Big Data

Lab 5

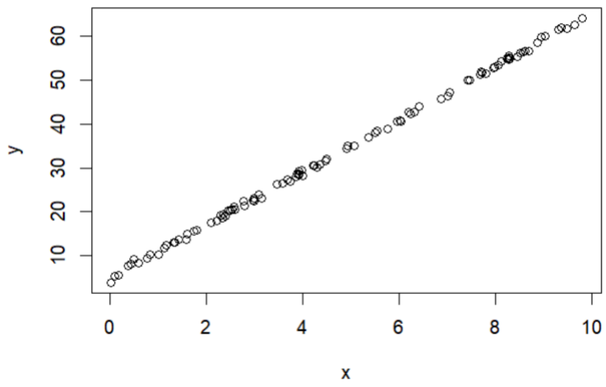
(Linear Regression)

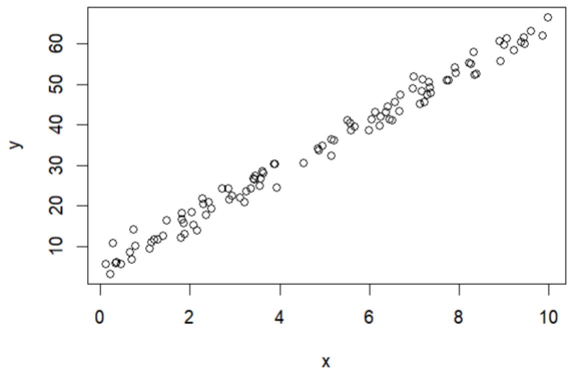
Name	ID	SEC	BN
Norhan Reda Abdelwahed Ahmed	9203639	2	31
Hoda Gamal Hamouda Ismail	9203673	2	33

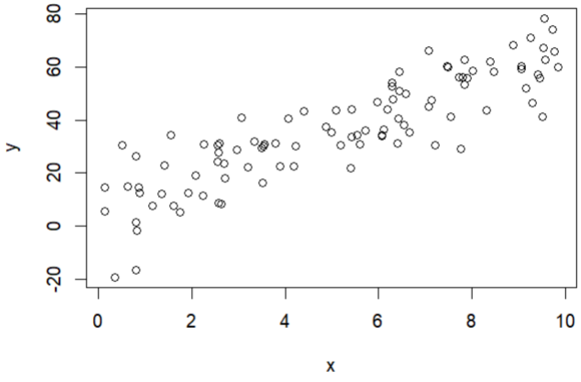
Supervisor: Eng. Omar Samir

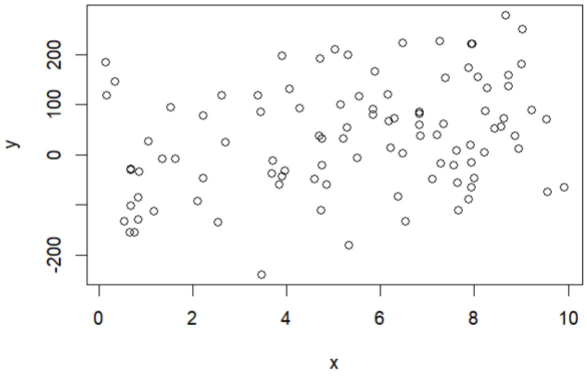
(Q1)Try changing the value of standard deviation (sd). How do the data points change for different values of standard deviation?

From the equation, it seems like we add noise with random distribution to the data. So with increasing the standard deviation of this noise, the data points become more scattered and further from the linear model.


std=0.5


std=2


std=10


std=100

(Q2)How are the coefficients of the linear model affected by changing the value of standard deviation in Q1?

With increasing the standard deviation of this noise, the coefficients become further from the coefficients of the actual linear model which are a=5, b=6.

```
Std = 0.5
Coefficients:
(Intercept)            x
      4.888        6.028
```

```
std=2
Coefficients:
(Intercept)            x
      4.924        6.013
```

```
std=10
Coefficients:
(Intercept)            x
      5.827        5.950
```

```
std=100
Coefficients:
(Intercept)            x
     -28.57        11.89
```

(Q3)How is the value of R-squared affected by changing the value of standard deviation in Q1?

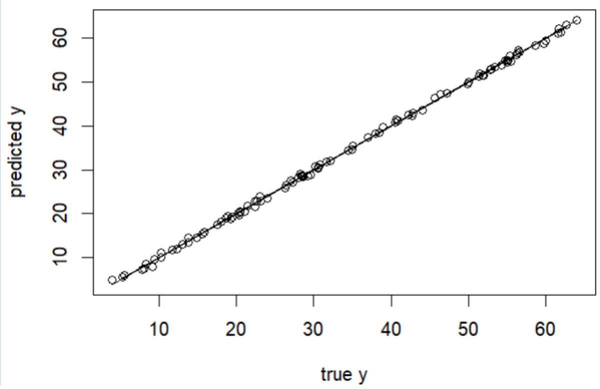
With increasing the standard deviation of this noise, R-squared value becomes lower.

