Multilevel and Longitudinal Analysis: Assessment 1

Test with open answers

Student ID: 11484265

2024-10-02

Contents

Descriptive Analysis

There are 54 schools and 29506 pupils. The average number of pupils per school is 546.
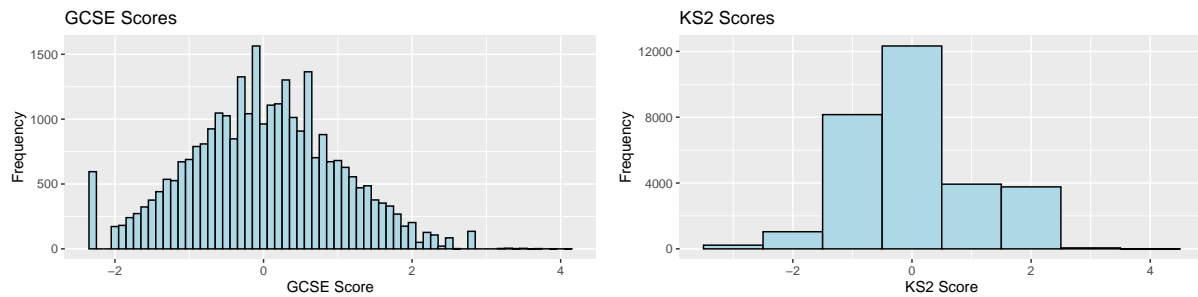
Figure 1: Histograms of GCSE and KS2 scores.

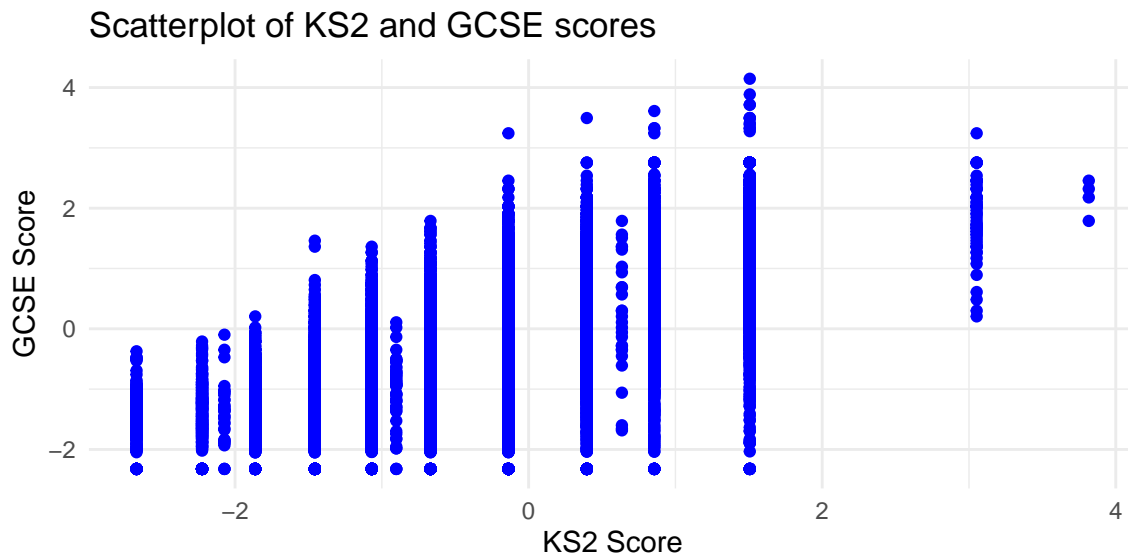There is a slight right skew in both histograms, suggesting a greater mean than median.

Figure 2: Scatterplot of KS2 and GCSE 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.

