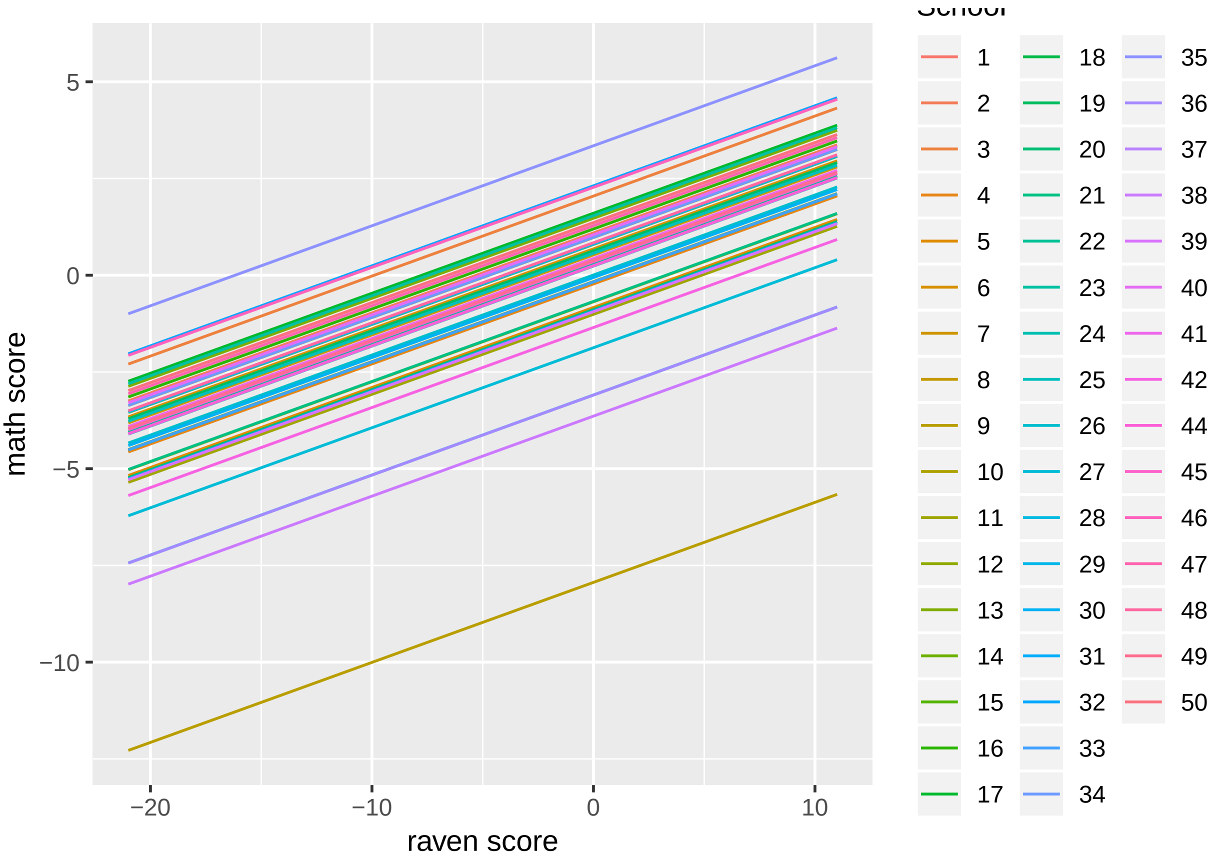
```
library(lme4)
gg1 <- plot_model(fitRI2, type = "diag", prnt.plot = FALSE, geom.size = 1)
gg2 <- plot_model(fitRI2, type = "re", sort.est = "(Intercept)", y.offset = 0.4, dot.size = 1.5) +
  theme(axis.text.y = element_blank(), axis.ticks.y = element_blank()) +
  labs(title = "Random intercept (RI)", x = "school", y = "math")
gg3 <- ggplot(data = data.frame(x = lme4::ranef(fitRI2)$school[[1]]), aes(x = x)) + geom_density() +
  labs(x = "math", y = "density", title = "Density of RI") +
  stat_function(fun = dnorm, args = list(mean = 0, sd = attr(VarCorr(fitRI2)$school, "stddev")), col =
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")

gg1[[2]]$school + ggtitle("QQ-plot of random intercepts")
```



```
ggarrange(gg2, gg3, gg4, gg5)
```



