An Analysis into Differences in Educational Attainment within the UK and in comparison with France

Sian Brady s2144034 and Katy Reid s2069134

30 November 2023

Word count: 3990

Contents

1 Executive Summary

Section 4 of this report researches the differences in academic achievement of 15-year-olds between UK nations in 2015 and 2018. Data is extracted from the Programme for International Student Assessment (PISA) covering a range of variables such as sex, immigrant background, and school type. PISA tests are designed so that two-thirds of students achieve scores of between 400-600 marks [1]. Around 16,000 observations were obtained and modelled separately for reading, maths, and science. The following conclusions were found:

- English students are most successful in all three subjects, with Welsh students scoring an average of 24 marks lower, Scottish 38.5 marks lower, and Northern Irish students achieving the lowest marks (75 lower than English students), showing that region influences academic achievement.
- There is minimal difference between 2015 and 2018; year does not play a factor in academic achievement. Figure 1 displays average results in each nation in 2015 and 2018.
- Estimated scores are largely inaccurate for Northern Ireland, which suggests further factors or biases in the dataset.

Section 5 looks into the differences in achievement dependent on age when starting primary education between the UK and France. Separate data is again extracted from PISA, with different variables such as community size and age when starting early education introduced. After modelling around 27,000 observations separately for the same three subjects, these conclusions were obtained:

- Children starting primary education aged 5 achieved the best results in all subjects. Academic attainment gets worse for children starting primary education older than this; the optimum range is ages 4-7. Children starting primary education aged 9 or older achieve, on average, 83 marks less than those starting at age 5.
- In the UK, age when starting primary education affects attainment less than in France. Figure 2 shows average test scores against starting age in both countries. The differences for France are more obvious than for the UK, however overall average scores are very similar.

Figure 1: Average score over all subjects throughout the UK in 2015 and 2018.

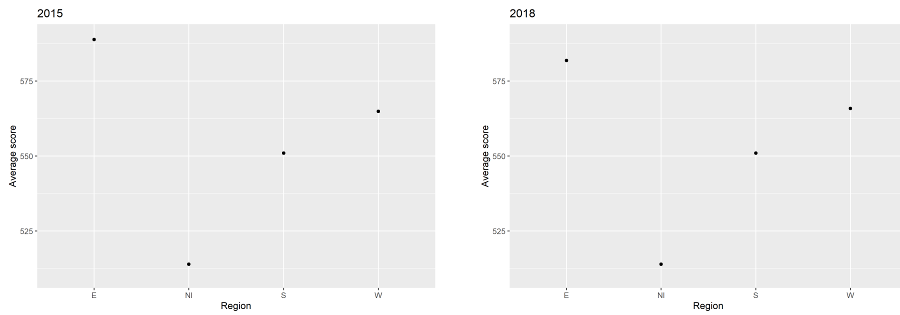
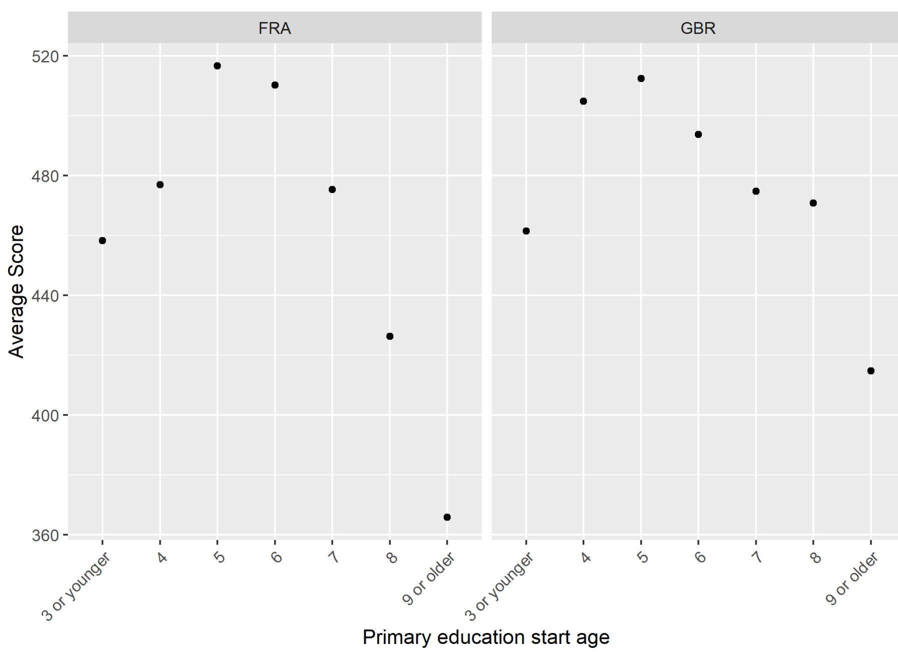


Figure 2: Average score for different primary education start ages.



2 Introduction

Differences in academic attainment within the UK and in comparison to France are investigated using the Programme for International Student Assessment (PISA) 2015 and 2018 data [2]. We will carry out a detailed analysis in order to answer the following:

Which factors are associated with educational attainment, and what is the nature of the associations?

How do the PISA results differ between the UK nations (England, Scotland, Northern Ireland, and Wales), and how have they changed between 2015 and 2018?

Is there a significant relationship between the age that a child starts primary education and their test score, and does this differ between France and the UK?

Factors such as school type (public or private), familial wealth, and gender are likely to influence educational attainment. The aim of this project is to analyse test score data and identify key factors associated with academic attainment. By identifying factors that influence results, future changes can be made to the education system that help maximise the potential of all students.

3 Sampling Design

The Organisation for Economic Co-operation and Development's (OECD) programme for comparing educational attainment in 15-year-old students worldwide, PISA, conducts a survey every three years covering scores in three main subjects and surrounding factors. [3] contains information about the sample design of the 2018 survey.

Systematic probability proportional to size (PPS) sampling is used. A two-stage stratification involves schools which have 15-year-old students. These are assigned to mutually exclusive groups based on characteristics (explicit strata) and sampled systematically from a comprehensive national list of all PISA-eligible schools (sampling frame). The probability of a school being selected is proportional to the estimated number of 15-year-olds enrolled in that school.

The second stage of stratification involves the students within the school. The sampling frame is a complete list of eligible students in the school, and a sample of around 42 participants are selected with equal probability. All students are chosen if the school has fewer than 42 eligible students.

Being group number 5, we used PV5MATH, PV5SCIE, and PV5READ as our response variables. As such, we immediately filtered the data-set to remove all columns PV#MATH, PV#SCIE, PV#READ where $\# \neq 5$. We ignored sampling weights and measurement errors associated with the response variables for these tasks.

4 Exploring differences in educational attainment between the UK nations

4.1 Missing Data

Our response variable was reading, maths, or science scores, and we focused on the 15 explanatory variables in Table 1.

We began with 27975 observations for the UK, however, this included observations with missing data as shown in Figure 3. There were no missing values for subject score, year, or region, the main variables we were investigating.

The continuous variables missing the most data were percentage of disadvantaged students and number of males/females enrolled in a school. We considered the correlation between these and ESCS, which had less missing data. The correlations were -0.3, 0.07 and 0.08 for disadvantaged students, number of females, and number of males, respectively. As very weak correlations, ESCS did not explain these variables well. We also evaluated box-plots against the categorical variables, and saw no strong association, so we removed all observations with NA values for disadvantaged and the number of enrolled males/females.

School type and mother's/father's education were also missing many data points, so we investigated if other variables could explain these. No strong association was found, so we removed all observations with NAs for these variables. This left us with 16196 observations, of which 498 were missing data. This was only 3% of the observations, so we removed these.

Our final data set had 15698 observations. Removing all observations with missing data can introduce bias, as missing data may not be random.

Figure 3: Missing Data.

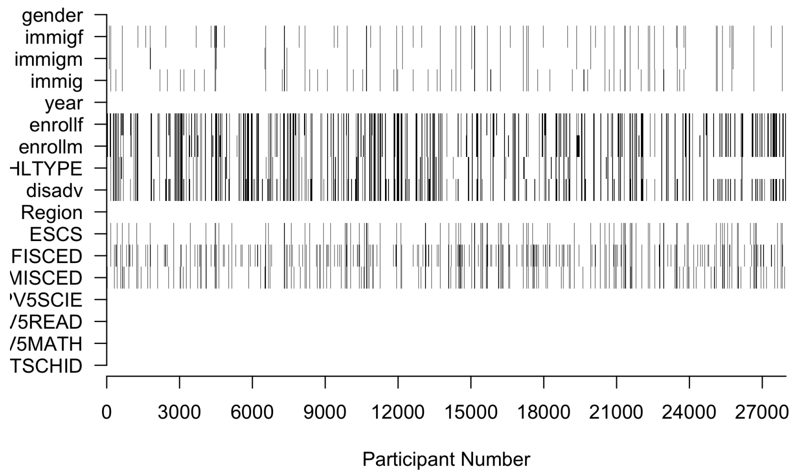


Table 1: Variable Definitions.

