TMA4315 Generalized Linear Models

Compulsory exercise 3: (Generalized) Linear Mixed Models

Håvard Fossdal* Rasmus Grødeland† Yawar Mahmood‡

08 November, 2024

In this project, we will study a simulated dataset with clustered data. The simulated dataset are based on the *Junior School Project* data frame from the `faraway`-package, in which test scores from students attending different primary schools in inner London have been collected.[1] The simulated dataset consists of $1,154$ observations of the following variables,

- `school`: integer in the range 1-49 encoding different schools
- `gender`: a factor with levels *boy*, *girl*.
- `social`: categorical variable indicating the social class of the father (original class 1-2 = S1, 3-4 = S2, 5-6 = S3 and 7-9 = S4)
- `raven`: the student's test score (centered around zero)
- `math`: the student's math score (centered around zero)

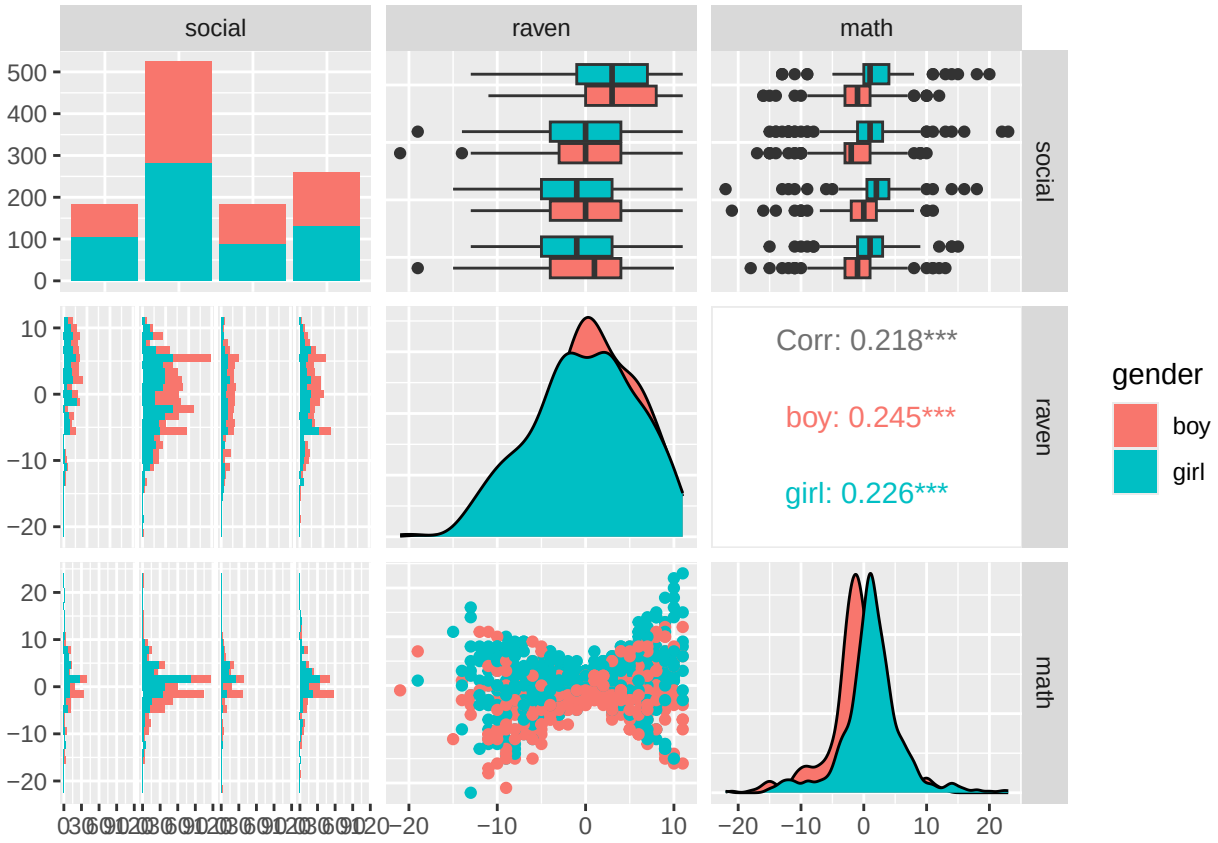
In the original dataset, a total of 50 schools were considered. However, school number 43 lacks measurements in the simulated data, so it has been removed from the subset we are studying. In the sections that follow, we will model `math` as the response, where we group the data by schools 1 through 49.

Basic Data Analysis

We start by visualizing the relationships between the covariates `social`, `raven`, and `math`, grouped by gender. We use `ggpairs()` from the GGally package to visualize the relationships.

*Group 2, Department of Mathematical Sciences, haafo@stud.ntnu.no
†Group 2, Department of Mathematical Sciences, rasmug@stud.ntnu.no
‡Group 2, Department of Mathematical Sciences, yawarm@stud.ntnu.no

[1] For an in-depth description of the data frame, the reader may refer to the documentation of the `faraway`-package.