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



QQ-plot: compares the distribution of the quantiles in the ordered list of random intercepts γ_{0i} to those of a normal distribution. Is used for telling how closely the set of random intercept resemble a normal distribution. Here, the one outlier at the low end of the γ_{0i}'s is the only substantial outlier - removing the data point may have visibly reduced the slope of the blue curve to the point that the fit onto the normal distribution would be even closer. The distribution is approximately normal, judging by the QQ-plot.

Random Intercept plot: the plot can again be used to check how close the γ_{0i} distribution is being normal, and whether there are substantial outliers in terms of predicted value and variance. Here, the some bottom outlier is found, and the graph of the mean math value resembles the CDF of the normal distribution, which is what we are looking for to confirm normal distribution in γ_{0i}.

Density of RI: univariate distribution of γ_{0i} plotted against a normal distribution $N(0, \tau_0^2)$. Can be used for checking closeness to normal distribution, in terms of mean, variance and more generally the shape of the curve. This γ_{0i}-distribution quite closely resembles the normal distribution. Removing the bottom outlier would reduce τ_0^2, making the black curve taller and thinner, which would improve the fit.

Residuals vs fitted values: used to check for biases (patterns in the dataset) in relation to the fitted regression, which here is a linear regression. Can also be used to find outliers. The red trend line shows some deviation from normalcy at the edges, but overall it's close fit to the normal distribution.

Normal QQ-plot: a standardised version of the first plot. Benefits: mean exactly at zero, variance approximately one, and the normal distribution would fall on the line y=x. Has the same uses as QQ-plot.

df-plot: here the differences in the random intercepts is clearly seen. Can be used again to check for outliers, visualise the overall distribution, as well as see how big an impact the random intercepts have on the math values. Our model predicts that attending the poorest scoring school means it's nigh impossible to achieve the lowest math scores of the highest scoring school. While there are a few outliers, most schools are normally distributed around the same linear math score curve.

d)

```
library(lme4)
fitRI3 <- lmer(math ~ raven + social + (1 | school), data = dataset)
anova(fitRI2, fitRI3)
```