R Variable	Description
CNTSCHID	School identifier
PV5MATH	Standardised score (maths)
PV5READ	Standardised score (reading)
PV5SCIE	Standardised score (science)
MISCED	Mother's ISCED education level
FISCED	Father's ISCED education level
ESCS	PISA index of economic, social and cultural status
Region	England, Northern Ireland, Wales, or Scotland
disadv	Percentage of students from socioeconomically disadvantaged homes
SCHLTYPE	Public, Private (independent) or Private (government-dependent)
enrollm	Number of males in the school
enrollf	Number of females in the school
year	2015 or 2018
immig	Immigration Status
immigm	Immigration Status (mother)
immigf	Immigration Status (father)
sex	Male or Female
mixedsex	Is the school mixed?

4.2 Graphical Analysis

We first explored association between test scores and various factors.

Figure 4 compares subject scores between sexes. The box-plots suggested very little difference in science, but females did better than males in reading and vice-versa for maths.

Differences in scores between regions were more obvious in 2015 than in 2018, as shown in Figure 5a. England and Wales saw a decrease between years, while Scotland and Northern Ireland saw an increase. England was by far highest in both years and all subjects, and Wales mainly lowest.

Figure 5b shows that for both sexes, average scores in all three subjects were higher in single-sex schools. Females had more variation than males and did slightly better in single-sex schools.

There were obvious differences in average scores between schools, as seen in Figure 5c. Private independent had the highest scores in all subjects, and public the lowest.

From Figure 5d, we saw that in both years, subject score had a negative association with percentage of disadvantaged students.

Figure 4: Average Subject Score by Sex.

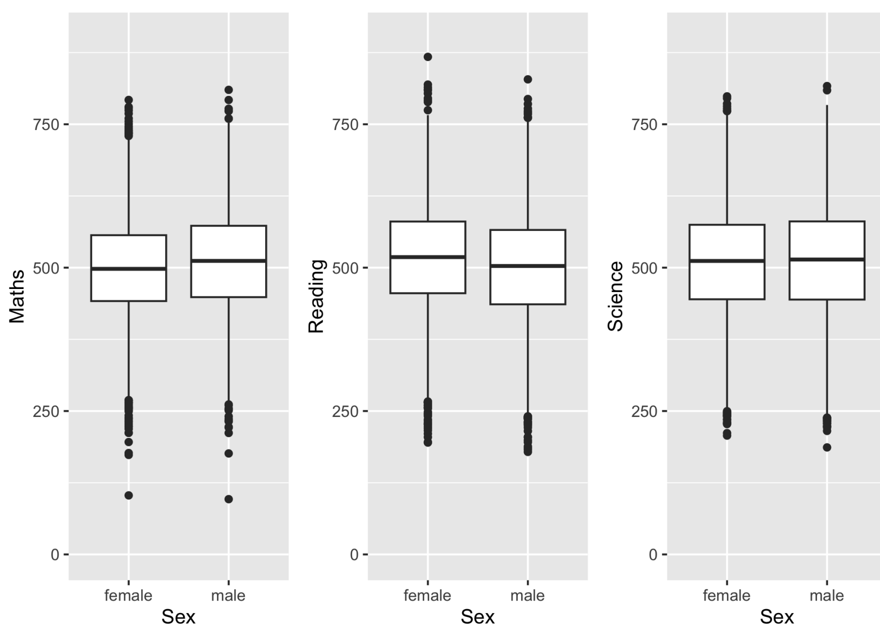
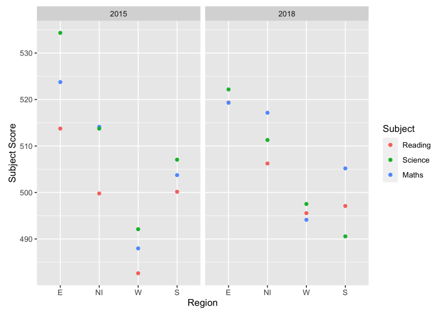
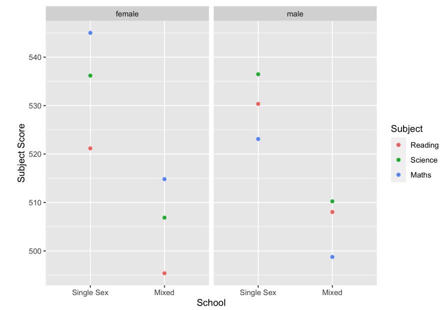


Figure 5: Exploring Subject Scores against Variables.

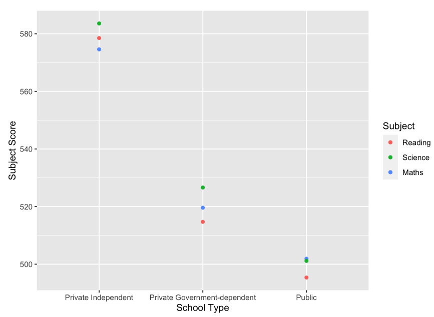
(a) Average Subject Score by Region.



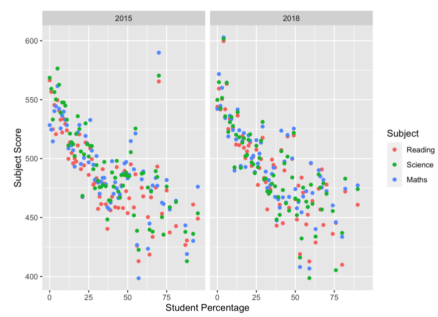
(b) Average Subject Score for Mixed and Single-sex Schools.



(c) Average Subject Score by School Type.



(d) Subject Score by percentage of Socioeconomically Disadvantaged Students.



These further graphs explore associations between the factors.

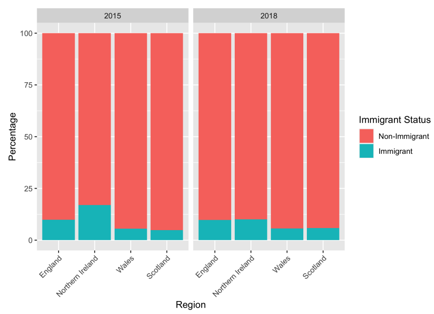
From Figure 6a, we saw in both years that a higher percentage of immigrants studied in England and Northern Ireland than in Wales and Scotland. There were fewer immigrants in Northern Ireland in 2018 than 2015, which could indicate an association between year and immigration in Northern Ireland.

Percentage of parents with immigrant status against education level is shown in Figure 6b. The graphs are U shaped and differ somewhat; the education level with the highest proportion of immigrants was level 3 for mothers and fathers, and the lowest at level 0.

Figure 7 strongly suggests that region influences school type. There were many more private schools in England than any other region; of these, the majority were government-dependent.

Figure 6: Exploring Immigrant Status of Students and Parents.

(a) Proportion of Immigrants to Non-immigrants by Region.



(b) Proportion of Parents with Immigrant and Non-immigrant Status at each ISCED Education Level.

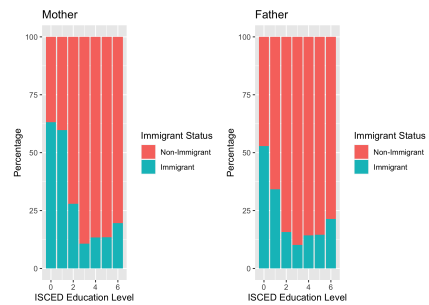
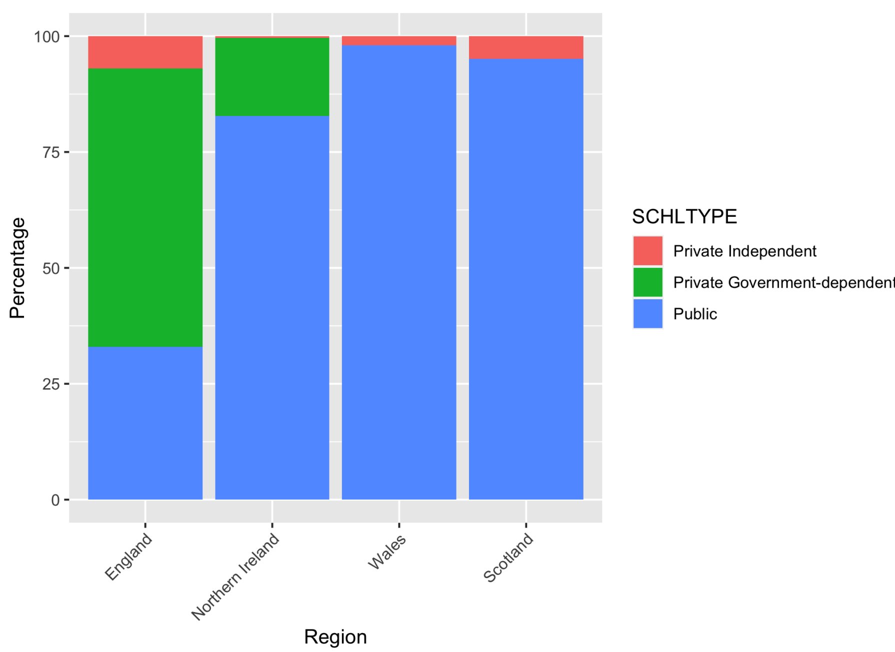


Figure 7: Proportion of Public and Private Schools by Region.