As R-squared represents the scatter of data points around the line and also indicates the correlation between the true and predicted outputs.

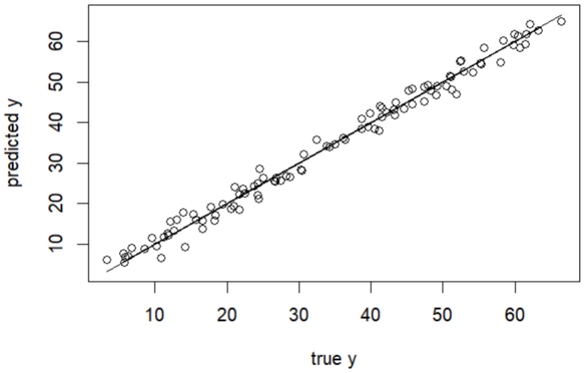
So with increasing the standard deviation of this noise, the data points become more scattered, and the model has less ability to predict the true output (the correlation becomes lower).

```
std=0.5
OLS gave slope of  6.028471 and an R-sqr of  0.9993349

std=2
OLS gave slope of  6.012782 and an R-sqr of  0.9867277

std=10
OLS gave slope of  5.94995 and an R-sqr of  0.7553887

std=100
OLS gave slope of  11.89283 and an R-sqr of  0.09200188
```

