Advanced Mathematical and Statistics (MTH – 522) Homework 1 Report – Linear Regression on Pearson's father-son data).

Rowan Raj Ignatius / 02169507

Source code:

```
#Part A:Loading the data from the library 'Using R' and specifying the data set.
library(UsingR);
data(father.son);

#Part B:Plotting the values of the given data set in the scatter plot.
plot(father.son$fheight, father.son$sheight,
    xlab="Father's height (in)",
    ylab="Son's height (in)",
    pch=1);

#Part C: Adding the Regression line on the scatter plot
sons_h <- father.son$sheight
fathers_h <- father.son$fheight
regression_line <- lm(sons_h ~ fathers_h, data = father.son)
abline(regression_line, col='red')

#Part D: Adding the SD line
slope_SF <- sd(sons_h)/ sd(fathers_h)
mean_S <- mean(sons_h)
mean_F <- mean(fathers_h)
#using the straight line equation y-y1 = m(x-x1)
x <- 0 #x by default will be zero if interception in concerned
intercept <- slope_SF * (x - mean_F )+ mean_S
abline(a= intercept, b = slope_SF, col='blue', lty=4, lwd=3)

#Part E: Marking the center point of regression
points(mean(fathers_h), mean(sons_h) ,
```

```
    col='yellow',
    pch= 20)


#Part F: Extending the X & Y axis line through the center of regression
abline(v=mean(fathers_h), col="green")
abline(h=mean(sons_h), col="green")


#Part G: Out put of Linear regression.
summary(regression_line)
```

Code Explanation:

A) Fetching the specified data

Since the ‘UsingR’ library has already been installed, beginning the code by calling the library and the specific data set.

library(UsingR);

data(father.son);

HW1- AMS.R* | father.son | Untitled1*

Filter

	fheight	sheight
1	65.04851	59.77827
2	63.25094	63.21404
3	64.95532	63.34242
4	65.75250	62.79238
5	61.13723	64.28113
6	63.02254	64.24221
7	65.37053	64.08231
8	64.72398	63.99574
9	66.06509	64.61338
10	66.96738	63.97944
11	59.00800	65.24451
12	62.93203	65.35102

Showing 1 to 13 of 1,078 entries, 2 total columns

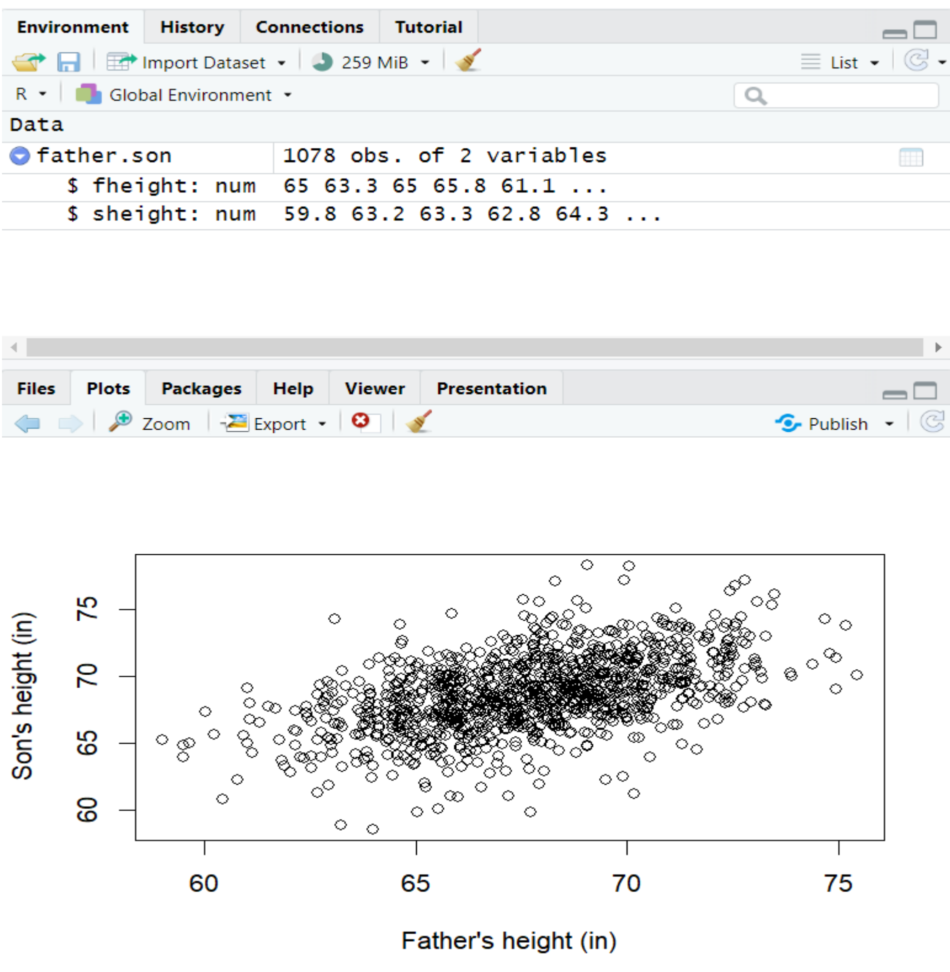
B) Plotting the given data on the scatter plot.

```
plot(father.son$fheight, father.son$sheight,
    xlab="Father's height (in)",
    ylab="Son's height (in)",
    pch=1);
```