Model Estimation

Model 0

Table 1: Model 0 Coefficients

	Dependent variable:
	gcsescore
Constant	−0.025
	(0.039)
	t = −0.638
Observations	29,506
Log Likelihood	−40,669.620
Akaike Inf. Crit.	81,345.240
Bayesian Inf. Crit.	81,370.120
Note:	*p<0.1; **p<0.05; ***p<0.01

grp	var1	var2	vcov	sdcor
school	(Intercept)	NA	0.0787218	0.2805741
Residual	NA	NA	0.9155908	0.9568651

The fixed effect shows a non-significant average GCSE score of -0.02462, suggesting no consistent baseline performance across schools. The random effects indicate variability in GCSE scores among schools, with a variance of 0.07872, highlighting the significance of school-level factors. Overall, the model fit suggests that while school differences are important, significant unexplained variability remains in students' scores.

Model 1

Table 3: Model 1 Coefficients

	Dependent variable:
	gcsescore
ks2score	0.718***
	(0.004)***
	t = 167.222***
Constant	−0.001
	(0.023)
	t = −0.023
Observations	29,506
Log Likelihood	−30,836.570
Akaike Inf. Crit.	61,681.140
Bayesian Inf. Crit.	61,714.320
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 4: Model 1 Random Effects

grp	var1	var2	vcov	sdcor
school	(Intercept)	NA	0.0275527	0.1659901
Residual	NA	NA	0.4704743	0.6859113

The fixed effects indicate that the average GCSE score averages to nearly zero when KS2 scores are 0, with each additional point in KS2 resulting in a significant increase of approximately 0.718 points in GCSE scores. The random effects show a variance of 0.02755 for school intercepts, highlighting variability in average scores across schools. Additionally, the residual variance of 0.47047 suggests that there is still considerable unexplained variability in students' GCSE scores.

Impact of KS2 Scores

Including ks2 in model 1 decreased variance in schools 0.05; residual went down by approximately 0.5 compared to model 0. The relationship between "ks2score" and "gcsescore" 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 1 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 gcsescore is assumed to be linear. The coefficient for ks2score is approximately 0.72, indicating that for every one-unit increase in ks2score, gcsescore increases by about 0.72, holding the school effect constant.

Model 2

Table 5: Model 2 Coefficients

	Dependent variable:
	gcsescore
ks2score	0.722***
	(0.004)***
	t = 167.208
ks2_sq	0.028***
	(0.003)***
	t = 8.436
Constant	−0.026
	(0.023)
	t = −1.137
Observations	29,506
Log Likelihood	−30,801.030
Akaike Inf. Crit.	61,612.060
Bayesian Inf. Crit.	61,653.520
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 6: Model 2 Random Effects

grp	var1	var2	vcov	sdcor
school	(Intercept)	NA	0.0276798	0.1663724
Residual	NA	NA	0.4693363	0.6850812

The fixed effects show a non-significant intercept of -0.0264 and a significant positive relationship between KS2 scores and GCSE scores, with each additional point in KS2 resulting in an increase of approximately 0.7219 points in GCSE scores, while KS2 squared indicates a progressively larger effect. The random effects reveal a variance of 0.02768 for school intercepts, indicating variability in average scores across schools, alongside a residual variance of 0.46934 reflecting considerable unexplained variability in GCSE scores.

Model 3

Table 7: Model 3 Coefficients

	Dependent variable:
	gcsescore
ks2score	0.718***
	(0.008)***
	t = 88.614
ks2_sq	0.024***
	(0.003)***
	t = 6.903
Constant	−0.027
	(0.023)
	t = −1.152
Observations	29,506
Log Likelihood	−30,767.200
Akaike Inf. Crit.	61,548.390
Bayesian Inf. Crit.	61,606.440
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 8: Model 3 Random Effects

grp	var1	var2	vcov	sdcor
school	(Intercept)	NA	0.0281393	0.1677476
school	ks2score	NA	0.0024816	0.0498152
school	(Intercept)	ks2score	0.0028789	0.3445107
Residual	NA	NA	0.4672604	0.6835645

The fixed effects indicate a non-significant intercept of -0.0270, while each additional point in KS2 scores results in a significant increase of approximately 0.718 points in GCSE scores, and the positive coefficient for KS2 squared (0.0237) suggests an increasing effect as KS2 scores rise. The random effects reveal a variance of 0.0281 for school intercepts and a variance of 0.0025 for the random slope of KS2 score, indicating moderate variability in school effects.

