Lab 5

Predictive Analysis ||

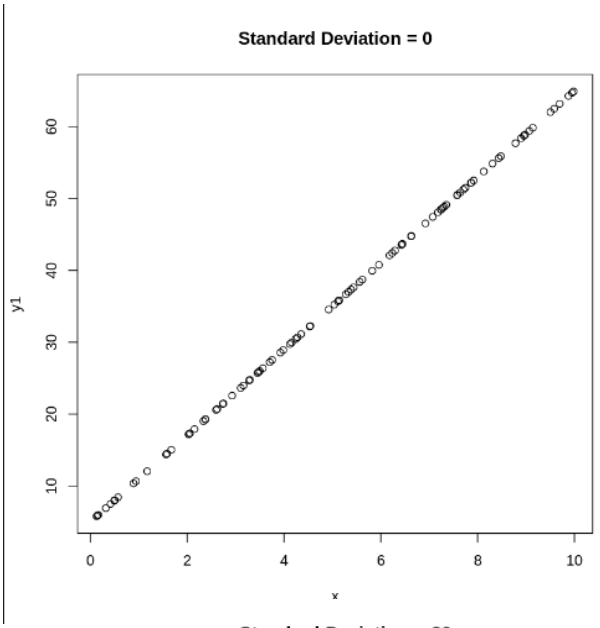
Karim Mahmoud Kamal	Sec: 2	BN: 12
Mustafa Mahmoud Hamada	Sec: 2	BN: 25

