NITHIN DAS, CWID: 10422784, Date: 10/15/2019

# Assignment 2

**6.9 From the depression data set described in Table 3.4 create a data set containing only the variables AGE and INCOME.**

## (a) Find the regression of income on age.

The summary of the regression model is as below:

```
Call:
lm(formula = Income ~ Age, data = depression_data)

Residuals:
    Min      1Q  Median      3Q     Max
-22.856 -10.315  -4.332   6.270  47.275

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 27.77274    2.32808  11.929  < 2e-16 ***
Age         -0.16206    0.04856  -3.337 0.000955 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.03 on 292 degrees of freedom
Multiple R-squared:  0.03674,  Adjusted R-squared:  0.03345
F-statistic: 11.14 on 1 and 292 DF,  p-value: 0.000955
```

Regression Equation:

Income= -0.1621 * Age + 27.77

**(b) Successively add and then delete each of the following points:**

**AGE INCOME**

**42 120**

**80 150**

**180 15**

**and repeat the regression each time with the single extra point. How does the regression equation change? Which of the new points are outliers? Which are influential?**

## (i)Adding (42,120) record

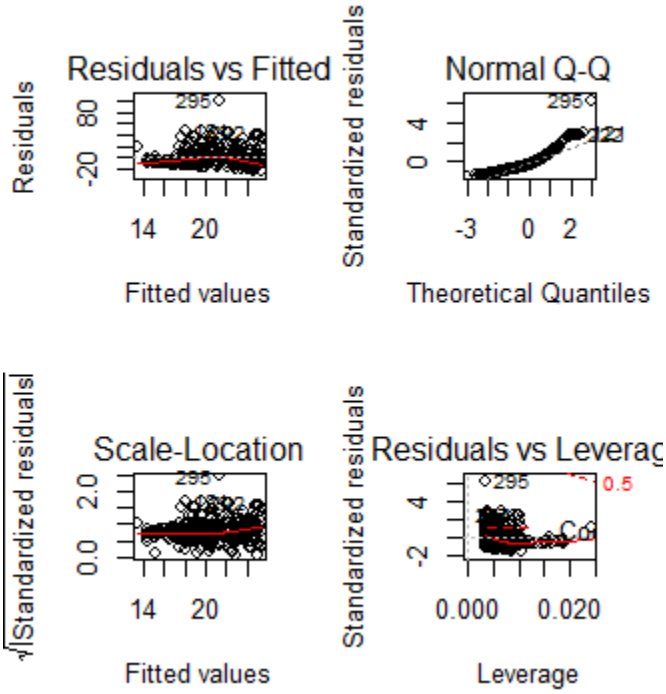
The summary of the regression model is as below:

```
Call:
lm(formula = Income ~ Age, data = depression_data)

Residuals:
    Min      1Q  Median      3Q     Max
-23.257 -10.555  -4.612   6.034  98.692

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 28.21889    2.48924  11.336   <2e-16 ***
Age         -0.16455    0.05194  -3.168   0.0017 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.08 on 293 degrees of freedom
Multiple R-squared:  0.03312,  Adjusted R-squared:  0.02982
F-statistic: 10.04 on 1 and 293 DF,  p-value: 0.001697
```



Regression Equation:

Income= -0.1645*(Age)+28.2189

**(ii)Adding (80,150) record**

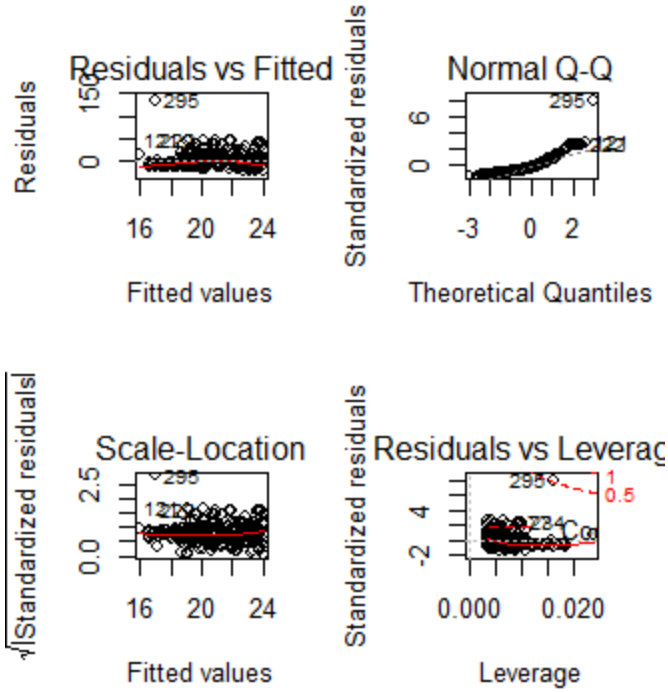
The summary of the regression model is as below**:**

```
Call:
lm(formula = Income ~ Age, data = depression_data)

Residuals:
    Min      1Q  Median      3Q     Max
-22.004 -11.116  -4.553   6.250 132.983

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 26.03193    2.61245   9.965   <2e-16 ***
Age         -0.11268    0.05433  -2.074   0.0389 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.93 on 293 degrees of freedom
Multiple R-squared:  0.01447,  Adjusted R-squared:  0.01111
F-statistic: 4.302 on 1 and 293 DF,  p-value: 0.03893
```