Impact of Squared KS2 Scores

With the introduction of squared KS2 scores in model 2, the relationship becomes quadratic. The coefficient for squared KS2 scores is positive (0.028), which suggests a U-shaped relationship. This indicates that the effect of ks2score on gcsescore 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 KS2 score is 0.002482, indicating some variability in how the effect of KS2 score on GCSE score differs across schools. This suggests that while there is a general positive trend (captured by the fixed effect of squared KS2 score), 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- and Within-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.

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 KS2 scores more effectively than those without. Recognizing that the relationship between KS2 and GCSE scores 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 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)

Model 0 and 1

	npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr..Chisq.
m0	3	81345.24	81370.12	-40669.62	81339.24	NA	NA	NA
m1	4	61681.15	61714.31	-30836.57	61673.15	19666.1	1	0

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

	npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr..Chisq.
m1	4	61681.15	61714.31	-30836.57	61673.15	NA	NA	NA
m2	5	61612.06	61653.52	-30801.03	61602.06	71.08612	1	0

The ANOVA results show that model 2 has a significantly better fit than model m1, as indicated by the Chi-squared statistic (71.0861247') 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 indicates that squared KS2 scores are highly significant. This means that the quadratic relationship contributes significantly to explaining the variation in GCSE score beyond what is captured by the linear term alone. The coefficient for squared KS2 scores represents the curvature in the relationship.

Since the coefficient is positive, it indicates a U-shaped relationship, suggesting that as KS2 score increases, the effect on GCSE score diminishes at first and then increases again.

The findings suggest that educational interventions or policies should consider the complexity of how prior scores relate to future performance. Strategies might differ for students with different levels of KS2 scores.

Model 3 (Q3e)

Likelihood Ratio Test (Model 2 and 3)

	npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr..Chisq.
m2	5	61612.06	61653.52	-30801.03	61602.06	NA	NA	NA
m3	7	61548.39	61606.44	-30767.20	61534.39	67.66731	2	0