4.3 Modelling Process

We assumed that responses Y_i were uncorrelated, with common variance σ^2 and expectations of the form $\mathbb{E}(Y_i|x_{i,1}, x_{i,2}, ..., x_{i,j}) = \beta_0 + \beta_1 x_{i,1} + ... + \beta_j x_{i,j}$. This was written as

$$\mathbb{E}(\mathbf{Y}|\mathbf{X}) = \mathbf{X}\beta$$

where $\mathbf{X}$ and β were given by

$$\mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,30} \\ 1 & x_{2,1} & \cdots & x_{2,30} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,2} \end{pmatrix}, \qquad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_30 \end{pmatrix}$$

The columns in $\mathbf{X}$ represent a variable or interaction as detailed in Table 2.

Table 2: Design matrix columns.

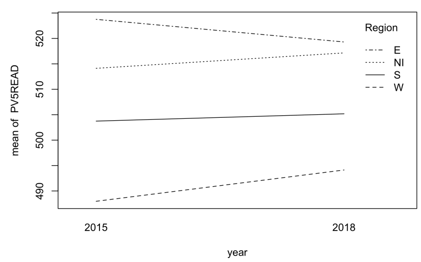
$x_{i,0}$	1	$x_{i,16}$	genderMale
$x_{i,1}$	MISCED	$x_{i,17}$	mixedsexTRUE
$x_{i,2}$	FISCED	$x_{i,18}$	RegionNI:year2018
$x_{i,3}$	ESCS	$x_{i,19}$	RegionW:year2018
$x_{i,4}$	RegionNI	$x_{i,20}$	RegionS:year2018
$x_{i,5}$	RegionW	$x_{i,21}$	RegionNI:SCHLTYPEPrivate Government-dependent
$x_{i,6}$	RegionS	$x_{i,22}$	RegionNI:SCHLTYPEPublic
$x_{i,7}$	disadv	$x_{i,23}$	RegionW:SCHLTYPEPublic
$x_{i,8}$	SCHLTYPEPrivate Government-dependent	$x_{i,24}$	RegionS:SCHLTYPEPublic
$x_{i,9}$	SCHLTYPEPublic	$x_{i,25}$	RegionS:genderMale
$x_{i,10}$	enrollm	$x_{i,26}$	RegionW:genderMale
$x_{i,11}$	enrollf	$x_{i,27}$	RegionNI:genderMale
$x_{i,12}$	year2018	$x_{i,28}$	RegionS:immigimmi
$x_{i,13}$	immigimmi	$x_{i,29}$	RegionW:immigimmi
$x_{i,14}$	immigfimmi	$x_{i,30}$	RegionNI:immigimmi
$x_{i,15}$	immigmimmi	$x_{i,31}$	CNTSCHID

4.4 Reading Results

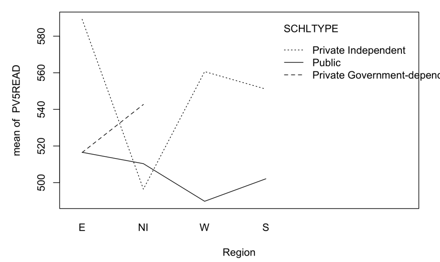
We began by considering variables that might interact. Figure 8a shows that England worsened from 2015 to 2018, and all other regions improved at different rates, implying an interaction between region and year. Conducting an ANOVA test concluded that the interaction between region and year was significant at the 5% level, hence we included this interaction in our model. Other logical interactions were school type with region, shown in Figure 8b, and sex with year, shown in Figure 8c. These were both significant interactions included in the model.

Figure 8: Interactions included in the Reading Model.

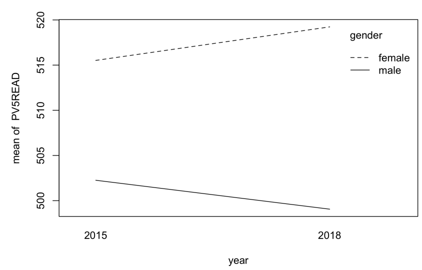
(a) Year and Region.



(b) Region and School Type.



(c) Year and Sex.



The initial model involved the variables detailed in Table 1 and the three interactions above. We used backward selection based on AIC to reduce the model and checked the model based on p-values. Two variables were removed due to having insignificant p-values - mother's immigration status ($x_{i,15}$) and school identifier ($x_{i,31}$). This resulted in a final model of:

$$\mathbb{E}(\mathbf{Y}|\mathbf{X}) = \mathbf{X}\hat{\beta}$$

where $\mathbf{X}$ was the design matrix as defined in Section 4.3 and $\hat{\beta}$ was the vector of regression coefficients, $\hat{\beta_{15}} = \hat{\beta_{31}} = 0$.

Modelling assumptions were verified (see Figure 9). The residual vs fitted plot showed an even spread above and below zero, so the assumption of independence of residuals was valid. A scale-location plot and Q-Q plot were further produced; no trend in scale-location suggested the assumption of constant variance held and the Q-Q plot had minimal deviation, so the assumption of normally distributed residuals was also valid. No observations had high residual and leverage that would disproportionately impact the model. From this, we concluded that a linear model was appropriate.

Figure 10 shows the regression estimates and confidence intervals (CI). The most accurate estimates were those with the smallest CI, namely number of males in the student's school (0.013, 95% CI [0.007, 0.019]) and number of females in the student's school (0.018, 95% CI [0.012, 0.024]). The estimate for students in Northern Ireland studying in private government-dependent schools was 88.303, with a 95% CI [35.446, 141.161]. This was the widest CI, suggesting that the estimate is imprecise with a wide margin of error.

Using interaction estimates we looked at the influence of nations between years. England in 2015 was the dummy variable and had the highest estimate. Northern Ireland was lowest, -77.417, while Scotland fell higher at -38.667 and Wales just higher at -24.763. England and Wales saw a decrease in predicted scores between 2015 and 2018, of -7.415 and -1.649, respectively, whereas Northern Ireland and Scotland saw improvements of 3.805 and 6.675 respectively. Despite these varied trends, the order of highest to lowest estimates in 2018 remained the same.

The adjusted R-squared 0.1401 tells us that this model explained 14.01% of the variability in the data.

Figure 9: Reading Modelling Assumptions.