Regression Equation:

Income= -0.1127*Age +26.0319

**(iii)Adding (180,15) record**

The summary of the regression model is as below**:**

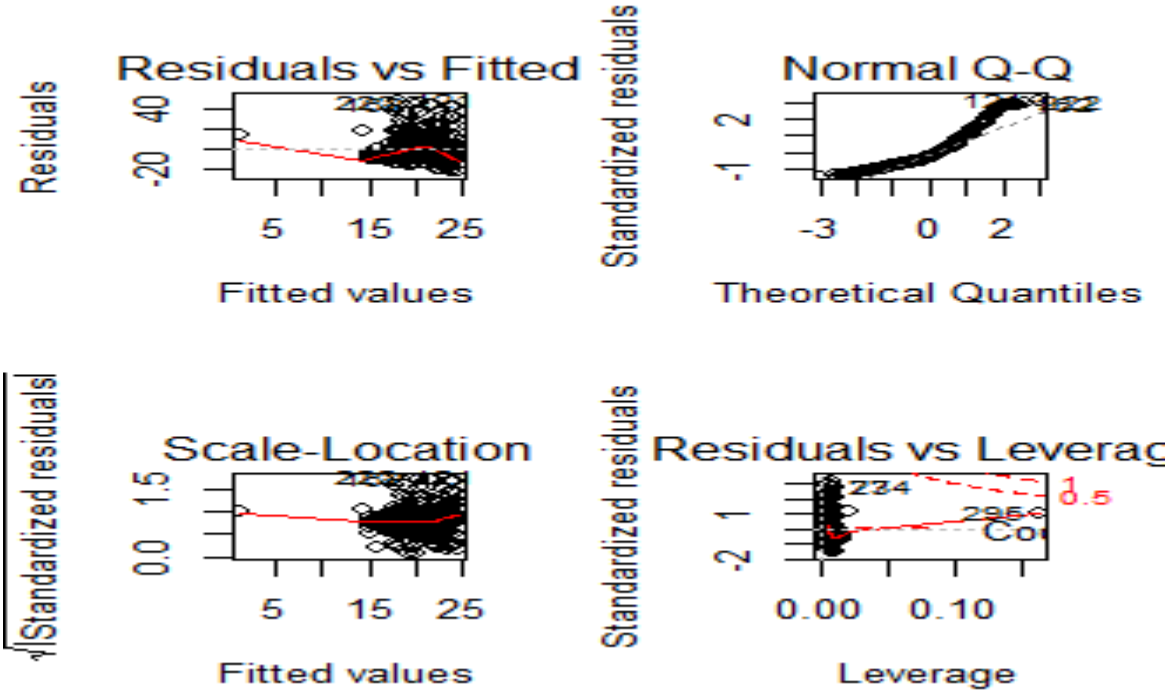
```
Call:
lm(formula = Income ~ Age, data = depression_data)

Residuals:
    Min      1Q  Median      3Q     Max
-22.390 -10.398  -4.248   6.463  46.887

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 26.95730    2.17992  12.366  < 2e-16 ***
Age         -0.14265    0.04449  -3.206  0.00149 **
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.03 on 293 degrees of freedom
Multiple R-squared:  0.0339,   Adjusted R-squared:  0.0306
F-statistic: 10.28 on 1 and 293 DF,  p-value: 0.001493
```