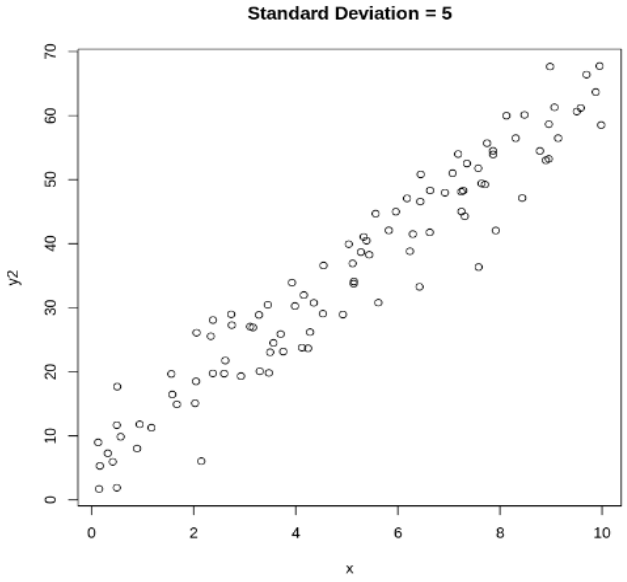
Eng. Omar Samir

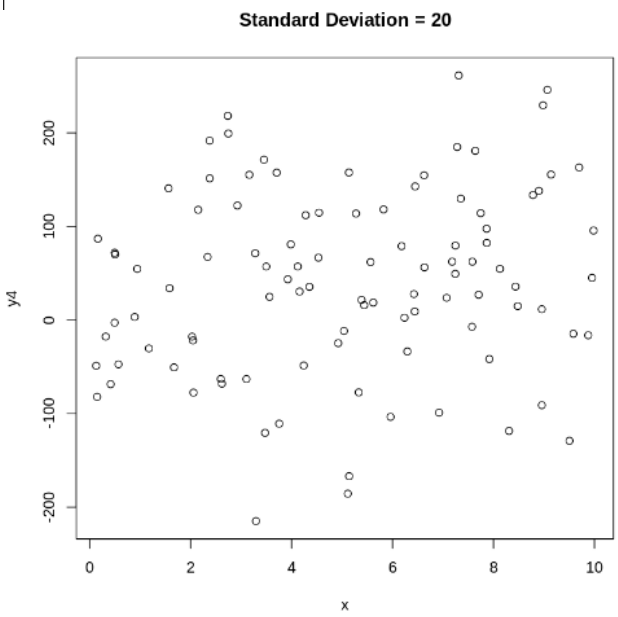
Part 1

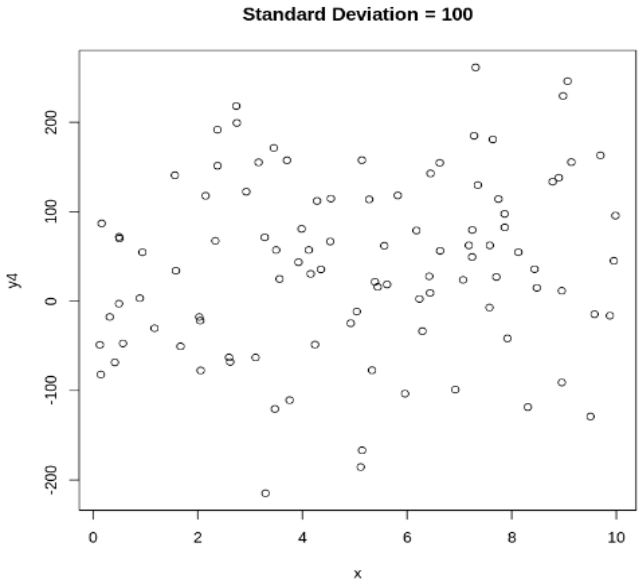
Q1) How do the data points change for different values of standard deviation?

As shown in the following figures ➔ The change in the standard deviation changes the variance of the noise in the input data so for **large standard deviation the points are more scattered** from the line and for small standard deviation the points are very close to a line i.e. the square error is minimal and will equal zero when standard deviation = 0.









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

The intercept coefficient represents the value of y when x=0 ➔ bias.

The coefficient for x represents the change in y for a unit change in x.

As the standard deviation increases, the deviation from the original line also increases.

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

When the standard deviation increases, the R-squared value decreases.

R-squared values range between 0 and 1, where 1 indicates that all variation in the dependent variable (y) can be explained by the independent variable (x) and the intercept.

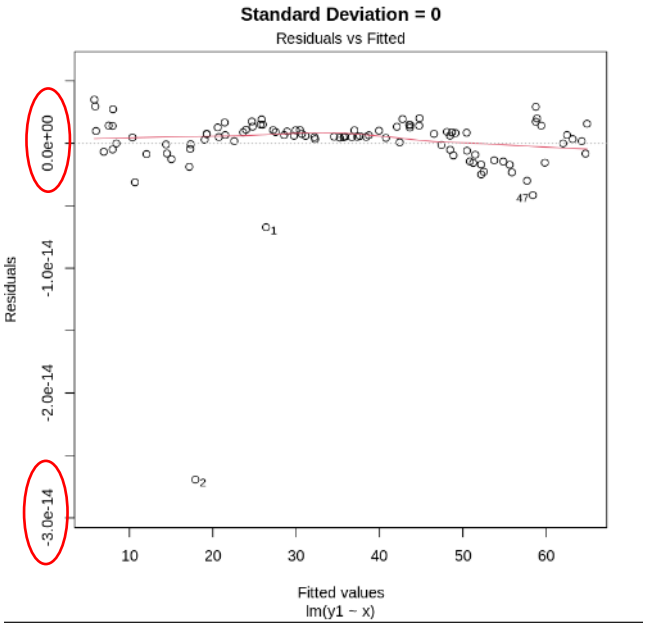
R-squared represents the dispersion or scattering of data points around the regression line; **the greater the scattering, the lower the R-squared value**.