The difference between model 2 and 3 is significant. Model 3 fits the data better than model 2 given the significant chi statistic (71.0861247). Allowing the slopes for KS2 scores to vary across schools enhances the model's predictive ability for GCSE scores, indicating that the relationship between KS2 and GCSE scores differs by school. However, as shown in Figure 3 and 4, the slopes exhibit minimal variation across schools.
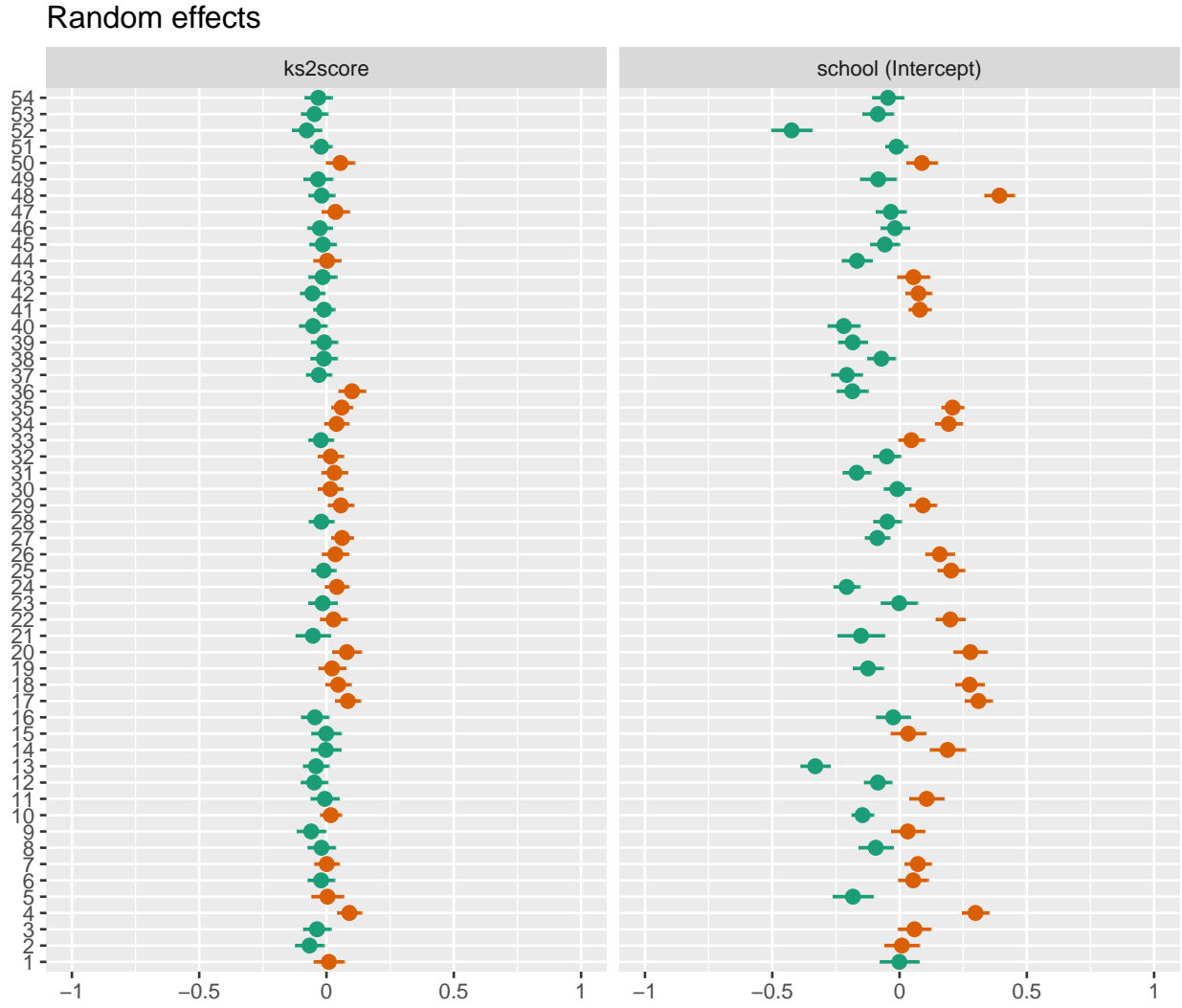

Figure 3: Model 3 Predicted Values

Figure 4: Model 3 random effects distribution.

The variance in the slopes of ks2score across schools in model 3 is 0.0024816. The influence of KS2 scores on GCSE outcomes varies slightly between schools, indicating that while prior achievement predicts future performance differently across schools, these differences are not substantial. The random slopes for KS2 scores show minimal variation, with most values concentrated near zero, suggesting a uniform effect of KS2 scores on GCSE results. In contrast, the random intercepts exhibit greater variability, indicating significant differences in baseline GCSE scores among schools, even after accounting for KS2 effects.