Using the **Plot(x,y)** function I have specified the data that is to be plotted on the X and Y axis. Here, I have plotted Father's height on the X axis and the son's height on the Y axis.

On the following lines, using the **xlab** and **ylab** arguments, I have labeled the X and Y axis as *Father's height (in)* and *Son's height (in)* respectively.

With **pch** as 1 I have used circles to plot the data points.



Rowan Raj Ignatius
Student ID: 02169507

C) Adding the Regression line

sons_h <- father.son$sheight

fathers_h <- father.son$fheight

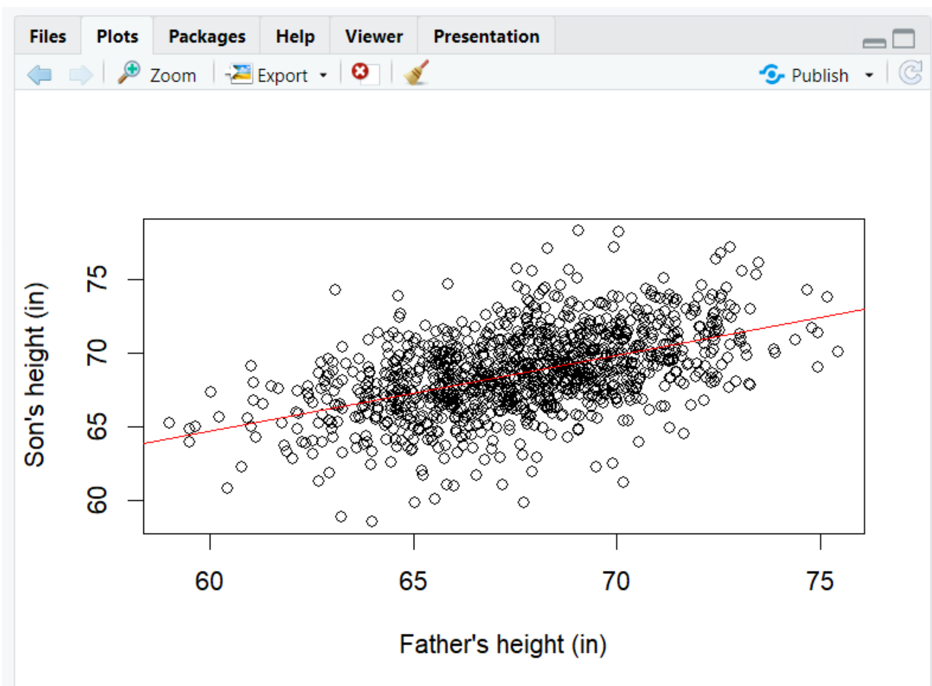
regression_line <- lm(sons_h ~ fathers_h, data = father.son)

abline(regression_line, col='red')

I have assigned variables '**sons_h'** and **'fathers_h'** for sons and fathers height from the data set (father.son).