The information provided by the figures can be summarized as follows:

- **Social vs. Math**. Math scores vary by social class and gender, with girls generally scoring higher than boys across several social classes. That being said, the `math` vs. `social` box-plot reveals that the influence of a father's social class on his child's math score is quite limited, countering the assumption that a higher social standing is synonymous with the ability to better support the child's learning, at least in the case of mathematics.
- **Raven vs. Math**. There is a positive correlation between raven and math scores for both genders. This suggests that higher raven scores (better cognitive performance) are associated with higher math scores.
- **Distribution Differences**. Girls tend to score higher in math on average. This coincides with recent studies showing that girls perform consistently better than boys in all areas of learning.[2]. The `ggpairs()` plot also suggests that the distributions of raven scores differ between genders, which might contribute to the observed difference in math scores.

Fitting a Linear Model

We now fit a linear model with `math` as the response variable and `raven` and `gender` as covariates. The model for the k-th student is then given by,

$$Y_k = X_k\beta + \epsilon_k.$$

[2]Carroll, M. 2023. Sex gaps in education. Cambridge University Press & Assessment.

Explanation of Model Components

- **Response Variable (Y_k)**. The response variable is `math`, representing the centered math score for each student. The response is stochastic, meaning it varies across observations.
- **Predictors (X_k)**. If we let p denote the number of explanatory variables (including the intercept), $\mathbf{X}_k$ becomes a $1 \times p$ matrix of covariates, and $\boldsymbol{\beta}$ becomes a $p \times 1$ vector of regression coefficients. Since we are regressing on `raven` and `gender`, $p = 3$ in this model. The linear combination $\mathbf{X}_k\boldsymbol{\beta}$ is often referred to as the *systematic* component in the context of generalized linear models (GLM), representing the fixed effects of the predictors on the response.
- **Systematic Component and Link Function**. In GLM terms, the systematic component is mapped to the expectation of the random component (response) through a *link* function. In this model, we assume errors ϵ_k have a mean of 0, meaning the model uses the *identity* link, so that the expectation of Y_k is directly given by $\mathbf{X}_k\boldsymbol{\beta}$.
- **Error Term (ϵ_k)**. This term accounts for the random deviation of a student's observed math score from the score predicted by the model, assuming errors are independently distributed with mean 0 and constant variance σ^2.

Model Fit and Analysis

```
##
## 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
```

- **(Intercept)**. The intercept is estimated at -1.3131, representing the baseline math score for boys (the reference level for gender) when the raven score of the student is zero.
- **Raven (Estimate = 0.1965)**. This coefficient indicates that for each additional unit-increase in raven score, the math score is expected to increase by approximately 0.1965, holding gender constant. The p-value for this coefficient is very small, suggesting a statistically significant positive relationship between raven score and math score.
- **Gender (girl) (Estimate = 2.5381)**. The coefficient for `gender` (girl) is 2.5381, indicating that on average, girls score approximately 2.54 points higher in math compared to boys. This difference is highly statistically significant.

The estimates suggest that both `raven` and `gender` are significant predictors of math performance. Both coefficients are positive, meaning that a higher raven score is associated with better math performance, and

that girls outperform boys with the same raven score. These observations are in line with the introductory data analysis in the `ggpairs()` plot.

- **Residual Standard Error**. The residual standard error of 4.76 provides an estimate for the typical deviation of observed math scores from the predictions made by the model.
- **R-Squared (0.1105)**. The R-squared value of 0.1105 indicates that around 11 % of the variability in math scores are explained by raven scores and gender. This is a relatively low R-squared, suggesting that other factors not included in this model may also influence math scores.

Purpose of the Model

In this model, we investigate how gender and raven scores affect the math scores of primary school students. We assume that the effects of gender and raven scores on a student's math performance are linear and independent, with no interaction between these predictors.

Accounting for School-Specific Differences

We now add a random intercept for `school` to our model. This adjustment is crucial, as grading systems might not be identical across schools. We aim to fit a random intercept model where `school` is treated as a random effect. The model is specified as follows:

$$Y_i = X_i\beta + 1\gamma_{0i} + \epsilon_i,$$

where

- Y_i: The vector of math scores for students in school i, with dimension $n_i \times 1$, where n_i is the number of students in school i.
- X_i: The design matrix for the fixed effects, with dimension $n_i \times p$, where p is the number of fixed-effect predictors, including the intercept.
- β: The fixed-effect parameter vector with dimension $p \times 1$.
- 1: A column vector of ones with dimension $n_i \times 1$.
- γ_{0i}: The random intercept for school i, representing the deviation of the intercept for school i from the overall mean intercept. This captures the school-specific effect on math scores.
- ϵ_i: The error term vector for students in school i, with dimension $n_i \times 1$.

In this model, the systematic component $X_i\beta$ represents the effects of fixed predictors (`raven` and `gender`), whilst the random intercept γ_{0i} accounts for unobserved factors that may lead to baseline differences in math scores across schools.

