

Multilevel and Longitudinal Analysis: Assessment 1

Test with open answers

Student ID: 11484265

2024-09-30

Descriptive Analysis

There are 54 schools and 29506 pupils. The average number of pupils per school is 546.

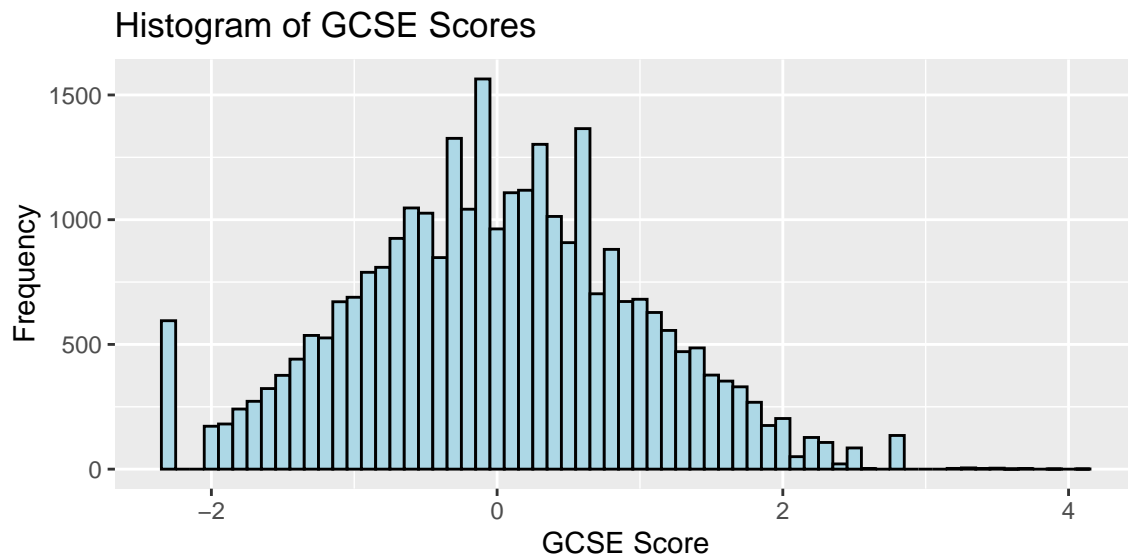


Figure 1: Histogram of GCSE point scores.

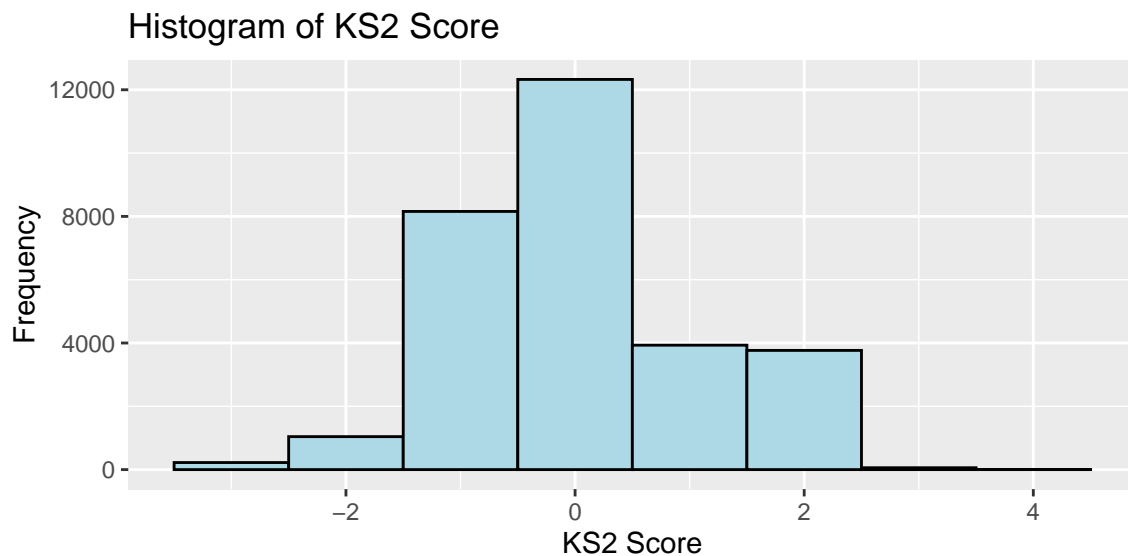


Figure 2: Histogram of KS2 scores.

