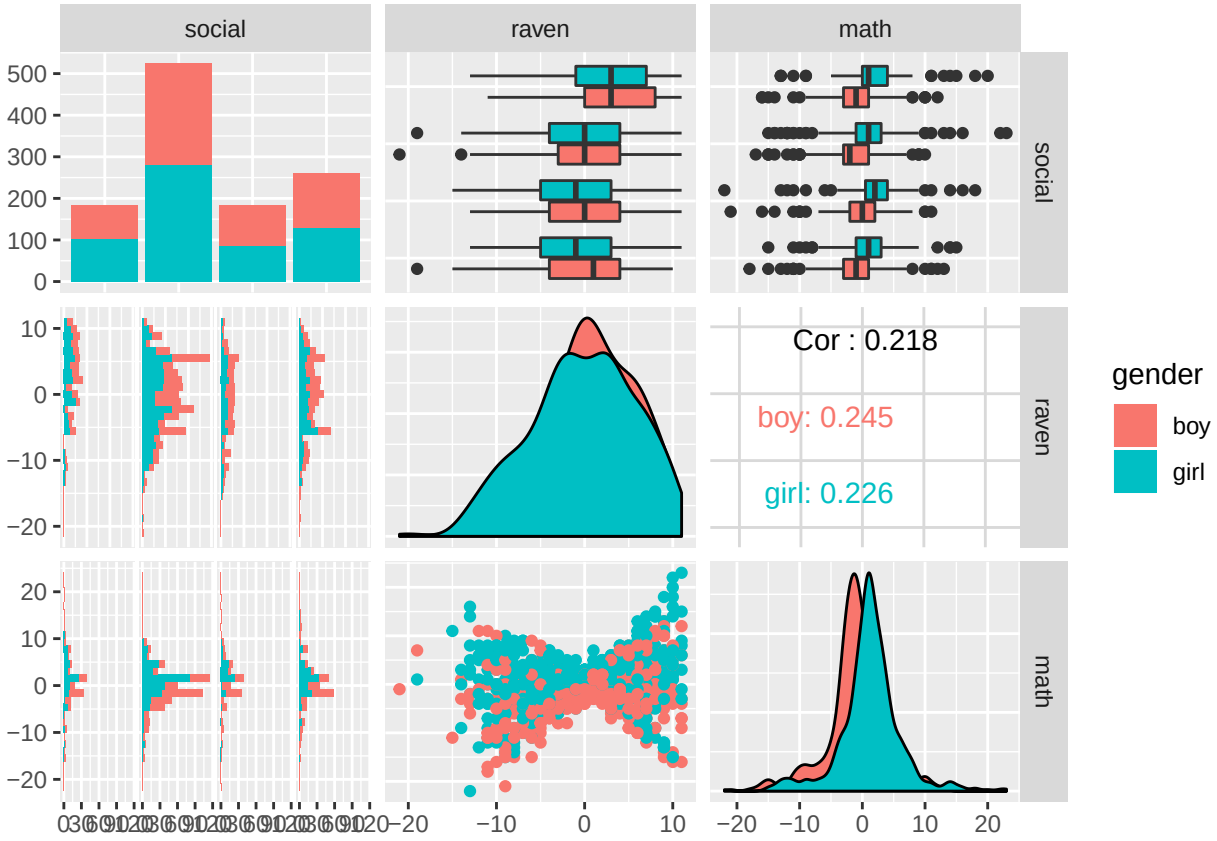
glm3

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.

```r
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:

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

* The 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 $n_i x 1$, 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 $n_i x p$, where p is the number of covariates (including intercept). β is the coefficients, and of dimension $px1$. $\mathbf{1}$ is of dimension $n_i x 1$, 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 $\mathbf{1\gamma i0}$ will be a $n_i x1$ 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 $n_i x1$.

- 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 $\beta_0 + \gamma_{0k}$.

```
library(lme4)
```

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

```
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 not affect the relationship between the covariates and the response. The estimate only describes the dependency in the model between the covariates and the response, and if there is no correlation between the covariates and which school a studen attends to (i.e. relationship between how much a certain covariate affects the respons and the school he/she attends), there will be no change. Thus there is a correlation between the intercept for the linear model and which school a student attends, so the intercept will be changed. The error may also have a lower variance as we explain more of the variance in our model.
- 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.
- The model is only asymptotically, and we can only approximate the t-distribution for the parameters and the model.

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

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

```r
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

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

```r
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
```

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

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