Distributional Assumptions

To fit this model, we make the following distributional assumptions,

- **Random Intercept (γ_{0i})**. We assume $\gamma_{0i} \sim \mathrm{N}(0, \tau_0^2)$, where τ_0^2 is the variance of the school-level random intercepts. This means that each school intercept is drawn from a normal distribution with a mean of zero and variance τ_0^2.
- **Error Term (ϵ_i)**. We assume $\epsilon_i \sim \mathrm{N}_{n_i}(0, \sigma^2 I_{n_i})$, where σ^2 is the residual variance, and I_{n_i} is the identity matrix of dimension n_i. The errors are independent and identically distributed.

We further assume that γ_{0i} and the components of ϵ_i are mutually independent.

Dependency Assumptions

- **Within-School Dependency**. The random intercept for each school introduces a dependency among observations within the same school. Students in the same school share the same random intercept γ_{0i}, creating a correlation. This within-school correlation is given by $\frac{\tau_0^2}{\tau_0^2+\sigma^2}$.
- **Between-School Independence**. Observations from different schools are assumed to be independent, and this is a reasonable assumption to be made.

Model Fit and Analysis of the Random Intercept Model

Fitting the model using the `lmer()`-function, we obtain the following result.

```
## 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
```

Random Effects

- **School-Level Random Intercept**.
  - **Variance**. The variance of the random intercept is estimated at 3.879. This reflects the variability in the average math score between schools.
  - **Standard Deviation**. The standard deviation of the random intercept is 1.969. This gives an average measure of how much each school's math score baseline deviates from the overall average baseline.
- **Residual Variance**. The within school variability is estimated at 19.220, with a standard deviation of 4.384. This represents the variation in math scores that is not explained by the model.

Fixed Effects

- **(Intercept)**. The intercept is estimated at -1.26915, representing the average math score for boys (the reference level for `gender`) with a raven score of 0.
- **Raven (Estimate = 0.21442)**. The positive coefficient for `raven` indicates that on average, a one-unit increase in raven score is associated with a 0.214 point increase in the math score, after accounting for school-level differences. The high t-value, 9.197, makes this effect statistically significant.
- **Gender (girl) (Estimate = 2.51119)**. The coefficient for `gender` (girl) is 2.51119, meaning that girls score, on average, 2.51 points higher compared to boys, controlling for raven scores and school-level differences. Again, the t-value is high, being estimated at 9.411, so the effect is significant.

Correlation of Fixed Effects

- The correlation between the intercept and `gender` (girl) is -0.404, indicating a moderate negative relationship. This may suggest that schools with higher intercepts tend to show a smaller gender difference.

Comparison of the Model to the Earlier Model

The estimates for `raven` and `gender` in the mixed effects model differ slightly from those in previous model.

- **Raven**. In the previous model, the estimate for `raven` was 0.1965, whereas here it is 0.21442. By accounting for school-level differences, we observe a slightly higher effect.
- **Gender**. The estimate for `gender` (girl) in the previous model was 2.5381, compared to 2.51119 in the mixed effects model. The estimate is slightly reduced, as a result of some of the variation between the genders now being explained by differences between schools.

The differences in parameter estimates arise because the mixed effects model accounts for school-specific heterogeneity, allowing for variations in math score averages across schools. For example, if we include a school known for handing out high math scores where the majority of the students are female, the previous model could overestimate the real effect of gender. Similarly, by including a school known for handing out high math scores to its students, of which a majority have below-average raven scores, the previous model may underestimate the real effect of this covariate. By including a random intercept for schools, we isolate these school-level effects, leading to adjusted estimates for `gender` and `raven`.

Both models still indicate that `gender` and `raven` have a significant effect on math scores, suggesting these relationships are genuine and not due to unobserved differences between schools. Additionally, the mixed model reduces uncertainty in the fixed effects, as seen by the lower standard errors for `gender` and `raven`.

The Lack of p-values

The `lmer()`-function does not provide p-values by default for fixed effects, because calculating them in mixed effects models is complex due to the presence of random effects. In these models, the assumption of homogeneity required for traditional linear models does not hold, as we account for school-specific heterogeneity. Hence the t-statistics and p-values are generally not t-distributed.

This issue also extends to general linear hypotheses of the form $H_0 : C\beta = d$ versus $H_1 : C\beta \neq d$, where the F-statistics are also not F-distributed under H_0. Although these statistics have known asymptotic distributions (normal and χ^2-distributions with degrees of freedon equal to the number of columns in C), assuming that the statistics follow these approximated distributions can sometimes lead to the wrong conclusions.

As a result, Dr. Douglas Bates, the creator of `lmer()`, chose not to include p-values for fixed effects in the summary output. For approximate p-values, alternative methods such as the Kenward-Roger or Satterthwaite approximations can be used.[3]

Hypothesis Testing Based on Asymptotic Distributions

Assuming that the null-distributions of the t-statistics computed by the `lmer()`-function are approximate normal, meaning that we can use the asymptotic properties of the t-statistic to do inference, we can for example test for the significance of regression of `raven`.