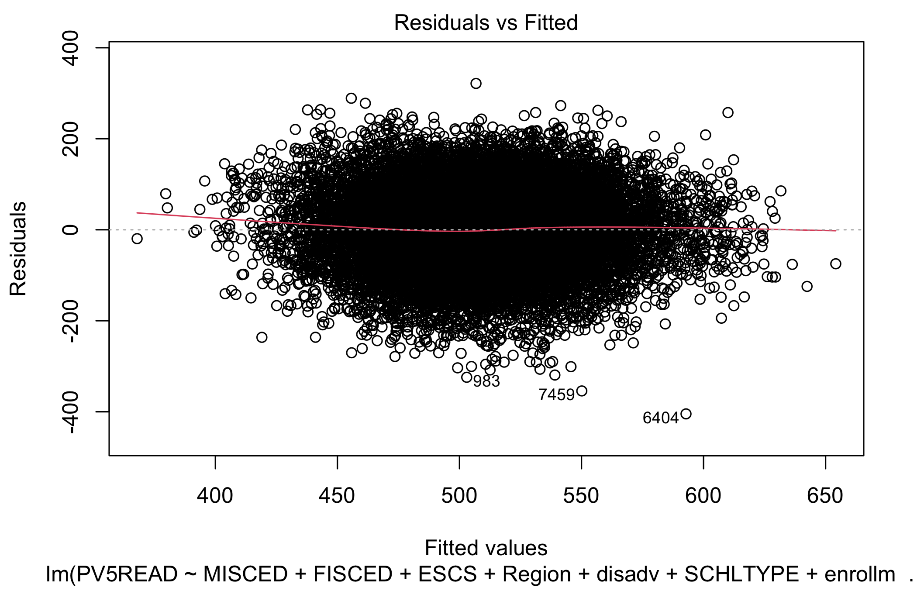
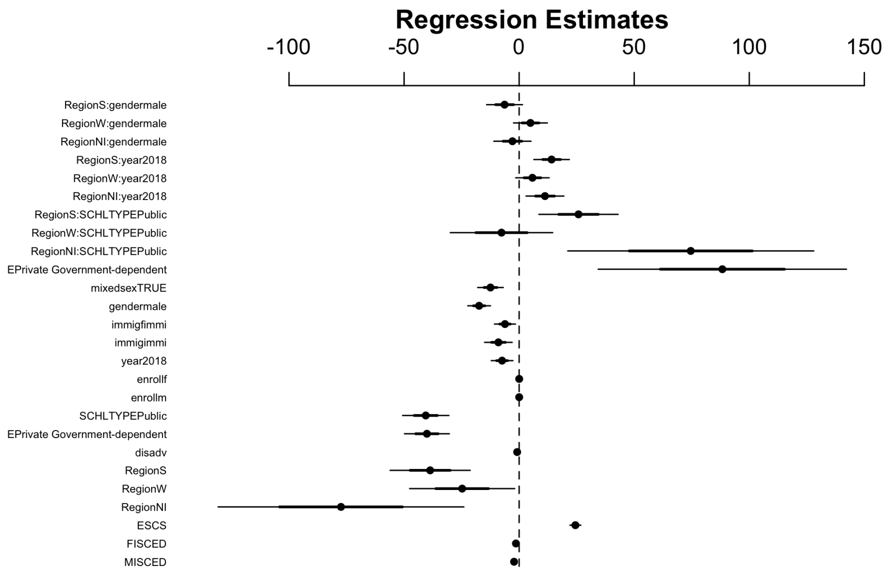


Figure 10: Reading Model Coefficient Estimates and Confidence Intervals.



4.5 Maths Results

The interactions in figure 11a show that all regions except Scotland had improved scores between 2015 and 2018, albeit at different rates. This implies that there was an interaction between region and year. Interaction plots for school type with region (Figure 11b) and immigration status with region (Figure 11c) also showed differing trends. Carrying out ANOVA tests supported both plots at the 5% level to conclude that the interactions were significant.

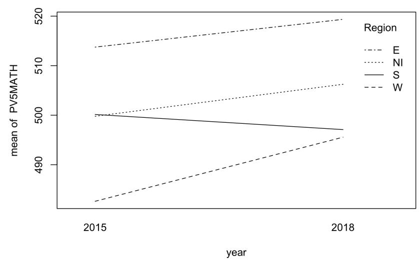
The initial model contained the response variables in Table 1 and the three interactions above. Backward selection was used to reduce the model. The interactions between region and year ($x_{i,18}$, $x_{i,19}$, $x_{i,20}$) and father's immigration status ($x_{i,14}$) were removed due to AIC comparisons, and school identifier was removed ($x_{i,31}$) based on an insignificant p-value. Our final model was

$$\mathbb{E}(\mathbf{Y}|\mathbf{X}) = \mathbf{X}\hat{\beta}$$

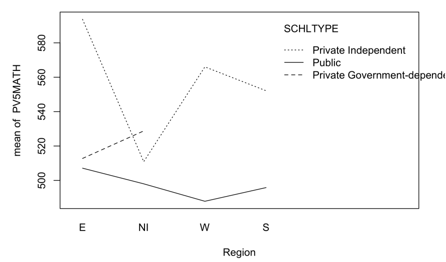
where $\mathbf{X}$ was the design matrix as defined in Section 4.3 and $\hat{\beta}$ was the vector of regression coefficients, $\hat{\beta_{14}} = \hat{\beta_{18}} = \hat{\beta_{19}} = \hat{\beta_{20}} = \hat{\beta_{31}} = 0$.

Figure 11: Interactions included in Maths Model.

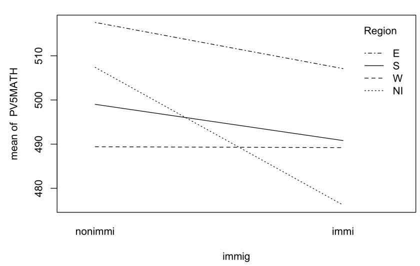
(a) Region and Year.



(b) School type and Region.



(c) Immigration Status and Region.



Assumptions were verified for the final model. Figure 12 displays the residuals against fitted values - the assumption of independent residuals was valid, as the red line aligns closely to the dashed line, and the spread of residuals tells us that homoscedasticity also held. A Q-Q plot was produced to confirm further that the residuals were approximately normally distributed.

Figure 13 displays the regression estimates and CI. As for the reading model, number of males in the student's school (0.018, 95% CI [0.013, 0.023]) and number of females in the student's school (0.009, 95% CI [0.003, 0.014]) were the most accurate estimates. Students in Northern Ireland studying in private government-dependent schools (65.944, 95% CI [17.796, 114.092]) had the widest CI.

There were similar trends to the reading model in terms of differences between nations. England had the highest estimate, Northern Ireland had the lowest estimate, -59.730, Scotland fell higher at -37.035 and Wales just higher at -21.665. Since this model had no interaction between region and year, all regions increased by an estimated 3.498 in 2018.

The adjusted R-squared value 0.1722 means this model explained 17.22% of the data's variability.

Figure 12: Maths Modelling Assumptions.

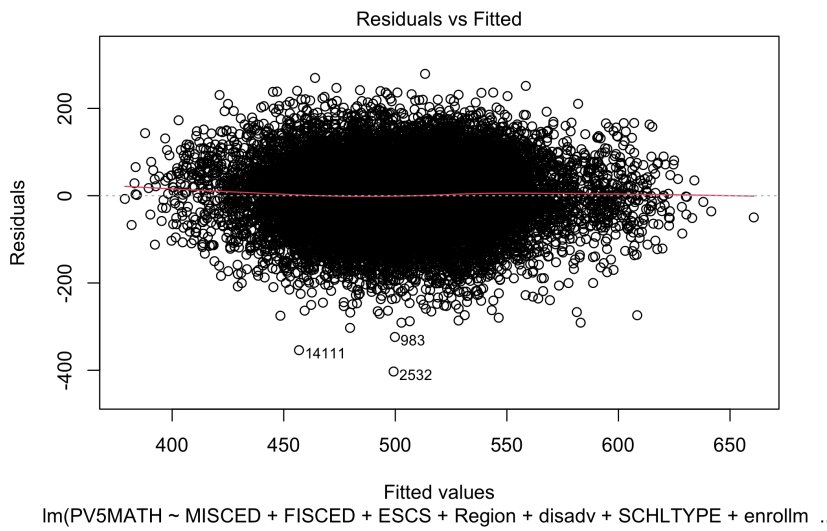
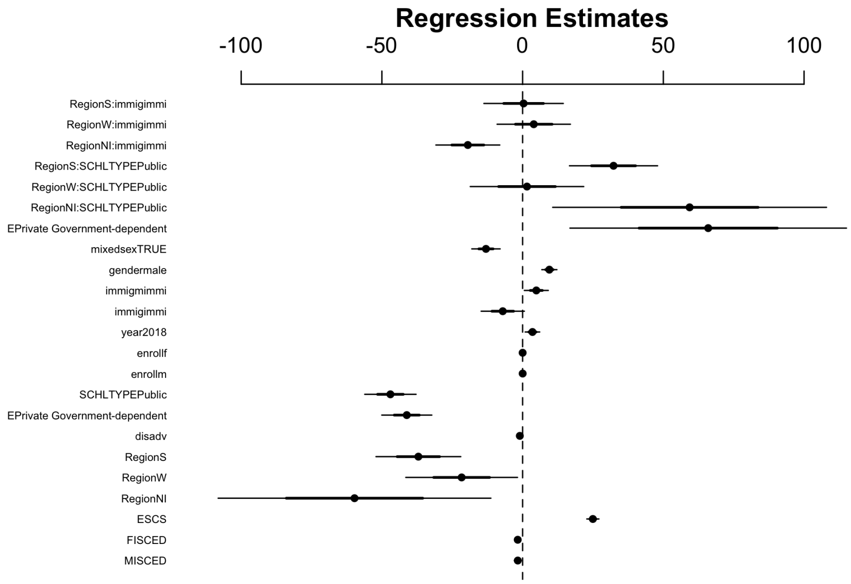


Figure 13: Maths Model Coefficients and Confidence Intervals.



