



The Department of Health Sciences

Electronic Assignment Submission Cover Sheet Anonymous Submission

This Cover Sheet should be the first page of your assignment.

Student Examination Number:	Y3874726
Module Code:	HEA00001M
Module Information: <i>Module Title and Module Assignment</i>	Introduction to Regression Analysis
Submission Deadline:	23 March 2020
Attempt: <i>* Delete as appropriate</i>	First attempt
Actual Word Count:	2,128 (Two thousand one hundred twenty-eight only)

I confirm that I have:

- checked that I am submitting the correct and final version of my assignment
- formatted my assignment in line with departmental guidelines
- conformed with University regulations on academic integrity
- included an accurate word count in line with departmental guidelines
- added my examination number on every page of the assignment
- not written my name anywhere in the assignment
- saved my assignment in Word or pdf format

PLEASE PLACE A CROSS IN THE BOX TO CONFIRM THESE STATEMENTS:

X

Note: if you have any questions please see the submission FAQs information

Introduction to Regression Analysis Assessment Report

Y3874726

Contents

1	Question 1	4
1.1	Birth weight(Bwght)	4
1.2	Friendly Maths Test Score at age 10 (Math_score)	4
1.3	Home ownership of parents at age 5 (Own_home)	7
1.4	Region	9
2	Question 2	11
2.1	Investigate mean Shortened Edinburgh Reading Test Score at 10 (Reading_score) by Region	11
2.1.1	Descriptive summary	11
2.1.2	ANOVA	13
2.1.3	ANOVA results	13
2.2	Difference Estimation	14
2.3	Non-parametric Test	14
3	Question 3	15
3.1	Reading Score and Math Score Regression analysis	15
3.1.1	Pre-Regression Statistical and Graphical Investigation	15
3.1.2	Simple Linear Regression	16
3.2	Adding Birth weight and Gender to the model	18
3.2.1	Assumptions Check	19
3.3	Family Income and Home ownership	20
3.3.1	Model 1: Home ownership	20
3.3.2	Assumptions check	21
3.3.3	Model 2: Family income	22
3.3.4	Assumptions checking	24
3.3.5	Model Comparison	25
4	Reading Support and Nations	25
4.1	Reading Support	25
4.2	Nations	26
4.3	Reading Support and Nations	28
4.4	Logistic Regression	28
	Appendices	30
5	Shapiro-Wilk Test Results	30
6	Tukey HSD	30

1 Question 1

1.1 Birth weight(Bwght)

Table I: Table presenting summary of birth weight by gender in the data. (Abbreviations used: CI= Confidence interval, gms=grams)

Statistic	Male	Female
Minimum (in gms)	200	397
Maximum (in gms)	5500	5188
Mean (in gms)	3321.12	3207.68
Median	3345	3232
(95% CI) of mean	(3303.44, 3338.79)	(3190.79, 3224.57)
Standard Deviation (in gms)	603.54	554.83
Valid cases	4480	4147
Missing	429	332
Total	4909	4479

According to Table I, mean Birth weight for males (3321.12gm [95%CI: (3303.44, 3338.79)], SD=603.54) is slightly more than that of females (3207.68gm [95%CI: (3190.79, 3224.57)], SD=554.83). Non-overlap of confidence intervals (CIs) indicates significant differences. Minimum value of Birth weight for males (200gm) is more than 5 standard deviation(SD) and the maximum (5500gm) is more than 3 SD away from the mean(3321gm). For females, the minimum Birth weight(397gm) is 5.06 SD and the maximum (5188gm) is 3.56 SD away from the mean. Fig. 1 and 2 indicate normality of Birth weight for Males and Females (also indicated by proximity of respective means and medians (Males: Mean=3321, Median=3345. Females: Mean=3207, Median=3232)) with slight left-skewness (Mean < Median) observed in both.

1.2 Friendly Maths Test Score at age 10 (Math_score)

Table II: Table presenting the summary of Math scores by gender in the data. (Abbreviations used: CI= Confidence interval)

Statistic	Male	Female
Minimum	1	1
Maximum	72	71
Mean	44.25	43.25
Median	45	44
(95% CI) of mean	(43.79, 44.70)	(42.82, 43.68)
Standard Deviation	12.74	11.77
Valid cases	2996	2875
Missing	1913	1604
Total	4909	4479

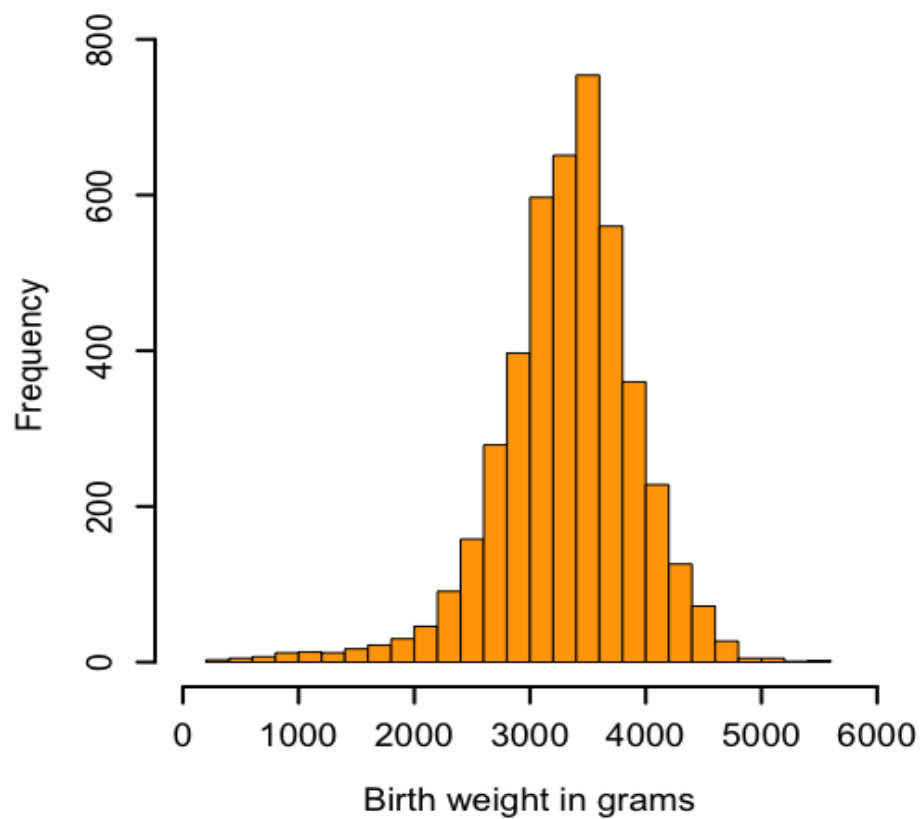


Figure 1: Histogram showing the distribution of birth weight in males.

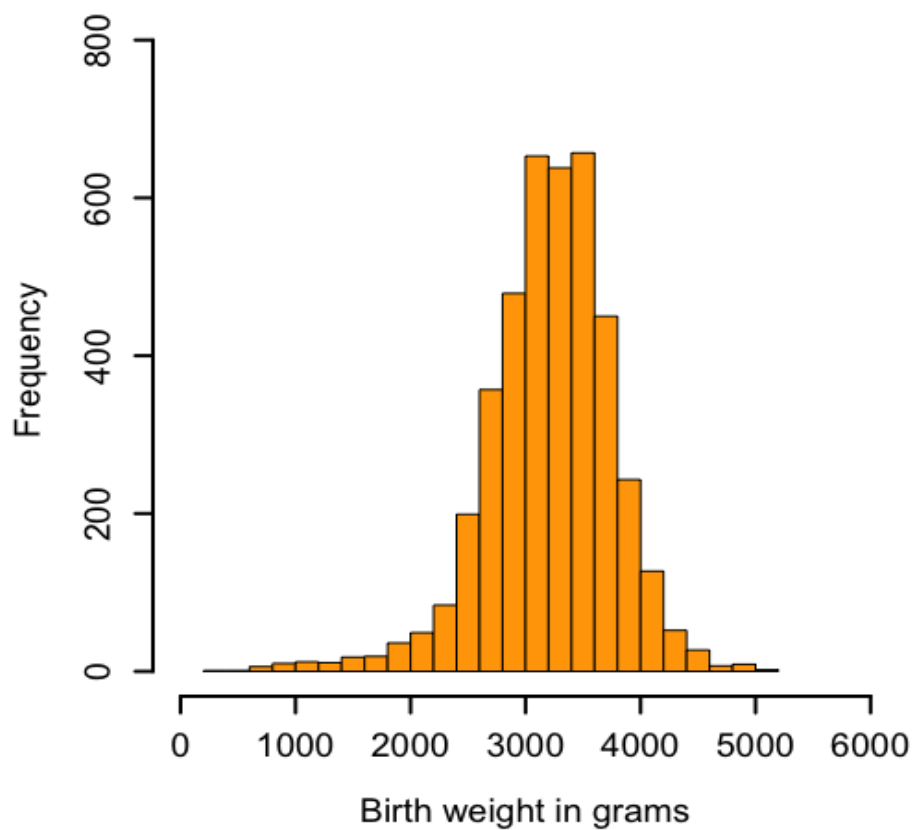


Figure 2: Histogram showing the distribution of birth weight in females.

Mean Math score for males (44.25 [95%CI: (43.79, 44.70)], SD=12.74) is higher than that of females (43.25 [95%CI: (42.82, 43.68)], SD=11.77). Proximity of mean and median Math score for both groups (Males: Mean=44.25, Median=45. Females: Mean=43.25, Median=44) indicates normality, confirmed in Fig. 2 and 3. Slight left-skewness (Mean < Median) observed in both groups.