- $H_0 : \beta_{\textbf{raven}} = 0$ **vs.** $H_1 : \beta_{\textbf{raven}} \neq 0$ **(t-value** $= 9.197$**)**. Under H_0, the probability of the t-statistic taking on a greater absolute value than 9.197 is assumed to be $2 \cdot (1 - \Phi(9.197))$, where Φ denotes the cdf of the standard normal (the 2 in front is a result of the distribution being symmetric around 0). This expression is approximately equal to 3.687×10^{-20}, meaning that the p-value corresponding to this hypothesis test is approximately 3.687×10^{-20}. Thus, we **reject** H_0 at a confidence level $\alpha = 0.05$, and the regression is deemed significant.

A similar hypothesis could be used to test for the significance of regression of `gender`.

Confidence Intervals for Fixed-Effects

We may compute confidence intervals for the estimated fixed-effects in a variety of ways. The *simplest* is to assume that the the t-statistics are approximately normal. Alternatively, one can for example use parametric bootstrapping to compute confidence intervals for the estimates, as preferred by Dr. Bates.[4] We will now compute a 95 %-confidence interval for the effect of the female gender on math scores, β_{gender}, using both of these methods.

- **Assuming normality**. By assuming that the t-statistics are approximately normal, a 95 %-confidence interval for β_{gender} becomes
$$\hat{\beta}_{\text{gender}} \pm z_{0.975} \cdot \text{SE}(\hat{\beta}_{\text{gender}}),$$
where $z_{0.975}$ is the 0.975-quantile of the standard normal, SE denotes standard error and $\hat{\beta}_{\text{gender}}$ the estimated effect of the female gender. Using R, we find that $\beta_{\text{gender}} \in (1.98819,\ 3.03418)$ with 95 % confidence.
- **Using parametric bootstrapping**. By creating $1,000$ bootstrapped samples based on the estimated $\hat{\beta}_{\text{gender}} = 2.51119$, we find that $\beta_{\text{gender}} \in (1.98395, 3.01288)$ with 95 % confidence. The bootstrapped distributions were sampled using the `confint()`-function.

We notice that the confidence intervals are quite similar, which suggests that the normal approximation might be justified in this case.

Excluding Gender From the Random Intercept Model

We now fit a reduced random intercept model for `school` with `raven` as the only fixed effect. The model can be stated in terms of the response of the jth student at the ith school,

$$Y_{ij} = \beta_0 + \beta_1 \cdot \text{raven}_{ij} + \gamma_{0i} + \epsilon_{ij},$$

where

[3] Dr. Bates explained this in a response to the R-help list in 2006.
[4] Dr. Bates explained this in a response to the R-help list in 2006.

- Y_{ij}: The math score for student j in school i.
- β_0: The overall intercept.
- β_1: The fixed effect coefficient for `raven`.
- γ_{0i}: The random intercept for school i, capturing school-specific deviations from the overall intercept. We assume that $\gamma_{0i} \sim \mathrm{N}(0, \tau_0^2)$.
- ϵ_{ij}: The residual error term for each student. We assume that $\epsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$ are independent and identically distributed.

Model Fit and Analysis

Fitting the model using the `lmer()`-function, we obtain the following result.

```
## 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
```

Random Effects

- **School-Level Random Intercept**.
  - **Variance**. The variance of the random intercept for `school` is estimated at 4.002, with a standard deviation of 2.001. This indicates variability in baseline math scores across schools.
- **Residual Variance**. The within-school variability is estimated at 20.711, with a standard deviation of 4.551. This represents the individual-level variability in math scores not explained by `raven` or school-level effects.

Fixed Effects

- **Intercept (Estimate = 0.03840)**. This estimate represents the average math score for students with a raven score of 0, adjusted for school-level differences. The t-value of 0.120 indicates that the intercept is not significantly different from 0.
- **Raven (Estimate = 0.20682)**. The coefficient for `raven` is positive, indicating that, on average, a one-unit increase in raven score is associated with a 0.20682 point increase in math score, holding `school` constant. The t-value of 8.554 suggests this effect is highly significant.

Covariance and Correlation Between Responses Within the Same School

Since we are using a random intercept model, there is a correlation between responses from students in the same school due to the shared random intercept γ_{0i}.

Covariance Between Students in the Same School

The covariance between two responses Y_{ij} and Y_{il} for students j and l in the same school i can be calculated as follows,

$$\text{Cov}(Y_{ij}, Y_{il}) = \text{Cov}(\beta_0 + \beta_1 \cdot \text{raven}_{ij} + \gamma_{0i} + \epsilon_{ij}, \beta_0 + \beta_1 \cdot \text{raven}_{il} + \gamma_{0i} + \epsilon_{il}).$$

Since the random intercept γ_{0i} is shared across students in the same school, the covariance is equal to the variance of γ_{0i}, which is estimated to be 4.002.