4.6 Science Results

Finally, we modelled the students' science scores. Figure 14a shows the interaction between year and region. Similarly to the reading and maths scores, we saw different trends for the four regions across the two years, and so this interaction was included. Results from an ANOVA test supported this, providing a p-value significant at the 5% level. We also chose to include the interaction between year with students immigration status (Figure 14c) and school type with region (Figure 14b) based on interaction plots and significant ANOVA p-values.

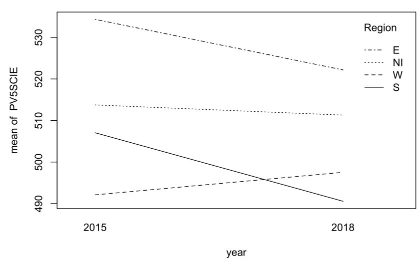
The first full model contained the 15 variables detailed in Table 1 and the three interactions above. As for the previous models, we used backward selection based on AIC and p-values to select our model. We removed the variables school identifier ($x_{i,31}$), mother's immigration status ($x_{i,15}$), and gender ($x_{i,16}$). This gave a final model:

$$\mathbb{E}(\mathbf{Y}|\mathbf{X}) = \mathbf{X}\hat{\beta}$$

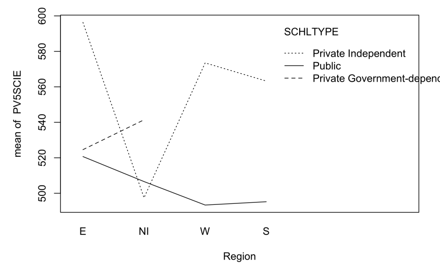
where $\mathbf{X}$ was the design matrix as defined in Section 4.3 and $\hat{\beta}$ was the vector of regression coefficients, $\hat{\beta_{15}} = \hat{\beta_{16}} = \hat{\beta_{33}} = 0$.

Figure 14: Interactions included in Science Model.

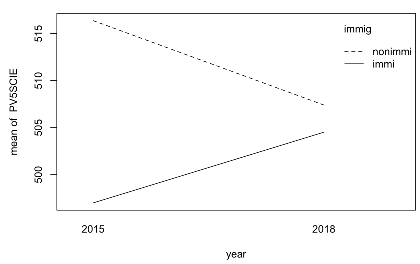
(a) Year and Region.



(b) School Type and Region.



(c) Immigration Status and Region.



We then checked the modelling assumptions in Figure 15. The assumptions of independent residuals, constant variance and of normally distributed residuals were valid. No one observation had a disproportionate impact on the model.

Figure 16 shows the model coefficients and CI. There are similar trends to the previous two models. Number of males in the student's school (0.012, 95% CI [0.007, 0.018]) and number of females in the student's school (0.011, 95% CI [0.006, 0.017]) were the most accurate estimates, and the estimate for students in Northern Ireland studying in private government-dependent schools (87.888, 95% CI [35.448, 140.327]) was least precise.

In 2015, England again had the highest estimate, and Northern Ireland had the lowest estimate (-89.641), with Scotland higher at -37.578 and Wales just higher at -25.182. All nations saw a decrease in 2018 but by different amounts; England -15.678, Northern Ireland -4.382, Wales -2.784, and Scotland -8.767.

The adjusted R-squared value was 0.1653, so 16.53% of the variability in the data was explained in this model.

Figure 15: Science Modelling Assumptions.

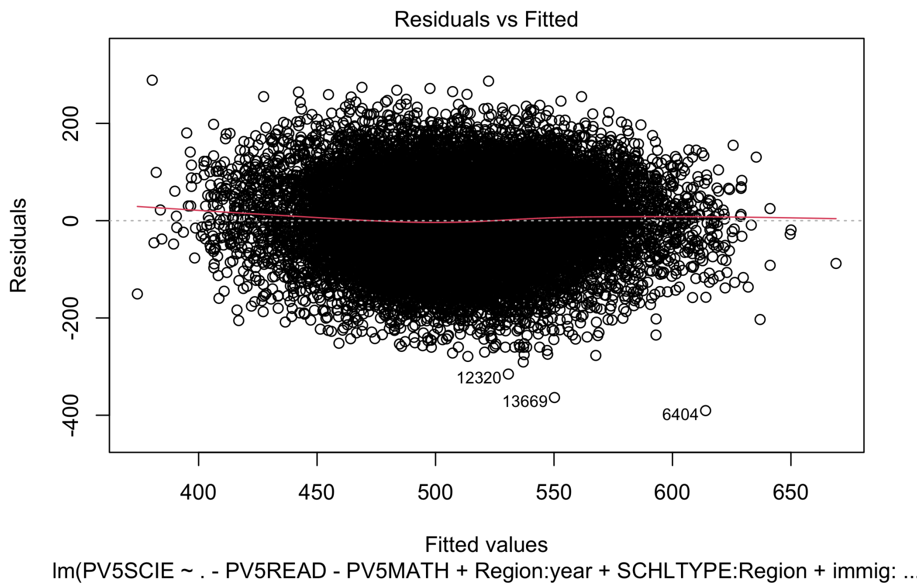
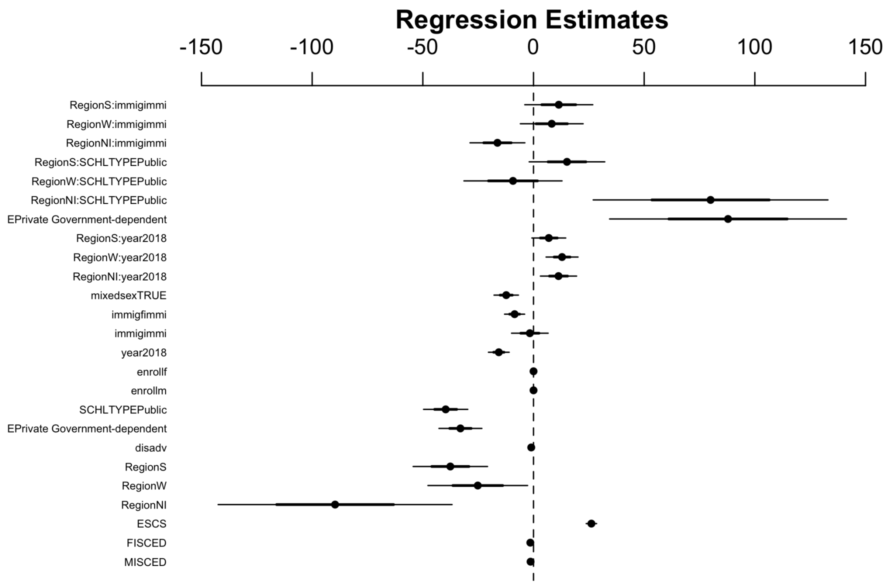


Figure 16: Science Model Coefficient Estimates and Confidence Intervals.



5 Exploring differences in educational attainment between France and the UK

5.1 Missing Data

We then investigated whether a significant relationship exists between the age at which a child starts primary education and their test score, and whether this differs between France and the UK. We collected data from the PISA database covering the UK and France, reusing some variables from part one and introducing others which were relevant. The variables are detailed in Table 3.

The total number of observations for the UK was 27975, and for France 12416, giving 40391 total observations. 3479 observations were missing data regarding the age at which a child starts primary education. As this was our main explanatory variable, we removed all these observations. Figure 17 shows the remaining missing data.

Most missing data points occurred for school type, community and percentage of disadvantaged students. School type and community are categorical variables, so we investigated box plots of these against other variables. These showed no clear associations, suggesting that other variables did not represent school type and community well and should be kept in our model. We therefore dropped all observations with missing values for school type and community.

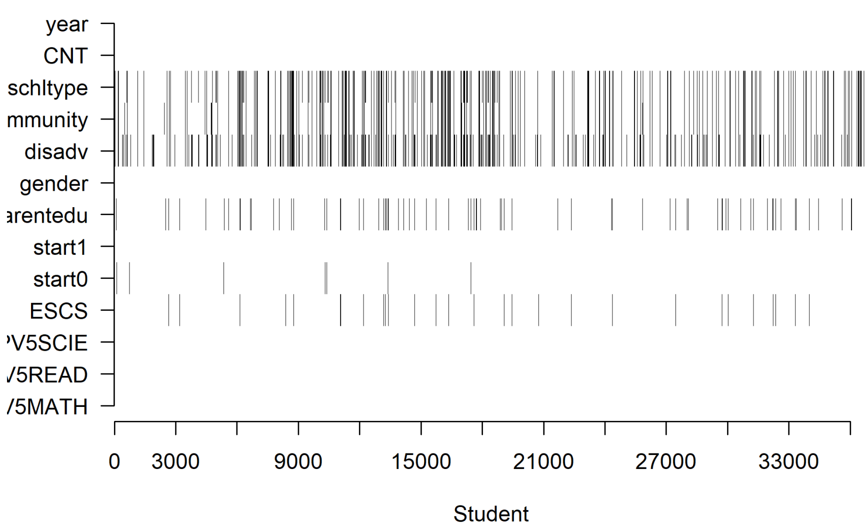
We also removed missing observations from percentage of disadvantaged students - a correlation of -0.34 with ESCS and box-plots against categorical variables showed that other factors did not explain it well. This left us with 1377 observations with missing data, which was approximately 5% of the observations, so we removed these.

