AFRICAN INSTITUTE FOR MATHEMATICAL SCIENCES

(AIMS RWANDA, KIGALI)

Name: Stephen Kiilu
Course: Statistical Machine Learning

Assignment Number: BDML1
Date: January 10, 2021

Question 1

1. $\chi = \mathbb{R}$ and dimension of $\mathbb{R} = 1$
2. Output space,$\Upsilon = \mathbb{R}$ and dimension of $\mathbb{R} = 1$.
3. $\mathbf{x} \in \mathbb{R}^{n\times 1}$ and dimension is $n \times 1$
4. $\mathbf{y} \in \mathbb{R}^{n\times 1}$ and dimension is $n \times 1$
5. $P(Y_i|X_i = x_i)$
   $\mathbf{y}_i = f(\mathbf{x}_i) + \epsilon, \forall_i \sim N(\theta_i x_i, \sigma^2)$
   The distribution of $P(\mathbf{y}|\mathbf{x} = x)$ is N $\sim (\mathbf{x}\theta, \sigma^2)$.
6. $\mathbf{y}_i = f(\mathbf{x}_i) + \epsilon_i$
   $\mathbf{y}_i = \theta\mathbf{x}_1 + \epsilon_i$ with $\epsilon_1 \sim N(0, \sigma^2), \forall_1 = 1, \cdots, n$, Which is a Homoscedastic Gaussian noise model.
7. Regression because the output space is continuous.
8. $\frac{\partial SSE_n(a)}{\partial a}$

$$\begin{aligned} SSE_n(a) &= (\mathbf{y} - a\mathbf{x})^T(\mathbf{y} - a\mathbf{x}) \\ &= \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{x}a - (\mathbf{x}a)^T\mathbf{y} + (ax)^T\mathbf{x}a \\ \frac{\partial SSE_n(a)}{\partial a} &= a(\mathbf{x}^T\mathbf{x}) - \mathbf{x}^T\mathbf{y} \end{aligned}$$

9. $\frac{\partial SSE_n(a)}{\partial a}$

$$
\begin{aligned}
\frac{\partial SSE_n(a)}{\partial a} &= a(\mathbf{x}^T\mathbf{x}) - \mathbf{x}^T\mathbf{y} = 0 \\
(\mathbf{x}^T\mathbf{x})a &= \mathbf{x}^T\mathbf{y} \\
a &= \frac{\mathbf{x}^T\mathbf{y}}{\mathbf{x}^T\mathbf{x}} \\
\theta &= \frac{\mathbf{x}^T\mathbf{y}}{\mathbf{x}^T\mathbf{x}}
\end{aligned}
$$

10.

$$
\begin{aligned}
\mathbb{E}(\hat{\theta}) &= (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbb{E}(\mathbf{y}) \\
&= (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbb{E}(\mathbf{x}\theta + \epsilon) \\
&= (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{x}\theta \\
&= \theta
\end{aligned}
$$

11.

$$
\begin{aligned}
\mathbb{V}(\theta) &= (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T Var(\mathbf{y})\mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1} \\
&= (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}\sigma^2 I_n\mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1} \\
&= \sigma^2(\mathbf{x}^T\mathbf{x})^{-1}
\end{aligned}
$$

12.

$$
\begin{aligned}
\mathbb{E}(Y_i|X_i) &= \mathbb{E}(\theta\mathbf{x}_1 + \epsilon) \\
&= \theta\,\mathbb{E}(\mathbf{x}_i) + \mathbb{E}(\epsilon_i)\, X_i \in \mathbb{R} \text{ and } \epsilon_i \sim N(0, \sigma^2). \\
&= \theta\mathbf{x}_i \\
\mathbb{E}(\mathbf{y}|\mathbf{x}) &= \theta\mathbf{x}\ , \mathbf{x} \in \mathbb{R}^{n\times 1}\ , \theta \in \mathbb{R}
\end{aligned}
$$

13.

$$
\begin{aligned}
\mathbf{y} &= f(\mathbf{x}) + \epsilon \\
\mathbf{y} &= \theta\mathbf{x} + \epsilon, \quad \epsilon \sim N(0, \sigma_I^2)
\end{aligned}
$$

14.

$$
\begin{aligned}
\mathbf{y} &= \theta\mathbf{x} + \epsilon \\
\hat{\mathbf{y}} &= \theta\mathbf{x} \\
&= \mathbf{x}((\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{y})
\end{aligned}
$$