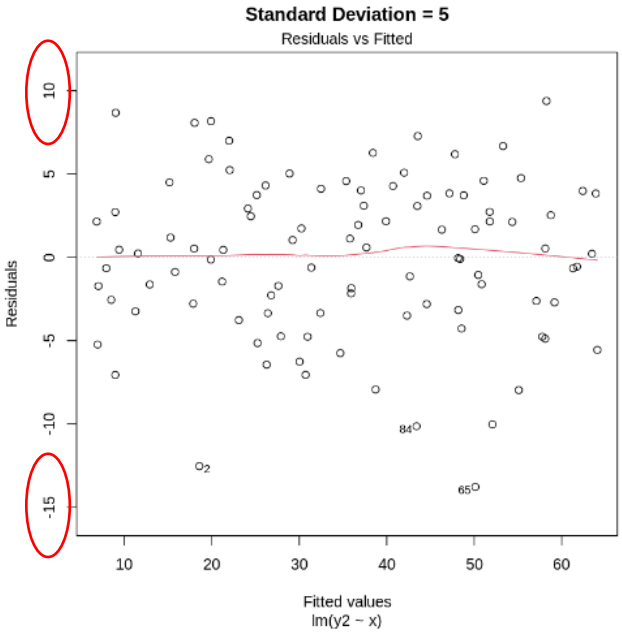
Q4) What do you conclude about the residual plot? Is it a good residual plot?

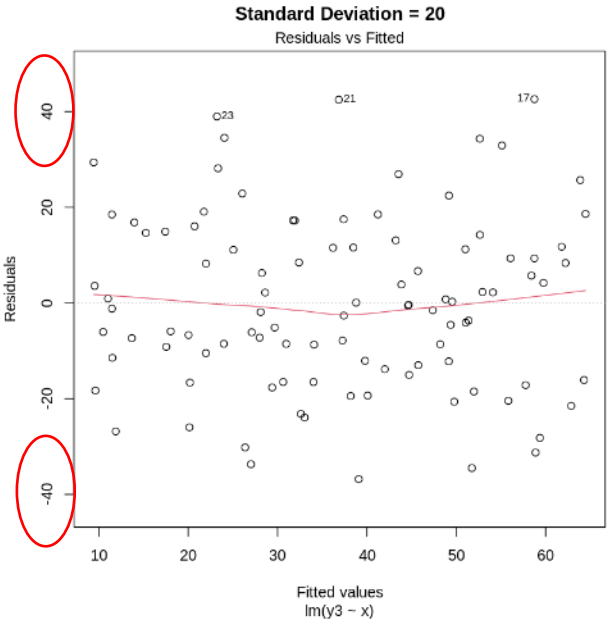
The residual values offer a measure of the goodness of fit, indicating how well the line fits the data.

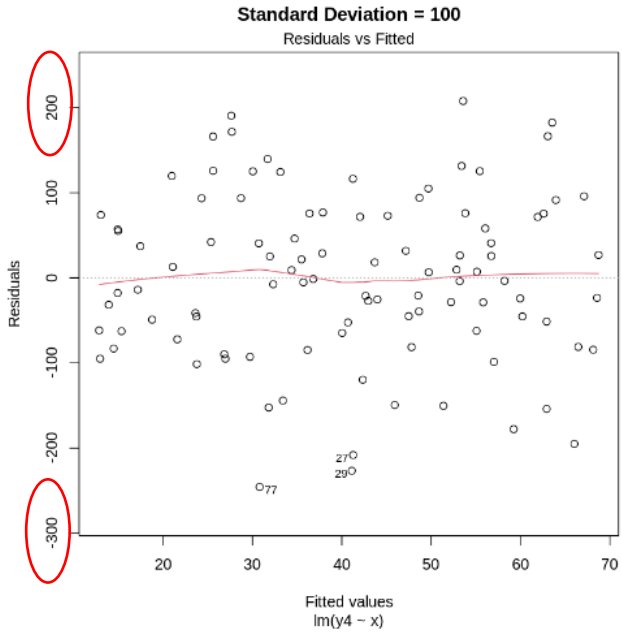
Despite observing a linear pattern in the plot and generating the data with constant variance, we observe significant residual values when the standard deviation is large.