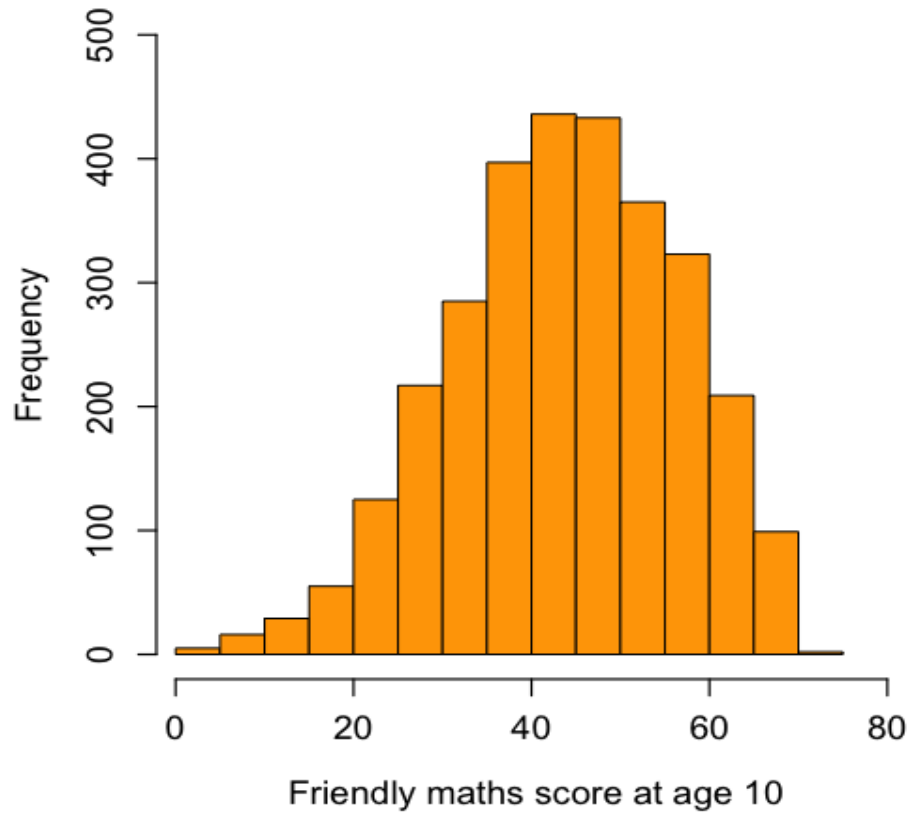


Figure 3: Histogram showing the distribution of friendly maths test score at age 10 in males.

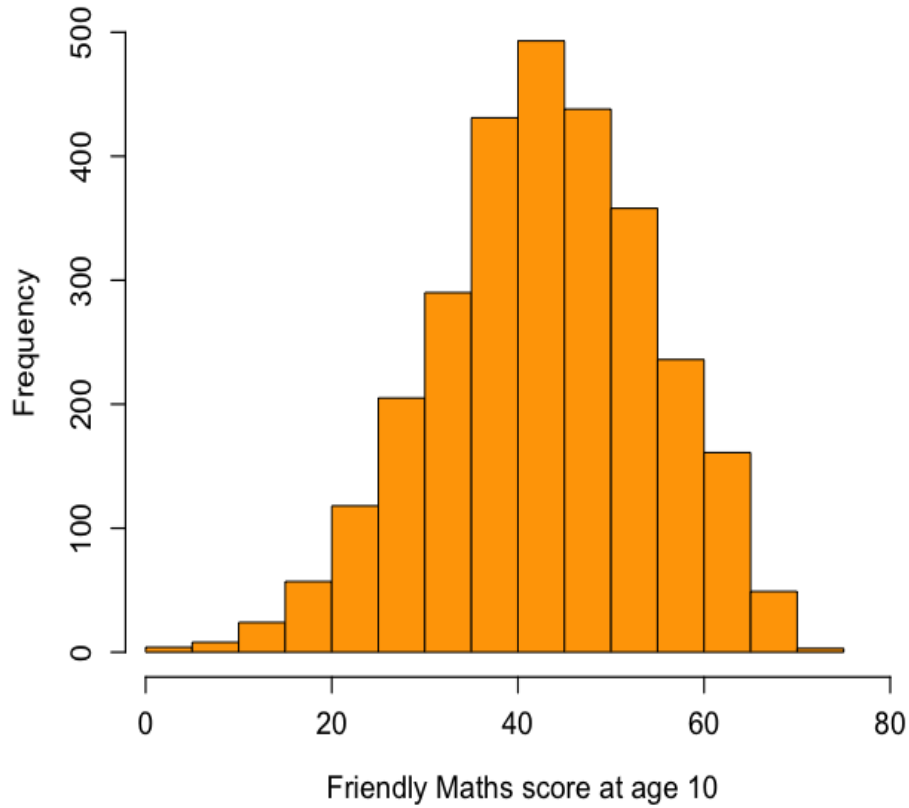


Figure 4: Histogram showing the distribution of friendly maths test score at age 10 in females.

1.3 Home ownership of parents at age 5 (Own_home)

Own_home is a categorical variable with three levels : No Information, Yes and No. Table III describes its summary statistics.

Table III: Contingency Table showing home ownership of parents of children at age 5 by gender of their children.

Category	Male	Female
Yes	1864 (54.74%)	1809 (57.68%)
No	1535 (45.08%)	1350 (42.60%)
No Information	6 (0.18%)	10 (0.32%)
Cases with values in three levels	3405 (100%)	3169 (100%)
Missing values	1504	1310
Total	4909	4479

According to Table III, the percentage point difference between those reporting home ownership with male children is 9 (Yes=54.74%, No=45.08%) while for female children it is 15 (Yes=57.68%, No=42.60%), indicating a bigger difference in the latter group. Higher proportion of home ownership seen among parents of females than males(57.68% vs 54.74%).

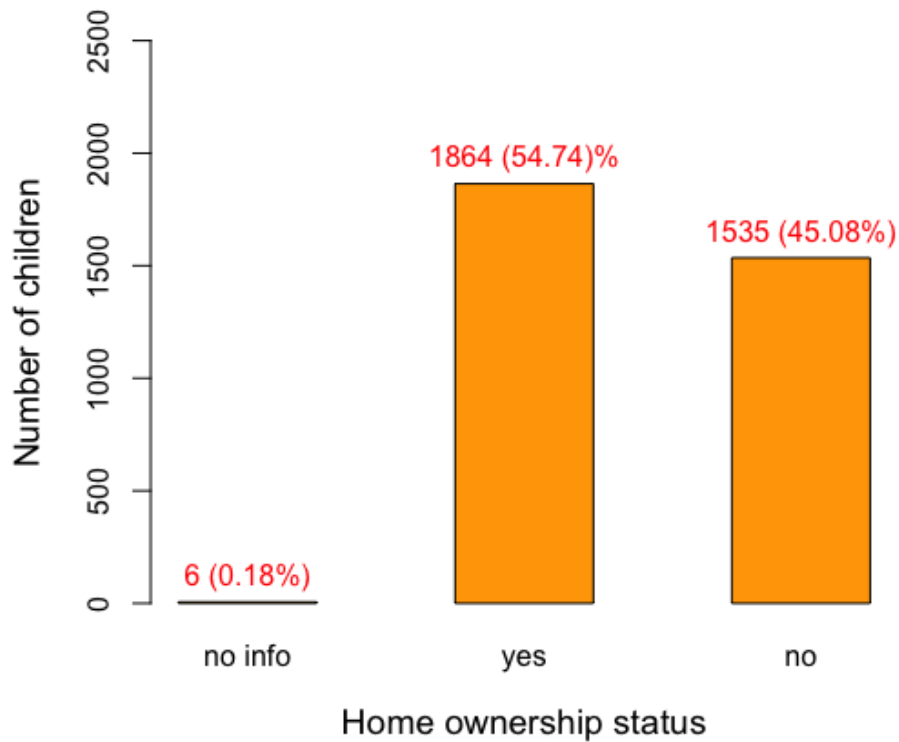


Figure 5: Barchart showing the number of people with home ownership in Males. The numbers in bracket indicate percentages.

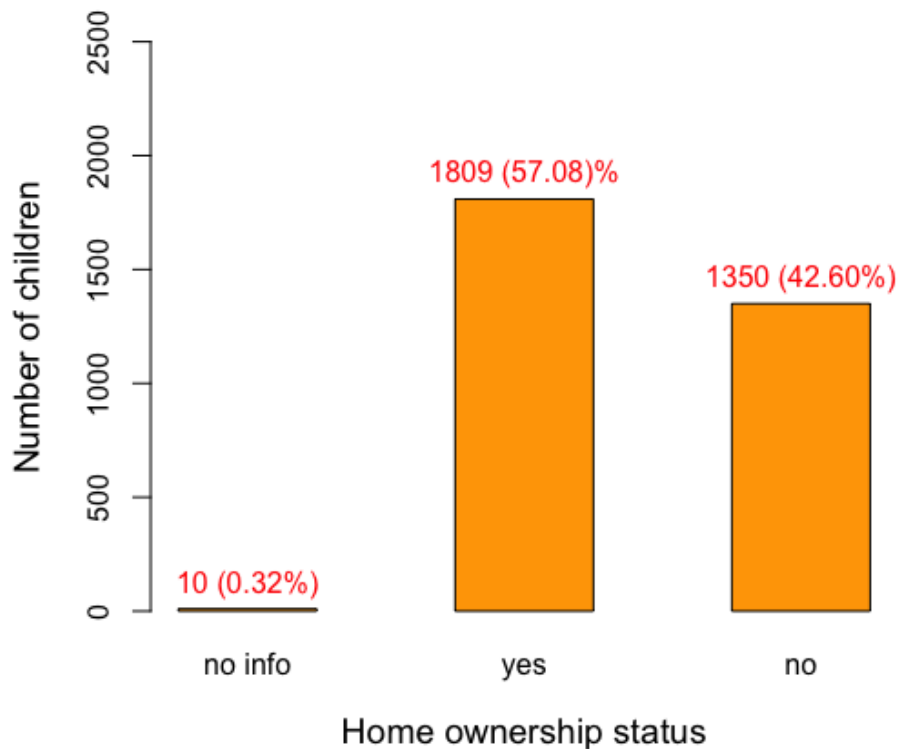


Figure 6: Barchart showing the number of people with home ownership in Females. The numbers in bracket indicate percentages.