While there are pupils who score low on the GCSE on a wide range of KS2 scores, the general trend is that higher GCSE scorers generally have higher KS2. Those who scored extremely high on the KS2 score may not always have the highest GCSE, but are all above average.

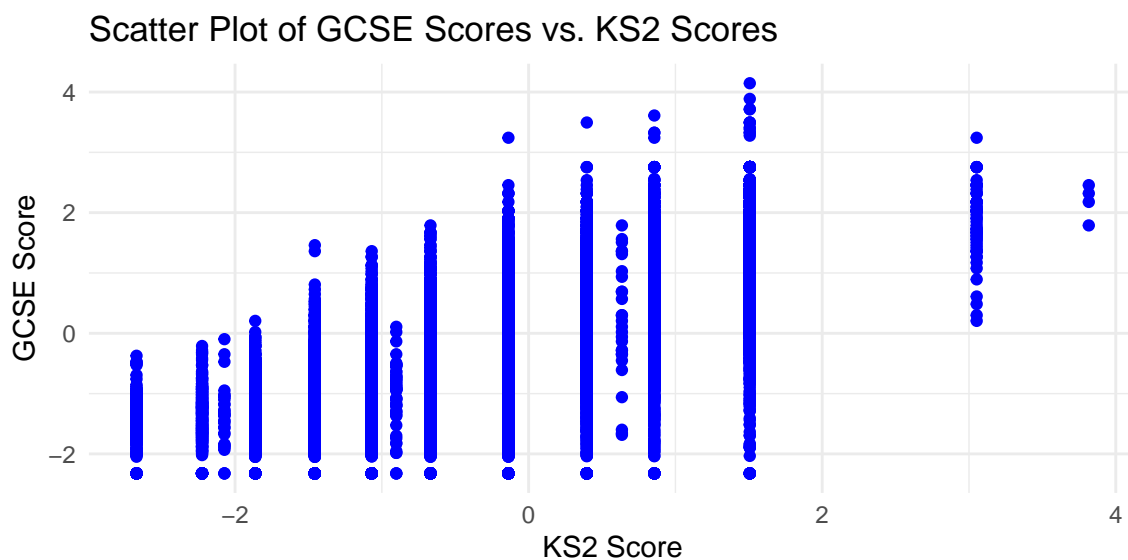
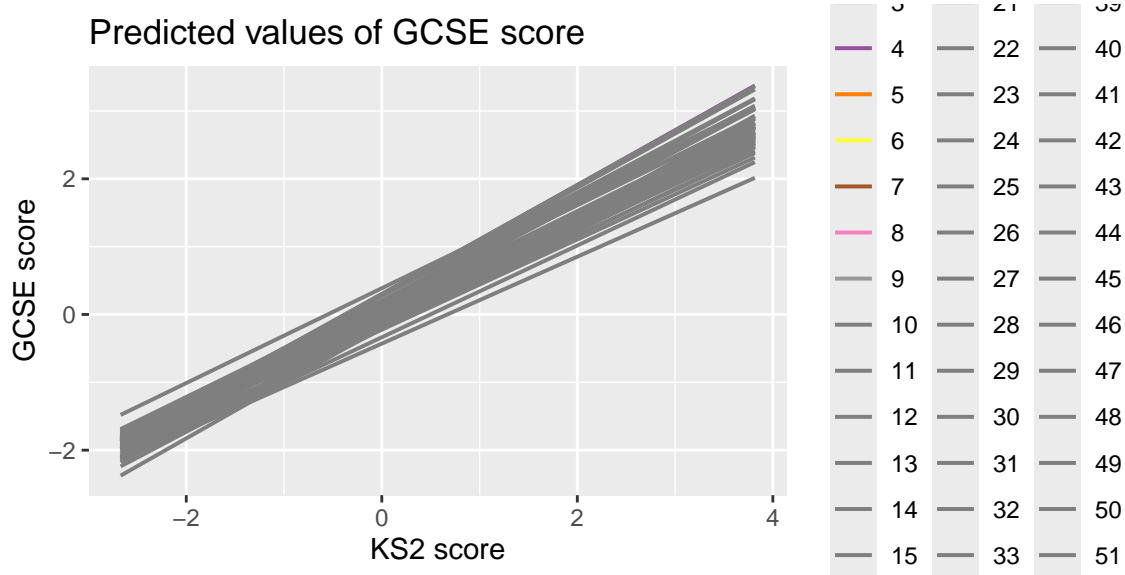


