report

Student ID: 11484265

2024-09-25

Descriptive Analysis

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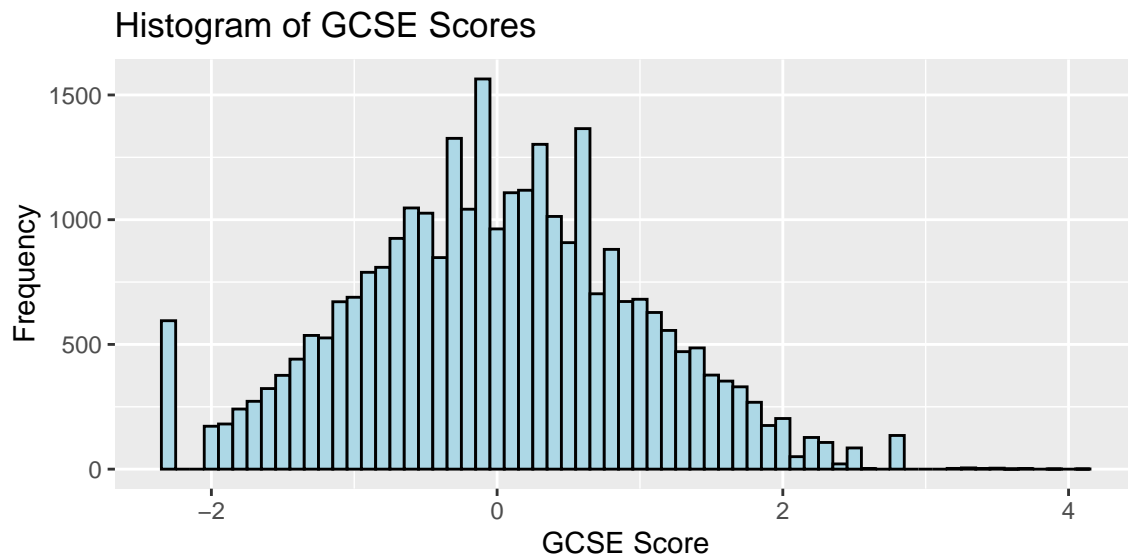
Histogram of GCSE Scores

Figure 1: Histogram of GCSE point scores.
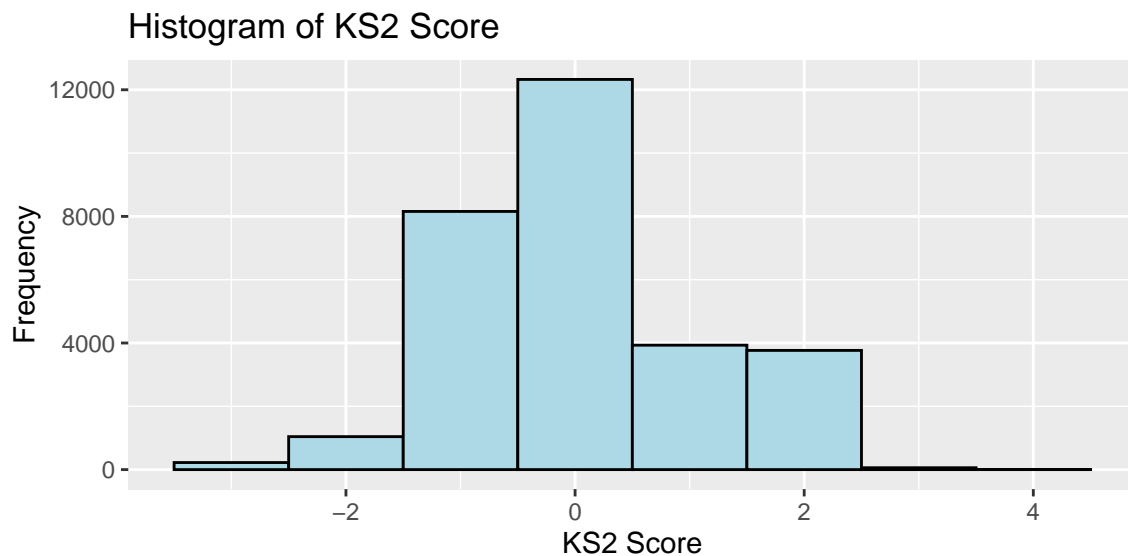
Histogram of KS2 Score