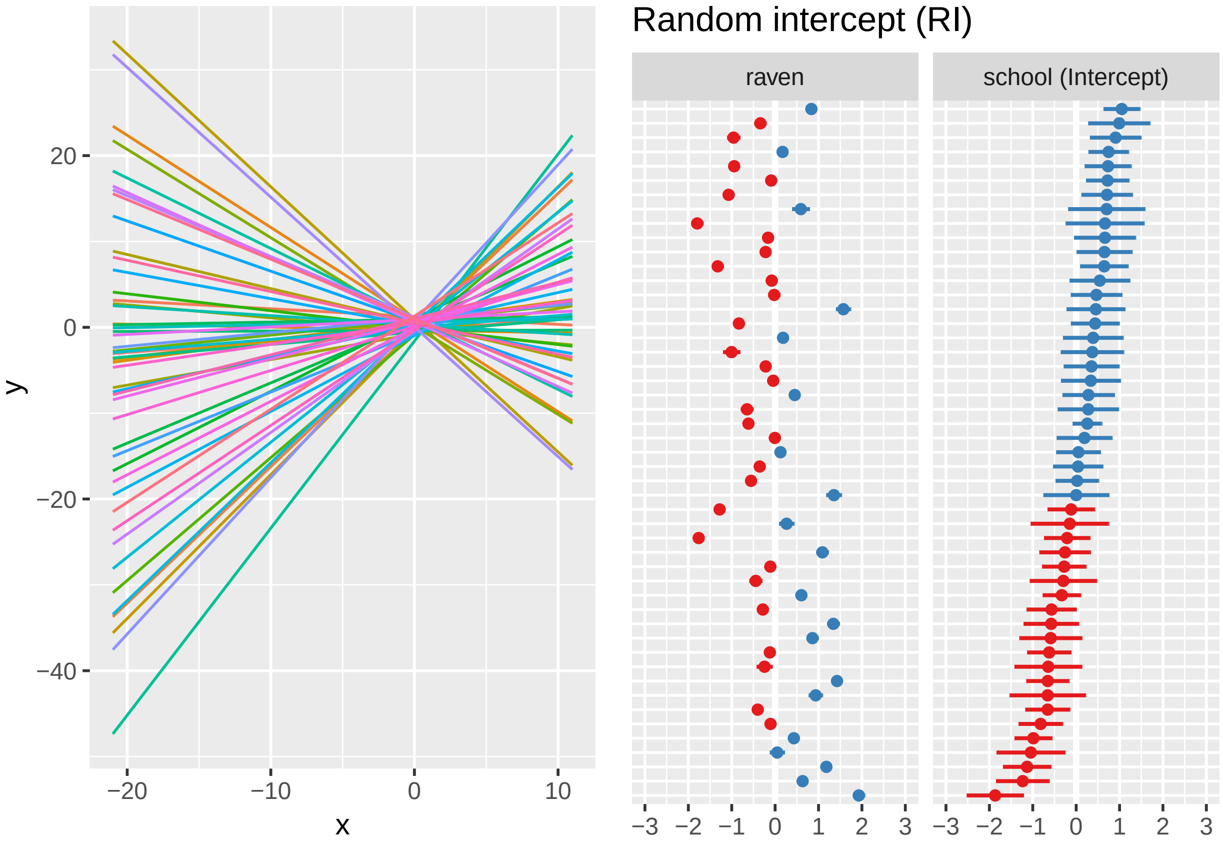
```
## refitting model(s) with ML (instead of REML)

## Data: dataset
## Models:
## fitRI2: math ~ raven + (1 | school)
## fitRI3: math ~ raven + social + (1 | school)
##        Df    AIC    BIC  logLik deviance  Chisq Chi 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 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Looking at the χ^2-value and the p-value for the alternative model (fitRI3), we can abolish the H_0-hypothese (fitRI2 is the best model) with significance level 0.05. However the p-value is close to the significance level, so it can be usefull to use the AIC and BIC to conclude if we want to abolish the H_0-hypothese.
- REML should only be used when the random effects change between the models. This is due to the fact that the REML likelihood depends on which fixed effects are in the model, and thus will change between the models. Therefore ML is prefered in this case.
- We se that the AIC is a little bit lower for fitRI3, which means that there is less information loss in that model. The BIC on the other hand is much lower for the fitRI2 model. However together with the p-value we can conlude that fitRI3 is the best model, because 2 out of 3 test support this model as the best.

```
library(GGally)
library(sjPlot)
library(lme4)
library(ggpubr)
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
## -2.87462 -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 (
ggarrange(gg1, gg2, ncol = 2, legend = FALSE)
```



$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + \gamma_{0i} + \gamma_{1i} x_{ij} + \epsilon_{ij} \epsilon_{\mathbf{i}} \sim \mathbf{N}(\mathbf{0}, \sigma^{\mathbf{2}}\mathbf{I}), \qquad \gamma_{\mathbf{i}} = \begin{pmatrix} \gamma_{0i} \\ \gamma_{1i} \end{pmatrix} \sim \mathbf{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \mathbf{Q} = \begin{pmatrix} \tau_0^2 & \tau_{01} \\ \tau_{01} & \tau_1^2 \end{pmatrix} \right).$$

- From the plots we can see that schools with high school intercept, γ_{0i}, has a lower γ_{1i} for the raven test. Thus schools with higher intercept will generaly have negativ steep slopes, where as schools with intercept close to zero will have less steep or close to zero slope, and schools with low intercept will have steep slopes.