This suggests that the model doesn't fit the data very well and the noise is high.









Part 2

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

The residual values serve as a measure of the goodness of fit, indicating how well the line fits with the data.

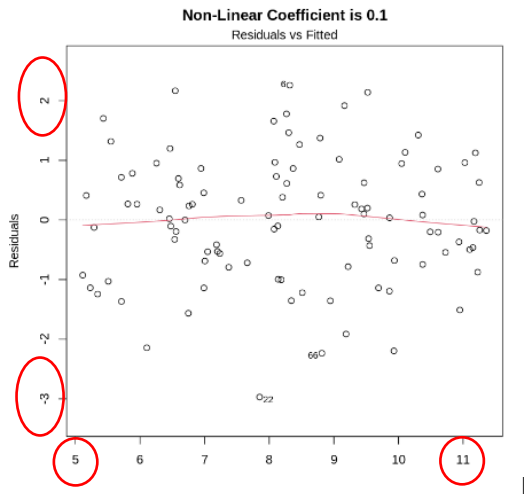
In this case, we observe **a linear pattern** in the plot, given the **negligible coefficient associated with the non-linear term**, and the data was generated with a constant variance, so it seems good.

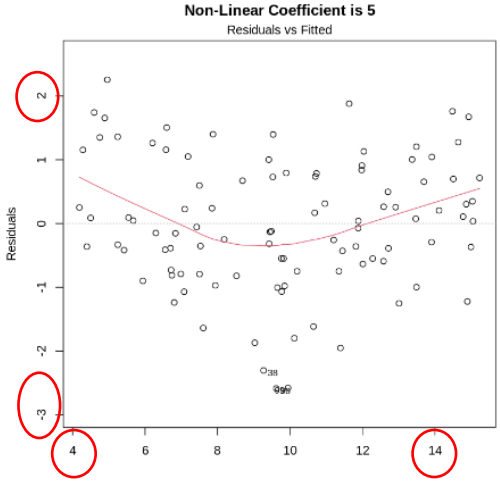
Q6) What do you notice about the residual plot?

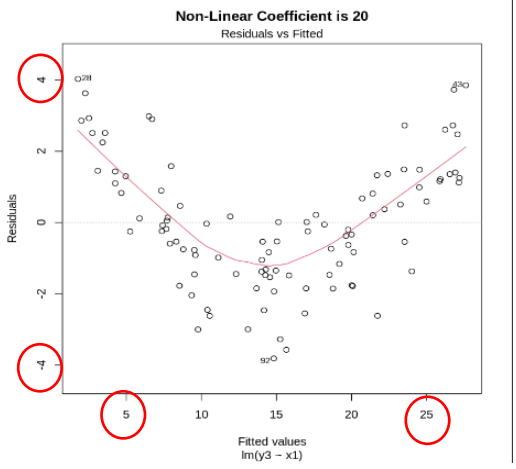
The residual values serve as a measure of the goodness of fit, indicating how well the line fits with the data.

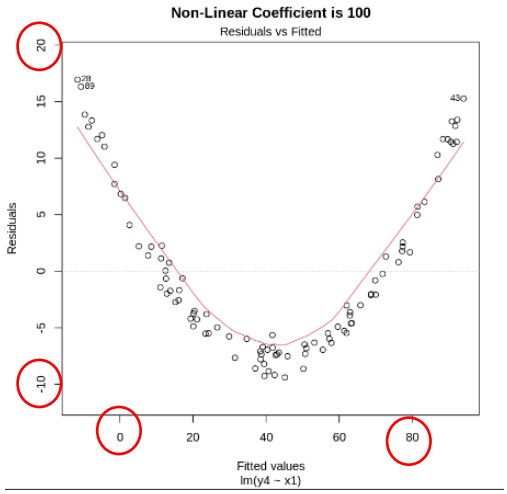
In this case, we observe **a non-linear pattern** in the plot, evidenced by the **increasing coefficient associated with the non-linear** component.