As per Fig. 5 and 6, a higher proportion of people own home in both gender groups (54.74% in males, 57.68% in females).

1.4 Region

The variable *Region* in the dataset is characterized by 14 categories. Table IV provides the descriptive summary of *Region*.

Table IV: Table showing the geographical area of the mother's residence at the time of birth segregated by gender. The percentages in the bracket are computed from the total non-missing cases.

Regions	Males	Females
Overseas birth	3 (0.07%)	1 (0.02%)
North Wales	66 (1.47%)	58 (1.40%)
East Anglia	125 (2.79%)	133 (3.20%)
South Wales	158 (3.52%)	142 (3.42%)
Northern Ireland	165 (3.68%)	168 (4.05%)
South West	254 (5.66%)	254 (6.12%)
East Midlands	269 (6%)	250 (6.02%)
North	292 (6.51%)	233 (5.61%)
York and Humbrside	383 (8.54%)	384 (9.25%)
Scotland	432 (9.63%)	386 (9.30%)
West Midlands	466 (10.39%)	417 (10.05%)
London	554 (12.35%)	497 (11.98%)
North West	556 (12.39%)	511 (12.31%)
South East	764 (17.03%)	716 (17.25%)
Non-missing	4487 (100%)	4150 (100%)
Missing	422	329
Total	4909	4479

According to Table IV approximately 51% of people in both gender groups reside in 4 Regions (West Midlands, London, North West and South East) and less than 5% in North Wales and East Anglia combined.

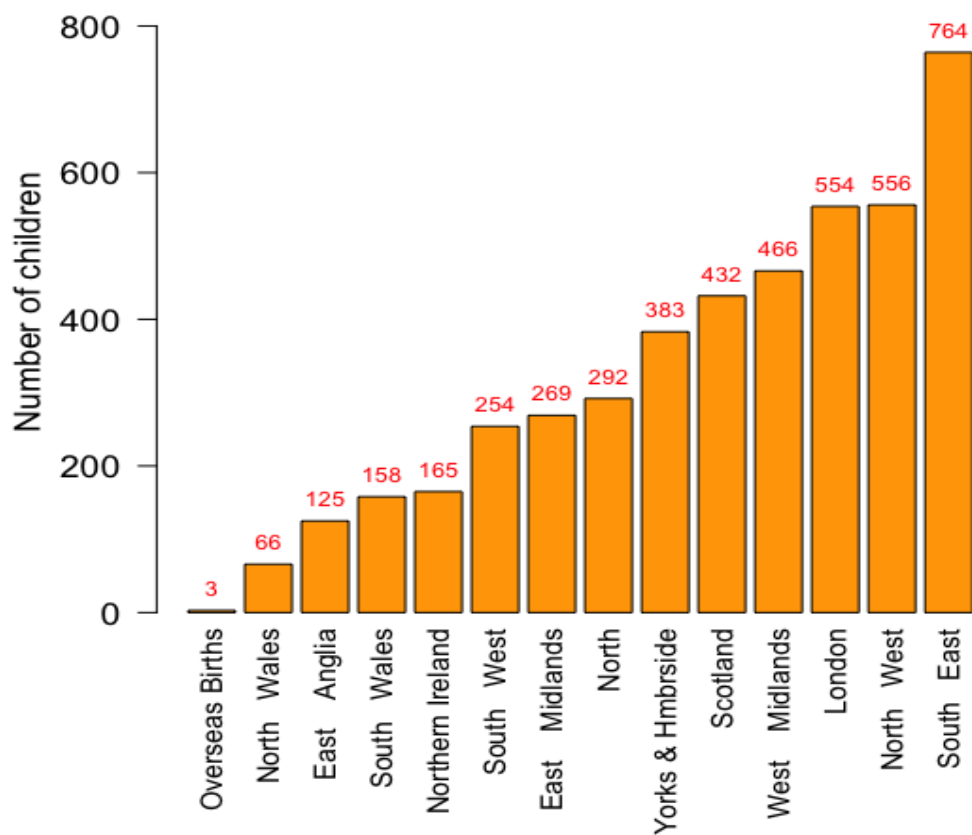


Figure 7: Barchart showing geographic residence of mothers of male babies at birth. The numbers above bars indicate absolute figures.

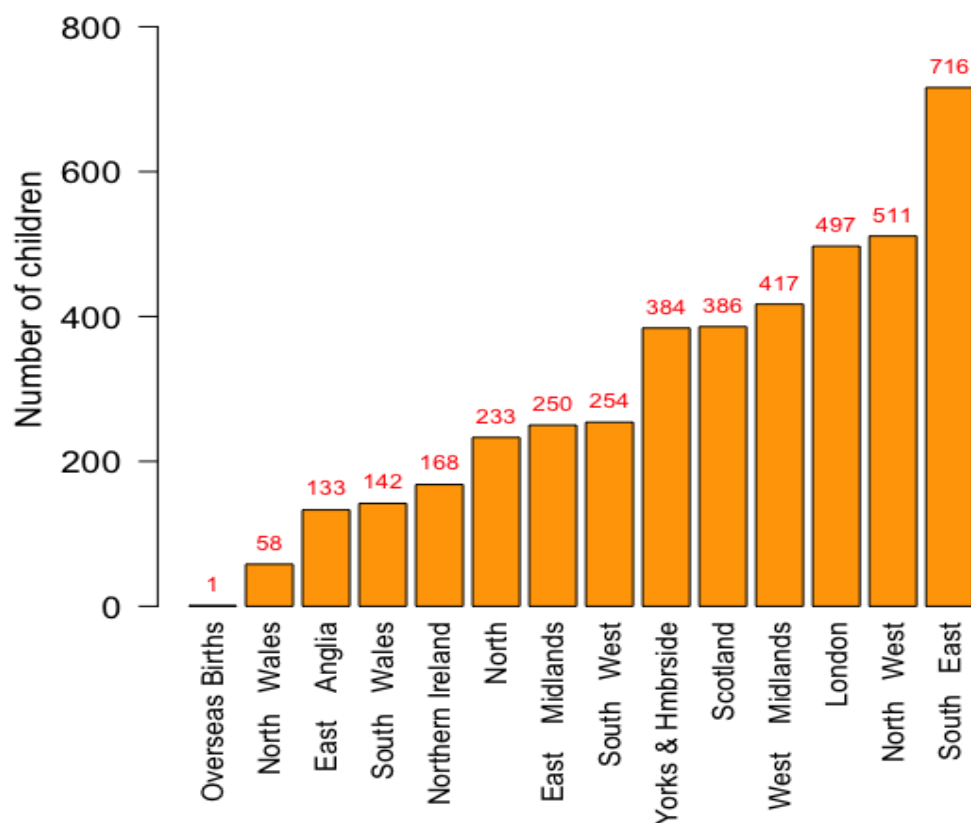


Figure 8: Barchart showing geographic residence of mothers of female babies at birth. The numbers above bars indicate absolute figures.

2 Question 2

2.1 Investigate mean Shortened Edinburgh Reading Test Score at 10 (Reading_score) by Region

2.1.1 Descriptive summary

Table V describes the summary statistics for Reading scores.

Table V: Table presenting the summary of Reading Score in the data.

Statistic	Value
Minimum	0
Maximum	65
Mean	40.01 (39.68, 40.33)
Median	41
Standard deviation	12.84

Mean Reading score of respondents is 40.01 ([95%CI: 39.68, 40.33], SD=12.84). The minimum value of Reading score is more than 3 SD and the maximum is 1.95 SD away from the mean indicating outlier presence. Proximity of Mean(40) and median(41) indicates normality, confirmed in in Fig. 9 and 10; minor deviations observed.

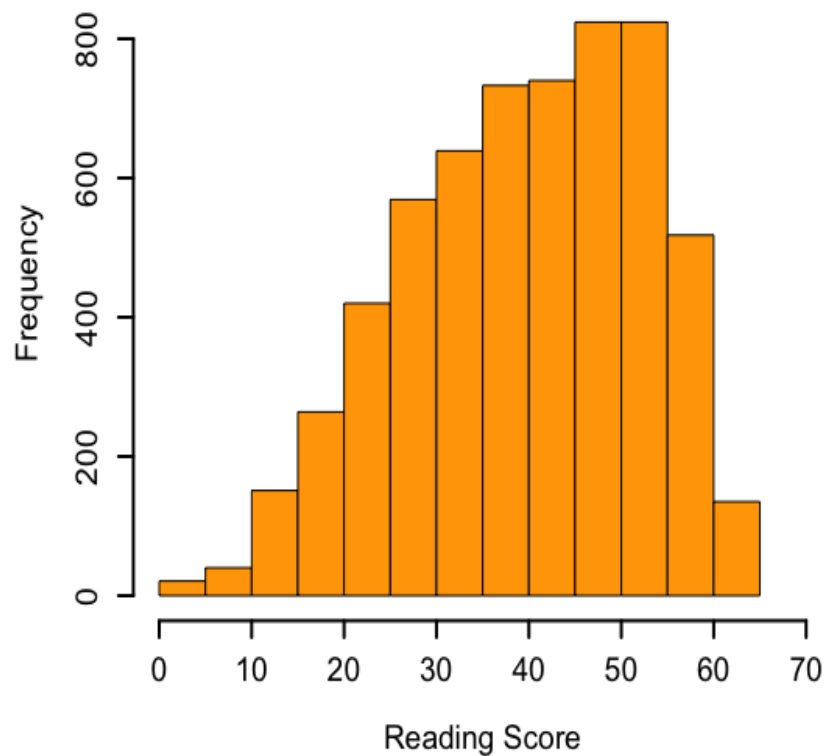


Figure 9: Histogram showing the distribution of Reading Scores in the data.