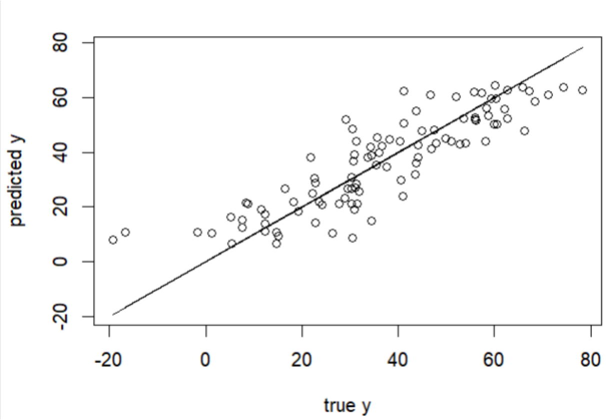
std=0.5



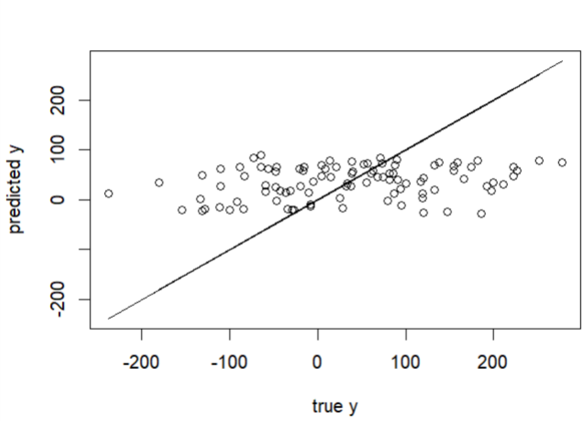
std=2



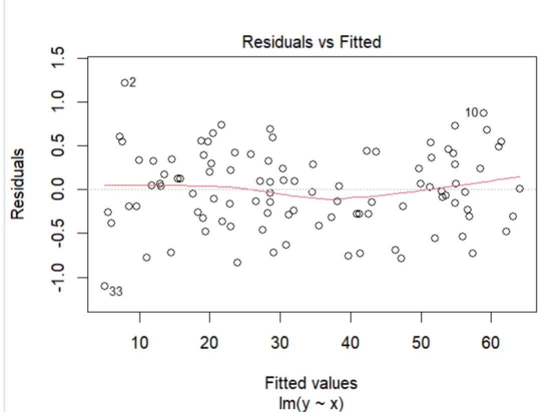
std=10

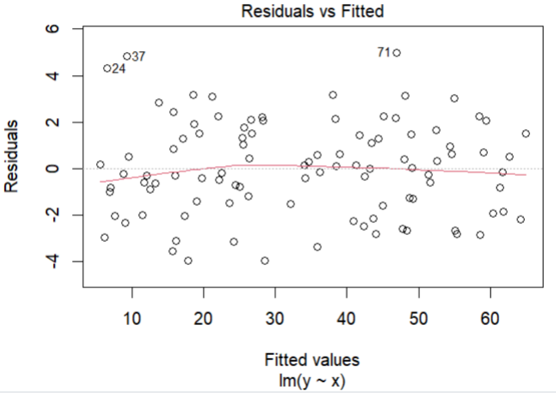


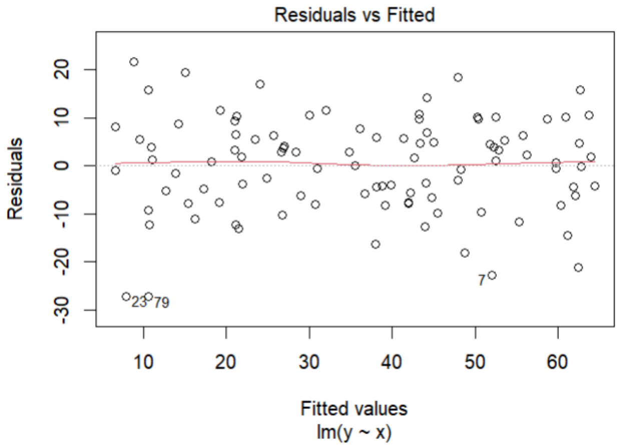
std=100

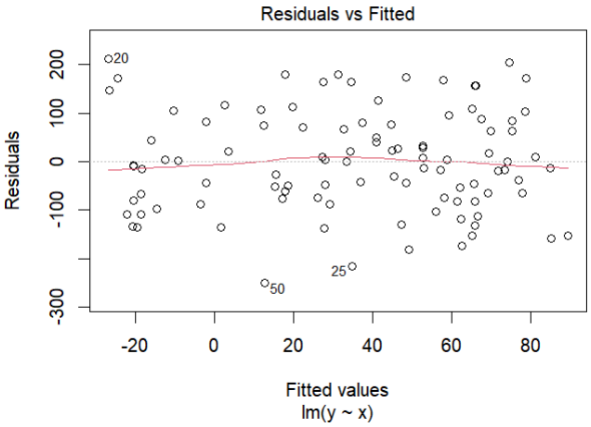


(Q4)What do you conclude about the residual plot? Is it a good residual plot?