On the following the line using the '**lm()**' function, I have calculated the linear regression between son's height (dependent variable) with the father's height (independent variable) and have stored it value in the variable 'regression line'.

With the help if '**abline'** function, I have added the regression line and specified the color of the line as Red



d) Add the SD line (with blue color, different from the regression line) to the same plot

slope_SF <- sd(sons_h)/ sd(fathers_h)

mean_S <- mean(sons_h)

mean_F <- mean(fathers_h)

x <- 0

intercept <- slope_SF * (x - mean_F)+ mean_S

abline(a= intercept, b = slope_SF, col='blue' *lty=4, lwd=3*)

I have found the standard deviation of the father_h and son_h using the **sd(father_h)** and **sd(son_h)** function respectively and the mean of the father_h and son_h using the **mean(father_h)** and **mean(son_h)** function respectively.

I have assigned variables slope_SF, mean_S and mean_F for slope (standard deviation of son_h divided by the standard deviation of father_h).

The SD line passes through the center of regression, (mean(X), mean(Y))=(67.6871, 68.68407)

So the equation is given as:
y-y1=slope*(x-x1)

y-68.68407=slope*(x-67.6871),

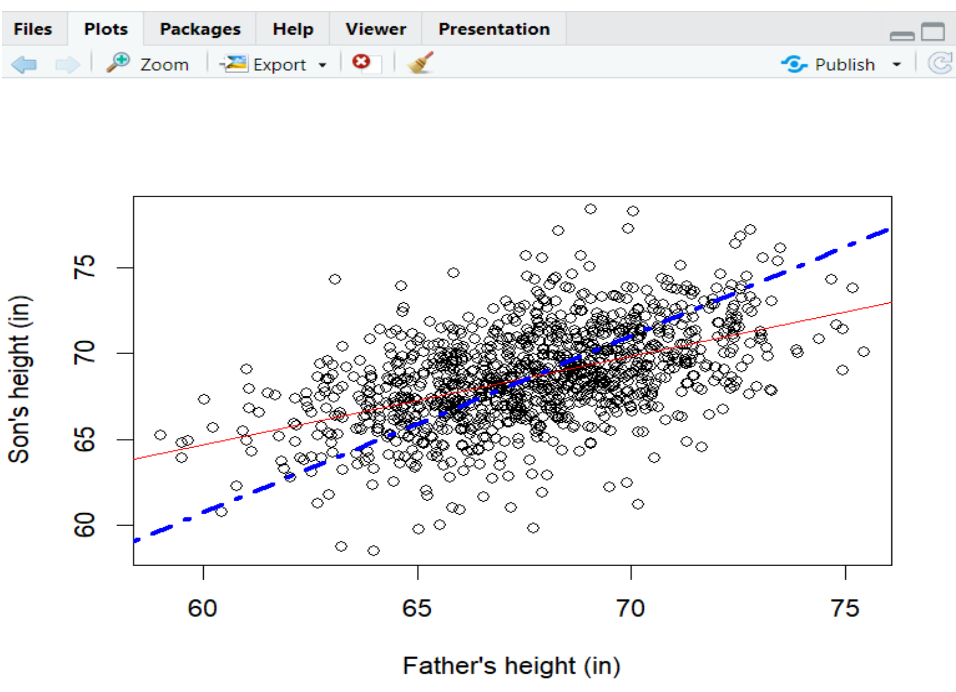
which by re-arranging yields:

y=1.025441*x + (68.68407-67.6871*1.025441).