Diagnostic Checks and Residual Analysis

Quantile-Quantile plots (Q4a)

QQ Plot for Random Intercepts

QQ Plot of KS2 Scores
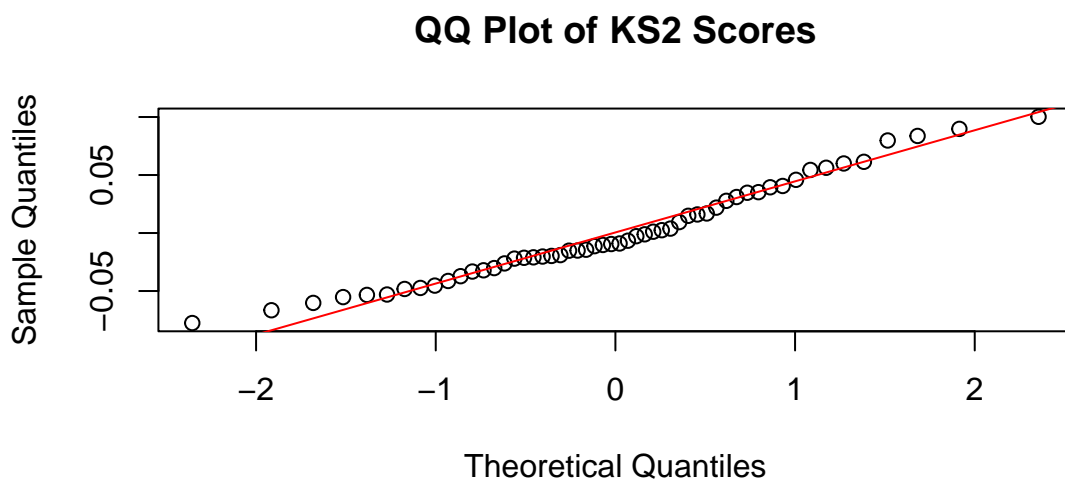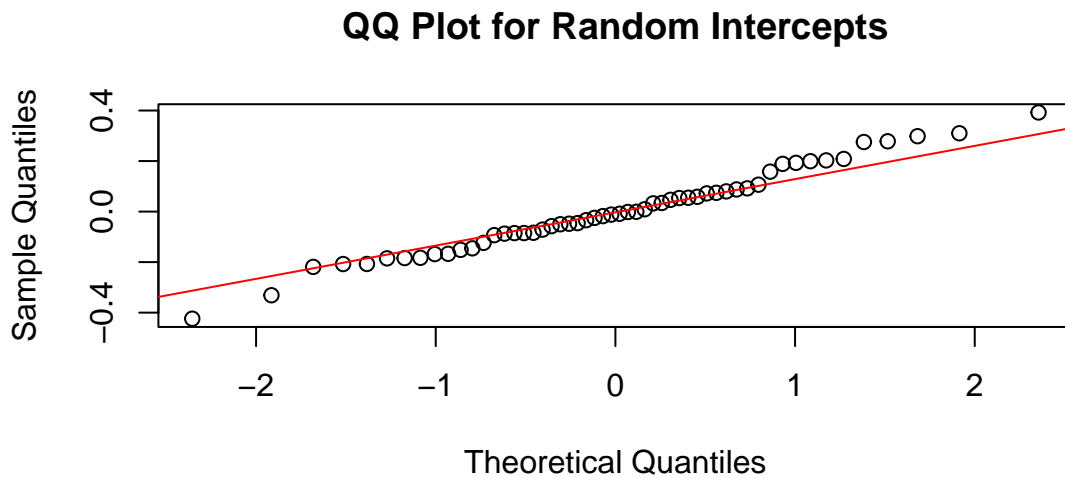
QQ plots for the random effects look normally distributed as they follow the theoretical distribution with no notable curvature or outliers.

Residuals against Fitted Values for Model 3 (Q4b)