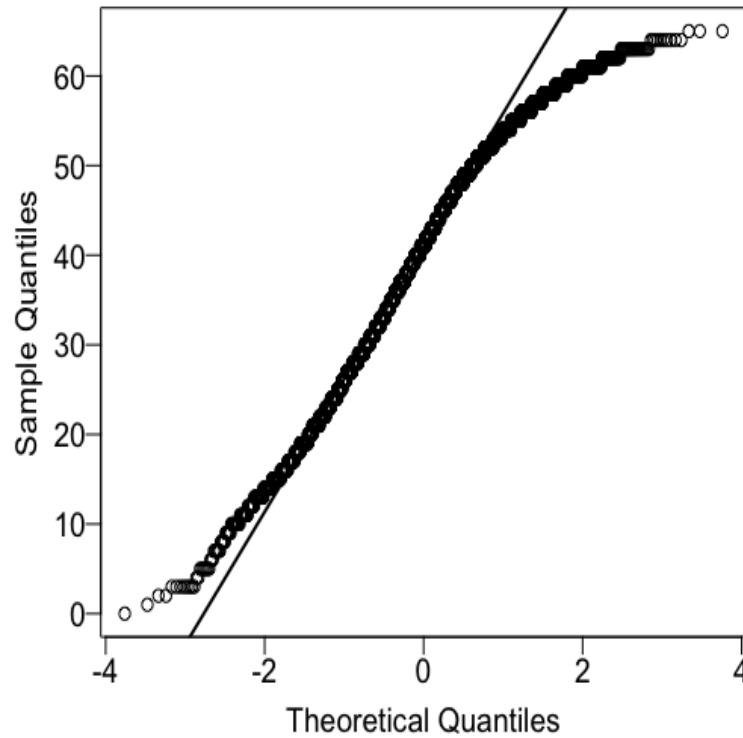


Figure 10: Quantile-Quantile plot for Reading scores in the dataset.

Preliminary side-by-side boxplots are used to explore the relationship between Reading scores (continuous) and Region (categorical). Fig. 11 shows large within-group variations in the Reading scores across all Regions. To further investigate the relationship One-way ANOVA is used.

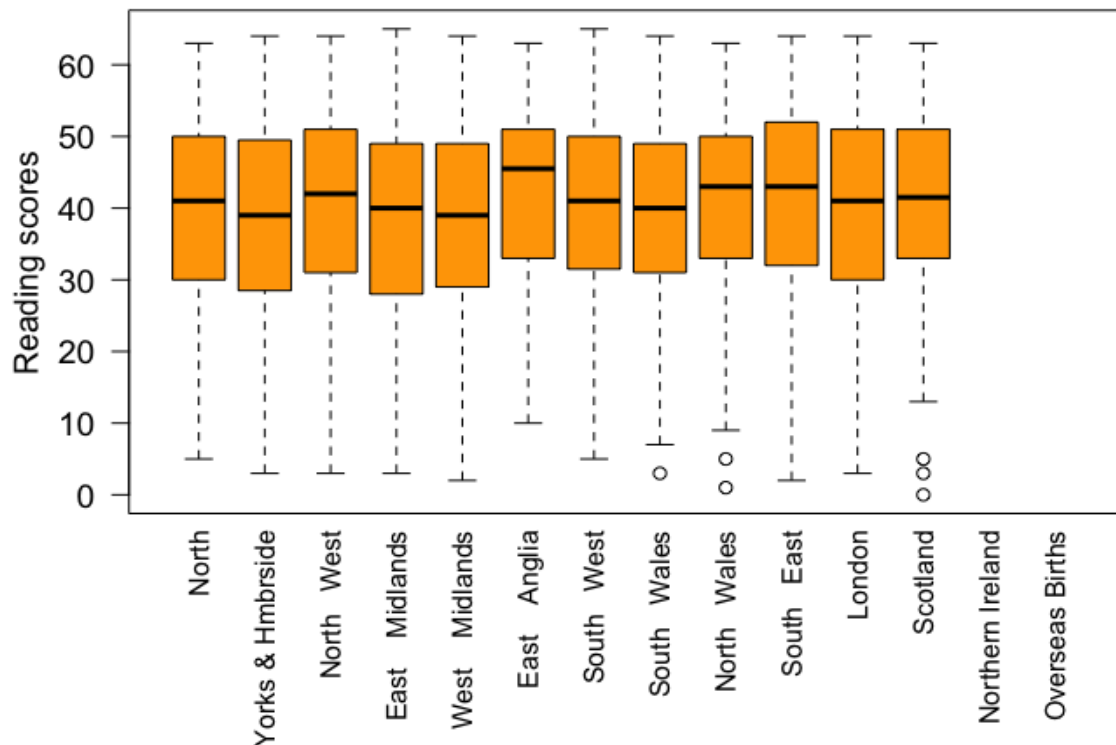


Figure 11: Side-by-side boxplot showing reading scores for the 14 regions. Northern Ireland and Overseas cases do not have any valid cases for Reading scores hence indicate no boxplot.

2.1.2 ANOVA

Assumptions for ANOVA are tested.

1. Reading scores are normally distributed (Section 2.1.1).
2. Shapiro-Wilk test for Reading scores in each group provides $p < 0.05$ indicating normality. (Appendix)
3. *highest SD* $\leq 2 \times$ *lowest SD* rule is used to check similarity of SD across the groups. From Table VI it is clear that the highest value of SD (13.60) $\leq 2 \times$ lowest SD ($2 \times 11.93 = 23.86$).

All assumption hold.

Table VI: Table presenting summary of Reading scores across regions in the data.

Region	Number of people with Reading Scores	Mean Reading Score	Standard deviation of Reading Score
Scotland	614	41.13	11.93 (Lowest)
South Wales	225	39.13	12.30
South West	356	39.75	12.30
North West	762	40.39	12.61
London	627	39.86	12.80
North	425	39.69	12.81
West Midlands	650	38.78	12.88
East Midlands	390	38.98	12.98
South East	980	41.43	13.04
East Anglia	176	41.41	13.26
North Wales	89	40.73	13.41
Yorks and Hmbrside	575	38.31	13.60 (Highest)
Northern Ireland	0	NA	NA
Overseas Birth	0	NA	NA
Total	5869		

2.1.3 ANOVA results

H_0 : Mean Reading scores of all groups are the same.

H_a : Not all of the mean Reading scores are same.

Table VII: Table showing the results of the ANOVA test for Reading Score and Region in the dataset.

Source	DF	Sum of Squares	Mean Square	F-value	Pr>F
Model	11	6575.52	597.77	3.65	< 0.0001
Error	5857	960614.15	164.01		
Total	5868	967189.67			

Table VII shows F-statistic is 3.65($p < 0.0001$), H_0 is rejected; at least one of the group means differs from the other. R^2 value is $\frac{\text{Sum of squares Model}}{\text{Sum of squares total}} = \frac{6575.52}{967189.67} = 0.006$. About 0.6% of the variation in Reading scores can be explained by Region, remaining 99.4% variation is due to person-to-person variation within each of the 12 groups.

2.2 Difference Estimation

Post rejection of H_0 , Tukey Honest Significant Difference (HSD) is used to identify differing pairs because no prior hypotheses were established. The test performs hypothesis testing for all possible pairs (in this case $\binom{12}{2} = 66$) and reports the difference in means and the associated p-values. Assumptions for the test verified in Section 2.1.2. Only statistically significant results are presented in Table VIII due to a high number of comparisons. (Complete results in Appendix)

Table VIII: Table showing the results of Tukey Honest significant Different test for Reading Scores and Region.

Pairs	Mean Difference	Lower Limit	Upper Limit	p-value
South East-Yorks and Hmbrside	3.12	0.92	5.32	<0.0001
South East-West Midlands	2.66	0.54	4.77	0.002
Scotland-Yorks and Hmbrside	2.82	0.39	5.25	0.008
Scotland-West Midlands	2.36	0.00	4.71	0.05
South East-East Midlands	2.45	-0.05	4.96	0.06

Table VIII shows that:

1. The mean Reading scores of children in South East is higher than that of Yorks and Hmbrside (difference=3.12, [95%CI: 0.92,5.32], p-value <0.0001) and West Midlands (difference=2.66, [95%CI: 0.54, 4.77], p-value=0.002). It appears that children in South East have a better reading ability at age 10 than the two regions.
2. The mean Reading score of children in Scotland is higher than that of Yorks and Hmbrside (difference=2.82, [95% CI: 0.39, 5.25], p-value=0.008) indicating that children in Scotland have better reading ability at age 10 than those in Yorks and Hmbrside.
3. Marginal statistical significance observed for Scotland-West Midlands (difference=2.36, [95%CI: 0, 4.71], p=0.05) and South East-East Midlands (difference=2.45, [95%CI: -0.05, 4.96], p=0.06). Upper CI value for both indicates that Scotland and South East children may have a higher reading ability compared to the respective comparison groups.

2.3 Non-parametric Test

Kruskal-Wallis test is used. It is the non-parametric analogue of One-way ANOVA test.

H_0 : All 12 populations have the same median Reading score.

H_a : Not all 12 median scores are equal.