Figure 3: Scatterplot of KS2 and GCSE scores.

Model Estimation



Impact of KS2 Scores

Including ks2 in model 1 decreased variance in schools 0.05; residual went down by approximately 0.5. The relationship between “ks2score” and “gcse score” is positive, suggesting every unit increase in ks2 scores increases GCSE score by approximately 0.72. The slightly lower AIC and BIC values and higher log-likelihood and reduced deviance in model 2 indicate a slightly better fit. This suggests that the addition of the squared term provides a significant improvement in the model. The relationship between ks2score and gcse score is assumed to be linear. The coefficient for ks2score is approximately 0.72, indicating that for every one-unit increase in ks2score, gcse score increases by about 0.72, holding the school effect constant.

Impact of Squared KS2 Scores

With the introduction of squared KS2 scores ($ks2_sq$) in model 2, the relationship becomes quadratic. The coefficient for $ks2_sq$ is positive (0.028), which suggests a U-shaped relationship. This indicates that the effect of $ks2score$ on $gcse$ score is not constant. As $ks2score$ increases, the effect on GCSE score initially may diminish before it begins to rise again. This suggests that while higher scores generally lead to better outcomes, the benefit may be more pronounced for students with moderate scores rather than those at the extremes.

Impact of Random Slopes of KS2 Scores

The variance for the random slope of $ks2score$ is 0.002482, indicating some variability in how the effect of $ks2score$ on $gcse$ score differs across schools. This suggests that while there is a general positive trend (captured by the fixed effect of $ks2score$), the strength of that trend is not uniform but quite similar.

The correlation between the random intercept and the random slope for $ks2score$ is 0.34. This positive correlation indicates that schools with better baseline performance, indicated by higher intercepts, tend to also have steeper slopes, suggesting that schools that generally perform well also see greater gains from higher KS2 scores.

Between-School Variance:

By allowing the slope to vary, model 3 captures differences in how the effectiveness of $ks2score$ varies across schools. This means that some schools may have students whose GCSE score improves significantly with increases in $ks2score$, while others may not see as much improvement. The introduction of random slopes thus helps explain why some schools are more effective at translating prior scores into future performance.

Within-School Variance:

The model still captures within-school variance through the residual term. However, by accommodating differences in the slopes, model m3 can provide more precise estimates of expected outcomes for students within schools. This allows for a better understanding of individual student performance, as it accounts for the unique context of each school.

Implications for Educational Processes

The variation in slopes reflects contextual influences on student performance, such as teaching quality and resources. For instance, schools with strong tutoring programs may utilize $ks2score$ more effectively than those without. Recognizing that the relationship between $ks2score$ and $gcse$ score varies by school can guide targeted policies. Schools can adopt successful practices from higher-performing peers to enhance student outcomes. The model emphasizes that educational outcomes are shaped not just by individual scores but also by the school context, highlighting the need for a holistic approach that considers both student and school characteristics.

Interpretation and Model Comparison

Model 0 (Q3a)

The Intraclass Correlation Coefficient (ICC) for model 0 is 0.0791721. This means that 7.9172097% of the variability in student performance is explained by the fact that students attend different schools. The

remaining 92.1% of the variance is due to differences within schools, including individual factors such as student characteristics, teaching quality, or other unobserved factors. Alternatively, this can be interpreted as there being a 7.9172097% correlation between the GCSE scores of two randomly chosen students from the same school.