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.
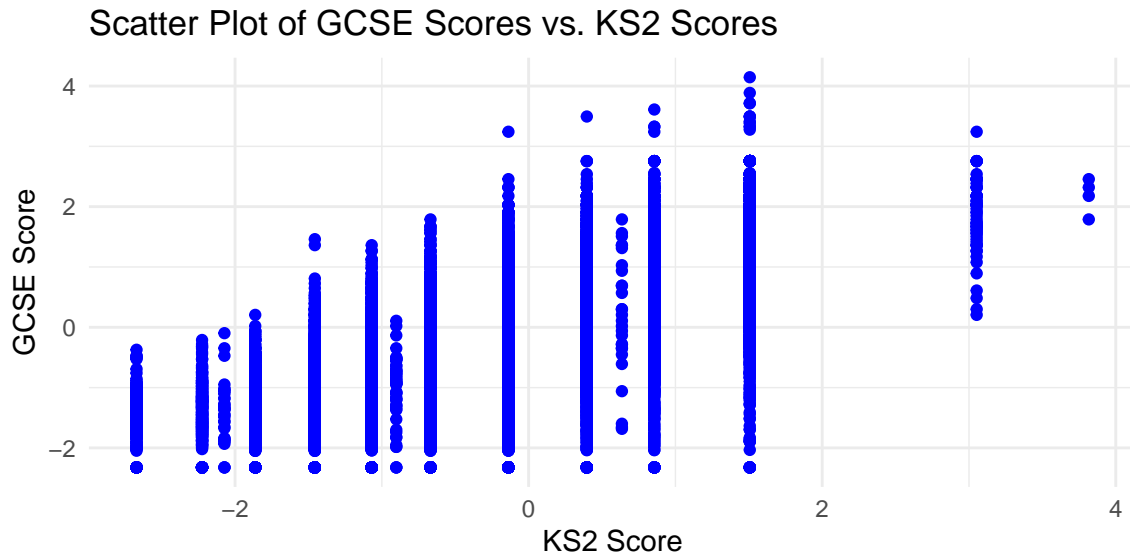
Scatter Plot of GCSE Scores vs. KS2 Scores

Figure 3: Scatterplot of KS2 and GCSE scores.

Model Estimation

Predicted values of GCSE score

-Model 0 (m0): An empty model with "gcsescore" as the dependent variable and a random intercept at the school level. -Model 1 (m1): m0 + a fixed effect for "ks2score". -Model 2 (m2): m1 + "ks2score" squared to capture non-linear effects. -Model 3 (m3): m2 + a random slope for "ks2score" to allow the relationship between "ks2score" and "gcsescore" to vary by school.

Compare how the inclusion of "ks2score" and its squared term "ks2score2" affects the model fit and the interpretation of the relationship between "ks2score" and "gcsescore". Specifically, analyse how the introduction of non-linearity in Model 2 improves the model fit over Model 1.

Including ks2 in model 1 decreased variance in schools 0.05; residual went down by apprximately 0.5.

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 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 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.

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 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.

evaluate the significance and impact of allowing the slope of "ks2score" to vary by school in Model 3. Discuss how these modifications influence the explanation of variance between schools and within schools and how they reflect the underlying educational processes.

The variance for the random slope of ks2score is 0.002482, indicating some variability in how the effect of ks2score on gcsescore 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.

Contextual Factors: 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.

Educational Interventions: Recognizing that the relationship between ks2score and gcsescore varies by school can guide targeted policies. Schools can adopt successful practices from higher-performing peers to enhance student outcomes.

Reflection of Underlying Educational Processes Individual Differences: 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.

Question 3: Interpretation and Model Comparison (40%)

-Q3a: Calculate the Intraclass Correlation Coefficient (ICC) for m0 and interpret it. Provide both interpretations (variance explained by schools and the average correlation between pupils within the same school).

-Q3b: What proportion of variation is explained by m1 compared to m0? Calculate and interpret R2.

m1 explains about ~0.5 more of the variation. This suggests that prior achievement (as measured by ks2score) plays a significant role in predicting students' future exam scores. This strong relationship suggests that interventions aimed at improving ks2score could lead to meaningful improvements in gcsescore.

Q3c: Conduct a Likelihood Ratio Test (LRT) comparing m1 and m0. What is the score, degrees of freedom, and which model does the test suggest choosing? Perform a similar test for m2 versus m1 and explain the reasoning behind the model choice.