Correlation Between Students in the Same School

The correlation between two responses Y_{ij} and Y_{il} in the same school are given by,

$$\text{Corr}(Y_{ij}, Y_{il}) = \frac{\text{Cov}(Y_{ij}, Y_{il})}{\sqrt{\text{Var}(Y_{ij}) \cdot \text{Var}(Y_{il})}}.$$

With $\text{Var}(Y_{ij}) = \sigma^2 + \tau_0^2 = 20.711 + 4.002$, we can calculate this correlation as,

$$\text{Corr}(Y_{ij}, Y_{il}) = \frac{4.002}{20.711 + 4.002} = \frac{4.002}{24.713} \approx 0.162.$$

A correlation of 0.162 suggests a moderate degree of similarity in math scores among students within the same school, attributable to shared experiences in the schools.

Derivation of the Predicted Random Intercept $\hat{\gamma}_{0i}$

We now want to find a formula for the the predicted random intercept of school i, $\hat{\gamma}_{0i}$. Grouping the observations by `school`, we can express the marginal distribution of the responses $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in_i})^T$ in terms of a multivariate normal,

$$Y_i \sim \text{N}_{n_i}(X\beta, V := \sigma^2 I_{n_i} + \tau_0^2 11^T).$$

Using the fact that $\gamma_{0i} \sim N(0, \tau_0^2)$ is normally distributed as well, the joint distribution of Y_i and γ_{0i} can be expressed as

$$\begin{pmatrix} Y_i \\ \gamma_{0i} \end{pmatrix} = \mathrm{N}_{n_i+1}\left(\begin{pmatrix} X_i\beta \\ 0 \end{pmatrix}, \begin{pmatrix} V & \tau_0^2 1 \\ \tau_0^2 1^T & \tau_0^2 \end{pmatrix} \right).$$

The *best linear unbiased predictor (BLUP)* $\hat{\gamma}_{0i}$ is then determined by solving for the expectation of $\gamma_{0i}|Y_i$ and inserting estimates for β and τ_0^2 in the resulting expression. Using the joint distribution of the variables and known properties of the multivariate normal, we find that $\mathrm{E}(\gamma_{0i}|Y_i) = 0 + \tau_0^2 1^T \cdot V^{-1}(Y_i - X_i\beta)$. Since the covariance matrix of Y_i, denoted by V, is non-singular *Toeplitz*, it has a known inverse. By substituting for the inverse and inserting the parameter estimates, we find that

$$\hat{\gamma}_{0i} = \frac{n_i\hat{\tau}_0^2}{\hat{\sigma}^2 + n_i\hat{\tau}_0^2}\left(\frac{1}{n_i}\sum_{j=1}^{n_i}\left(Y_{ij} - (\hat{\beta}_0 + \hat{\beta}_1 \cdot \mathrm{raven}_{ij}) \right) \right) =: \frac{n_i\hat{\tau}_0^2}{\hat{\sigma}^2 + n_i\hat{\tau}_0^2} e_i,$$

after some simplifications.

The term in the outer parentheses e_i is known as *the average raw residual* at the student level (level 0), and we observe that for random intercept models, the BLUP is a scaling of this average residual. Notice also that the weight will always be between 0 and 1, so intuitively the BLUP can be thought of as shrinking the school-specific average raw residual, where the scaling factor depends on the size of the dataset, i.e. number of students at the ith school n_i, and the residual- and random-effect variances, σ^2 and τ_0^2. Asymptotically, we have that $\hat{\gamma}_{0i} \to e_i$ as $n_i \to \infty$, i.e. the BLUP tends to the average raw residuals for large samples, and $\hat{\gamma}_{0i} \to 0$ as $\hat{\sigma}^2 \to \infty$ for τ_0^2 fixed, i.e the BLUP tends to zero when the residual variance is much greater than the random-effect variance.

Furthermore, as the name would imply, $\hat{\gamma}_{0i}$ is unbiased and minimizes the expectation of the squared error $\mathrm{E}(\gamma_{0i} - \hat{\gamma}_{0i})^2$ in the sense that $\mathrm{E}(\gamma_{0i} - \hat{\gamma}_{0i})^2 \leq \mathrm{E}(\gamma_{0i} - \sum_j^{n_i} \alpha_{ji}Y_j)^2$ for all possible choices of α_{ji}, i.e. the squared error is minimized among all possible linear predictors of γ_{0i}.

Explanation of Diagnostic Plots

- **Random Intercept Plot**. This plot shows the estimated random intercepts for each school, highlighting differences in baseline math scores across schools. Schools with larger positive or negative intercepts have math scores significantly higher or lower than the overall average.
- **Quantile Plot for Random Intercepts**. This plot displays quantiles of the random intercepts, helping to check if the random intercepts follow a normal distribution. Points falling along a straight line support the normality assumption.
- **Density Plot of Random Intercepts**. This density plot visualizes the distribution of random intercepts for schools, helping to check if the distribution of random effects aligns with the assumed normal distribution. A bell-shaped curve would suggest normality.
- **Residuals vs. Fitted Values**. This plot displays residuals as a function of fitted values. Ideally, residuals should be randomly scattered around zero without a distinct pattern. Any systematic pattern might indicate that the model is not fair to use.
- **Normal Q-Q Plot of Standardized Residuals**. This plot compares the quantiles of standardized residuals to the theoretical quantiles of a normal distribution. If the residuals are normally distributed, they will lie along the line. Deviations from the line may suggest issues with the normality assumptions.
- **School-Level Prediction Lines**. This plot shows the predicted math scores for each school as a function of raven scores, illustrating how each school's intercept affects math scores. This plot highlights both the within-school effect of `raven` on math scores and the between-school differences in intercepts.