Since we are looking to find the y intercept the value of x=0
Thus I have assigned the variable **'intercept'** to perform the calculation '**slope_SF * (x - mean_F)+ mean_S'**

To add the SD line, using R command **abline(a= intercept, b = slope_SF, col='blue', *lty=4, lwd=3*)**

The command **col** specifies the color of the line and the **lty** specifies the appearance of the line in this case '4' (a long dashed line), **lwd** specifies the width of the line.

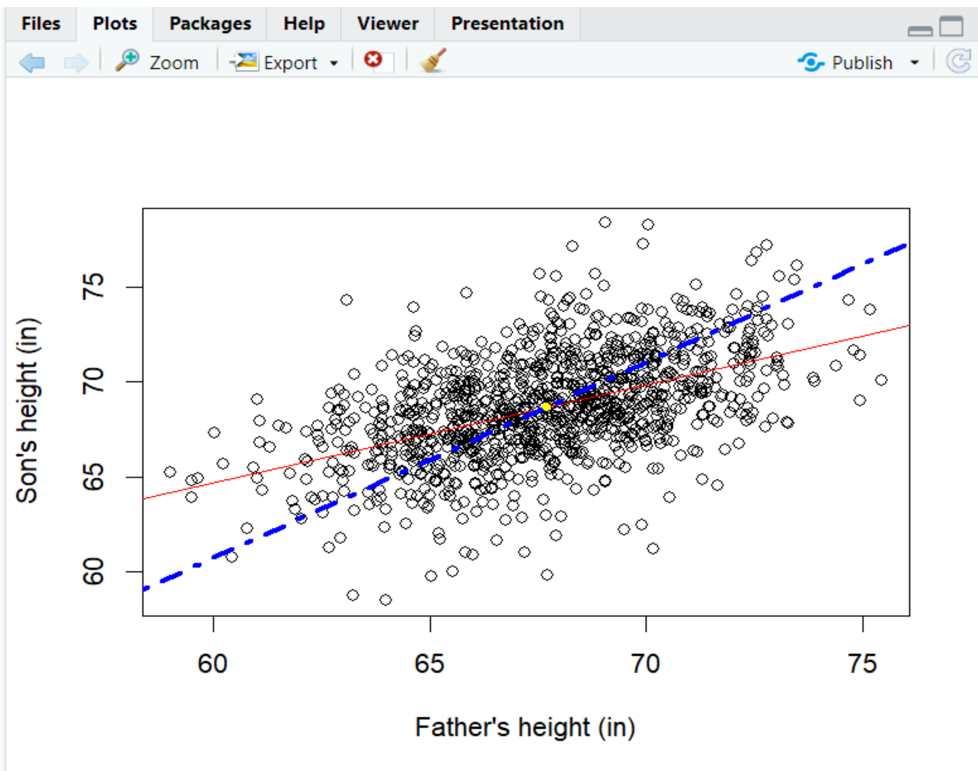


Rowan Raj Ignatius
Student ID: 02169507

e) Mark the center of regression

```
points(mean(fathers_h), mean(sons_h) ,
    col='yellow',
    pch= 20)
```