15

$$
\begin{aligned}
\mathbb{V}(Y_i|X_i) &= \mathbb{V}(f(\mathbf{x}_i) + \epsilon_i) \\
&= \mathbb{V}f(\mathbf{x}_i) + \mathbb{V}(\epsilon_i) \\
&= \sigma^2
\end{aligned}
$$

$$\begin{aligned}\mathbb{V}(\mathbf{y}|\mathbf{x}) &= \mathbb{V}(f(\mathbf{x}) + \epsilon)\\ &= \mathbb{V}f(\mathbf{x}) + \mathbb{V}(\epsilon)\\ &= \sigma^2 I_n\end{aligned}$$

16

$$\begin{aligned}\mathbb{E}(\hat{Y}_i|X_i) &= \mathbb{E}(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{y}\mathbf{x}_i)\\ &= \mathbf{x}_i((\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbb{E}(\theta\mathbf{x} + (\epsilon)\\ &= \theta\mathbf{x}_i\end{aligned}$$

17

$$\begin{aligned}\mathbb{V}(\hat{Y}_i|X_i) &= \mathbb{V}(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{y}\mathbf{x}_i)\\ &= \mathbb{V}(\mathbf{x}_i(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\epsilon\\ &= \mathbf{x}_i^2\sigma^2(\mathbf{x}^T\mathbf{x})^{-1}\\ \mathbb{V}(\hat{\mathbf{y}}|\mathbf{x}) &= (\mathbf{x}^T\mathbf{x})\sigma^2(\mathbf{x}^T\mathbf{x})^{-1}\\ &= \sigma^2.\end{aligned}$$

18

$$\begin{aligned}\mathbb{E}(\theta &= \theta\\ \mathbb{V}(\theta) &= \sigma^2(\mathbf{x}^T\mathbf{x})^{-1}\\ &\theta \sim N(\theta, \sigma^2(\mathbf{x}^T\mathbf{x})^{-1})\end{aligned}$$

19

$$\begin{aligned}\mathbf{y}_i\hat{|}\mathbf{x}_i \sim N(\theta\mathbf{x}_i, \mathbf{x}_i^2\sigma^2(\mathbf{x}^T\mathbf{x})^{-1})\\ \mathbf{y}_i\hat{|}\mathbf{x}_i \sim N(\theta\mathbf{x}, \sigma^2 I_n)\end{aligned}$$

20

$$\hat{\sigma^2} = \frac{SSE}{n-p-1} = \sum_{i=1}^{n}\frac{(\mathbf{y}_i - \mathbf{y}_i\hat{)}^2}{n-p-1}$$

Question 2

1. Thid data is obtained from a DNA microarray survey. The goal is to use different DNA gene expression to predict if one has prostrate cancer or not.
   A set of 500 different DNA gene combination were used to model if one has a probability of having prostrate cancer. Y is the response variable, 0 means that one has no prostrate cancer and 1 means one has prostrate cancer. All other 500 are different DNA gene combination to explain Y.

2. We plot barplot of the response variable.

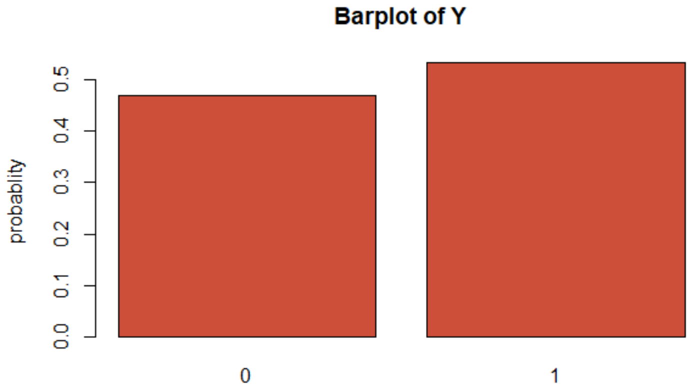

Figure 1: histogram

The probability of response is almost equal, that response 0 and response 1 is approximately 0.5.

3. We can make some the following comments about the data set.
The dimension of input space χ is 500, n, the sample size is 79.
n is small and p is large i.e p>>>n, this implies that we have a high dimensional data.This is called poverty in data.