Comment on Findings

- **Random Intercepts**. The random intercept plot and density plot show the variability in baseline math scores across schools. Some schools are consistent in scoring higher or lower - showing that school effects are meaningful to consider.
- **Residuals**. The Residuals vs. Fitted plot suggests no major pattern, indicating that the model fits the data well. A random scatter around zero supports the linearity assumption.
- **Normality of Residuals**. The Q-Q plot suggests that the residuals mostly follow a normal distribution.

Adding Social Status to our model

We start by adding the social status of the father, `social`, as a fixed effect in our model. The model is specified as,

$$Y_{ij} = \beta_0 + \beta_1 \cdot \text{raven}_{ij} + \beta_2 \cdot \text{social}_{ij} + \gamma_{0i} + \epsilon_{ij}$$

where,

- Y_{ij}: Math score for student j in school i.
- β_0: Overall intercept.
- β_1: Fixed effect of `raven`.
- β_2: Fixed effect of `social`, with levels S2, S3, and S4 (S1 as reference).
- γ_{0i}: Random intercept for school i.
- ϵ_{ij}: Residual error term for each student.

Fitting model and Comment on output

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: math ~ raven + social + (1 | school)
##    Data: dataset
##
## REML criterion at convergence: 6849.2
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.4649 -0.4599  0.0069  0.4475  4.5581
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  school   (Intercept)  4.007   2.002
##  Residual             20.615   4.540
## Number of obs: 1154, groups:  school, 49
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  0.13101    0.46036   0.285
## raven        0.20823    0.02452   8.491
## socialS2    -0.44154    0.41189  -1.072
## socialS3     0.69108    0.50049   1.381
## socialS4    -0.06411    0.46785  -0.137
##
```

```
## Correlation of Fixed Effects:
##          (Intr) raven  soclS2 soclS3
## raven    -0.129
## socialS2 -0.671  0.155
## socialS3 -0.566  0.161  0.627
## socialS4 -0.612  0.150  0.678  0.569
```

- **Intercept (Estimate = 0.13101)**. This is the estimated math score for a student in social class S1 (reference level) with a centered raven score of 0, adjusted for school effects.
- **Raven (Estimate = 0.20823)**. The coefficient for `raven` is positive, indicating that, on average, a one-unit increase in `raven` score is associated with a 0.208 increase in math score, holding `social` and `school` constant. This effect is highly significant.
- **Social S2 (Estimate = -0.44154)**. Students in social class S2 score approximately 0.442 points lower than students in social class S1, but this effect is not statistically significant ($t = -1.072$).
- **Social S3 (Estimate = 0.69108)**. Students in social class S3 score approximately 0.691 points higher than students in social class S1, but this effect is not statistically significant ($t = 1.381$).
- **Social S4 (Estimate = -0.06411)**. Students in social class S4 score about 0.064 points lower than students in social class S1, but this effect is not statistically significant ($t = -0.137$).
- **School-Level Random Intercept (Variance = 4.007)**. The variability in baseline math scores across schools is reflected in the random intercept variance of 4.007.
- **Residual Variance (Variance = 20.615)**: This is the within-school variability in math scores that remains unexplained by the fixed effects.

Model Comparison Using Likelihood Ratio Test

To assess whether adding `social` as a fixed effect significantly improves the model, we perform a likelihood ratio test,

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

Results of the Likelihood Ratio Test

- **Chi-Squared Test**. The likelihood ratio test yields a Chi-squared statistic of 8.1175 with 3 degrees of freedom and a p-value of 0.04364. This indicates that adding `social` as a fixed effect provides a statistically significant improvement in model fit ($p < 0.05$) - even though the coefficients themself are not significant.

AIC and BIC Comparison

- **AIC**. `fitRI2` has an AIC of 6858.9, while `fitRI3` has an AIC of 6856.8. A lower AIC in `fitRI3` suggests a *better* model fit - but it is only by a little.
- **BIC**. `fitRI3` has a slightly higher BIC (6892.1) compared to `fitRI2` (6879.1), but it is only by a little. Similar to the AIC, the *better* model is the one with the lowest BIC, in this case `fitRI2`.

Explanation of REML vs. ML in Model Comparison

The output states "refitting model(s) with ML (instead of REML)" because,

- **REML (Restricted Maximum Likelihood)**. REML is typically used for estimating variance components in mixed models.
- **ML (Maximum Likelihood)**. ML is used when comparing models with different fixed effects. As we have different fixed effects (with and without `social`), ML is used to ensure a valid comparison.

Fitting a Random Intercept and Slope Model

Finally, we fit a model with both a random intercept and a random slope for `raven` at each school,