The plot exhibits a fan-shaped pattern, where the variance of residuals increases with larger fitted values. indicating potential heteroscedasticity. Clusters of points on the right side indicate possible outliers or influential data points that may be impacting the model. Additionally, the residuals do not appear randomly distributed around zero, showing systematic bias and suggesting that the linearity assumption may be violated, as indicated by the nonlinear boundary in the extreme clusters.
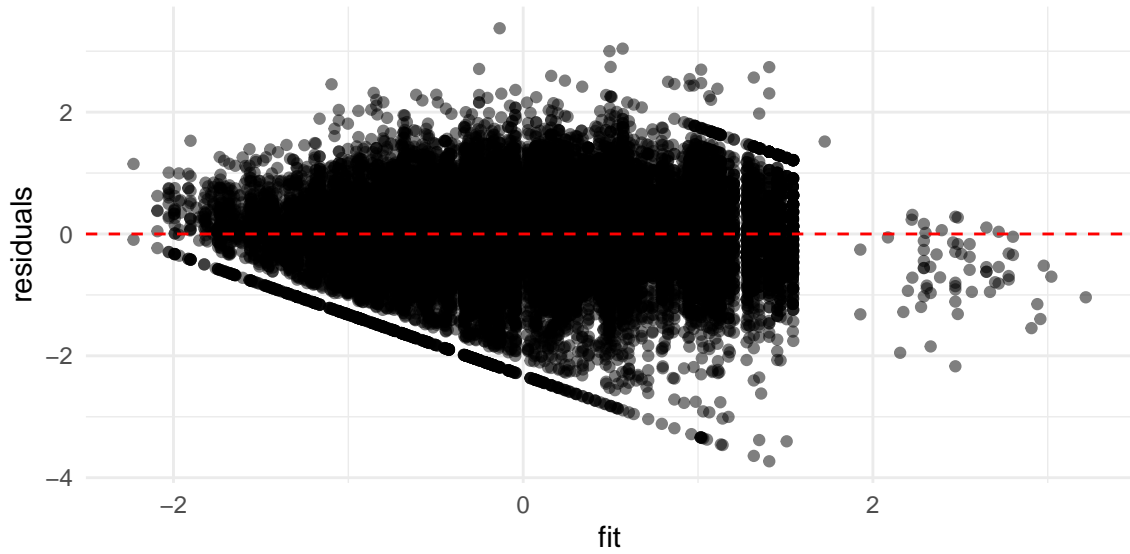
Figure 5: Model 3 fitted values residual plot.

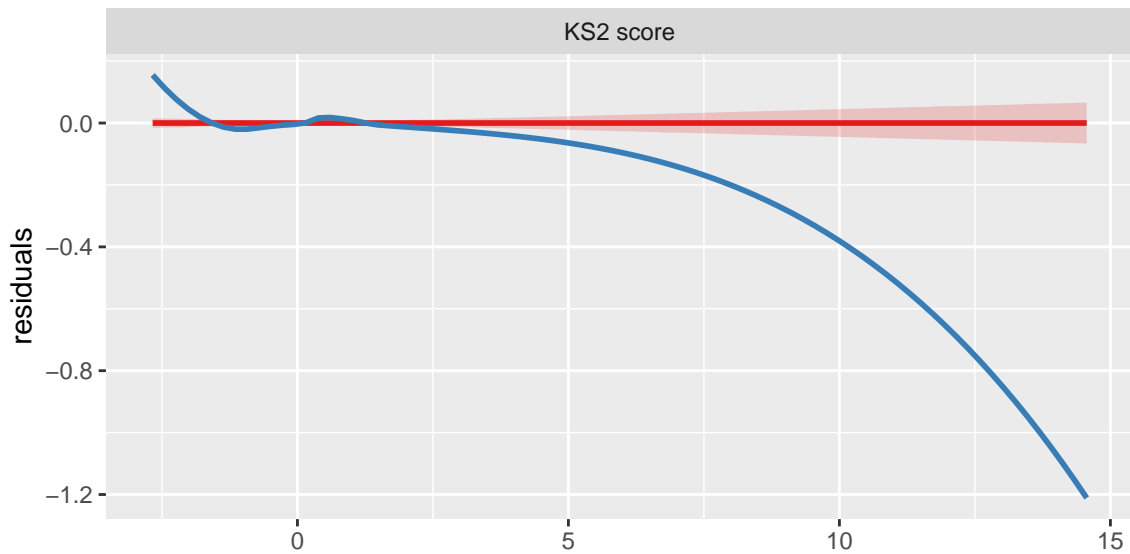## Residuals against KS2 Scores for Model 3 (Q4c)

Figure 6: Model 3 KS2 score residual plot.