R Squared (Q3b)

Model 1 explains about 49.9124356% more of the variation compared to model 0. Prior achievement, measured by KS2 scores, strongly predicts GCSE performance, suggesting that improving KS2 scores could significantly boost GCSE outcomes.

Likelihood Ratio Tests (Q3c)

```
## Data: dataaw
## Models:
## m0: gcsescore ~ (1 | school)
## m1: gcsescore ~ ks2score + (1 | school)
##      npar   AIC   BIC logLik deviance Chisq Df Pr(>Chisq)
## m0      3 81345 81370 -40670    81339
## m1      4 61681 61714 -30837    61673 19666  1 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model 0 and 1

The degrees of freedom for the comparison between model 1 and 0 is 1, since model 1 has one additional parameter compared to the simpler model 0. The Chi-squared statistic from the Likelihood Ratio Test is 19666. This value reflects the improvement in model fit going from model 0 to model 1. The p-value is extremely low, which suggests that the additional parameter in model m provide a significantly better fit to the data.

Model 1 and 2

```
## Data: dataaw
## Models:
## m1: gcsescore ~ ks2score + (1 | school)
## m2: gcsescore ~ ks2score + ks2_sq + (1 | school)
##      npar   AIC   BIC logLik deviance Chisq Df Pr(>Chisq)
## m1      4 61681 61714 -30837    61673
## m2      5 61612 61654 -30801    61602 71.086  1 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA results show that model 2 has a significantly better fit than model m1, as indicated by the Chi-squared statistic (71.086) and the significant p-value. This suggests that the additional complexity of model m2, which includes the quadratic term, captures important relationships in the data that model m1 does not.

Significance of Squared KS2 Scores (Q3d)

The low p-value ($p < 2.2\text{e-}16$) from the ANOVA results indicates that the `ks2_sq` term is highly significant. This means that the quadratic relationship contributes significantly to explaining the variation in `gcsescore` beyond what is captured by the linear term `ks2score` alone.

The coefficient for `ks2_sq` represents the curvature in the relationship between `ks2score` and `gcsescore`. If the coefficient is positive, it indicates a U-shaped relationship, suggesting that as `ks2score` increases, the effect on `gcsescore` diminishes at first and then increases again. Conversely, a negative coefficient indicates an inverted U-shape, meaning that after a certain point, higher `ks2score` values lead to diminishing returns on `gcsescore`.

The findings suggest that educational interventions or policies should consider the complexity of how prior scores relate to future performance. Depending on the sign of the `ks2_sq` coefficient, strategies might differ for students with low, medium, or high `ks2score`.

Model 3 (Q3e)

The variance in the slopes of `ks2score` across schools in model 3 is 0.0024816. The influence of KS2 scores on GCSE scores slightly varies between schools, though the variation is quite small. This suggests that while schools may differ in how strongly prior achievement predicts future performance, the differences are not substantial.