Table IX: Table showing the results of Kruskal Wallis Test Results for Reading score and Region in the dataset.

Kruskal-Wallis statistic	41.833
Degree of Freedom	11
p-value	< 0.0001

$p < 0.0001$ provides sufficient evidence to reject the null hypothesis of similar median Reading scores (Table IX).

3 Question 3

3.1 Reading Score and Math Score Regression analysis

Refer Table V for descriptive summary of Reading scores. Mean Math score is 43.76 ([95%CI: 43.45, 44.08], SD=12.29) as seen in Table X. Minimum value of Math scores is more than 3 SD away from the mean and the maximum is more than 2 SD away making them as outliers. Mean (43.76) and Median (44) lie close to each other indicating normality.

Table X: Table summarizing Math scores in the data.

Statistic	Value
Minimum	1
Maximum	72
Mean	43.76 (43.45, 44.08)
Standard deviation	12.29
Median	44

3.1.1 Pre-Regression Statistical and Graphical Investigation

Scatter-plot in Fig. 12 shows a fairly positive linear relationship between the two variables. Five outliers (colored in red) are detected based on distance from the cluster of observations. Individual inspection of these cases confirms validity of values and informs retention. Correlation of 0.76 ($p < 0.0001$) indicates a strong positive linear relationship; it appears that higher Reading scores are associated with higher Math scores.

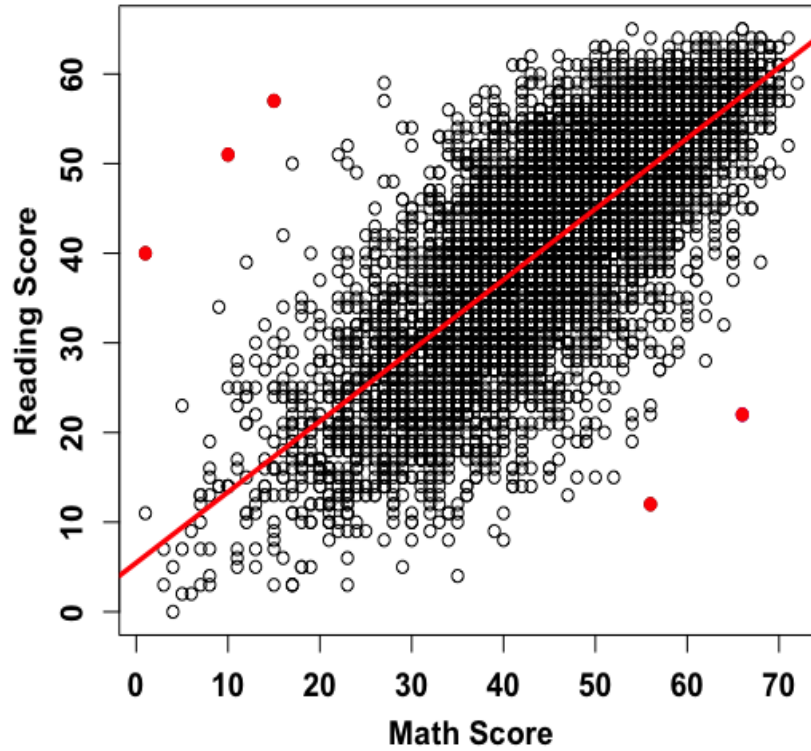


Figure 12: Scatter plot of Reading Score against Math score. Data points colored in red indicate potential outliers.

3.1.2 Simple Linear Regression

The model equation is :

$$Reading_score = \beta_0 + \beta_1 Math_score + \varepsilon \quad (1)$$

where β_0 is the constant and β_1 is the coefficient for Math score and ε is assumed to be independent and normally distributed $N(0, \sigma^2)$.

Table XI presents the result of the analysis. R^2 is 0.57 and the F-statistic = 7833 ($p < 0.0001$).

Table XI: Table showing the results of simple linear regression analysis of Reading score on Math Score in the dataset.

Predictor variable	Regression coefficient	95% CI	P-value
Math Score	0.79	(0.77, 0.81)	< 0.0001

For every unit increase in Math score the Reading score increases by 0.79 (95% CI: [0.77, 0.81], $P < 0.0001$) on average. R^2 indicates that an impressive 57% of the variation in Reading scores can be explained by Math scores.

Post-regression assumptions of normality of residuals and homoscedasticity are checked. Distribution of residuals fairly normal as seen in Fig. 13. Absence of any pattern in residuals and their uniform distribution around the line of regression at each predicted value of Reading scores confirms homoscedasticity (Fig. 14). Both assumptions hold.

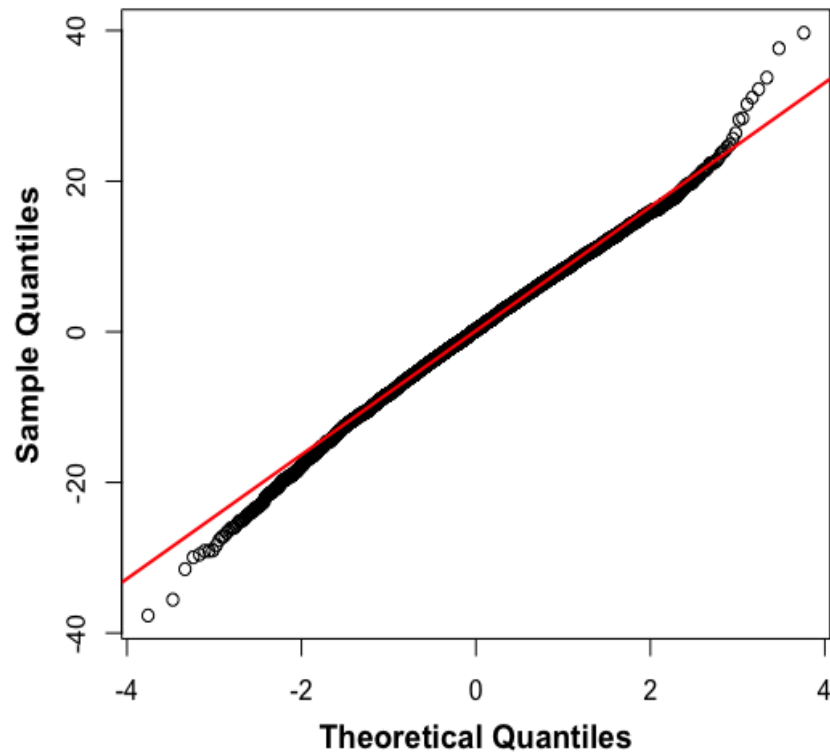


Figure 13: Normal quantile plot of the residuals for the Regression of Reading scores on Math scores.

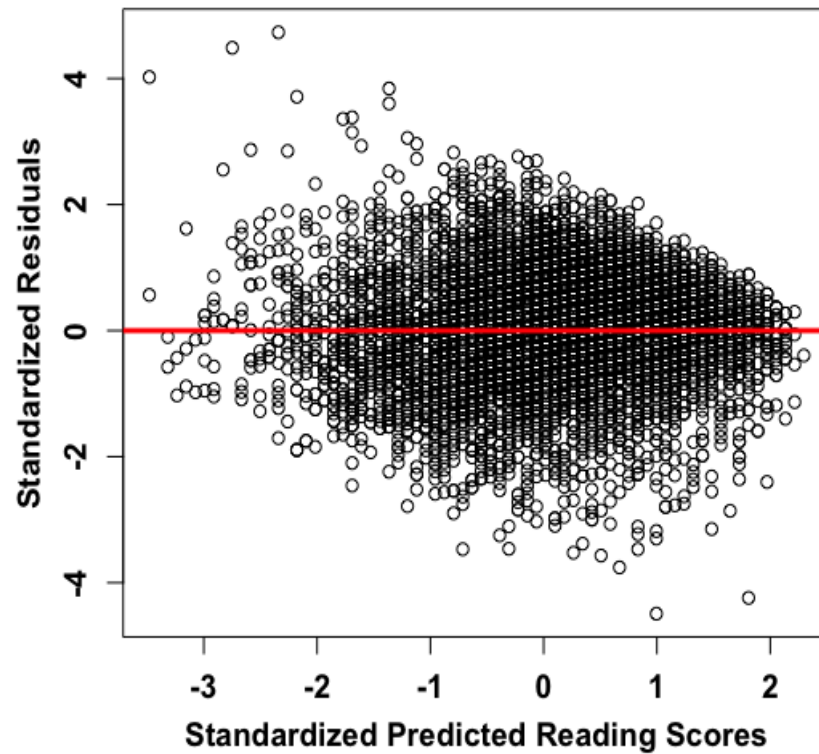


Figure 14: Plot of standardized residuals versus standardized predicted Reading scores for the Regression of Reading scores on Math scores.

3.2 Adding Birth weight and Gender to the model

Descriptive summaries of birth weight and Gender are provided in Table XII and XIII.

Table XII: Table describing birth weight of respondents in the data. Unit is grams.

Minimum	Maximum	Mean (95%CI)	Median	Standard deviation	NAs
200	5500	3267 (3254.27, 3278.90)	3289	583.36	764

Table XIII: Table showing number of males and female in the data.

Not reported	Male	Female
3 (0.03%)	4909(52.27%)	4479 (47.69%)

Gender is a categorical variable with two levels (Male and Female) and Birth weight, a continuous quantitative variable in the dataset. Cases lacking information on *Gender* and Birth weight are removed. Birth weight in grams is transformed to kgs to facilitate interpretability. The regression equation is:

$$Reading_score = \beta_0 + \beta_1 Math_score + \beta_2 Birth_weight + \beta_3 Gender + \varepsilon \quad (2)$$

where β_0 represents the constant, β_1 , β_2 and β_3 are the coefficient for variables Math score, Birth weight and *Gender* respectively and ε is the error term assumed to be independent and normally distributed with $N(0, \sigma^2)$.

