TMA4315 Generalized Linear Models

Compulsory exercise 3: (Generalized) Linear Mixed Models

Deadline: Friday, November 23rd, 2018 at 16.00

To be handed in on Blackboard.

Students may work on the assignment alone, or in groups of two or three, and are also encouraged to collaborate across groups. You do not have to use the same group as for the first/second exercise. Each group needs to hand in an [R Markdown] (http://rmarkdown.rstudio.com/) document with answers to the questions as well as all source code. You find a template here: https://www.math.ntnu.no/emner/TMA4315/2018h/template_glm.Rmd. Hand in the file as a pdf file (`.pdf`) only, no zipped folders necessary.

You shall deliver a document that answers all questions, using print-outs from `R` when asked for/when it seems necessary. It should include calculations, theory and fomulas you needed to answer the questions, and enough information so the teaching assistant can see that you have understood the theory (but do not write too much either, and exclude code not directly necessary for the answers). You will often be asked to interpret the parameters. This does not mean that you should just say what the numeric values are or state how you interpret the parameter in general, but rather explain what these values mean for the model in question.

As for the previous two compulsory exercises this exercise should be handed in on Blackboard and *please write the names of the group members and the group number you have on Blackboard on top of your document!*

Guidance by the teaching assistant (Ingeborg) will be given Tuesdays and Thursdays from 10:15-12:00 in room 922 at the 9th floor of Sentralbygg 2. She will be there from the beginning, but leaves if no one shows up. If you know you will be arriving very late, send her an email (ingeborg.hem@ntnu.no) to ensure that she will be in 922. *Note: No exercise hours Tuesday, November 20th!*

In this exercise, we will focus on (generalized) linear mixed models. Each problem (there are 5 problems) has a maximum score of 2 points. Please provide *short* answers (preferably one-two sentences) to all questions. We have included bullet points in each problem to make it easier to see what we ask about, but the points are not necessarily weighted equally.

We will use a SIMULATED (thus, not a real dataset) with clustered data. The R-package `faraway` contains a dataset called `jsp` is used as an example, and we have fitted a model to this dataset and then used the result to generate a new dataset. We do this because the original dataset does not give interesting results (you can ask Ingeborg if you are interested in how the new dataset is made). Assume we have the following variables:

- `school`: 50 schools, with code 1-50
- `gender`: a factor with levels boy, girl
- `social`: social class of the father, categorical (original class 1-2 = S1, 3-4 = S2, 5-6 = S3 and 7-9 = S4)
- `raven`: test score (centered around 0)
- `math`: math score (centered around 0)

We will use `math` as response, and group the data by school. The following commands load the dataset:

```
dataset <- read.table("https://www.math.ntnu.no/emner/TMA4315/2018h/jsp2.txt",
    header = TRUE)
# dataset <-
# read.table('/Users/ingebogh/Documents/TMA4315/GLM_h18/exercise3/jsp2.txt',
# header = TRUE)
```

Note that the number of schools in the subset of the dataset is 49. We only use a subset of the full dataset, which consists of 50 schools, but school 43 has no measurements in the subset we use.

a)

First we want to get to know the dataset. Use the `ggpairs` function from `GGally` on the data, but you should group the data using `gender` (`mapping = aes(col = gender)`), and you should only include the covariates `social`, `raven` and `math` (`column = c("social", "raven", "math")`). The legend can be added using the argument `legend = 1`. Note that the histograms are stacked; the histogram for one gender is made, and then the histogram for the other is put on top of each bar. The code required for this plot is:

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

- Comment briefly on the plot you have created.

First fit a *linear model* with `math` as response, and `raven` and `gender` as covariates. Model for the kth student:

$$Y_k = \mathbf{X}_k\beta + \varepsilon_k$$

where we assume that the ε_ks are independent (between students), and have mean 0 and variance σ^2 for all students.

- Explain what the different parts of this model are called.
- Comment briefly on the parameter estimates you have found.
- What are we investigating with this model?

b)

We think that the `math` grades might be different between schools and want to add school as a factor to our model. However, with 49 schools we don't want to add school as a fixed effect. We will instead fit a *random intercept model* with `school` as *random intercept*. For school i we study the measurement model:

$$\mathbf{Y}_i = \mathbf{X}_i\beta + \mathbf{1}\gamma_{0i} + \varepsilon_i$$

where $\mathbf{1}$, $\mathbf{Y}_i$ and ε_i have dimension $n_i \times 1$.