Despite this non-linearity, the data was generated with a constant variance.







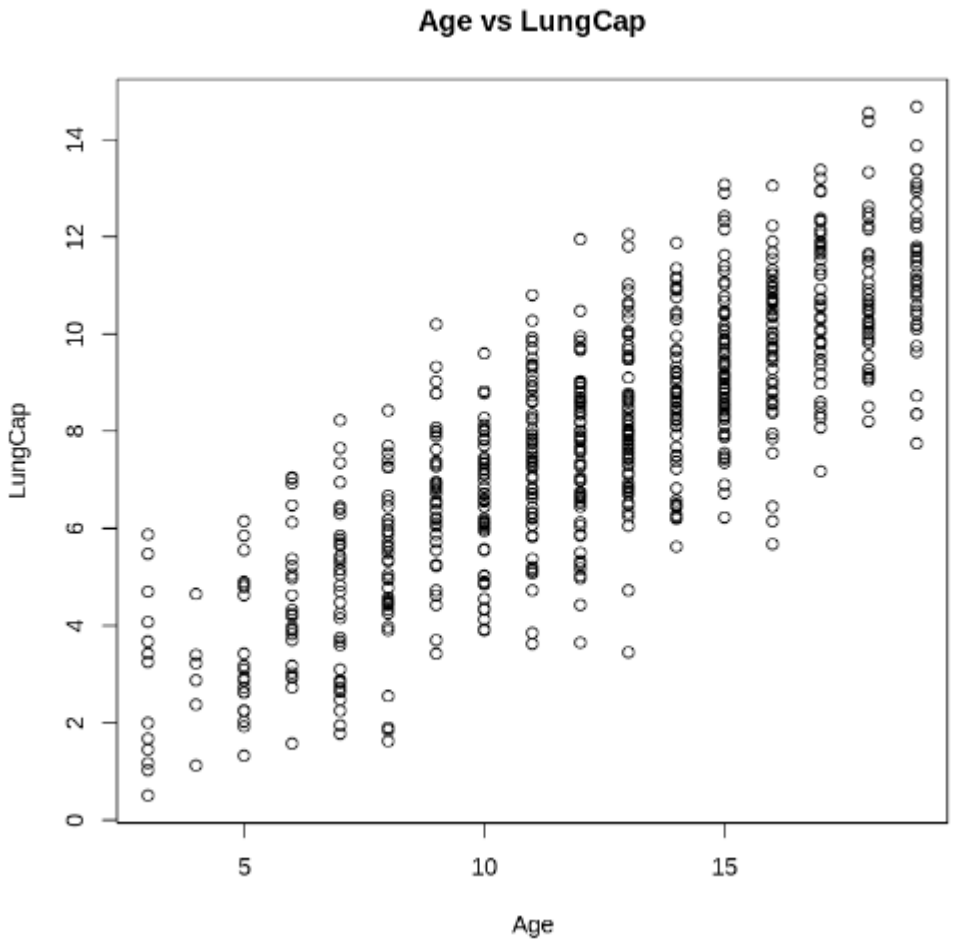


Part 3

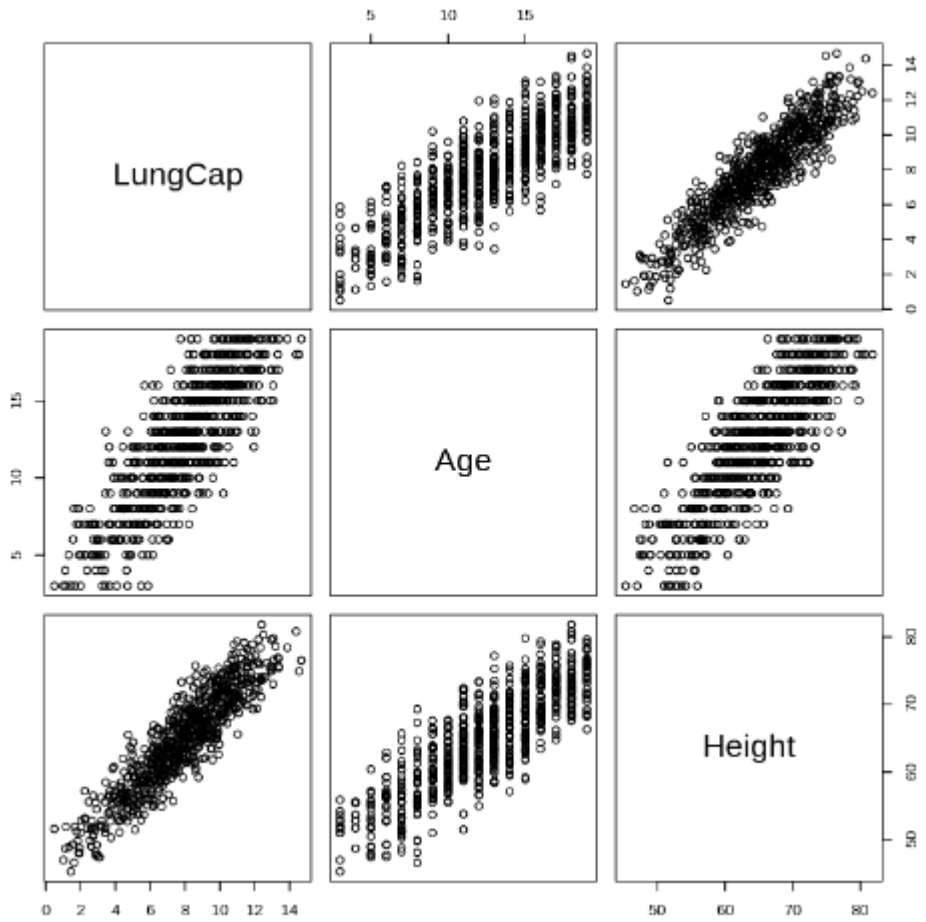
Q7) What are the variables in this dataset?

'LungCap' , 'Age' , 'Height' , 'Smoke' , 'Gender' , 'Caesarean'

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



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.

A matrix: 3 × 3 of type dbl

	LungCap	Age	Height
LungCap	1.0000000	0.8196749	0.9121873
Age	0.8196749	1.0000000	0.8357368
Height	0.9121873	0.8357368	1.0000000

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

Height

There is a stronger positive correlation between Height and LungCap compared to Age and LungCap.

This suggests that as Height increases, LungCap tends to increase as well.

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

Yes, they are correlated due to having a large correlation coefficient, which is very close to 1.

This indicates that an increase in height corresponds to an increase in lung capacity, and vice versa.

Additionally, we verify this hypothesis using Pearson's product-moment correlation.

```
        Pearson's product-moment correlation

data:  dfm$Height and dfm$LungCap
t = 59.856, df = 723, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8990911 0.9236522
sample estimates:
      cor
0.9121873
```

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

```
fit_lungCap <- lm(LungCap ~ Age + Height + Smoke + Gender + Caesarean, data = dfm)
```