We then had 27589 observations, a large sample size. Removing missing values can introduce bias, as missing data may not be random.

Table 3: Variable Definitions.

R Variable	Description
PV5MATH	Standardised score (maths)
PV5READ	Standardised score (reading)
PV5SCIE	Standardised score (science)
ESCS	PISA index of economic, social and cultural status
start 0	How old were you when you started ISCED level 0?
start 1	How old were you when you started ISCED level 1?
parentedu	Highest education level of parents
gender	Student gender
disadv	Percentage of students from socioeconomically disadvantaged homes
community	Community in which the school is located
schltype	Public, Private (independent) or Private (government-dependent)
CNT	Country Code
year	2015 or 2018

Figure 17: Missing Data.



5.2 Graphical Analysis

Figure 18 displays the mean subject scores in each country between years. The UK had higher scores in 2018, but there was lots of variation in 2015.

We then plotted subject scores by gender in Figure 19. Males did better than females in maths and science, and by a wider margin in maths. The plots suggested greater consistency for females, as the quartiles were narrower.

The plots of subject scores against percentage of disadvantaged students (Figure 20) show differing trends between countries. For France, the plot suggested a linear relationship and variation was fairly wide throughout. The plot for GBR resembled more of a concentrated curve, with variation increasing with percentage of disadvantaged students.

Figure 21 displays the proportions of ages when starting early/primary education by country. The majority of students in France started level 0 at age 3, while in the UK, starting age was more of a bell curve between ages 1-5. For starting level 1, France had a majority at age 6; the plot for the UK peaks at age 5. These plots tell us that most children in GBR started primary education a year earlier than in France.

We looked at average subject score separated by age when starting primary education in Figure 24. France followed a curve, peaking at age 5, with the lowest score for those starting aged 9 or older. There was a less distinct shape for GBR; still a curve, but more even across ages 4-8. These plots suggest that country affects average score, but this may not differ between subjects.

Figure 18: Average Subject Score by Country.

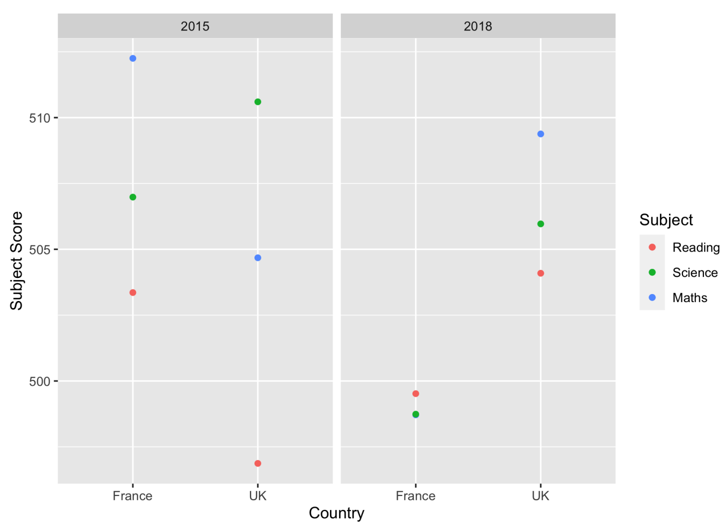


Figure 19: Average Subject Score by Sex.

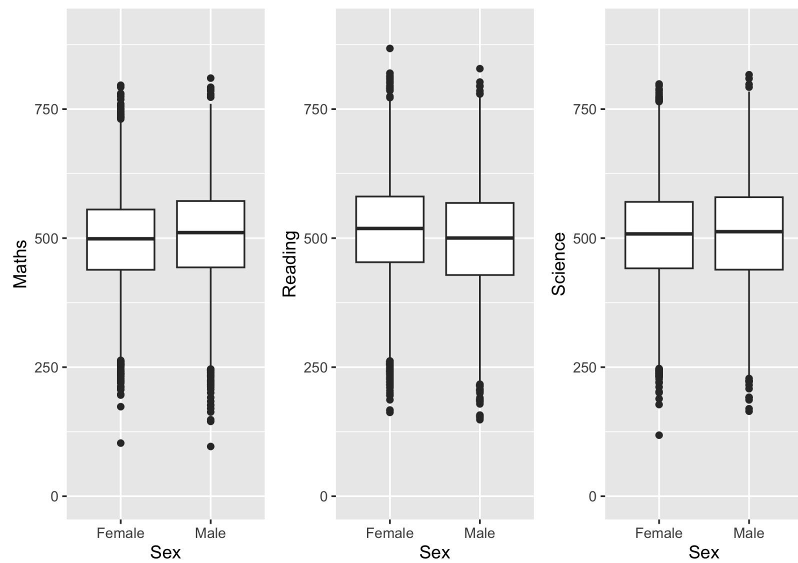


Figure 20: Subject Score by percentage of Students from Socioeconomically Disadvantaged Homes in the School.

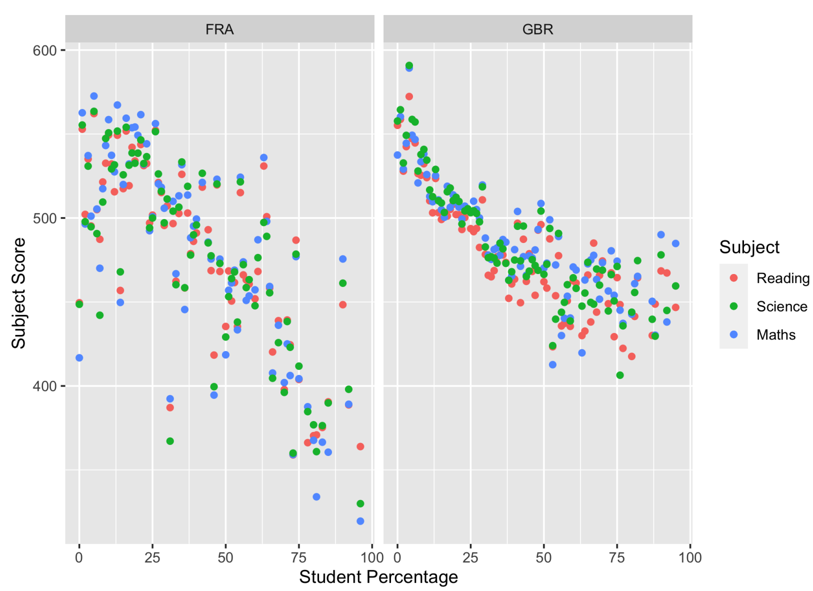
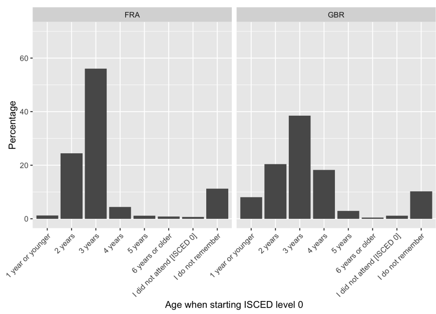


Figure 21: Comparing Ages at Starting School in the UK and France.

(a) Age when starting ISCED level 0.



(b) Age when starting ISCED level 1.

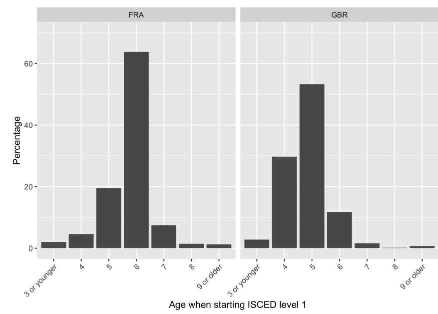
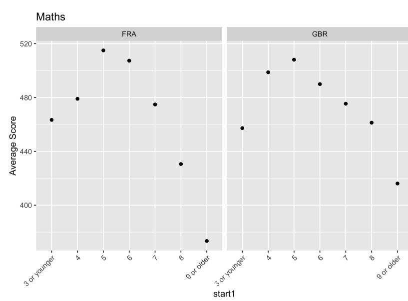
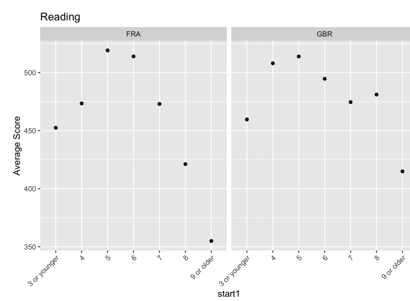


Figure 22: Average Age when starting ISCED level 1 against Subject Score.