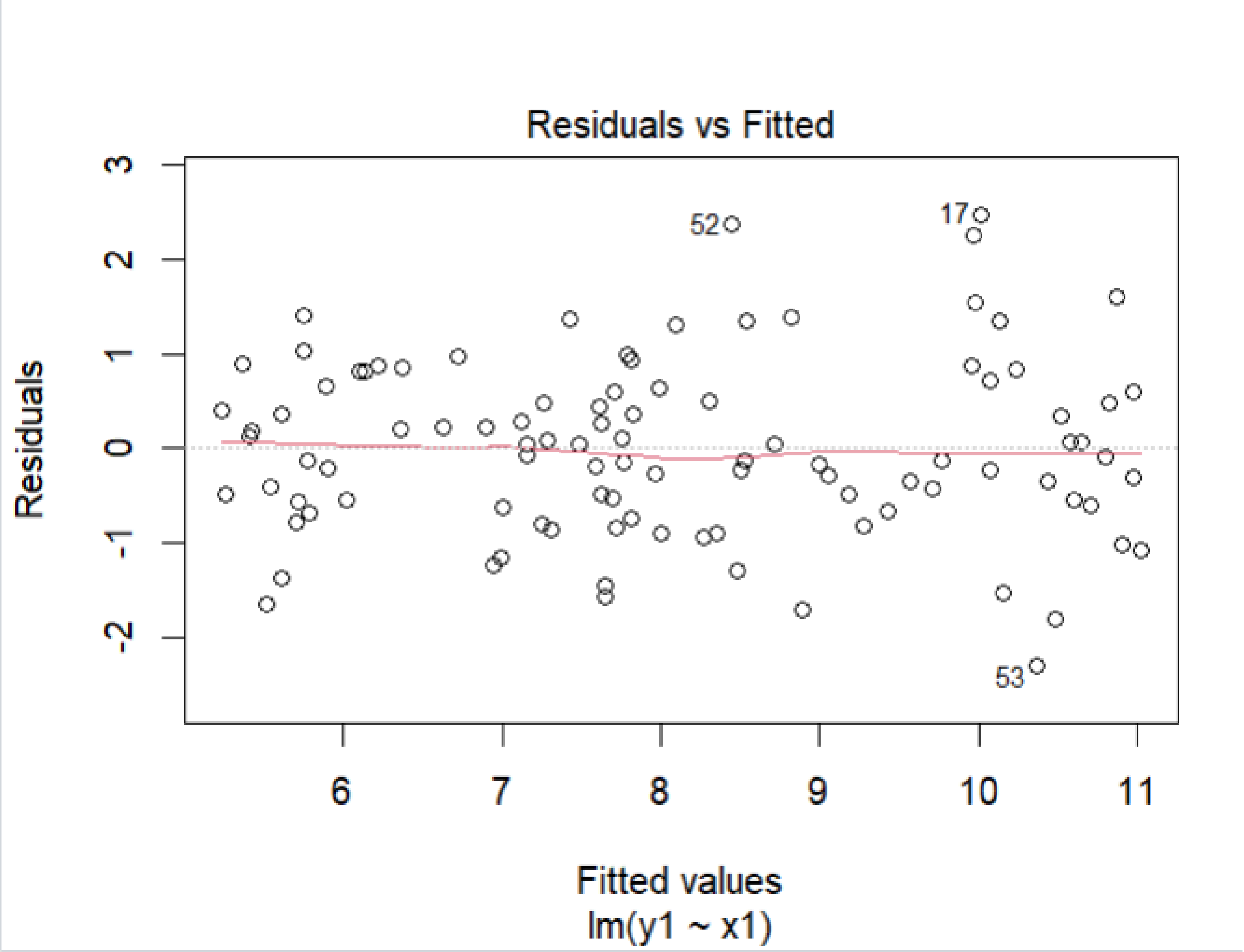




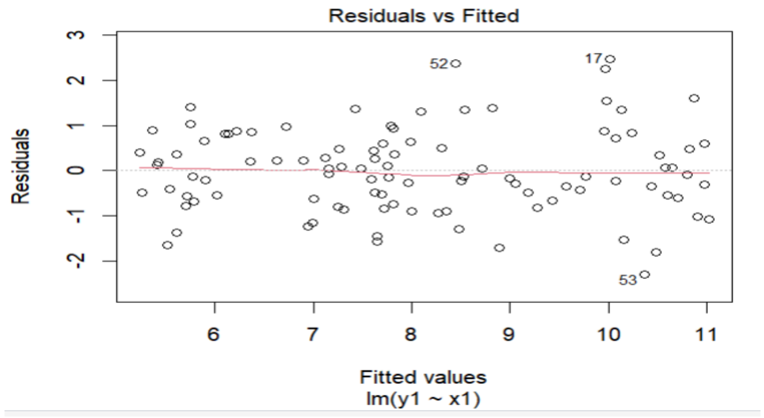
It seems to be a good residual as the data has no specific pattern. It appears randomly scattered

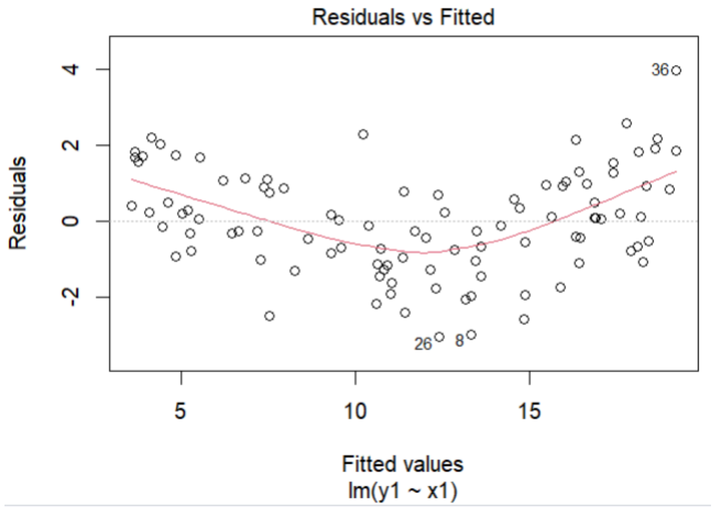
(Q5)What do you conclude about the residual plot? Is it a good residual plot?

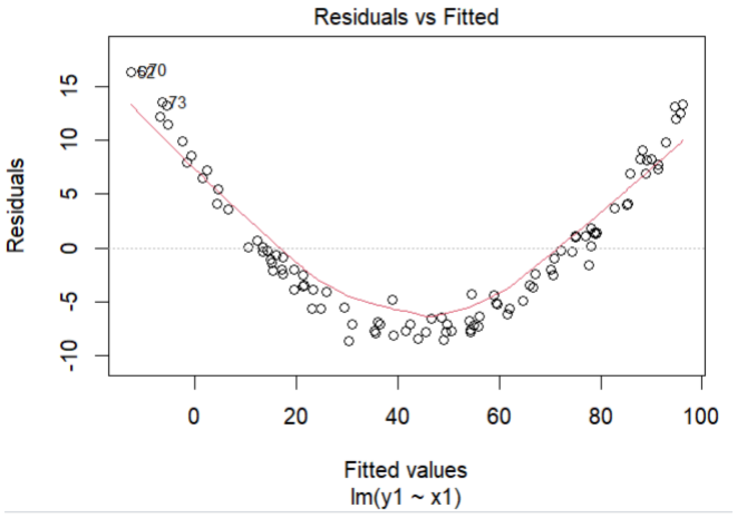
Yes , it is a good residual plot as the data has no specific pattern. It appears randomly scattered



(Q6)Now, change the coefficient of the non-linear term in the original model for (A)training and (B) testing to a large value instead. What do you notice about the residual plot?