Table XIV: Table showing the results of multiple regression analysis of Reading score on Math score, Birth weight and Gender.

Predictor variable	Regression coefficient	95% CI	P-value
Math Score	0.79	(0.77, 0.81)	<0.0001
Birth weight (in kgs)	0.91	(0.51, 1.33)	<0.0001
Gender (Female - Male)	2.76	(2.34, 3.19)	<0.0001

It can be noted that:

1. Adjusting for Birth weight and gender, one unit increase in Math score is associated with 0.79 ([95%CI: 0.77, 0.81], $p < 0.0001$) point increase in Reading score on average.
2. Adjusting for Math score and gender, a unit increase in birth weight (1 kg) is associated with 0.91 ([95%CI : 0.51, 1.33], $p < 0.0001$) point increase in Reading score on average.
3. Females have on average 2.76 ([95%CI: 2.34, 3.19], $p < 0.0001$) points higher on Reading score than males after adjusting for Birth weight and Math score.
4. The R^2 value is 0.58, slight increase from the previous ($R^2 = 0.57$) indicating that new model can explain 1% extra variation in the output variable post Birth weight and gender inclusion.

3.2.1 Assumptions Check

Fig. 15 shows that residuals lie fairly along the straight line without serious deviations indicating normal distribution. Fig. 16 shows that residuals are uniformly distributed along the regression line(residuals=0) and no specific pattern is observed. Barring few outliers no gross deviations are observed.

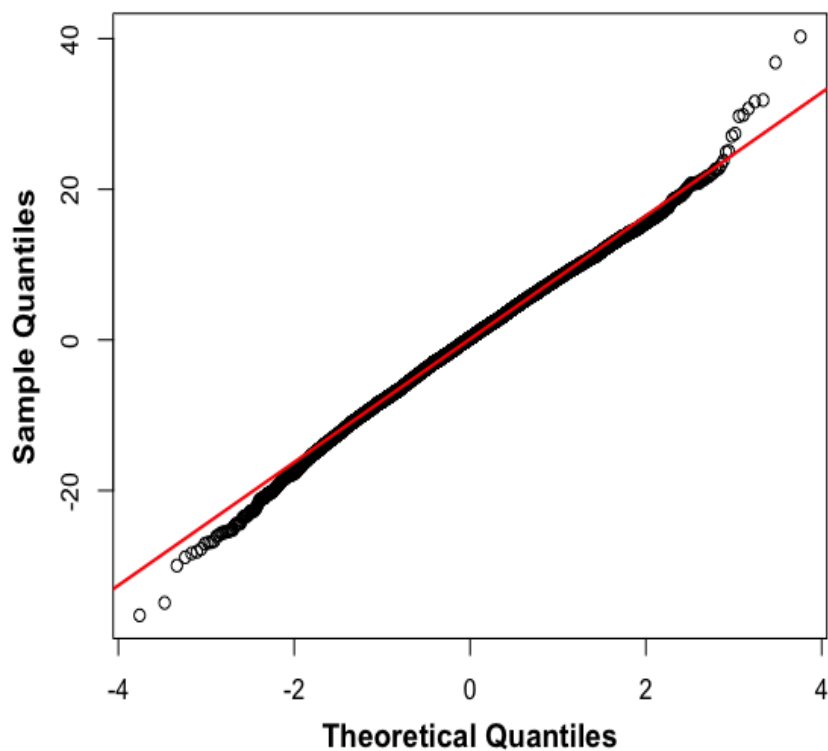


Figure 15: Normal quantile plot of the residuals for the Regression of Reading Score on Math Score, Birth weight and Gender.

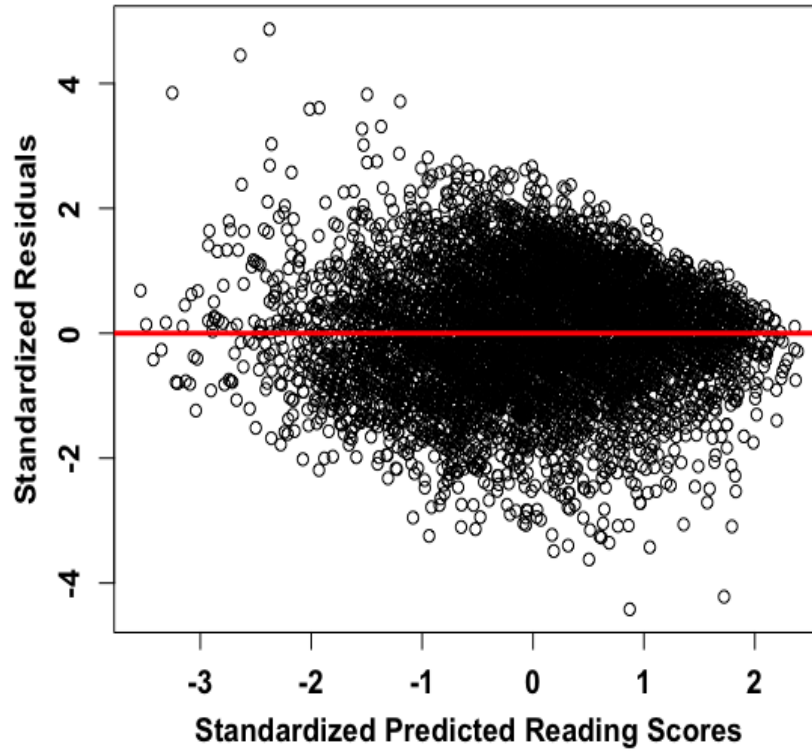


Figure 16: Plot of standardized residuals versus standardized predicted Reading score for the Regression of Reading Score on Math Score, Birth weight and Gender.

3.3 Family Income and Home ownership

3.3.1 Model 1: Home ownership

Descriptive summary of *Own_home* is presented in Table XV.

Table XV: Table summarizing home ownership status in the data.

No information	Yes	No
16 (0.25%)	3566 (55.92%)	2795 (43.83%)

Cases without information regarding home ownership are excluded. Equation 3 shows model to be fitted.

$$Reading_score = \beta_0 + \beta_1 Math_score + \beta_2 Birth_weight + \beta_3 Gender + \beta_4 Own_home + \varepsilon \quad (3)$$

where β_4 is the coefficient for home ownership. Other coefficients have explanations similar to presented for equation (2). Home ownership is a categorical variable with two levels (yes = has home ownership and no = no home ownership). Table XVI shows the results of multiple regression analysis.

Table XVI: Table showing results of multiple regression analysis of Reading score on Math score, Birth weight, Gender and Home ownership.

Predictor variable	Regression coefficient	95% CI	P-value
Math Score	0.77	(0.75, 0.79)	< 0.0001
Birth weight (in kgs)	0.83	(0.40, 1.26)	<0.0001
Gender (Female - Male)	2.62	(2.16, 3.08)	<0.0001
Home ownership(No - Yes)	-1.38	(0.90, 1.86)	<0.0001

3.3.2 Assumptions check

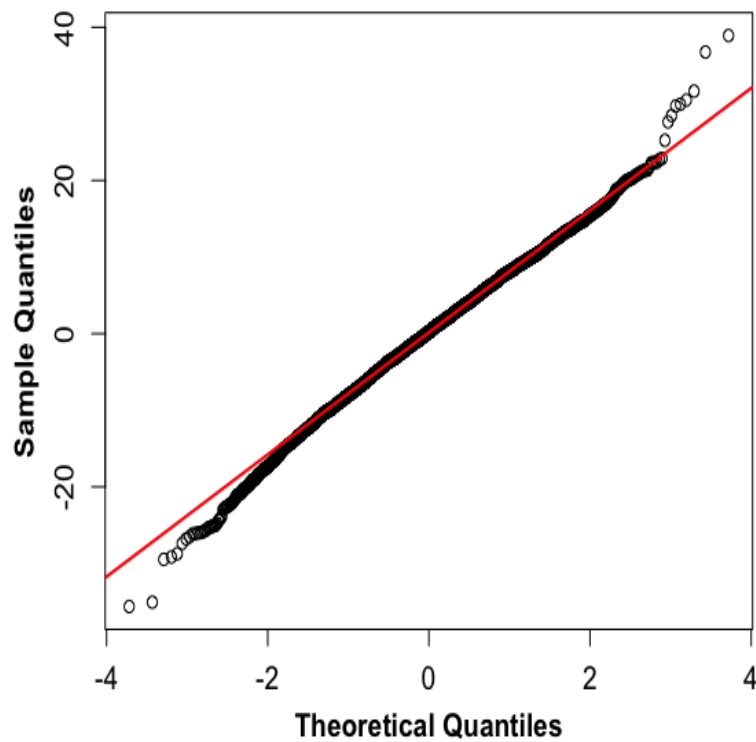


Figure 17: Normal quantile plot of the residuals for the Regression of Reading score on Math score, Birth weight, Gender and Home ownership.