- Explain what the different parts of this model are called and what dimensions the model components have.
- Write down distributional assumptions for γ_{0i} and ε_i.
- What do we assume about the dependency between the responses at school i, $\mathbf{Y}_i$, and school k, $\mathbf{Y}_k$?

Fit the model using `lmer` with the following code:

```
# install.packages('lme4') # install the package
library(lme4)
fitRI1 <- lmer(math ~ raven + gender + (1 | school), data = dataset)
summary(fitRI1)
```

- Compare the parameter estimates for `raven` and `gender` with the estimates from the linear model in a), and discuss.
- How do gender and raven score affect the math scores?
- In the print-out from `summary(fitRI1)` there are no p-values. Why is this?
- Test the null-hypothesis $H_0 : \beta_{\text{raven}} = 0$ against $H_0 : \beta_{\text{raven}} \neq 0$ and provide a p-value for the test. (Yes, we have many observations and believe that we can calculate a p-value even though the `lmer` package not by default want to report such a number.)
- Also provide a 95% confidence interval for the effect of the female gender on the math score.

c)

We now continue with a random intercept model (school) with only `raven` as fixed effect (remove `gender` from our model).

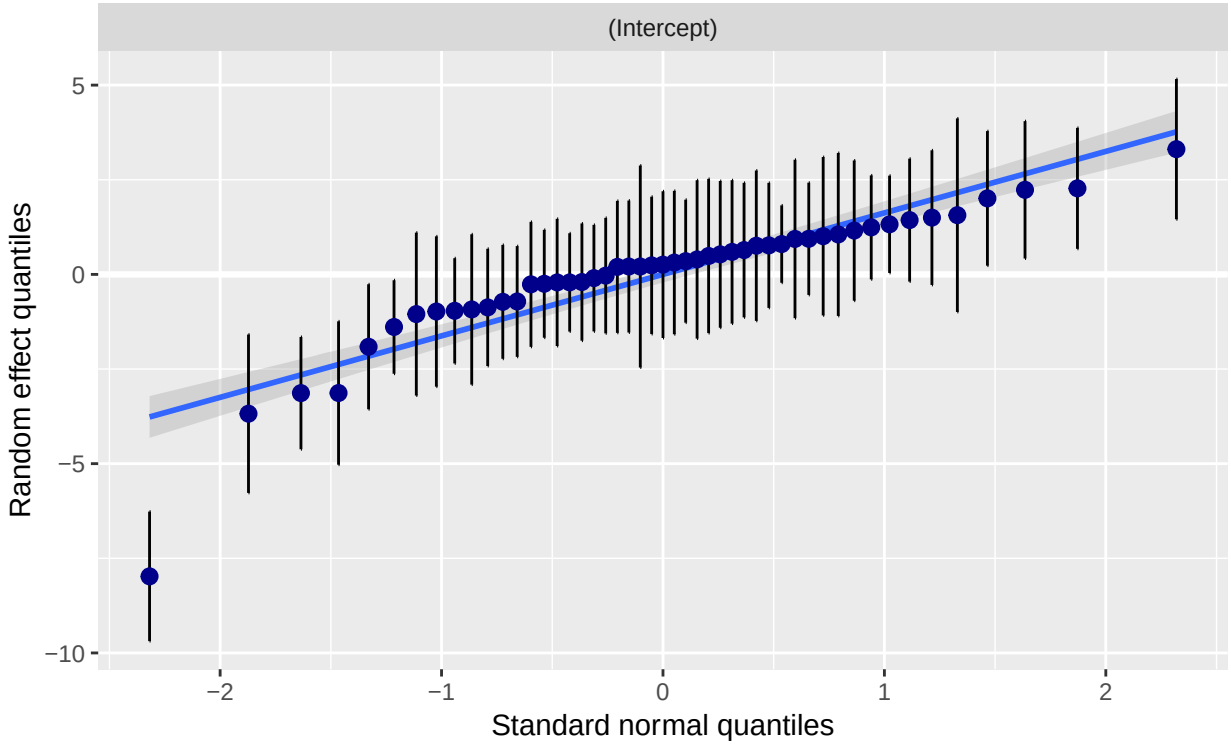
```
fitRI2 <- lmer(math ~ raven + (1 | school), data = dataset)
```

- Write down the mathematical formula for the covariance and correlation between response Y_{ij} and Y_{il} from school i.
- What is this correlation for our fitted model `fitRI2`? Comment.