4. We carry out missing value analysis and find out that our data has no missing values. The we randomly select variables and draw some descriptive statistics; we can visualize our data graphically

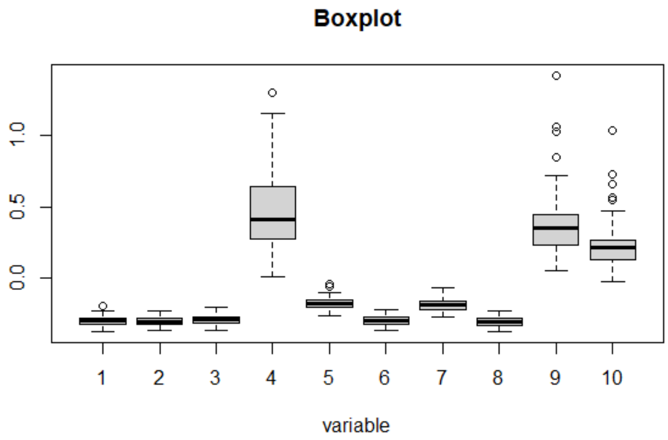

Figure 2: boxplot

Some variables e.g variable 9 and 10 show existence of outliers in our data set
The measurement scales are different, some variables are taking negative values. This

implies that our data set is non homogeneous.

The median values is very different across most of the variables and some of the median assumes negative values.

The spread from the median in some of variables is different e.g in variables 4, 9 and 10 the variance is large, while some variables like variables 1, 2, 3 show almost the same spread from the median.

Some variables like variable 1, variable 2, variable 3 have median at the center, this shows that they follow a normal distribution, while other variables like variable 4 and variable 10 do not follow a normal distribution.

```
Y             X206212_at         X207075_at        X215872_at        X201876_at
Min.   :0.0000   Min.   :-0.27572   Min.   :-0.3700   Min.   :-0.3745   Min.   :-0
1st Qu.:0.0000   1st Qu.:-0.22092   1st Qu.:-0.3189   1st Qu.:-0.3273   1st Qu.: 0
Median :1.0000   Median :-0.19108   Median :-0.2912   Median :-0.3019   Median : 0
Mean   :0.5316   Mean   :-0.18635   Mean   :-0.2945   Mean   :-0.3027   Mean   : 0
3rd Qu.:1.0000   3rd Qu.:-0.16036   3rd Qu.:-0.2733   3rd Qu.:-0.2783   3rd Qu.: 0
Max.   :1.0000   Max.   :-0.07154   Max.   :-0.2076   Max.   :-0.1972   Max.   : 1
  X211935_at        X206788_s_at       X216441_at        X209290_s_at       X219877
Min.   :0.04825   Min.   :-0.26716   Min.   :-0.3771   Min.   :0.00699   Min.   :-
1st Qu.:0.22764   1st Qu.:-0.20338   1st Qu.:-0.3304   1st Qu.:0.27076   1st Qu.:-
Median :0.35430   Median :-0.18122   Median :-0.3080   Median :0.40936   Median :-
Mean   :0.37536   Mean   :-0.17665   Mean   :-0.3074   Mean   :0.46183   Mean   :-
3rd Qu.:0.44529   3rd Qu.:-0.15582   3rd Qu.:-0.2827   3rd Qu.:0.64654   3rd Qu.:-
Max.   :1.42433   Max.   :-0.03853   Max.   :-0.2293   Max.   :1.30156   Max.   :-
```

From the above summary statistics we can conclude that;

The mean value of most variables lies between negative -1 and 1. This is also true to the median value.