<table>
<tr>
<td><b>Coefficient = 0.1</b></td>
<td>


</td>
</tr>
<tr>
<td><b>Coefficient = 10</b></td>
<td>


</td>
</tr>
<tr>
<td><b>Coefficient = 100</b></td>
<td>


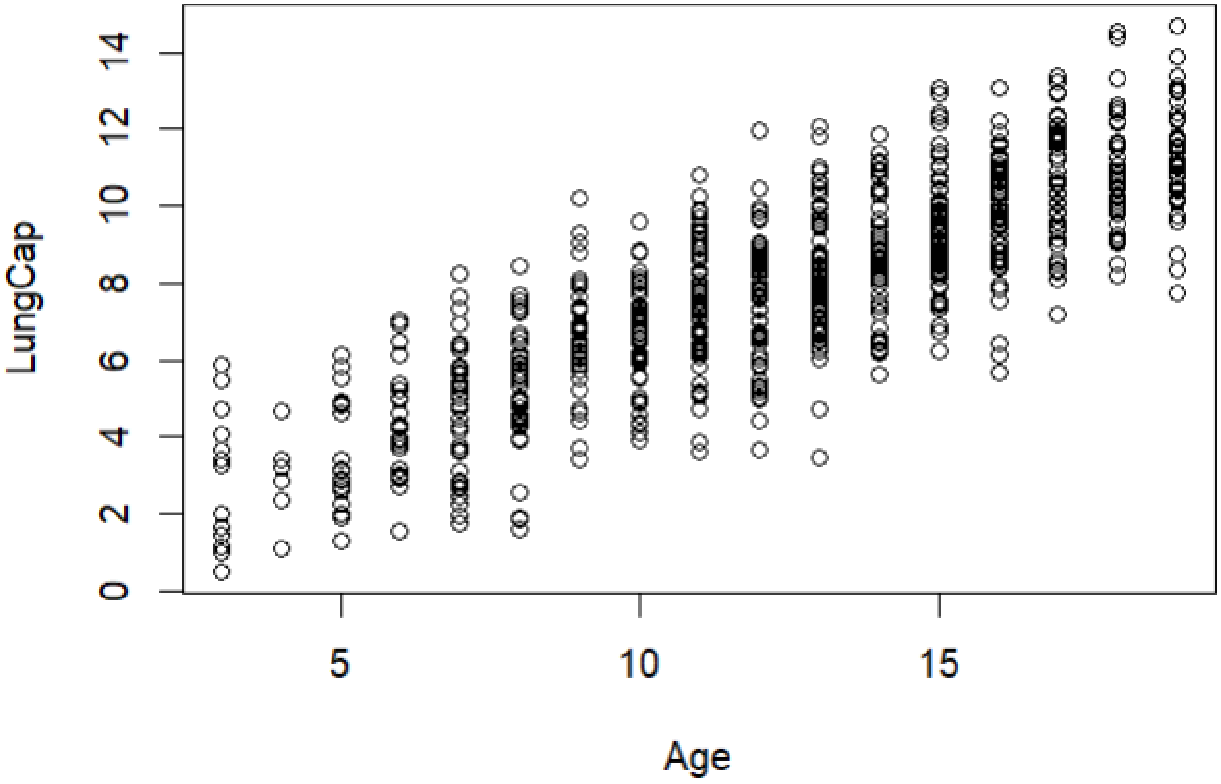
</td>
</tr>
</table>

It seems that as we increase the the coefficient of the non-linear term the residual plot will become more curved which implies it will be a non-linear pattern so it will not be a good residual plot

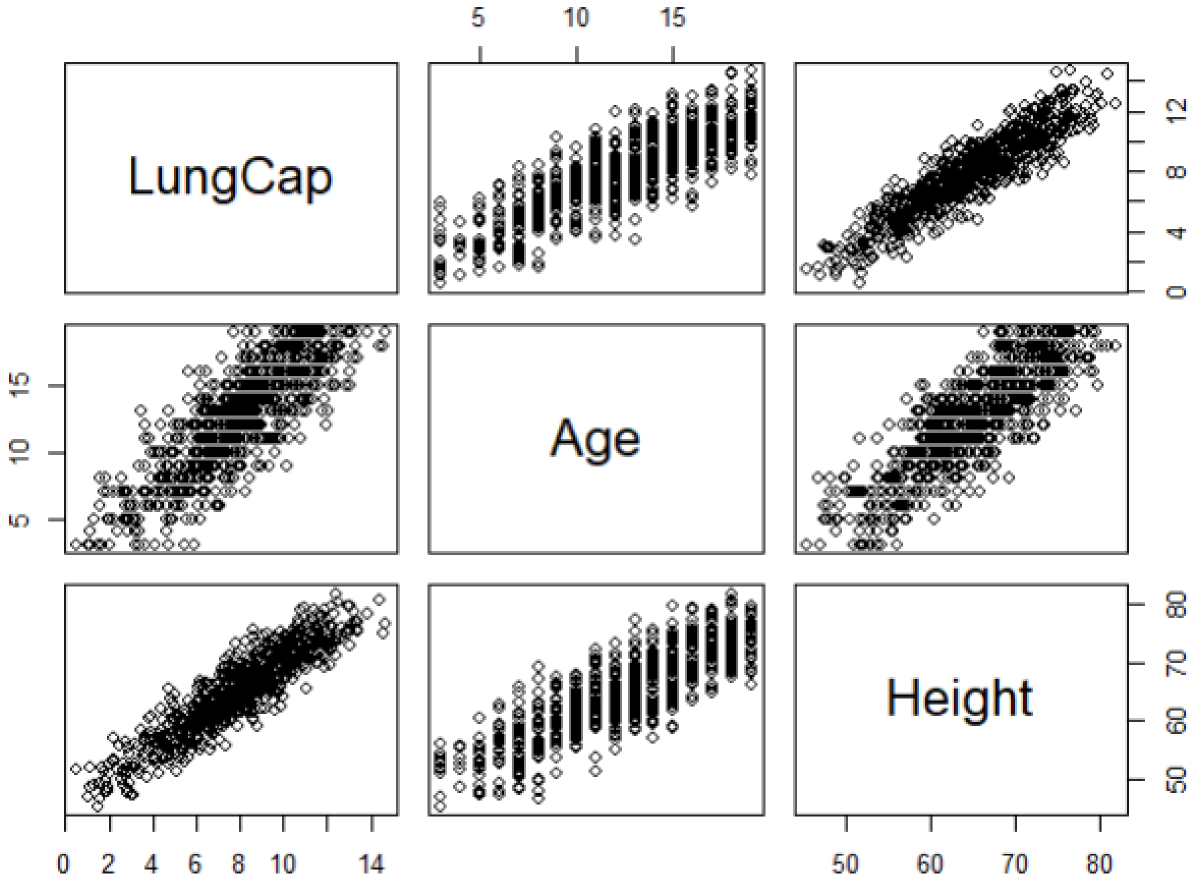
(Q7)What are the variables in this dataset?

```
"LungCap"   "Age"       "Height"    "Smoke"     "Gender"    "Caesarean"
```