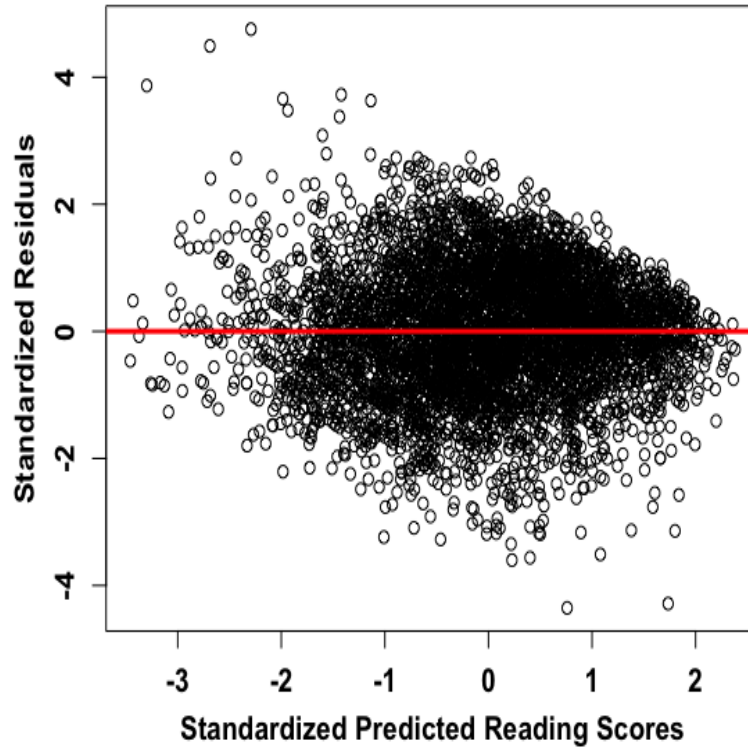


Figure 18: Plot of standardized residuals versus standardized predicted Reading score for the Regression of Reading Score on Math Score, Birth weight, Gender and Home ownership.

Both normality of residuals (Fig. 17) and homoscedasticity hold (Fig. 18) for the fitted model. Detailed explanation is similar to that in Section 3.2.1.

3.3.3 Model 2: Family income

Variable *Family_income* is considered as numeric with 7 income levels (Table XVII). Cases without any information are excluded from the analysis. Descriptive summary is provided in Table XVII. Integer notation is chosen to convey an intuitive sense of higher income associated with larger integer value.

Table XVII: Table showing number of respondents in the seven income groups.

Integer Notation	Gross weekly family income (in pounds)	Number of people (%)
1	Less than 35	87 (1.39%)
2	35-49	348(5.56%)
3	50-99	1840 (29.38%)
4	100-149	2173 (34.70%)
5	150-199	1041 (16.62%)
6	200-249	399 (6.37%)
7	More than 250	375 (5.99%)
		Total=6263 (100%)
		Missing =1174
		NA=1954
		Grand Total=9391

The regression equation to be fitted is:

$$Reading_Score = \beta_0 + \beta_1 Math_Score + \beta_2 Birth_weight + \beta_3 Gender + \beta_5 Family_income + \varepsilon \quad (4)$$

where β_5 is the coefficient for variable *Family_income*. Other coefficients have explanations similar as for equation (3). Table XVIII presents multiple regression results.

Table XVIII: Table showing the results of Multiple Regression of Reading score on Math score, Birth weight , Gender and Family income.

Predictor variable	Regression coefficient	95% CI	P-value
Math Score	0.77	(0.75, 0.79)	< 0.0001
Birth weight (in kgs)	0.75	(0.31, 1.19)	0.0035
Gender (Female - Male)	2.85	(2.39, 3.31)	<0.0001
Family income (Integer Levels)	0.69	(0.50, 0.89)	<0.0001

3.3.4 Assumptions checking

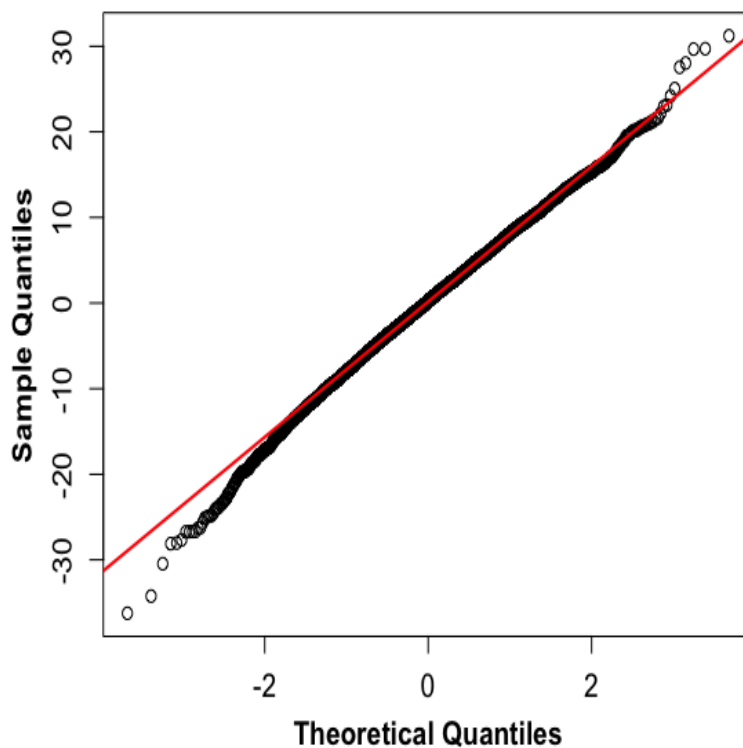


Figure 19: Normal quantile plot of the residuals for the Regression of Reading Score on Math Score, Birth weight, Gender and Family income.

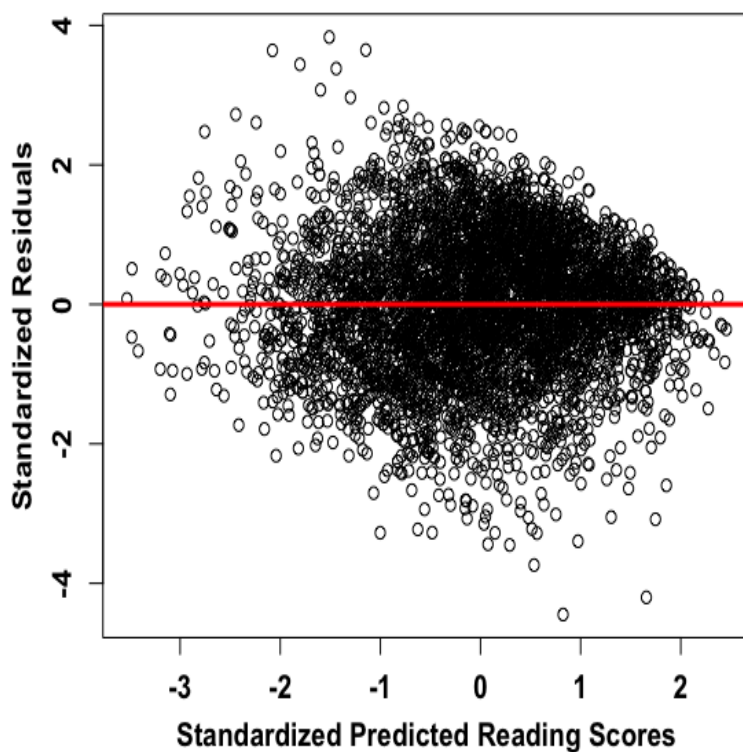


Figure 20: Plot of standardized residuals versus standardized predicted Reading score for the Regression of Reading Score on Math Score, Birth weight, Gender and Family income.

Both normality of residuals (Fig. 19) and homoscedasticity hold (Fig. 20). Detailed explanation similar to that in Section 3.2.1.

3.3.5 Model Comparison

Table XIX: Table describing the coefficients of explanatory variables for Model 1 and Model 2.

Predictor variables	Coefficient value for Model 1 (p-value)	Coefficient value for Model 2 (p-value)
Math score	0.77 (<0.0001)	0.77 (<0.0001)
Birth weight	0.83 (<0.0001)	0.69 (0.0035)
Gender (Female - Male)	2.62 (<0.0001)	2.78 (<0.0001)
Home ownership (No - Yes)	-1.38 (<0.0001)	NA
Family income (Integer Levels)	NA	0.61 (<0.0001)

Table XIX shows the comparison of Model 1 and Model 2. It is observed that:

1. Addition of home ownership and family income separately in the model does not affect the coefficient of Math score (coefficient=0.77, $p < 0.0001$).
2. Both home ownership (coefficient=-1.38, $p < 0.0001$) and family income (coefficient=0.61 $p < 0.0001$) are associated with a higher Reading score. Family income has a gradient effect on Reading scores, every one unit increase improves Reading score by 0.61 adjusting for other variables.

Table XX: Comparison of two models based on model statistics.

Model statistic	Model 1	Model 2
R-Squared	0.58	0.58
Mean Squared Error	67.03	66.43
Mean Absolute Error	6.45	6.43

As per Table XX the R-squared value for the two models is comparable and slight difference between the mean squared error and mean absolute error exists. It appears that Model 2 (with Family_income) provides a better fit.



4 Reading Support and Nations

4.1 Reading Support

$$Reading_Support = \begin{cases} 0 & \text{Reading Score more than or equal to 30} \\ 1 & \text{Reading Score less than 30} \end{cases}$$

Commands used to create Reading_support are shown in Fig. 21.

```

1
2 IF (Reading_score < 30) Reading_Support=1.
3 EXECUTE.
4
5 IF (Reading_score >= 30) Reading_Support=0.
6 EXECUTE.
7
8  FREQUENCIES VARIABLES=Reading_Support
9  /ORDER=ANALYSIS.
10 ►

```

Figure 21: SPSS Syntax used to compute the new variable Reading_Support.

Alternatively below steps may be followed in SPSS:

1. Transform → Compute Variable .
2. Put *Reading_support* in Target Variable. Numeric Expression=1 and at the bottom of the box click 'If (optional case selection condition)'.
3. Select 'Include if case satisfies condition': and put 'Reading_Score < 30'. Click Continue. Click OK.
4. Repeat same steps with Numeric Expression changed to 0 and condition in step 3 as 'Reading_Score >= 30'.

Table XXI reports the descriptive summary of Reading support.

Table XXI: Table describing requirement of Reading support in the data.