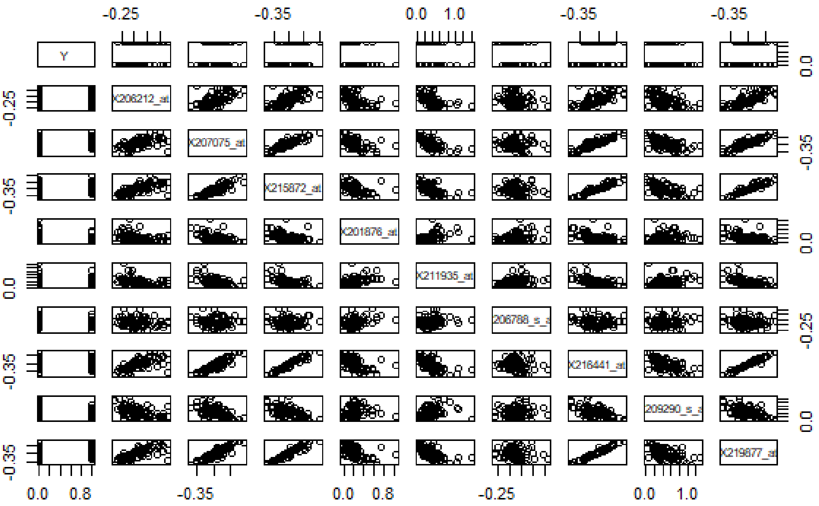

Figure 3: scatter plot

From the above scatter plot, the explanatory variables very weak or no linear relation with the response variable y. The variables indicate strong linear relationship among themselves, this is what we refer to as multicollinearity in data.

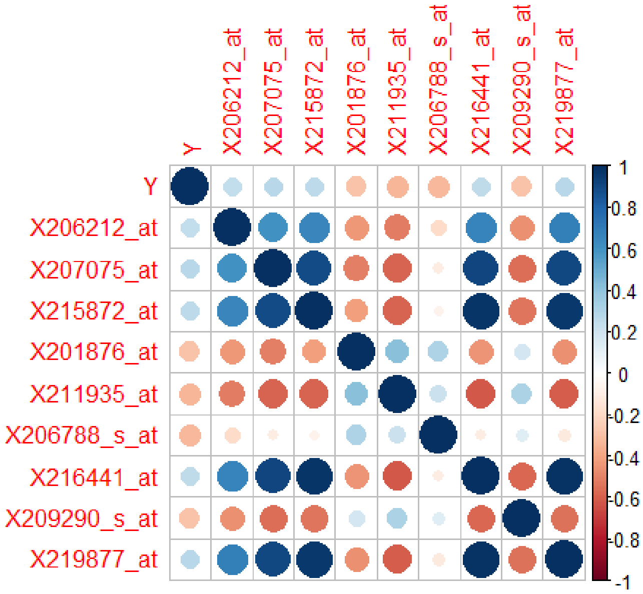

Figure 4: corrplot

The cor plot visualizes the very weak linear relationship between the explanatory variables and the response variables y.

Question 3

1.

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 1 & -2 & 4 \\ -2 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & -2 \\ -2 & 1 \\ 4 & 1 \end{bmatrix} = \begin{bmatrix} 21 & 0 \\ 0 & 6 \end{bmatrix}$$

2. $\mathbf{X}^T\mathbf{X} \in \mathbb{R}^{2\times 2}$ and the shape is 2×2

3.

$$(\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{26} \begin{bmatrix} 21 & 0 \\ 0 & 6 \end{bmatrix}$$

4.

$$\hat{\theta} = \frac{1}{26} \begin{bmatrix} 21 & 0 \\ 0 & 6 \end{bmatrix} \begin{bmatrix} 1 & -2 & 4 \\ -2 & 1 & 1 \end{bmatrix} \begin{bmatrix} -5 \\ 4 \\ -3 \end{bmatrix} = \begin{bmatrix} -1.19 \\ 1.83 \end{bmatrix}$$

5.

$$\begin{aligned} \hat{\mathbf{Y}} &= \mathbf{X}\hat{\theta} \\ &= \begin{bmatrix} 1 & -2 \\ -2 & 1 \\ 4 & 1 \end{bmatrix} \begin{bmatrix} -1.19 \\ 1.83 \end{bmatrix} = \begin{bmatrix} -4.86 \\ 4.21 \\ -2.92 \end{bmatrix} \end{aligned}$$

6.

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \begin{bmatrix} -0.143 \\ -0.214 \\ -0.071 \end{bmatrix}$$

7.

$$\begin{aligned} SSE(\hat{\theta}) &= \sum_{i=1}^{n} (\mathbf{Y}_i - \hat{\mathbf{Y}}_i)^2 \\ &= (-5 + 4.86)^2 + (4 - 4.21)^2 + (-3 + 2.92)^2 \\ &= 0.0701 \end{aligned}$$

8.

$$\begin{aligned} \frac{SSE(\hat{\theta})}{n-2} &= \frac{0.0701}{1} \\ &= 0.0701 \end{aligned}$$

9.

$$\hat{\sigma^2}(\mathbf{X}^T\mathbf{X})^{-1} = \begin{bmatrix} 0.0033 & 0.0000 \\ 0.0000 & 0.0117 \end{bmatrix}$$

10.