The blue line illustrates a non-linear relationship between the residuals and the predictor (KS2 score), initially fluctuating around zero before sharply declining as the KS2 score increases. This indicates that the model fails to adequately capture the non-linear relationship, suggesting that the linear nature of the term may be inappropriate. Additionally, residuals become more negative at higher KS2 scores, indicating heteroscedasticity, and that model errors vary with the level of the KS2 score.

Appendix A: Code

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

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



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 = "GCSE Scores", x = "GCSE Score", y = "Frequency")

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


# Scatter plot of ks2 and gcse
ggplot(dataw, aes(y = gcsescore, x = ks2score))+
```

```r
  geom_point(color = "blue") +
  labs(title = "Scatterplot of KS2 and GCSE 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)


# Report results

# M0 coefficients
stargazer(m0, header=F, report = "v*c*s*t*", title = "Model 0 Coefficients")

# M0 variances
kable(VarCorr(m0), booktabs = T, title = "Model 0 Random Effects",
  longtable = T) %>%
  kable_styling(position = 'center', font_size = 10,
                latex_options = c('striped',
                                  'hold_position'))


# M1 coefficients
stargazer(m1, header=F, report = "v*c*s*t*", title = "Model 1 Coefficients")

# M1 variances
kable(VarCorr(m1), booktabs = T, caption = "Model 1 Random Effects",
  longtable = T) %>%
  kable_styling(position = 'center', font_size = 10,
                latex_options = c('striped',
                                  'hold_position'))


# M2 coefficients
stargazer(m2, type = "latex", header=F, report = "v*c*s*t", title = "Model 2 Coefficients")

# M2 variances
kable(VarCorr(m2), booktabs = T, caption = "Model 2 Random Effects",
  longtable = T) %>%
  kable_styling(position = 'center', font_size = 10,
                latex_options = c('striped',
                                  'hold_position'))
```

```r
# M3 coefficients
stargazer(m3, type = "latex",  header=F, report = "v*c*s*t", title = "Model 3 Coefficients")

# M3 variances
kable(VarCorr(m3), booktabs = T, caption = "Model 3 Random Effects",
  longtable = T)  %>%
  kable_styling(position = 'center', font_size = 10,
                latex_options = c('striped',
                                  'hold_position'))


# 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_comp, "sc")^2

r2 <- 1 - (res_var1 + school_var1) / (res_var0 + school_var0)
anv01 <- data.frame(anova(m1, m0))
kable(anv01, booktabs = T) %>%
  kable_styling(position = 'center', font_size = 10,
                latex_options = c('striped',
                                  'hold_position'))
anv12 <- data.frame(anova(m1, m2))
chi12 <- anv12$Chisq[2]

kable(anv12, booktabs = T) %>%
  kable_styling(position = 'center', font_size = 10,
                latex_options = c('striped',
                                  'hold_position'))

var_ks2_m3 <- attr(VarCorr(m3)$school, "stddev")[2]^2


anv23 <- data.frame(anova(m2, m3))
chi23 <- anv12$Chisq[2]

kable(anv23, booktabs = T) %>%
  kable_styling(position = 'center', font_size = 10,
                latex_options = c('striped',
                                  'hold_position'))
```

```r
# Plot m3 slopes
plot_model(m3, type="pred",
           terms=c("ks2score","school"),
           pred.type="re", ci.lvl = NA,
           colors = "Dark2") +
  theme(legend.position = "none")


plot_model(m3, type = "re", show.values = F, , colors = 'Dark2')


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

# QQ plot for random intercepts
qqnorm(random_eff3[, "(Intercept)"], main = "QQ 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")
```