Reading Support Value=1	Reading Support Value=0	Valid Values	Missing Items	Total (Valid + Missing)
1337 (22.75%)	4541 (77.25%)	5878 (100%)	3513	9391

4.2 Nations

To create the variable Nations the SPSS Syntax in Fig. 22 was created.

```

*If the Region variable value is 7 or 8, variable Nations=2 indicating South Wales .
IF (Region=7 | Region=8) Nations=2.
EXECUTE.

*If the Region Variable value is 20, variable Nations=3 indicating Scotland.
IF (Region=20 ) Nations=3.
EXECUTE.

*If the Region variable value is 30, variable Nations=4 indicating Northern Ireland.
IF (Region=30) Nations=4.
EXECUTE.

*If the Region variable value is 98, variable Nations=5 indicating Overseas resident.
*This is done to avoid merging Overseas with NAs and for identifying cases for Nations=1, London.
IF (Region=98 ) Nations=5.
EXECUTE.

*If the Region variable value is neither 7,8,20,30 and 98, Nations=1 indicating London.
IF (Region~=7 & Region~=8 & Region~=20 & Region~=30 & Region~=98) Nations=1.
EXECUTE.

*Descriptive Frequencies for the computed Variable Nations.
FREQUENCIES VARIABLES=Nations
/ORDER=ANALYSIS.

```

Figure 22: SPSS Syntax used to compute the new variable Reading_Support.

Overseas residents (Nations=5) is created to facilitate the creation of Nations=1, London. Steps to create Nations using SPSS user interface are exactly as described in section 5.1 with conditions of Fig. 22. Table XXII summarizes the Nations variable.

Table XXII: Table describing number of residents across Nations.

Nations Value	Nations Label	Number of cases
1	London	7061 (81.72%)
2	Wales	424 (4.91%)
3	Scotland	818 (9.47%)
4	Northern Ireland	333 (3.85%)
5	Overseas	4 (0.05%)
		8640(100%)

4.3 Reading Support and Nations

Table XXIII: Contingency table showing distribution of Reading support requirement across Nations.

Nation	Reading Support Required =No	Reading Support Required =Yes	Total Valid Cases	Population in Sample
London	3779 (76.48%)	1162 (23.52%)	4941 (100%)	7061
Wales	250 (79.62%)	64 (20.38%)	314 (100%)	424
Scotland	506 (82.41%)	108 (17.59%)	614 (100%)	818
Northern Ireland	0	0	0	333
UK(Total)	4535	1334	5879	8636

According to Table XXIII :

1. 24% of London respondents require Reading support, highest among Nations. The corresponding percentage is 20% for Wales and 17% for Scotland.
2. Scotland scores highest in terms of non-requirement of Reading support with 83% of residents scoring more than or equal to 30 in the Reading test indicating that children in Scotland might have higher reading abilities in comparison to other nations.
3. Almost one-fifth of the sample in Wales (20.38%) indicate the requirement for Reading support.

4.4 Logistic Regression

Reading_support is a binary variable, hence, Logistic Regression is used. The model equation can be written as:

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 \text{Math_score} + \beta_2 \text{Gender} + \beta_3 \text{Family_income} + \beta_4 \text{Nations} \quad (5)$$

where p is the probability of Reading support requirement, β_0 is constant, β_1 , β_2 , β_3 and β_4 are the coefficients of Math score, *Gender*, Family income and Nations variable respectively. Nations=4 has no data for Reading support (Table XXIII), hence, excluded from the analysis. Left with 3 values of Nations, two dummy variables are created: Nation2 and Nation3.

- Nation2=1 if Nations=2 (Wales) and 0 otherwise.
- Nation3=1 if Nations=3 (Scotland) and 0 otherwise.
- When Nation2 and Nation3 are both 0, Nation=1 (London) is the case.

Table XXIV shows the results of Logistic Regression.

Table XXIV: Adjusted odds ratio for effects of explanatory variables on Reading support requirement.

Variable	Odds ratio ¹	95%CI	P value
Gender (Female/Male)	0.52	(0.43, 0.62)	<0.0001
Nations			
London ²	1		0.15
Wales	0.67	(0.45, 1.0)	0.051
Scotland	0.96	(0.72, 1.30)	0.78
Family Income	0.82	(0.76, 0.89)	<0.0001
Math score	0.84	(0.83, 0.85)	<0.0001

¹ Adjusted for the other variables in the table.

² Reference Category.

The proportion of Reading support requirement was lesser for females than males (19% vs 25%). After adjustment, Female sex as associated with a 48% (OR=0.52 [95%CI: 0.43, 0.62]) decrease in odds of reading support requirement. Adjusting for all variables Wales showed a 33% (OR=0.67 [95%CI: 0.45, 1.0]) decrease in odds of requirement of Reading support. Scotland was found to be non-significant (OR=0.96 [95%CI: 0.72, 1.30]). Family income was associated with 18% (OR=0.82 [95%CI: 0.76, 0.89]) decrease in odds of Reading support requirement for a unit increase in the level of income and Math score a 16% (OR=0.84 [95%CI: 0.83, 0.85]) decrease in odds of Reading support requirement for a unit increase in score. Hosmer-and-Lemeshow Test p-value is 0.30 indicating that model is a good fit for the data.

Appendices

5 Shapiro-Wilk Test Results

Table XXV: Table showing the results of Shapiro-Wilk test result to check the normality of Reading Scores across the regions in the dataset.

Region	Shapiro-Wilk Statistic	P-value
North	0.98	p<0.0001
Yorks Hmbrside	0.98	p<0.0001
North West	0.97	p<0.0001
East Midlands	0.98	p<0.0001
West Midlands	0.98	p<0.0001
East Anglia	0.92	p<0.0001
South West	0.98	p<0.0001
South Wales	0.98	0.01
North Wales	0.94	0.001
South East	0.96	p<0.0001
London	0.97	p<0.0001
Scotland	0.98	p<0.0001

6 Tukey HSD

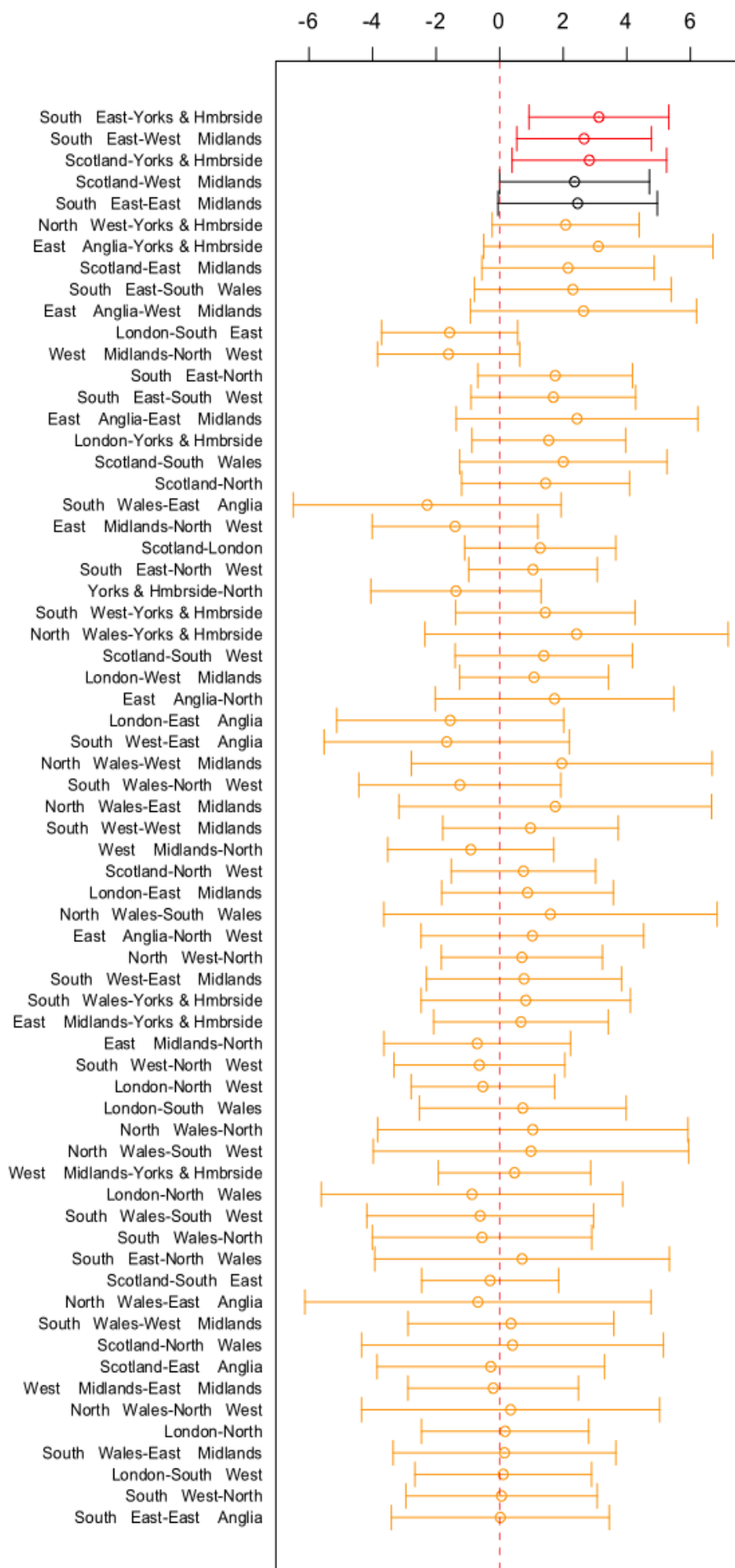


Figure 23: Figure displaying the result of Tukey Honest significant Differences on Reading Scores and Region. Red error bars indicate statistical significance, Black indicates marginal statistical significance and Orange indicates non-significance. Middle point in error bars indicates estimated mean difference.