(Q8)Draw a scatter plot of Age (x-axis) vs. LungCap (y-axis). Label x-axis "Age" and yaxis"LungCap



(Q9)Draw a pair-wise scatter plot between Lung Capacity, Age and Height.



(Q10)Calculate the correlation between Age and LungCap, and between Height and LungCap.

```
Correlation  (age ,lungcap)
0.8196749
correlation (height,lungcap)
0.9121873
```

(Q11)Which of the two input variables Age and Height are more correlated to the dependent variable LungCap?

```
Height is more correlated with LungCap
```

(Q12)Do you think the two variables Height and LungCap are correlated? Why?

Correlation (height,lungcap)

0.9121873

Height and LungCap are highly positive correlated

As correlation between them is near to 1

(Q13)Fit a linear regression model where the dependent variable is LungCap and use all other variables as the independent variables.

lm_model <- lm(LungCap ~ ., data = data)

(Q14)Show a summary of this model.

```
Residuals:
    Min      1Q  Median      3Q     Max 
-3.3388 -0.7200  0.0444  0.7093  3.0172 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -11.32249    0.47097 -24.041  < 2e-16 ***
Age           0.16053    0.01801   8.915  < 2e-16 ***
Height        0.26411    0.01006  26.248  < 2e-16 ***
Smokeyes     -0.60956    0.12598  -4.839 1.60e-06 ***
Gendermale    0.38701    0.07966   4.858 1.45e-06 ***
Caesareanyes -0.21422    0.09074  -2.361   0.0185 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.02 on 719 degrees of freedom
Multiple R-squared:  0.8542,	Adjusted R-squared:  0.8532 
F-statistic: 842.8 on 5 and 719 DF,  p-value: < 2.2e-16
```

(Q15)What is the R-squared value of this model? What does R-squared indicate?

R-squared value = 0.8542478

R-squared indicates the proportion of variance in the dependent variable explained by the independent variables in the model.Which indicates that 85.42478 % of the **variance in Lungcap is explained by The independent variable in the model**

(Q16)Show the coefficients of the linear model. Do they make sense? If not, which variables don't make sense to you? What should you do?