Likelihood Ratio Test (Model 2 and 3)

```
## Data: dataw
## Models:
## m2: gcsescore ~ ks2score + ks2_sq + (1 | school)
## m3: gcsescore ~ ks2score + ks2_sq + (ks2score | school)
##      npar    AIC    BIC logLik deviance  Chisq Df Pr(>Chisq)
## m2      5 61612 61654 -30801    61602
## m3      7 61548 61606 -30767    61534 67.667  2 2.024e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The difference is significant. Model 3 fits the data significantly better than model 2. This suggests that allowing the slopes for `ks2score` and `ks2_sq` to vary across schools improves the model's ability to predict GCSE scores. This indicates that the relationship between KS2 scores and GCSE scores does vary by school, even though the variation is small, as seen in the plot.

- Squared KS2 Scores

The variation in the random coefficients for `ks2_sq` across schools is minimal. All points are tightly clustered around zero, indicating that the squared term of KS2 scores has a relatively consistent effect across different schools. This suggests that the quadratic term of prior achievement (KS2 squared) does not vary much in its influence on GCSE scores between schools.

- KS2 Scores

Similar to the first panel, the random slopes for `ks2score` (the linear effect of KS2 scores) show little variation across schools. Most of the dots are concentrated near zero, meaning that the effect of prior achievement (KS2 scores) on GCSE scores is fairly uniform across schools. However, the small spread indicates that there are some differences in how KS2 scores affect GCSE outcomes in different schools, but these differences are quite limited.