$$(\mathbf{X}^T\mathbf{X}) = \begin{bmatrix} 1 & 2 \\ -2 & 1 \end{bmatrix} \begin{bmatrix} 1 & -2 \\ -2 & 1 \end{bmatrix} = \begin{bmatrix} 5 & 0 \\ 0 & 5 \end{bmatrix} = 5 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Question 4

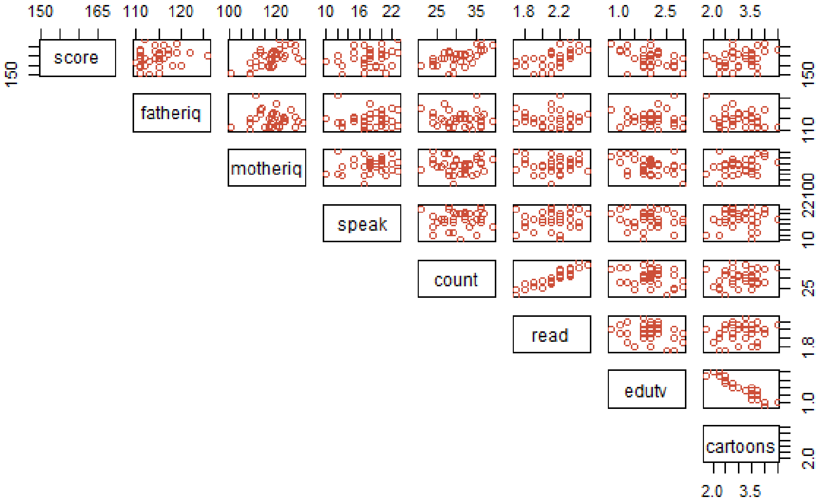

Figure 5: Scatterplot

1. From the above upper triangular pairwise scatterplot for this data, the explanatory variables which have the strongest linear relationship with the response are **motheriq**, **read** and **count**. You cannot tell from this plot the strongest related of all predictor variables.

2.
```
            score fatheriq motheriq   speak   count    read   edutv cartoons
score      1.0000   0.1881   0.5712  0.2679  0.5442  0.5252 -0.3703   0.2451
fatheriq   0.1881   1.0000  -0.0248 -0.0305 -0.0750 -0.0682  0.1162  -0.2484
motheriq   0.5712  -0.0248   1.0000  0.0722  0.0243 -0.0430 -0.3300   0.3384
speak      0.2679  -0.0305   0.0722  1.0000  0.0595  0.1851 -0.1545   0.1094
count      0.5442  -0.0750   0.0243  0.0595  1.0000  0.9103 -0.2157   0.1549
read       0.5252  -0.0682  -0.0430  0.1851  0.9103  1.0000 -0.1666   0.1257
edutv     -0.3703   0.1162  -0.3300 -0.1545 -0.2157 -0.1666  1.0000  -0.9234
cartoons   0.2451  -0.2484   0.3384  0.1094  0.1549  0.1257 -0.9234   1.0000
```

   From the correlation matrix above, among all explanatory variables **motheriq** has the strongest linear relationship with the response, count. The coefficient of correlation between **motheriq** and response variable **score** is 0.57. This is moderate linear relationship.

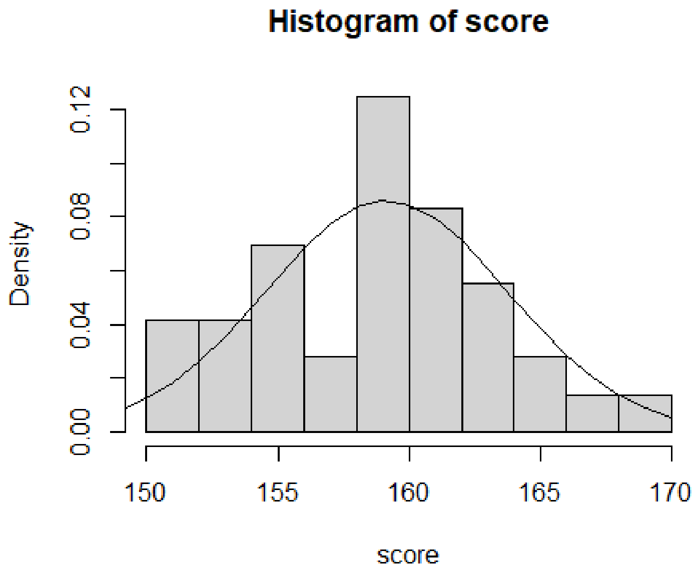

Figure 6: histogram

3. From the above histogram, the response variable follows a normal distribution. We can test the normality by using Shapiro test.
   Shapiro test of normality, the null hypothesis is, the null hypothesis is the data follow a normal distribution, and the alternative hypothesis is that the data is not normally distributed. From Shapiro test of normality the p-value is 0.76, this shows that the response variable follows a normal distribution.