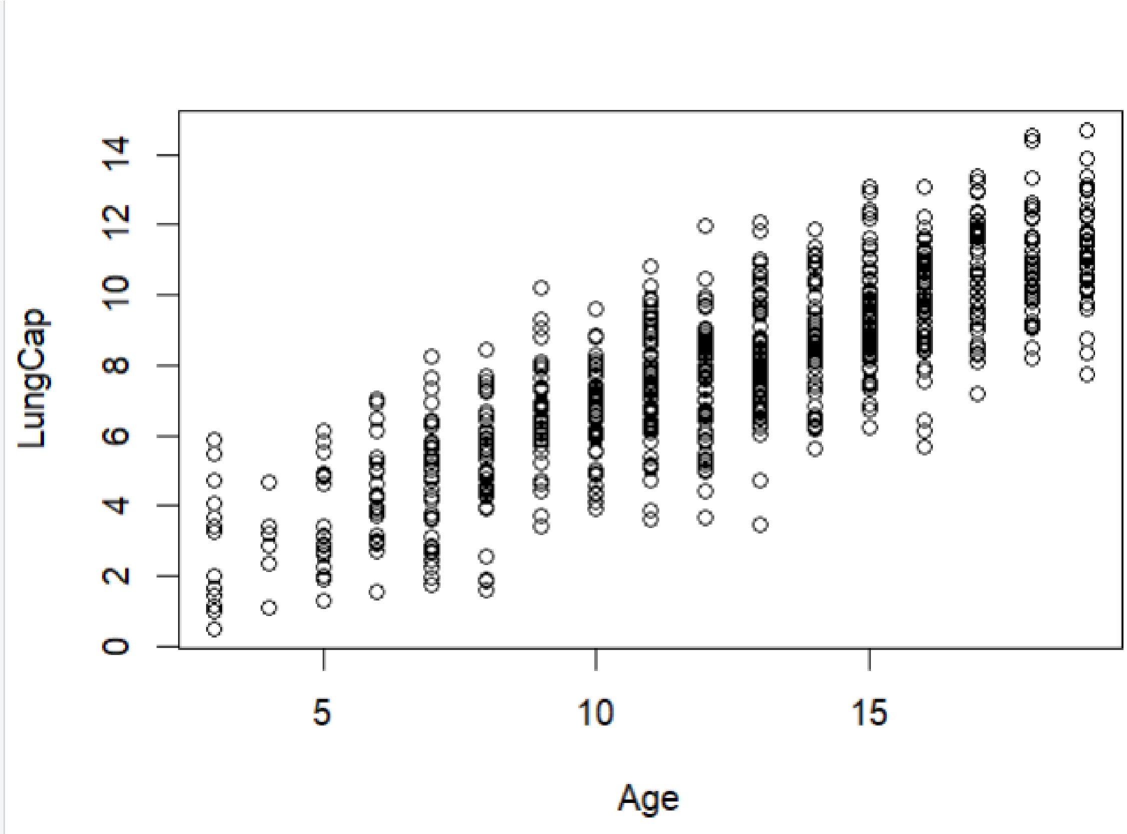
```
(Intercept)         Age      Height     Smokeyes  Gendermale  Caesareanyes 
 -11.3224856   0.1605296   0.2641128  -0.6095592   0.3870117  -0.2142182 
```

It seems that the value of the intercept doesn’t make sense as it is excessively large so it indicates strongly correlated inputs so we need to eliminate some of them

And also the sign of the intercept is negative which implies if all the variables are zero the lung capacity will be negative which not make sense

From the above information and statistics it seems that “Age” and “Height” are correlated so we can remove on of them for example “Height” and we can search for more correlated variables and remove one of them by doing more analysis to the data

(Q17) why red line ?

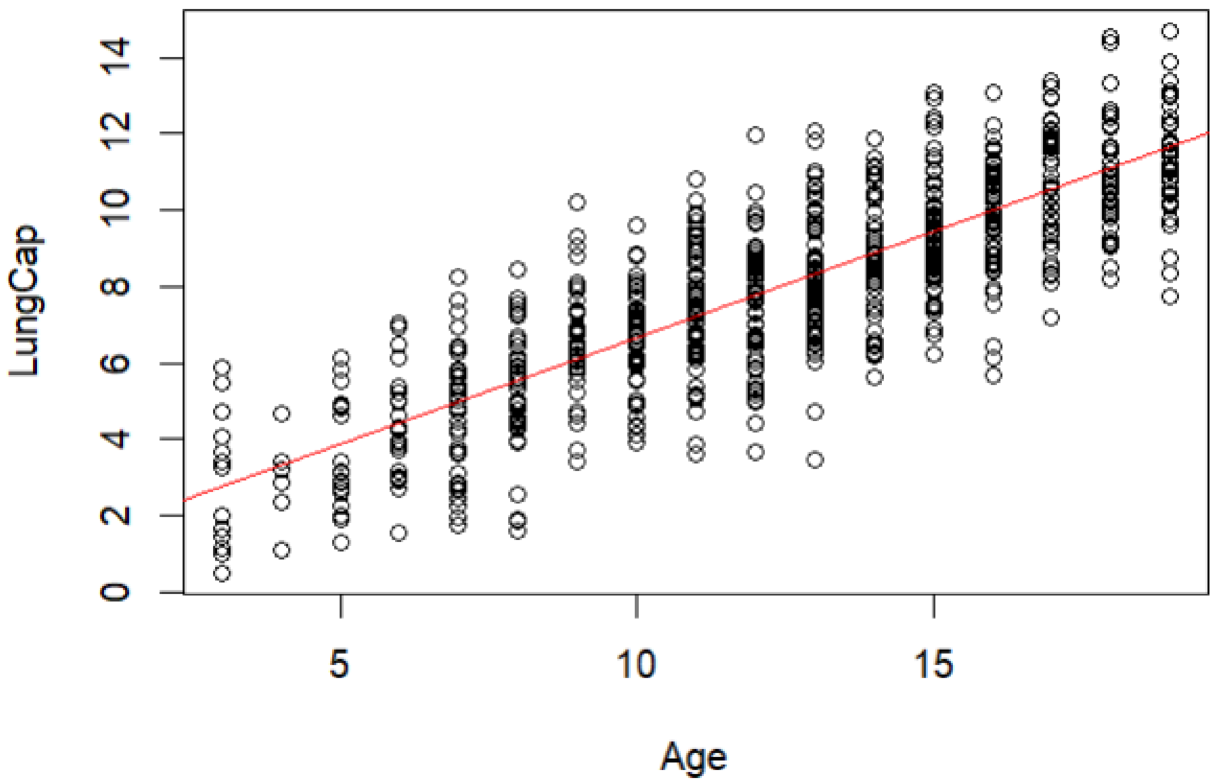


Because the model depend on variables that they are correlated and dependent and don’t make sense so it can’t fit data well