Random effects

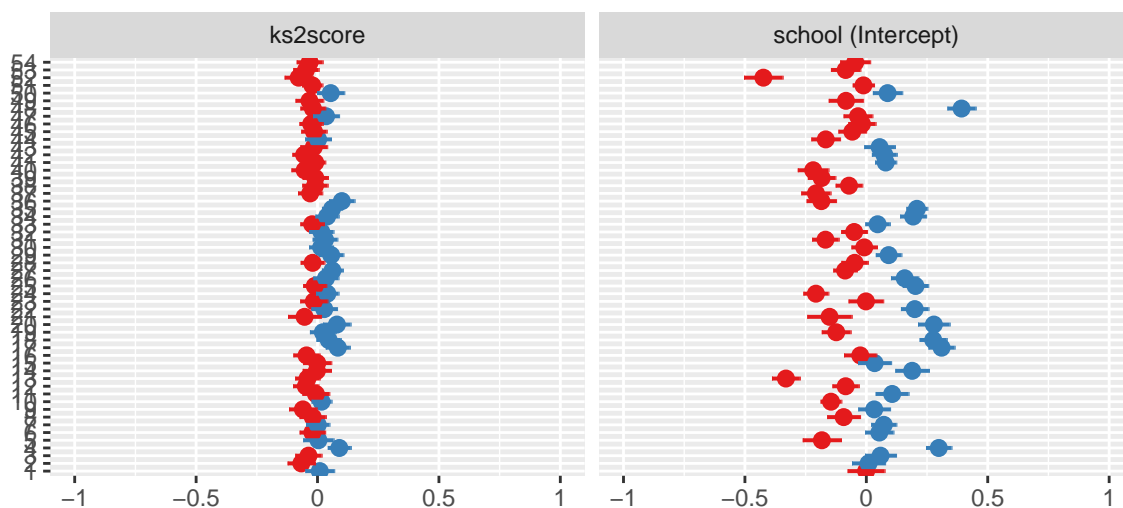
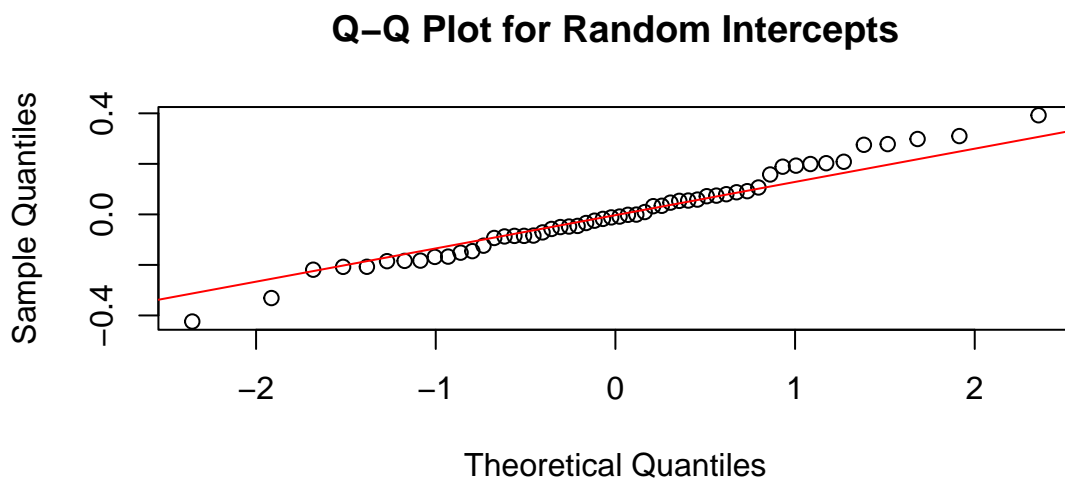


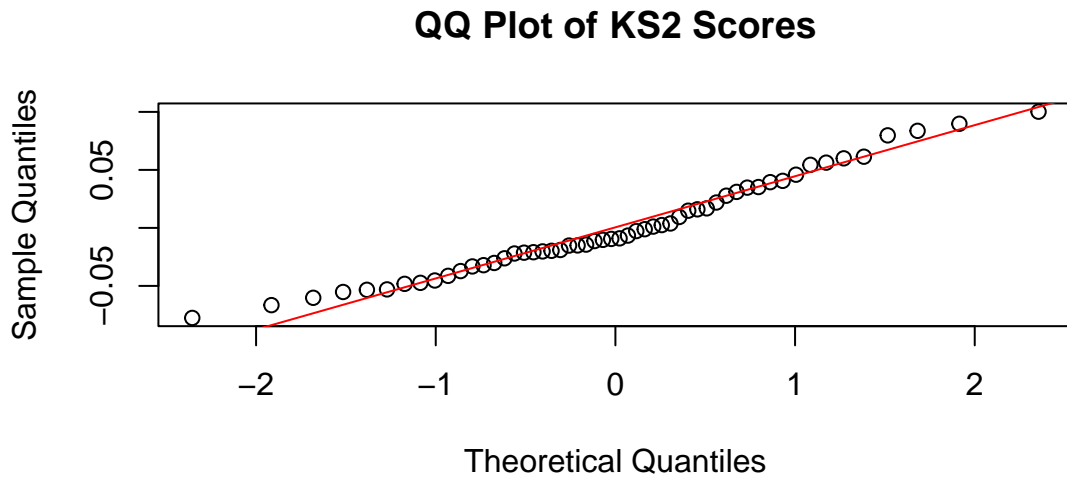
Figure 4: Random slopes and intercept from model 3.

- School (Intercept): The random intercepts for school show more variation compared to the other panels. The dots are spread more widely, indicating that the baseline GCSE scores (i.e., intercepts) differ substantially between schools. This suggests that some schools have higher or lower average GCSE scores compared to others, even after controlling for the effects of KS2 scores. The larger spread here reflects the greater variability in school-specific baseline performance.

Diagnostic Checks and Residual Analysis

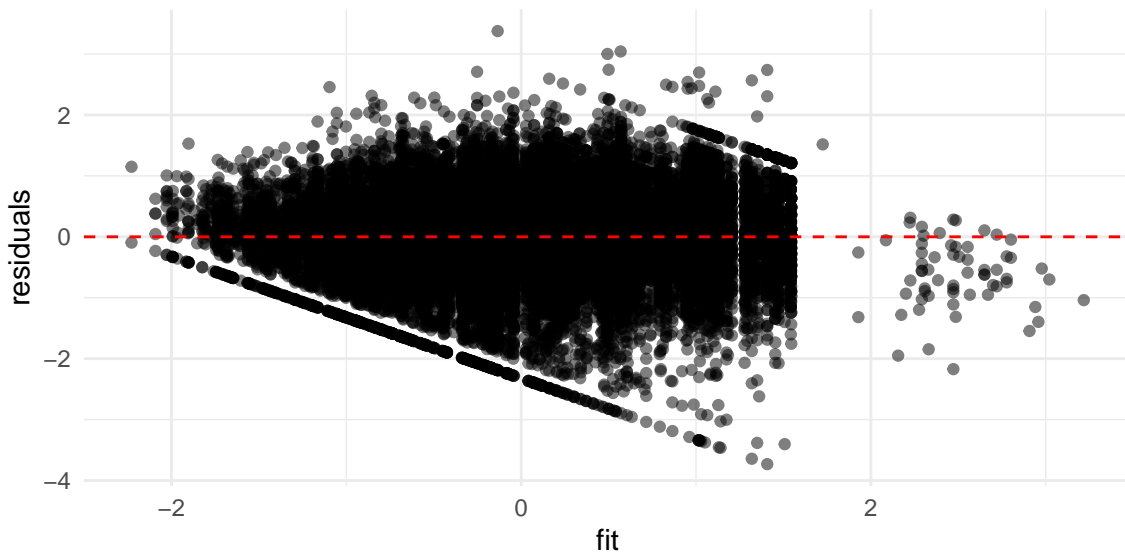
##Q-Q plots (Q4a)





QQ plots for the random effects look normally distributed as they follow the theoretical distribution with no notable curvature or outliers.

Residuals against Fitted Values for Model 3 (4b)



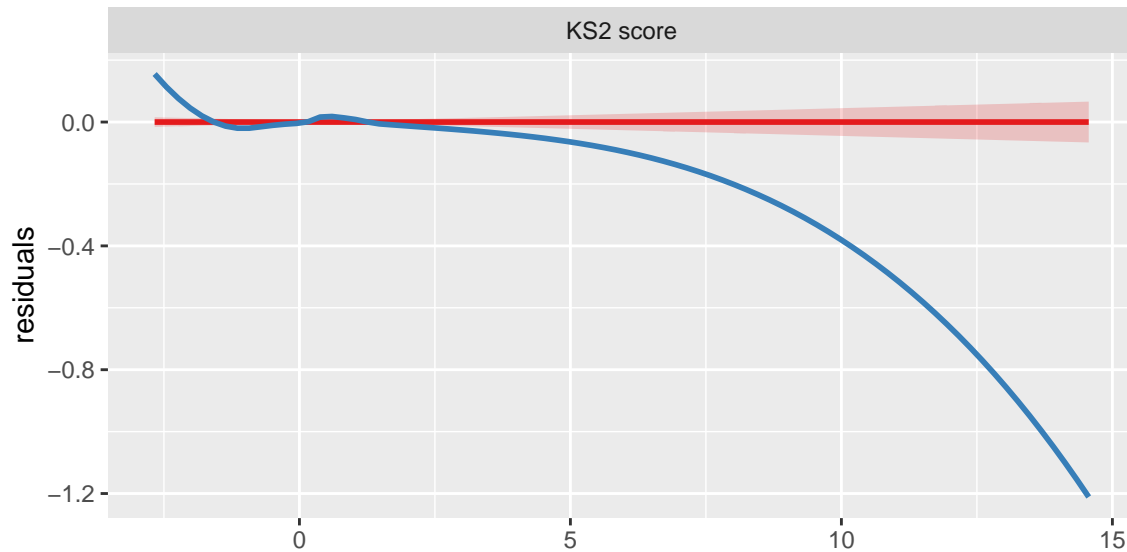
Fan-shaped pattern: The plot has a triangular or wedge-shaped structure, where the spread of residuals increases as the fitted values become larger. This indicates potential heteroscedasticity, meaning that the variance of residuals is not constant across the range of fitted values. It suggests that the model's errors are larger for certain levels of the fitted values.

Outliers: There are clusters of points that deviate from the general pattern, particularly on the right side of the plot. These may indicate outliers or influential data points that could be affecting the model.

Fitted vs. Residuals Relation: Residuals should ideally be randomly distributed around 0, but the spread on one side (both upper and lower boundaries) suggests some systematic bias or structure in the data that is not well captured by the model.

Non-linearity: The plot shows a nonlinear boundary, especially with the extreme lower-left and upper-right clusters. This could suggest that the linearity assumption may be violated, indicating that a more complex model might better capture the underlying relationships.

Residuals against KS2 Scores for Model 3 (4c)



The blue line indicates a non-linear relationship between the residuals and the predictor (KS2 score). The curve is not flat and does not follow the red line; it initially fluctuates around zero, then shows a sharp downward trend as the KS2 score increases.

This suggests that the model is not capturing the non-linear relationship between the KS2 score and the outcome variable well. A linear model might not be appropriate for this data, and a transformation of the predictor or a more flexible model (e.g., polynomial terms, splines) might improve the fit.

This funnel-shaped pattern, where residuals become more negative for larger KS2 scores, suggests heteroscedasticity. The model errors are not constant but vary depending on the level of the KS2 score.