```
## 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
```

Mathematical Model for Random- Intercept and Slope

The model formula is,

$$Y_{ij} = \beta_0 + \beta_1 \cdot \text{raven}_{ij} + \gamma_{0i} + \gamma_{1i} \cdot \text{raven}_{ij} + \epsilon_{ij}$$

where,

- Y_{ij}: Math score for student j in school i.
- β_0: Fixed effect intercept.
- β_1: Fixed effect slope for `raven`.
- γ_{0i}: School-specific random intercept.
- γ_{1i}: School-specific random slope for `raven`.
- ϵ_{ij}: Residual error term.

Model Output Analysis

Fixed Effects

- **Intercept (Estimate = 0.2603)**. The average math score for a student with a centered raven score of 0, accounting for both school-level variability in intercepts and slopes.
- **Raven (Estimate = 0.2498)**. This is the average effect of `raven` on math scores across schools. On average, a one-unit increase in raven score is associated with a 0.2498 point increase in math score.

Random Effects

- **Random Intercept (Variance = 0.5519)**. This represents the variability in baseline math scores between schools.
- **Random Slope for raven (Variance = 0.7293)**. This represents the variability in the effect of `raven` across schools, suggesting that the strength of the relationship between `raven` and math scores varies between schools.
- **Correlation Between Intercept and Slope (-0.40)**. The negative correlation between intercept and slope suggests that schools with higher baseline math scores tend to have a lower effect of `raven` on math scores, and the opposite is true aswell.

Interpretation of Plots

- **Left Panel** . This plot displays the fitted lines for each school across the range of raven scores. The varying slopes indicate that different schools have different relationships between raven- and math scores.
- **Right Panel** . This plot displays the predicted values for the random intercepts and slopes for each school, with 95 % confidence intervals.

Summary and Comments on Findings

- **Model Preference**. The likelihood ratio test and AIC comparison suggest that including `social` as a fixed effect improves model fit, as social status appears to influence math scores. However, the effect sizes are not statistically significant, and the BIC comparison favors the model that does not include `social` as a fixed effect. It is worth to mention, that BIC typically assigns a higher penalty to the addition of model parameters compared to AIC. Based on the ANOVA results alone, we cannot conclude that one of the models are "objectively" better than the other. We can however conclude that both `raven` and `gender` are better predictors of the math score compared to `social`. This coinsides with the some of the insights provided by the `ggpairs()` plot earlier in the report.

- **Random Intercept and Slope Model**. Allowing both the intercept and slope for `raven` to vary by school provides a model that captures both school-specific baseline differences and differential effects of `raven`. The negative correlation between intercept and slope suggests that schools with higher average math scores tend to see less impact from `raven`.

Why a Linear Mixed Effects Model is Not Suitable

A linear mixed-effects model is not suitable for modeling the probability of failure because,

- **Binary Outcome**. In this case, we are interested in a binary outcome (pass/fail) - not continuous. A linear model would predict values on a continuous scale, which could result in nonsensical probabilities.
- **Non-Normal Error Structure**. Linear models assume normally distributed residuals, which is not appropriate for binary outcomes, Bernoulli distribution.
- **Heteroscedasticity**. Binary data often have non-constant variance, which violates another assumption of linear models.

More Suitable Model Type: Generalized Linear Mixed Model (GLMM)

A generalized linear mixed model (GLMM) is more suitable for this type of binary outcome data. This model uses a logistic link function to ensure that the predicted probabilities lie within the [0, 1] range.

In a logistic mixed-effects model, the probability of failure is modeled as,

$$\text{logit}(P(Y_{ij} = 1)) = \beta_0 + \beta_1 \cdot \text{raven}_{ij} + \gamma_{0i},$$

where

- $Y_{ij} = 1$ indicates that student j in school i has failed.
- $\text{logit}(P(Y_{ij} = 1))$ is the log-odds of failure for student j in school i.
- β_0 and β_1 are fixed effects for the intercept and the effect of `raven`.
- γ_{0i} is a random intercept for school i, capturing school-level differences in failure rates.

Adding a Random School Intercept

To include a random school intercept in the logistic mixed-effects model, we add γ_{0i} to the linear predictor of the logit function,

$$\text{logit}(P(Y_{ij} = 1)) = \beta_0 + \beta_1 \cdot \text{raven}_{ij} + \gamma_{0i},$$

where

- $\gamma_{0i} \sim N(0, \sigma_\gamma^2)$ represents the school-specific random intercept, allowing each school to have a unique baseline log-odds of failure.

And all other points are as mentioned above.

Main Challenge with This Type of Model

The main challenge with logistic mixed-effects models is related to the interpretation of the marginal effects versus the conditional effects.

- **Marginal Model Interpretation**. Due to the non-linearity of the link function (logit), this is not straightforward. The fixed effects in GLMMs represent conditional effects given the random intercepts, which complicates the interpretation of overall effects.
- **Integration Over Random Effects**. To obtain the marginal probability of failure, we would need to perform numerical integration, which has numerous challenges.

Fitting a Logistic Mixed-Effects Model

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: fail ~ raven + (1 | school)
##    Data: dataset
##
##      AIC      BIC   logLik deviance df.resid
##    266.7    281.8   -130.3    260.7     1151
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.3844 -0.0482 -0.0136 -0.0070  7.4249
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  school (Intercept) 28.79    5.365
## Number of obs: 1154, groups:  school, 49
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.79907    1.83869  -4.786 1.71e-06 ***
## raven       -0.24086    0.04442  -5.423 5.86e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##       (Intr)
## raven 0.412
```

Interpretation of Fixed Effects

- **Intercept**. The intercept of -8.79907 represents the log-odds of failure for a student with a centered raven score of 0.