(Q18)Repeat Q13 but with these variables Age, Smoke and Cesarean as the only independent variables.

```
(Intercept)           Age     Smokeyes  Caesareanyes
  1.1086723     0.5561667   -0.6431029   -0.1460278
```

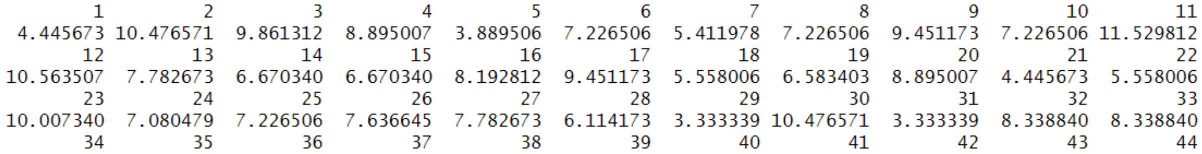
(Q19)Repeat Q16, Q17 for the new model. What happened?



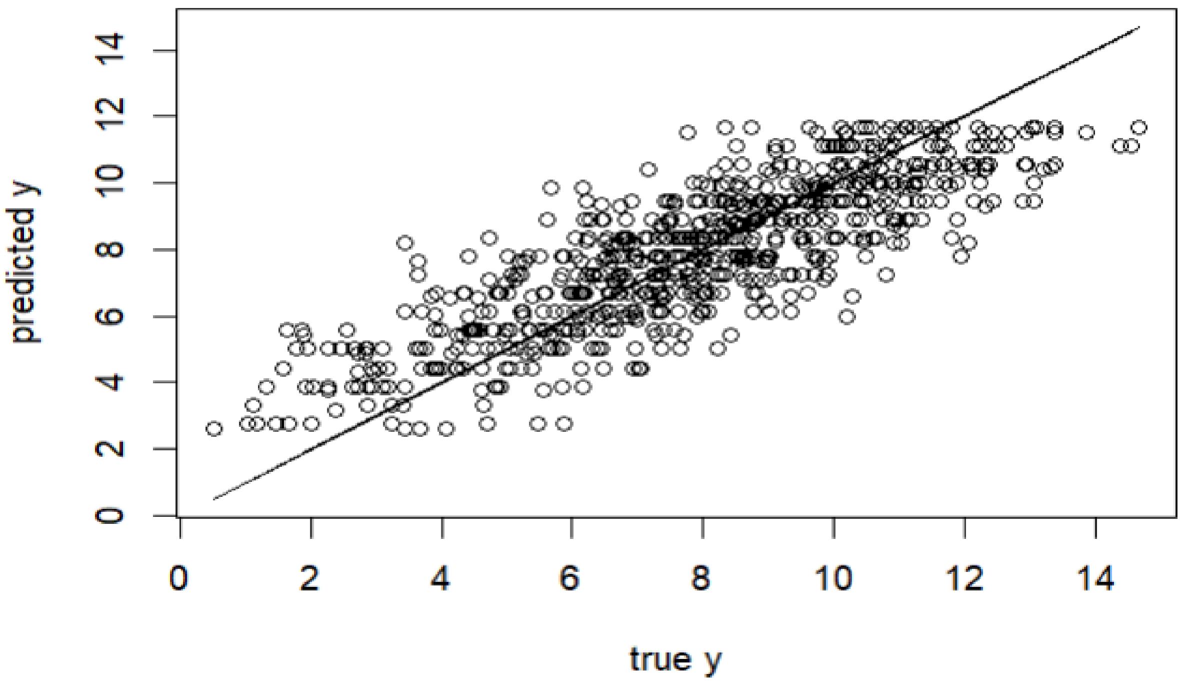
The red line appears because we have removed the variables are correlated with other variables and that doesn't make sense in the model fitting so the model can fit the data

(Q20)Predict results for this regression line on the training data.

A subset of the prediction values



```
        1         2         3         4         5         6         7         8         9        10        11 
 4.445673 10.476571  9.861312  8.895007  3.889506  7.226506  5.411978  7.226506  9.451173  7.226506 11.529812 
       12        13        14        15        16        17        18        19        20        21        22 
10.563507  7.782673  6.670340  6.670340  8.192812  9.451173  5.558006  6.583403  8.895007  4.445673  5.558006 
       23        24        25        26        27        28        29        30        31        32        33 
10.007340  7.080479  7.226506  7.636645  7.782673  6.114173  3.333339 10.476571  3.333339  8.338840  8.338840 
       34        35        36        37        38        39        40        41        42        43        44 
```



(Q21)Calculate the mean squared error (MSE) of the training data.

```
mean squared error (MSE) = 2.280169
```