```
knitr::opts_chunk$set(echo=FALSE, cache=FALSE,
  eval=TRUE, message=FALSE,
  warning=FALSE, fig.align='center',
  fig.dim=c(6,3))
```

```
library(rmarkdown)
library(tinytex)
library(checkpoint)
library(knitr)
library(kableExtra)
library(tidyverse)
library(haven)
library(sjPlot)
library(lme4)
```

```
#
```

```

# # source code from the files
# source(file="clean.R")
# source(file="model.R")
data <- read_dta("data/A-1-NPD-dataset2.dta")
# create copy of dataset to clean
dataw <- data
colconv <- c("school", "cohort", "pupil",
             "year2004", "year2005", "year2006",
             "time", "cons")
dataw[,colconv] <- lapply(dataw[,colconv], as_factor)

# Number of schools
nsch <- (length(unique(dataw$school)))

# Number of pupils
npup <- (length(unique(dataw$pupil)))

# Average number of pupils / school
navgpup <- as.integer(npup/nsch)

# GCSE histogram
ggplot(dataw, aes(x=gcsescore))+
  geom_histogram(binwidth = 0.1, fill = "lightblue", color = "black") +
  labs(title = "Histogram of GCSE Scores", x = "GCSE Score", y = "Frequency")

# KS2 histogram
ggplot(dataw, aes(x = ks2score)) +
  geom_histogram(binwidth = 1, fill = "lightblue", color = "black") +
  labs(title = "Histogram of KS2 Score", x = "KS2 Score", y = "Frequency")

# Scatter plot of ks2 and gcse
ggplot(dataw, aes(y = gcsescore, x = ks2score))+
  geom_point(color = "blue") +
  labs(title = "Scatter Plot of GCSE Scores vs. KS2 Scores",
       x = "KS2 Score",
       y = "GCSE Score") +
  theme_minimal()

# Building all 4 models
m0 <- lmer(gcsescore ~ (1 | school), data=dataw, REML=F)
m1 <- lmer(gcsescore ~ ks2score + (1 | school), data=dataw, REML=F)

# add squared ks2 scores
dataw$ks2_sq <- (dataw$ks2score)^2
m2 <- lmer(gcsescore ~ ks2score + ks2_sq + (1 | school), data=dataw, REML=F)
m3 <- lmer(gcsescore ~ ks2score + ks2_sq + (ks2score | school), data=dataw, REML=F)

```

```

# summary(m1)
# summary(m2)
# summary(m3)

# Plot m3 slopes
plot_model(m3, ,type="pred",
            terms=c("ks2score","school"),
            pred.type="re", ci.lvl = NA)

# store variance and corr
var_comp0 <- VarCorr(m0)

# random eff variance
school_var0 <- as.numeric(var_comp0[[1]][1])

# residual variance
res_var0 <- attr(var_comp0, "sc")^2

icc0 <- school_var0/(school_var0 + res_var0)

var_comp1 <- VarCorr(m1)
# random eff variance
school_var1 <- as.numeric(var_comp1[[1]][1])
# residual variance
res_var1 <- attr(var_comp1, "sc")^2

r2 <- 1 - (res_var1 + school_var1) / (res_var0 + school_var0)
anova(m1, m0)
anova(m1, m2)
var_ks2_m3 <- attr(VarCorr(m3)$school, "stddev")[2]^2
anova(m2, m3)
plot_model(m3, type = "re", show.values = F)

# Obtain coefficients of random effects from m3
random_eff3 <- ranef(m3)$school

# QQ plot for random intercepts
qqnorm(random_eff3[, "(Intercept)"], main = "Q-Q Plot for Random Intercepts")
qqline(random_eff3[, "(Intercept)"], col = "red")

# QQ plot for ks2
qqnorm(random_eff3[, "ks2score"], main = "QQ Plot of KS2 Scores")
qqline(random_eff3[, "ks2score"], col = "red")

# Build data frame to plot fitted vs residual plot
model_data <- data.frame(
  fit = fitted(m3),
  residuals = residuals(m3)
)

```

```
ggplot(model_data, aes(x = fit, y = residuals)) +  
  geom_point(alpha = 0.5) +  
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +  
  theme_minimal()  
  
# Plot ks2 scores vs residual plot  
plot_model(m3, type = "resid")
```