4. We perform simple linear regression using **motheriq**, because its most important predictor variable.

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 111.0930    11.8567   9.370 6.02e-11 ***
motheriq      0.4066     0.1002   4.058 0.000274 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.856 on 34 degrees of freedom
Multiple R-squared:  0.3263,Adjusted R-squared:  0.3065
F-statistic: 16.47 on 1 and 34 DF,  p-value: 0.000274
```

Our regression model is given by:

$$\hat{\mathbf{y}} = 0.04066\,\mathbf{motheriq} + 111.09$$

Where $\hat{\mathbf{y}}$ is the predicted **score**

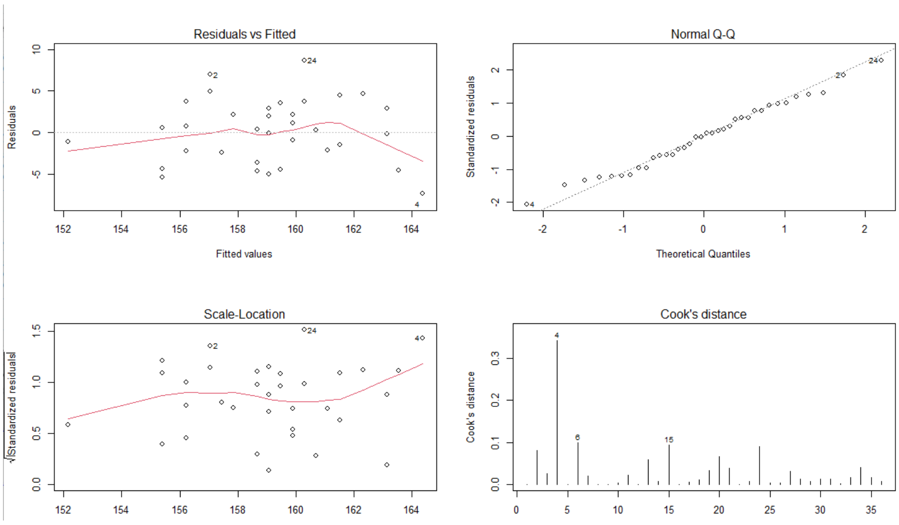

Figure 7: Residual analysis

5. **Linear association** - The first plot from left is a plot of residuals against fitted values, it shows residuals are not random. The assumption of linear association does not hold.

   Normality- for the assumption of normality to be true, the residuals should follow a straight line, from our second plot most data points follow a straight line which show that the residues are normally distributed, consequently the assumption of normality holds.

   Homoskedasticity - for this assumption, we expect that the residuals to have a constant variance, from our third plot the residues have a same variance. Homoskedasticity assumption is not violated.

   Check for outliers - we observe from our fourth plot that are significant outliers in our data. The outliers need to be investigated and a decision reached whether to remove or retail them in our data.
   Consequently the assumption of linear association and present of outliers affect the

suitability and quality of our model. Our model is not the best and should be improved.

6.

$$\hat{\mathbf{y}} = 0.04066\,\mathbf{motheriq} + 111.09$$

For every unit increase in **motheriq**, the expected **score** increase by 0.4066.

7.

```
       fit      lwr      upr
1 151.7524 147.8297 155.6752
```

From the R output, the fitted **score** for **motheriq** of 100 is 151.75. The confidence interval (147.83,155.67) is the range of true value of **score** at 95% confidence interval according to our model.