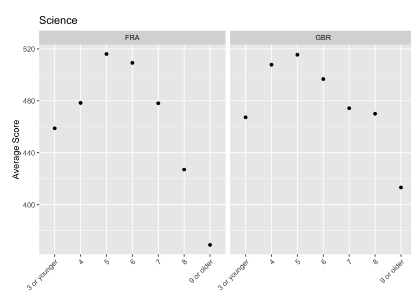
(a)



(b)



(c)



5.3 Modelling Process

A model was built for each subject. We began by investigating interaction plots and running ANOVA tests to decide which interactions should be included in our models. A significant interaction was found between country and age when starting primary education. Figures 23a, 23b, and 23c display these interactions.

We assumed that responses Y_i were uncorrelated, with common variance σ^2 and expectations of the form $\mathbb{E}(Y_i|x_{i,1}, x_{i,2}, ..., x_{i,j}) = \beta_0 + \beta_1 x_{i,1} + ... + \beta_j x_{i,j}$. This was written as

$$\mathbb{E}(\mathbf{Y}|\mathbf{X}) = \mathbf{X}\beta$$

where $\mathbf{X}$ and β were given by

$$\mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,30} \\ 1 & x_{2,1} & \cdots & x_{2,30} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,30} \end{pmatrix}, \qquad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{30} \end{pmatrix}$$

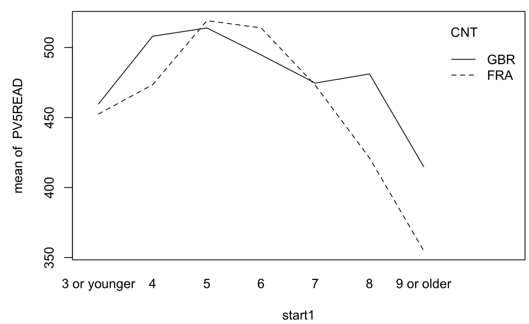
Each column in $\mathbf{X}$ represented a variable or interaction as detailed in Table 4.

Table 4: Design matrix columns.

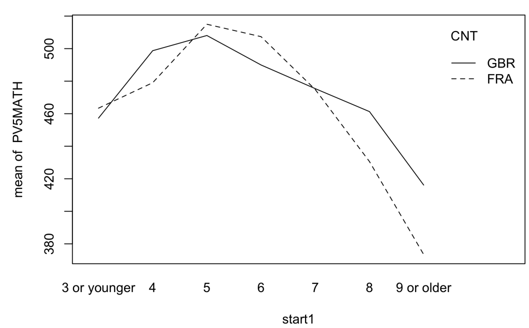
$x_{i,0}$	1	$x_{i,16}$	parentedu
$x_{i,1}$	start19 or older:CNTGBR	$x_{i,17}$	start19 or older
$x_{i,2}$	start18:CNTGBR	$x_{i,18}$	start18
$x_{i,3}$	start17:CNTGBR	$x_{i,19}$	start17
$x_{i,4}$	start16:CNTGBR	$x_{i,20}$	start16
$x_{i,5}$	start15:CNTGBR	$x_{i,21}$	start15
$x_{i,6}$	start14:CNTGBR	$x_{i,22}$	start14
$x_{i,7}$	year2018	$x_{i,23}$	start0I do not remember
$x_{i,8}$	CNTGBR	$x_{i,24}$	start0I did not attend [ISCED 0]
$x_{i,9}$	SCHLTYPEPrivate	$x_{i,25}$	start06 years or older
$x_{i,10}$	communityLarge City	$x_{i,26}$	start05 years
$x_{i,11}$	communityCity	$x_{i,27}$	start04 years
$x_{i,12}$	communityTown	$x_{i,28}$	start03 years
$x_{i,13}$	communitySmall Town	$x_{i,29}$	start02 years
$x_{i,14}$	disadv	$x_{i,30}$	ESCS
$x_{i,15}$	genderMale		

Figure 23: Interaction between Country and Age when starting Primary Education.

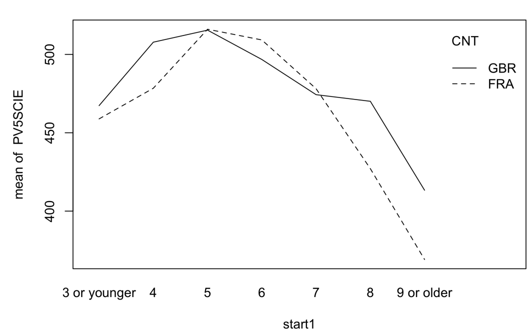
(a) Reading.



(b) Maths.



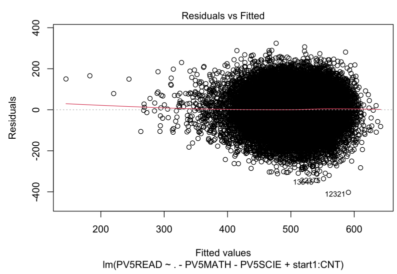
(c) Science.



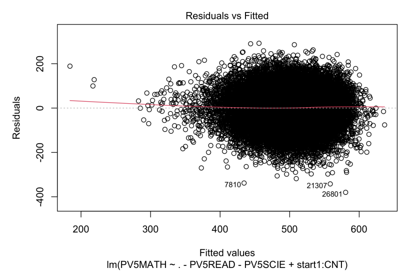
We checked the residual plots (Figures 24a, 24b, 24c) for these initial models to ensure that a linear model was appropriate. From these plots, we saw an even spread above and below zero on the fitted vs residual plot. Further residuals-leverage and Q-Q plots were produced; no trend in the residual vs leverage plot and minimal deviation on the normal Q-Q plot implied independence and normally distributed residuals with a constant variance. A linear model was therefore appropriate for this data. We then used backward selection, with both AIC comparisons and by checking P-values, to reduce the linear model for each subject. These methods resulted in no variable removals for all three subjects, giving the final models as detailed below.

Figure 24: Modelling Assumptions.

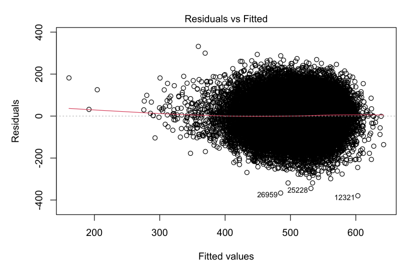
(a) Reading.



(b) Maths.



(c) Science.



5.4 Reading Results

Figure 25: Reading Model Coefficient Estimates and Confidence Intervals.

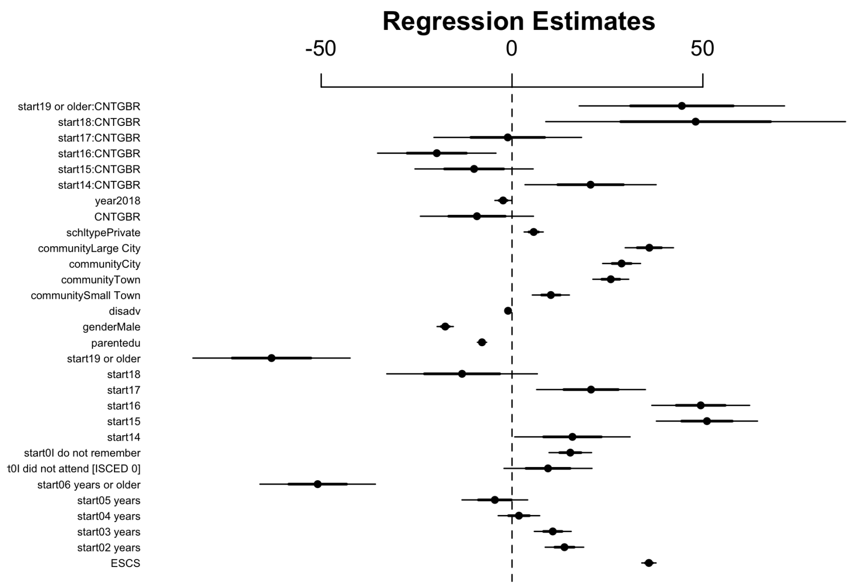


Figure 25 shows the regression estimates and their CI. Estimates for percentage of disadvantaged students (-1.043, 95% CI [-1.102, -0.985]) and highest ISCED education level of parents (-7.867, 95% CI [-9.053, -6.682]) were most accurate. The estimate for students in the UK starting primary education aged 8 was 48.104, with a 95% CI [9.623, 86.585]. As the widest CI, this estimate was the least precise, with a wide margin of error.