The degrees of freedom for the comparison is 1. This indicates that the more complex model (m1) has one additional parameter compared to the simpler model (m0). The Chi-squared statistic is 19666. This value reflects the improvement in model fit when moving from the simpler model (m0) to the more complex model (m1). The p-value ($\Pr(>\text{Chisq})$) is $< 2.2e\text{-}16$, which is extremely low. This suggests that the additional parameter(s) in model m1 provide a significantly better fit to the data than model m0.

```
## Data: dataw
## 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
```

Q3d: Interpret the significance and meaning of the "ks2score2" coefficient in m2. How does this inform your choice between m2 and m1?

The very low p-value ($p < 2.2e\text{-}16$) 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 ANOVA results show that model m2 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.

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.

```
## Data: dataw
## 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
```

Q3e: Discuss the variation in the slope of "ks2score" across schools in m3. Perform an LRT comparing m3 and m2. Is the difference significant? Plot the random coefficients and interpret the results.

The difference is significant.

```
## 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
```
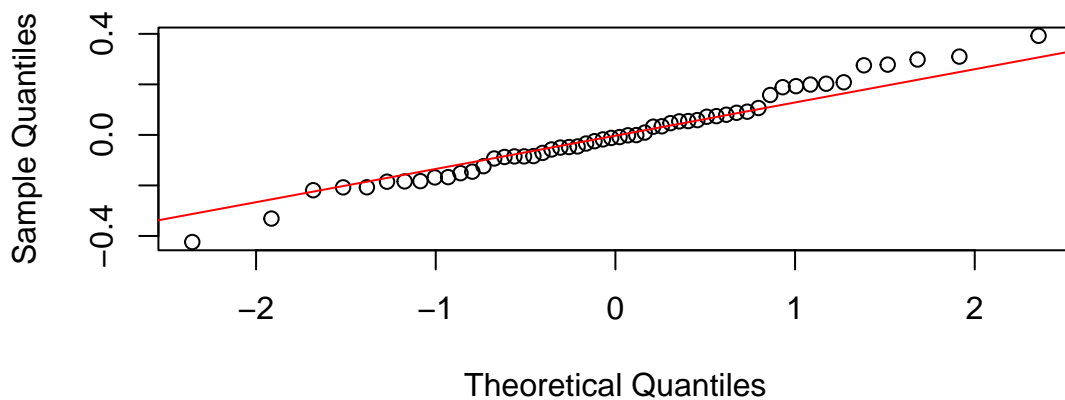