```
       fit      lwr      upr
1 151.7524 142.9896 160.5153
```

The fitted **score** for **motheriq** remains the same at 151.75. The predicted interval is (142.9896, 160.515). This means, according to our model 95% of children with a **motheriq** of 100 have a score between 142.98 and 160.52.

We can also plot confidence interval ;

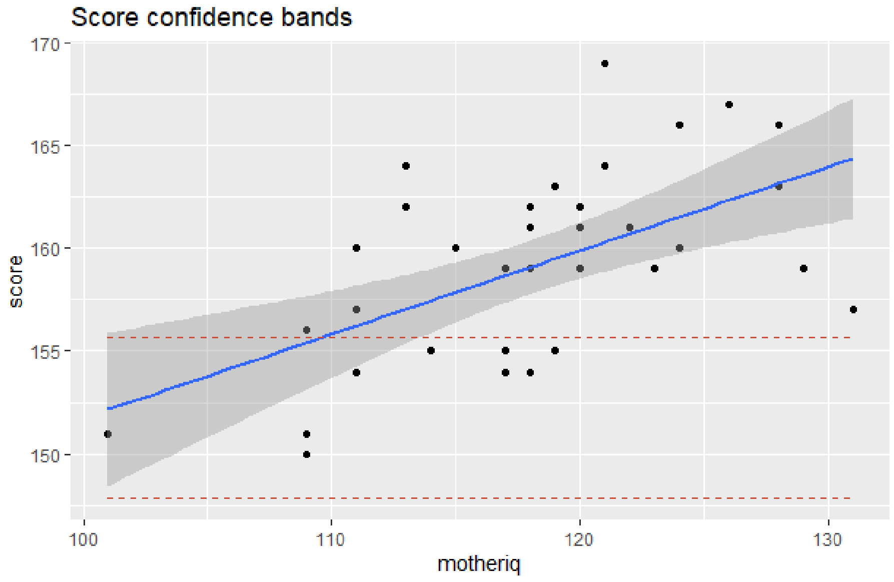

Figure 8: confidence bands

8. The multiple linear regression indicates that among all explanatory variables only **motheriq**, which is significant to our model, with a p-value of < 0.001. The predicted bands can plotted as below;

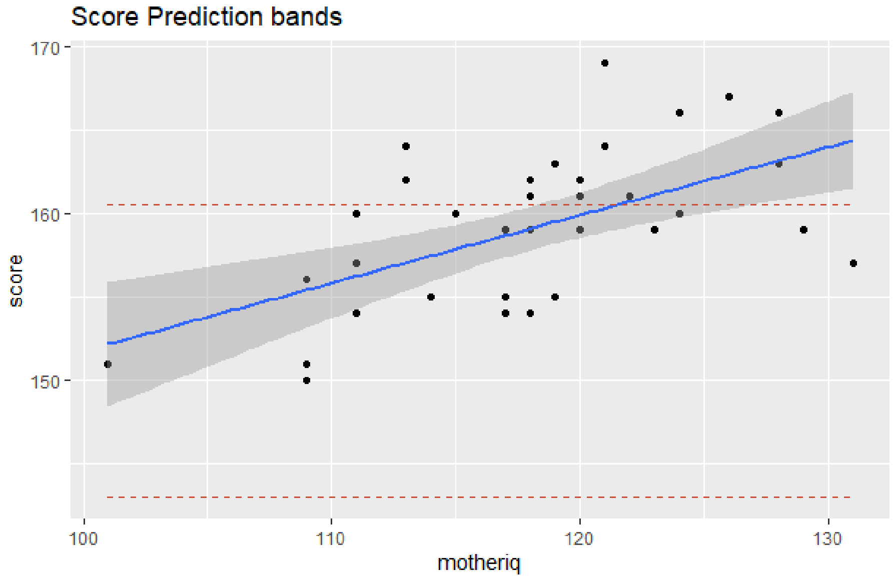

Figure 9: predicted bands

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 75.50849   24.02618   3.143  0.00393 **
fatheriq     0.25249    0.13756   1.835  0.07707 .
motheriq     0.40007    0.07291   5.488 7.33e-06 ***
speak        0.18764    0.14767   1.271  0.21429
count        0.20649    0.26631   0.775  0.44462
read         7.54405    5.58640   1.350  0.18769
edutv       -4.20244    2.24503  -1.872  0.07170 .
cartoons    -3.33899    2.01808  -1.655  0.10919
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.591 on 28 degrees of freedom
Multiple R-squared:  0.7496,Adjusted R-squared:  0.687
F-statistic: 11.97 on 7 and 28 DF,  p-value: 5.803e-07
```

Our multiple regression model is given by;

$\hat{y}$=0.252 **fatheriq** + 0.4000 **motheriq** + 0.188 **speak** + 0.206 **count** +7.54 **read** - 4.202 **edutv** - 3.339 **cartoons**.

$\hat{y}$ is the predicted score.