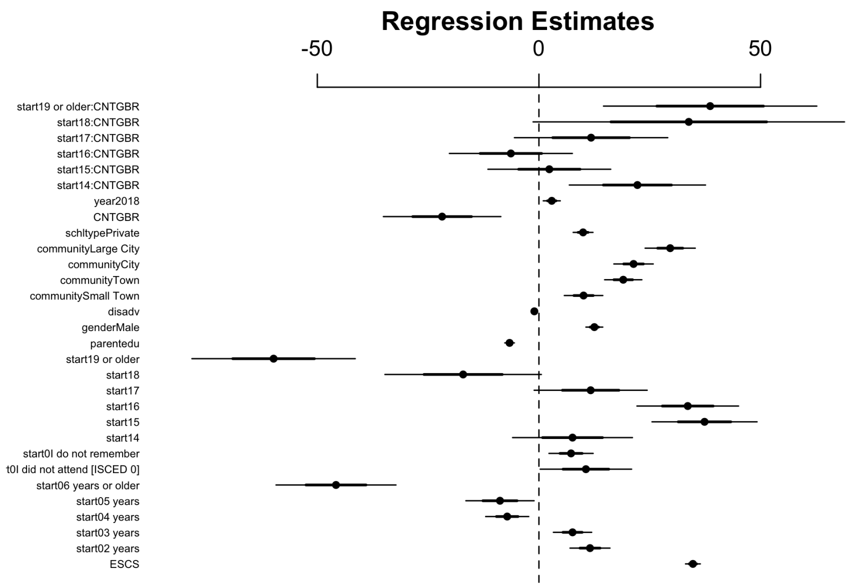
Our dummy variable was students in France who start primary education at age 3 or younger. In France, reading scores and primary education starting age were positively correlated up to age 5 (estimate 51.071). The correlation was negative beyond this, with students starting school at 9 or older having the lowest coefficient (-63.010).

For students starting primary education aged 3 or younger in the UK, a coefficient of -9.192 suggested lower reading scores than equivalents in France. The highest reading scores were achieved by those starting age 5 (31.936). Again, the lowest expected reading score came from those starting primary education aged 9 or older (-27.703).

The adjusted R-squared was 0.1968, meaning this model explained 19.68% of the variability in the data.

5.5 Maths Results

Figure 26: Maths Model Coefficient Estimates and Confidence Intervals.



The model coefficients are shown in Figure 26. As in the reading model, the most precise estimates were for percentage of disadvantaged students (-1.034, 95% CI [-1.086, -0.981]) and highest ISCED education level of parents (-6.624, 95% CI [-7.685, -5.563]), and the least precise estimate for students in the UK starting primary education aged 8 (estimate 33.809, 95% CI [-0.621, 68.239]).

In France, maths scores and primary education starting age were positively correlated between ages 3-5, with a peak of 37.363. Again, this correlation turned negative, and students starting school at age 9 or older had the lowest coefficient at -59.865.

A coefficient of -21.847 was lower than in France for students starting primary education aged 3 or younger in the UK. Estimated score peaked at age 5 (17.861). Lowest expected maths score came from those starting primary education aged 9 or older (-43.082).

The adjusted R-squared value 0.223 tells us that this model explained 22.3% of the variability in the data.

5.6 Science Results

Figure 27: Science Model Coefficient Estimates and Confidence Intervals.

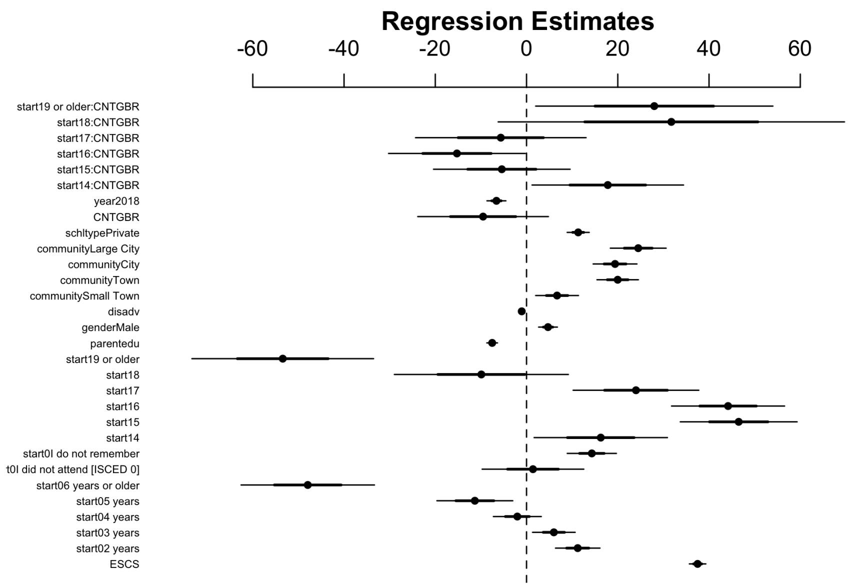


Figure 27 shows the regression estimates and CI. Similarly to the reading and maths models, the most accurate estimates were for percentage of disadvantaged students (-1.052, 95% CI [-1.108, -0.995]) and highest ISCED education level of parents (-7.532, 95% CI [-8.679, -6.386]). Again, the estimate for students in the UK starting primary education aged 8 (31.736, 95% CI [-5.462, 68.933]) was least accurate.

In France, science scores and primary education starting age were positively correlated between ages 3-5 (age 5 had the highest estimate of 46.504). Beyond age 5, the correlation was negative, however, coefficients were greater than zero for age 6 (44.173) and age 7 (23.986). Students starting school at 9 or older had the lowest coefficient of -53.434.

For students starting primary education aged 3 or younger in the UK, a coefficient of -9.525 suggested lower science scores than in France. Students starting primary education aged 5 had the highest estimate, (31.556). As in the previous models, correlation was negative for those over age 5, with lowest expected science score from those starting primary education aged 9 or older (-34.961).

The adjusted R-squared 0.2102 shows that this model explained 21.02% of the data's variability.

6 Conclusion

6.1 Differences in educational attainment within the UK

The results verify that test scores differ between the UK nations. Students in England achieve the highest scores in all subjects, followed by Wales, then Scotland, with Northern Ireland being the lowest.

Differences between nations changed minimally between 2015 and 2018, subject dependent. For maths and reading there was little variation; the most significant difference between years was seen for science. Accounting for these different trends, the order of nations in terms of best to worst results did not change.

Modelling estimates were least precise for Northern Ireland - observations for Northern Ireland make up 18% of the analysed data, which is greater than the proportion of Northern Irish inhabitants in the UK population [4], and therefore does not explain the inaccuracy. The models also do not support the graphical analysis. Bias and dataset quality may be the cause of this. Use of a larger sample size could improve estimation accuracy, as could alterations to account for bias.

All three models obtained very small estimates for number of males/females in the student's school. These factors could be removed from future models. 14-17% of the variation in the data was explained, which is fairly low. More variables from the PISA data-set could be added to the model to improve this, however, this could lead to over-fitting and a deceptively high R-squared value. Therefore, this must be done with caution. Due to missing observations, 44% of the original data-set was removed, which is likely to have created bias and affected variation in the model.

6.2 Differences in educational attainment between France and the UK

There is a significant relationship between the age at which a child starts primary education and their test score; this relationship differs slightly between France and the UK. The models are similar across all three subjects in both countries. To optimise test results in all three subjects, students should start primary education at age 5. The effect of starting at 5 in comparison to other ages is more obvious in France, while there is variation of only around 20 marks between ages 4-8 in the UK.

The models for science and reading show a spike in the UK for those starting primary education aged 8 - this was the least precise estimate in all the models, so we cannot draw a conclusion from this. Only 0.1% of the observations involved students in this category. This is extremely low, and may explain the inaccuracy of the estimate. Using a larger data-set could improve this.

Our models explain 19-22% of the variation in the data. As in part one, this is relatively low, however, methods to improve it, for example, adding more variables, risk over-fitting and increasing inaccuracy.

References

[1] https://www.oecd.org/pisa/pisafaq/#:~:text=In%20each%20test%20subject%2C%20there,deviations%20around%20100%20score%20points

[2] https://www.oecd.org/pisa/data/

[3] https://www.oecd.org/pisa/data/pisa2018technicalreport/PISA2018%20TecReport-Ch-04-Sample-Design.pdf

[4] https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/annualmidyearpopulationestimates/mid2021