Income= -0.1427*Age + 26.9573

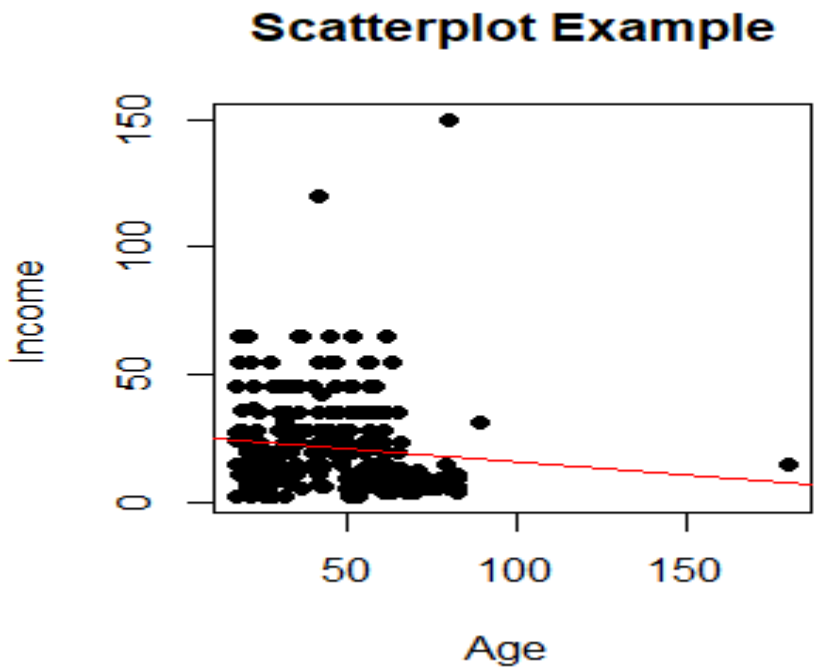
**Analysis:**

From the above regression outputs, we can see from the residual plots that records Age, Income=(42,120) and Age, Income=(80,150) are outliers.

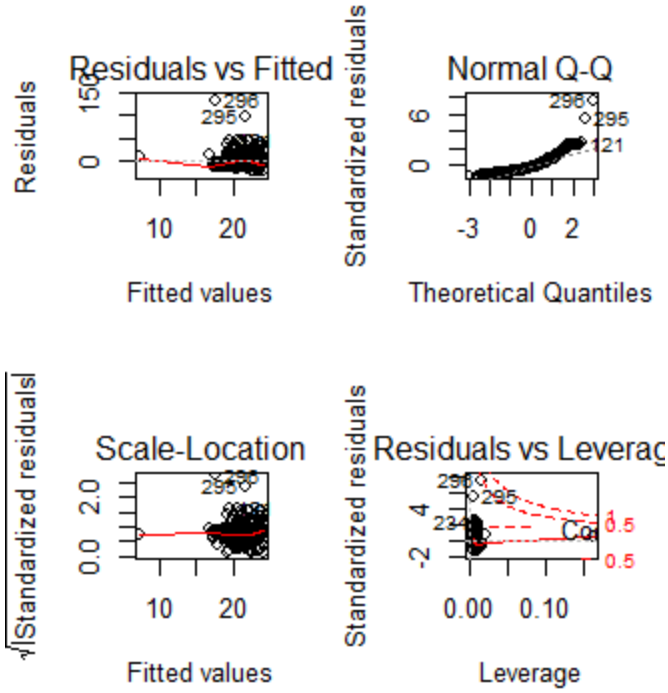
But the slope for Record (80,150) has become more flatter compared to the original regression slope. Original slope =-0.1621, Slope with outlier (80,150)= -0.1427. Also, the original R squared value was 0.036, after adding (80,150), R squared value decreased to 0.014. Therefore, the influential point is (Age, Income)=(80,150).

3) **In Problem 6.9 of your text book, add the three additional points at the same time and establish a regression line. Identify High leverage points and High influential points**

Scatter Plot with the regression line is shown below:




The output of linear regression is shown below:




Sample number 295 and 296 are close to cook's distance of 1, making them the high leverage and influential points.

Sample 295 – Age, Income :(42,120)

Sample 296- Age, Income : (80,150)

6.10 For the oldest child, perform the following regression analyses: FEV1 on weight, FEV1 on height, FVC on weight, and FVC on height. Note the values of the slope and correlation coefficient for each regression and test whether they are equal to zero. Discuss whether height or weight is more strongly associated with lung function in the oldest child.

# Solution:

FEV1 on Height:

Slope=14.1451

Intercept=-588.04

Correlation Coefficient= 0.9234

FEV1 on Weight:

Slope=2.48

Intercept=6.57

Correlation Coefficient= 0.893

FVC on Height:

Slope=16.5

Intercept=-687.34

Correlation Coefficient= 0.904

FVC on Weight:

Slope=2.94

Intercept=1.675

Correlation Coefficient= 0.88

None of the correlation coefficients and slopes are 0.

The correlation coefficient of Height on FEV1 and FVC is slightly higher than that of coefficients of Weight. This shows that height is strongly associated with lung function. However, lets look at the result of regression analyses.

| Regression Analyses | Multiple R Squared | Residual Standard Error |
|---|---|---|
| **FEV1 V/S Height** | 0.85 | 40.85 |
| **FEV1 V/S Weight** | 0.79 | 47.85 |
| **FVC V/S Height** | 0.81 | 54.11 |
| **FVC V/S Weight** | 0.78 | 58.19 |

Height has better R Squared values than Weight for both FEV1 and FVC. R square value of 0.85 for regression analysis between FEV1 and Height means that approximately 85% of variation in FEV1 values can be explained by the Height variable.

Similarly, Residual Standard error for height is better than weight for both FEV1 and FVC. Residual error of FEV1 v/s Height model is the difference between observed values of FEV1 and predicted/fitted values of FEV1.Since the residual values of height is less than weight for both the cases, height is strongly associated with the lung function.

Therefore, Height is strongly associated with the lung function in the oldest child.