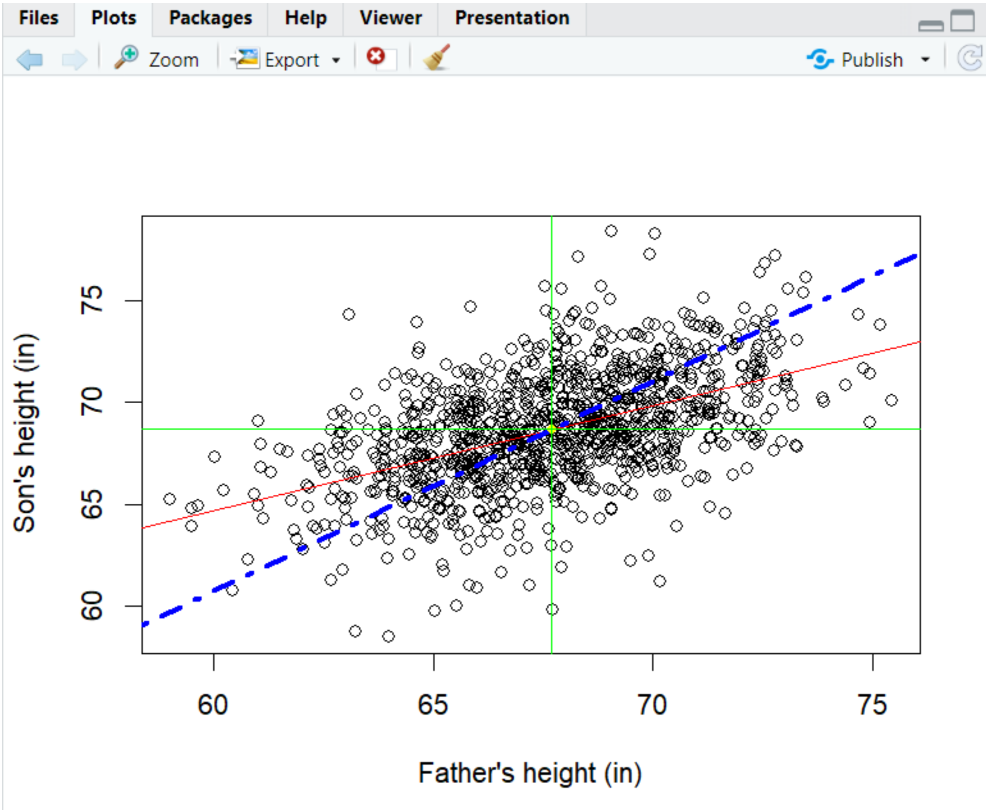
I have used the function **points()** to plot the center of regression (mean of father_h and mean of son_h) and **col** specifies the color of the point and the **pch** specifies the appearance of the point in this case '20' (filled circle).



f) Add horizontal and vertical lines (green color) through the center of regression

```
abline(v=mean(fathers_h), col="green")
abline(h=mean(sons_h), col="green")
```

With the help if **'abline'** function, I have added the vertical and horizontal lines each passing through the center of regression (mean of father_h and mean of son_h) and specified the color of the line as green.



g) Report the linear regression output (including R^2 etc)

summary(regression_line)

Using the summary() function I have generated the complete report of the regression model including the R^2 , residuals, coefficients etc.,

The following is the data included in the report:

Call: Specifies the formula for the linear regression used in the model

Residuals: The difference between the observed and predicted values is specified under residuals

Coefficients: Lists the coefficients in the model like Estimate, Std. Error, t value, Pr(>|t|)

Signif. codes: Specifies the predictors significance in the regression model

Residual standard error: Models with higher residual value gives a more accurate model as it gives the average on how the predications differ from the actual values.

Multiple R-squared: Specifies how the model handles the variation of the data. When the value is nearly one, it means the model has handled the data better.

Adjusted R-squared: Based on the high value of the Adjusted R-squared we can decide how good the model is in handling the predictors

F-statistic: Specifies the significance level of the complete model. A model with higher F-status and low p- value is considered to be significant one.

p-value: Highlights the significance level of the predictors and a p-value<0.5 specifies that at least one predictor is significant.

```
Call:
lm(formula = sons_h ~ fathers_h, data = father.son)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.8772 -1.5144 -0.0079  1.6285  8.9685 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 33.88660    1.83235   18.49   <2e-16 ***
fathers_h    0.51409    0.02705   19.01   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.437 on 1076 degrees of freedom
Multiple R-squared:  0.2513,	Adjusted R-squared:  0.2506 
F-statistic: 361.2 on 1 and 1076 DF,  p-value: < 2.2e-16
```

Rowan Raj Ignatius
Student ID: 02169507