  To interpret this in terms of probability,

$$\text{Probability of Failure} = \frac{1}{1 + e^{-(-8.79907)}} \approx 0.00015.$$

  This means that students with a centered raven score of 0 have a very low probability of failing.

- **Raven**. The coefficient for `raven` is -0.24086, which means that as `raven` increases, the log-odds of failure decreases. In terms of probability, higher raven scores are associated with a lower probability of failure.

Specifically, a one-unit increase in raven score multiplies the odds of failure by

$$e^{-0.24086} \approx 0.785.$$

This indicates a reduction in the odds of failure by about 21.5 % for each one-unit increase in `raven`.

Interpretation of Random Effects

- **Random Intercept for School**. The random intercept variance for `school` is estimated to be 28.79, with a standard deviation of 5.365. This large variance suggests significant variability in baseline failure rates across schools. This means that certain schools are more likely to have students fail, independent of individual raven scores. Perhaps due to school grading practices, or other factors.

Comparison with Linear Mixed Model

- **Interpretation of Effects**. In the original model, the effect of `raven` was on the actual math score. In the logistic model, the effect of `raven` is on the log-odds of failure. This changes the interpretation.
- **Model Inferences**. In the logistic model, the fixed effects are interpreted in terms of log-odds rather than changes in mean math scores. For example, a negative coefficient for `raven` would suggest that as `raven` increases, the probability of failure decreases. This makes the model more suitable for answering questions about pass/fail, and not so suitable for predicting math scores.
- **Random Intercept Variability**: The random intercept for `school` in the logistic model captures variability in failure probabilities across schools, which can differ from the variability in average math scores across schools.

Code used as part of the exercise

```
#-----------#
# Problem a #
#-----------#

# loading the dataset into the global environment
dataset <- read.table(
  "https://www.math.ntnu.no/emner/TMA4315/2018h/jsp2.txt",
  header = TRUE
)
dim(dataset)[1]  # number of observations in the simulated dataset

# preliminary plot of the observations
ggpairs(
  data = dataset,
  mapping = aes(col = gender),
  columns = c("social", "raven", "math"),
  legend = 1
```

```
)

# fitting the linear model
linear_model <- lm(
  formula = math ~ raven + gender,
  data = dataset
)
summary(linear_model)  # printing a summary of the model fit

#-----------#
# Problem b #
#-----------#

# fitting the random intercept model
fitRI1 <- lmer(
  formula = math ~ raven + gender + (1 | school),
  data = dataset
)
summary(fitRI1)  # printing a summary of the model fit

# observed t-statistic for raven parameter
T_raven <- (fixef(fitRI1)/sqrt(diag(vcov(fitRI1))))[2]
# p-value, assuming normality
p_raven <- 2 * pnorm(q = T_raven, lower.tail = FALSE)

# upper and lower CI limits, assuming normality
ci_gender_lower <- fixef(fitRI1)[3] - qnorm(0.975) * sqrt(diag(vcov(fitRI1)))[3]
ci_gender_upper <- fixef(fitRI1)[3] + qnorm(0.975) * sqrt(diag(vcov(fitRI1)))[3]

# bootstrapped CI
ci_gender <- confint(  # resulting 95 % CI: (1.983948, 3.012880)
  object = fitRI1,
  parm = 'beta_',
  level = 0.95,
  method = 'boot',
  nsim = 1000,
  seed = 1000
)['gendergirl',]

#-----------#
# Problem c #
#-----------#

# fitting the random intercept model
fitRI2 <- lmer(
  formula = math ~ raven + (1 | school),
  data = dataset
)
summary(fitRI2)  # printing a summary of the model fit

#-----------#
# Problem d #
#-----------#
```

```r
# fitting the random intercept model
fitRI3 <- lmer(
  formula = math ~ raven + social + (1 | school),
  data = dataset
)
summary(fitRI3)  # printing a summary of the model fit

anova(fitRI2, fitRI3)  # print ANOVA summary of fitRI2 vs. fitRI3

# fitting the random slope model
fitRIS <- lmer(
  formula = math ~ raven + (1 + raven | school),
  data = dataset
)
summary(fitRIS)  # printing a summary of the model fit

#-----------#
# Problem e #
#-----------#

# dummy variable indicating whether or not students fail math
dataset$fail <- ifelse(dataset$math < -10, 1, 0)

# fitting the generalized linear mixed model
fitRI1_logistic <- glmer(
  formula = fail ~ raven + (1 | school),
  data = dataset,
  family = binomial(link = "logit")
)
summary(fitRI1_logistic)  # printing a summary of the model fit
```