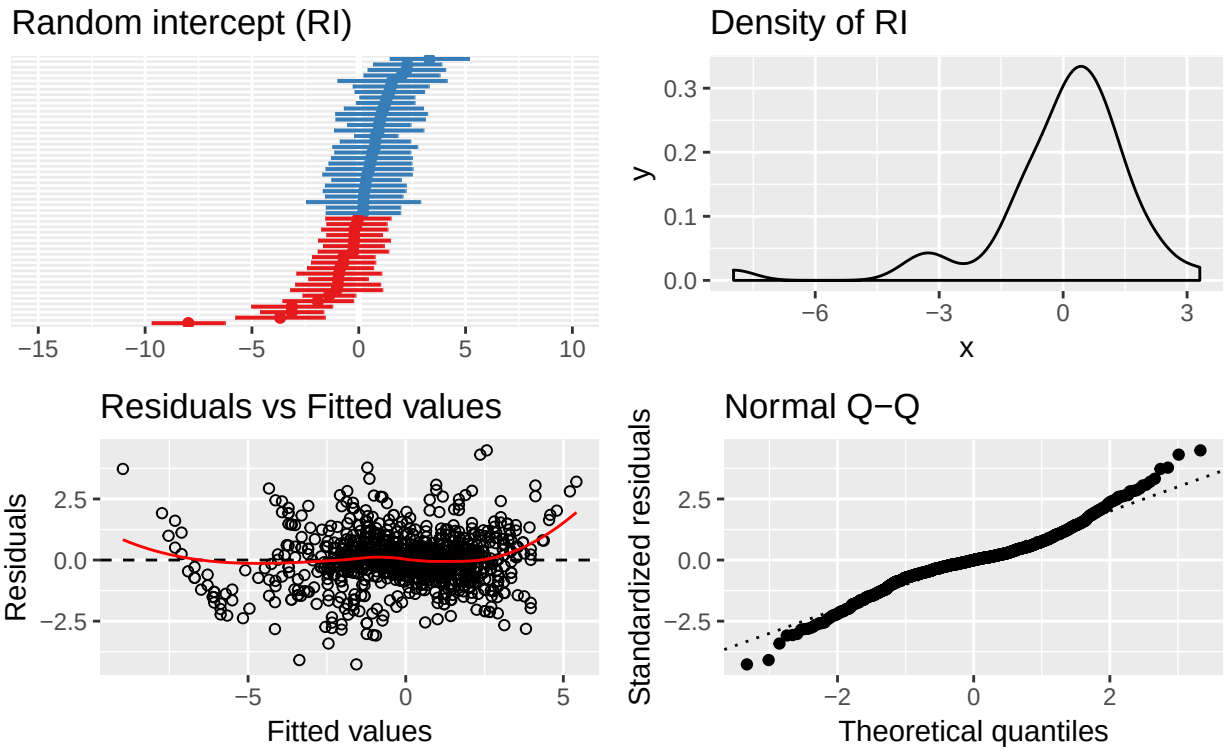
We want to study the predicted value for the random intercept for each school.

- Write down the mathematical formula for $\hat{\gamma}_{0i}$ for your random intercept model and explain what the different elements in the formula means.
- Explain what each of the six plots produced and displayed below can be used for (that is, why are we asking you to make these plots).
- Comment on your findings.