Random effects
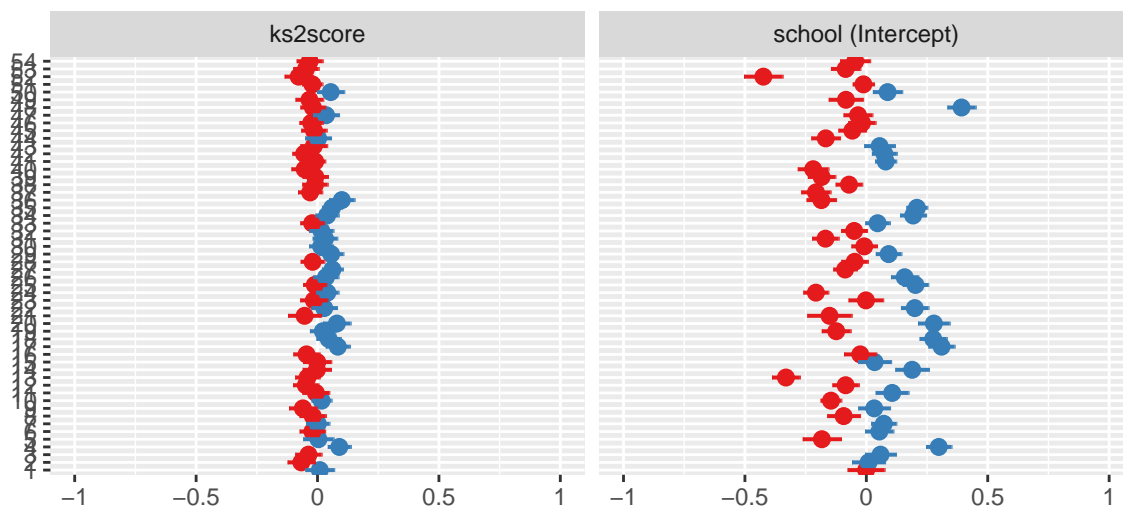
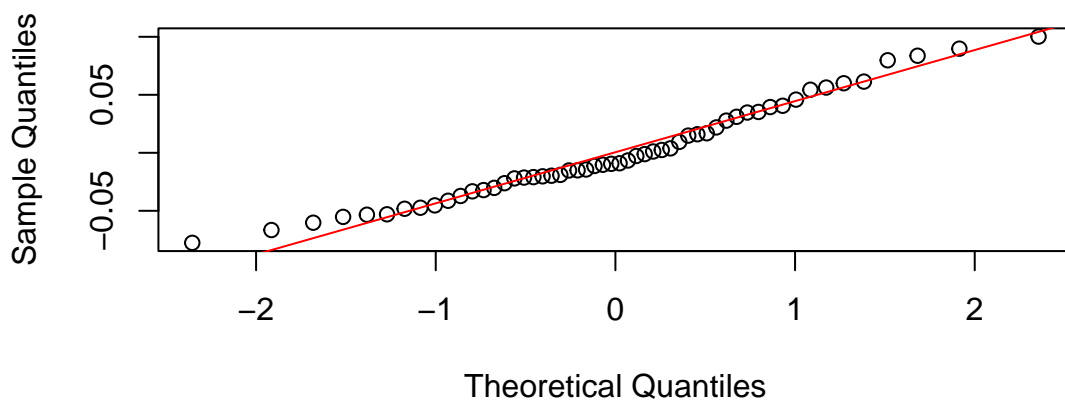# Question 4: Diagnostic Checks and Residual Analysis (20%)

4a: Create Q-Q plots for the random effects in m3 and interpret them. Are the random effects normally distributed?

They look normally distributed as they follow the qq line.

Q–Q Plot for Random Intercepts

## Q–Q Plot for Random Slopes (ks2score)

4b: Plot residuals against fitted values for m3. What does the plot suggest about the model fit?

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.
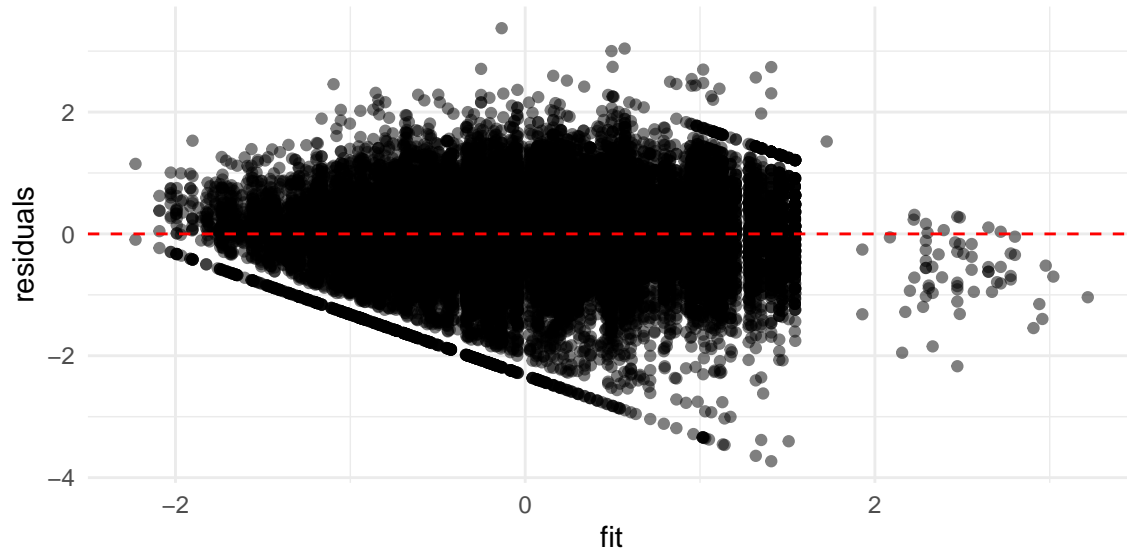
4c: Plot level 1 residuals against "ks2score". What does this plot reveal about the model, particularly regarding linearity and homoscedasticity?

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.