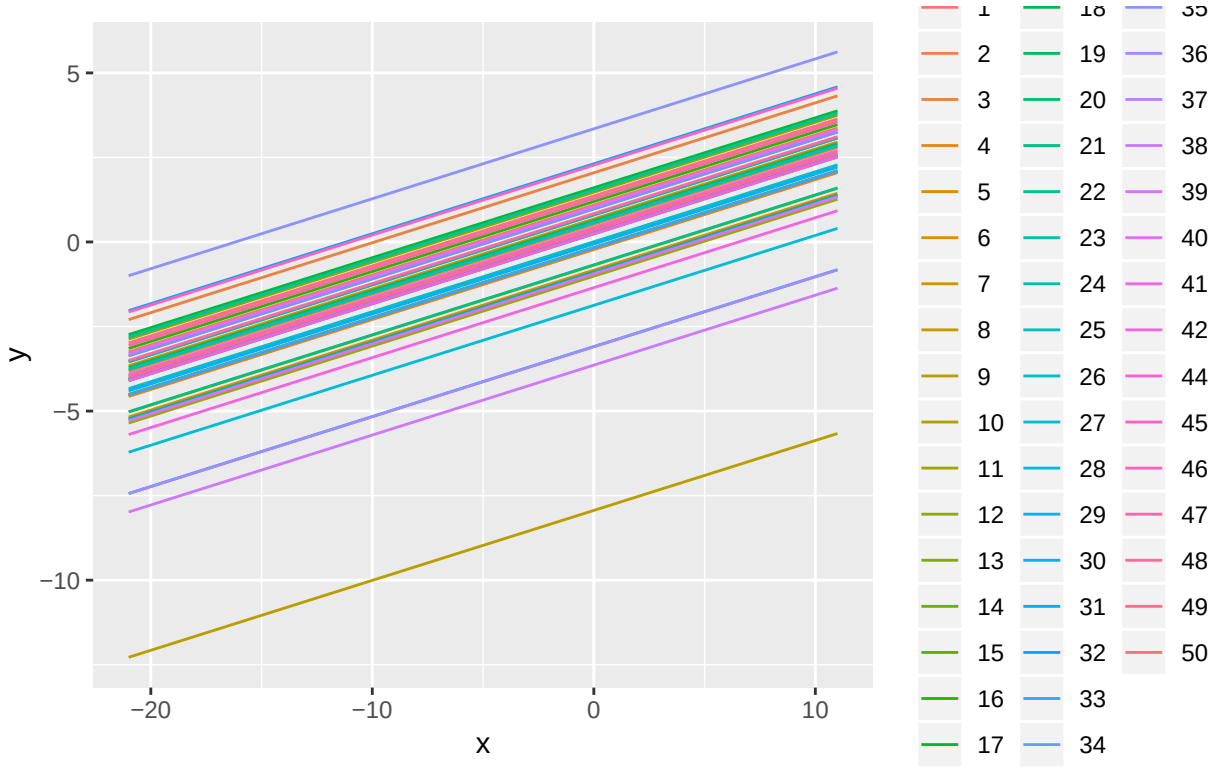
```
library(ggplot2)
library(sjPlot)
gg1 <- plot_model(fitRI2, type = "re", sort.est = "(Intercept)", y.offset = 0.4, dot.size = 1.5) +
  theme(axis.text.y = element_blank(), axis.ticks.y = element_blank()) + labs(title = "Random intercept
gg2 <- plot_model(fitRI2, type = "diag", prnt.plot = FALSE, title = "Quantile plot", geom.size = 1)
gg3 <- ggplot() + geom_density(aes(x = ranef(fitRI2)$school[[1]])) + labs(x = "x", y = "y", title = "De
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
```




```
ggarrange(gg1, 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()
```



d)

Now include the social status of the father in the model, as a fixed effect.

- Compare the model with and without the social status of the father using hypothesis test from the `anova` below (which is a likelihood ratio test - no, you need not look at the column called `deviance` since we have not talked about that). Which of the two models do you prefer?
- Also comment on the AIC and BIC of the two models (automatically added in the print-out from `anova`).
- Why does the print-out say "refitting model(s) with ML (instead of REML)" (i.e. why do we not want REML when comparing models with the same random terms but with different fixed terms)?

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

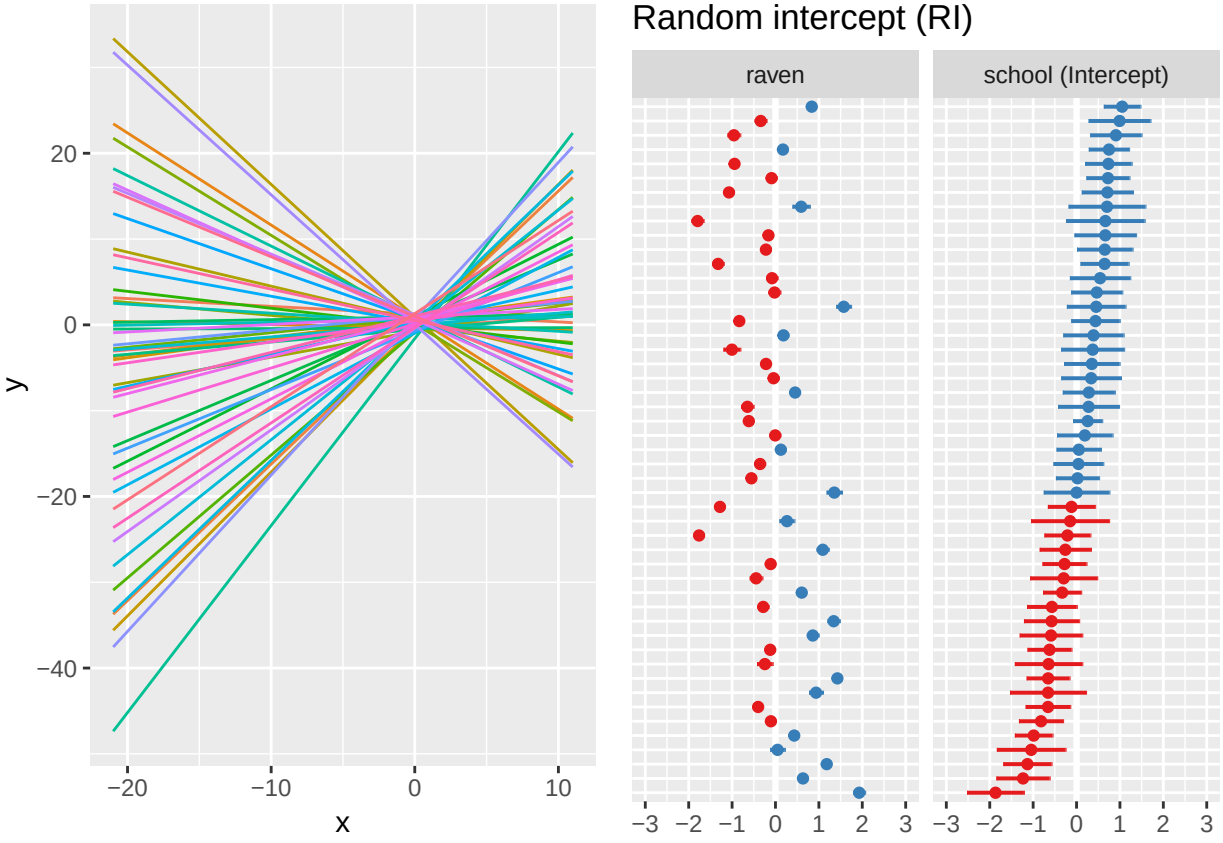
The last model we want to consider is a model with a random intercept and a *random slope* for the `raven` score at each school. It can be fit as follows:

```
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
                 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)
```



Plots:

1) Left panel (gg1) gives the fitted model- one line for each school (x-axis: score for "raven test" and y-axis: $\hat{\beta}_0 + \hat{\beta}_1 \text{raven} + \hat{\gamma}_{i0} + \hat{\gamma}_{1i} \text{raven}$ where i denotes the school, $i = 1, \ldots, 50$, and not 43).
2) Right panels (gg2) shows the predicted values for the random intercepts and slopes for the 49 schools. X-axis: predicted value with interval from 2.5% to 97.5% quantile of the distribution for the predicted value. Y-axis is only school, sorted by the random intercept, red lines for predicted values below 0 and blue above.

Finally:

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

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.

- Why is it not suitable to use a linear mixed effects model?
- What type of model would be more suitable? (hint: IL module 7)
- How would we add a random school intercept into this model (in which part of the model)?
- What is the main challenge with this type of models? (hint: marginal model)