Q14) Show a summary of this model.

```
Call:
lm(formula = LungCap ~ Age + Height + Smoke + Gender + Caesarean,
    data = dfm)

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 here ? What does R-squared indicate?

The R-squared value is 0.8542478.

R-squared represents the dispersion or scattering of data points around the regression line: **the higher the dispersion or scattering, the lower the R-squared value**.

In our case, the value is very close to 1, indicating that a line can fit the data well.

Q16) Show the coefficients of the linear model. Do they make sense?

Yes, the signs of the coefficients make sense.

The large intercept is logical because zero lies outside the observed data range. It's expected that lung capacity cannot be negative, and a normal newborn typically can't be shorter than 45 cm, resulting in a positive value for lung capacity.

Regarding the p-values, we need to remove the caesarean variable from the model since it has a significantly larger p-value compared to other variables.

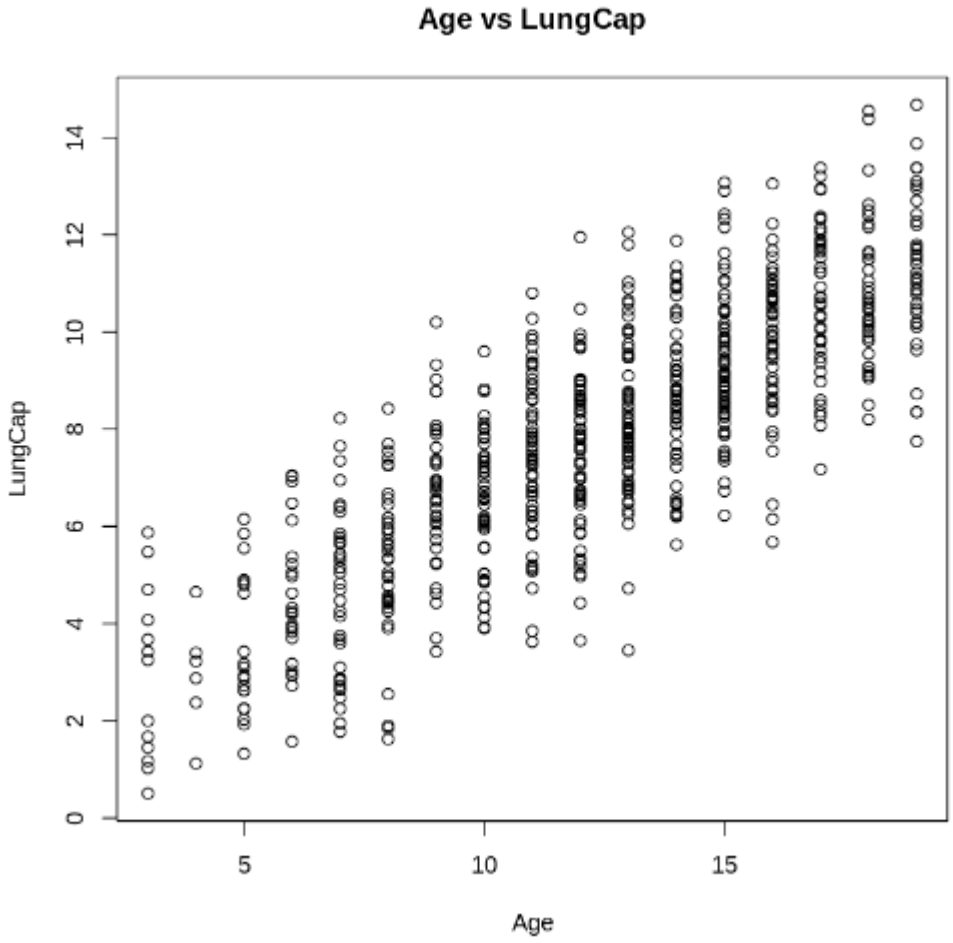
A matrix: 6 × 4 of type dbl

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-11.3224856	0.47097384	-24.040583	3.518644e-94
Age	0.1605296	0.01800726	8.914715	3.988466e-18
Height	0.2641128	0.01006215	26.248142	5.049675e-107
Smokeyes	-0.6095592	0.12597708	-4.838652	1.600357e-06
Gendermale	0.3870117	0.07965729	4.858459	1.452909e-06
Caesareanyes	-0.2142182	0.09073684	-2.360873	1.849787e-02

Q17) Redraw a scatter plot between Age and LungCap. Why the line is not displayed?

The intercept is at -11, and **the slope is 0.16**, so **the change in age alone cannot accurately capture the line.**

For example, when we compute -11 + 10 * 0.16, the result is -9.4, which lies outside the range of this plot.



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

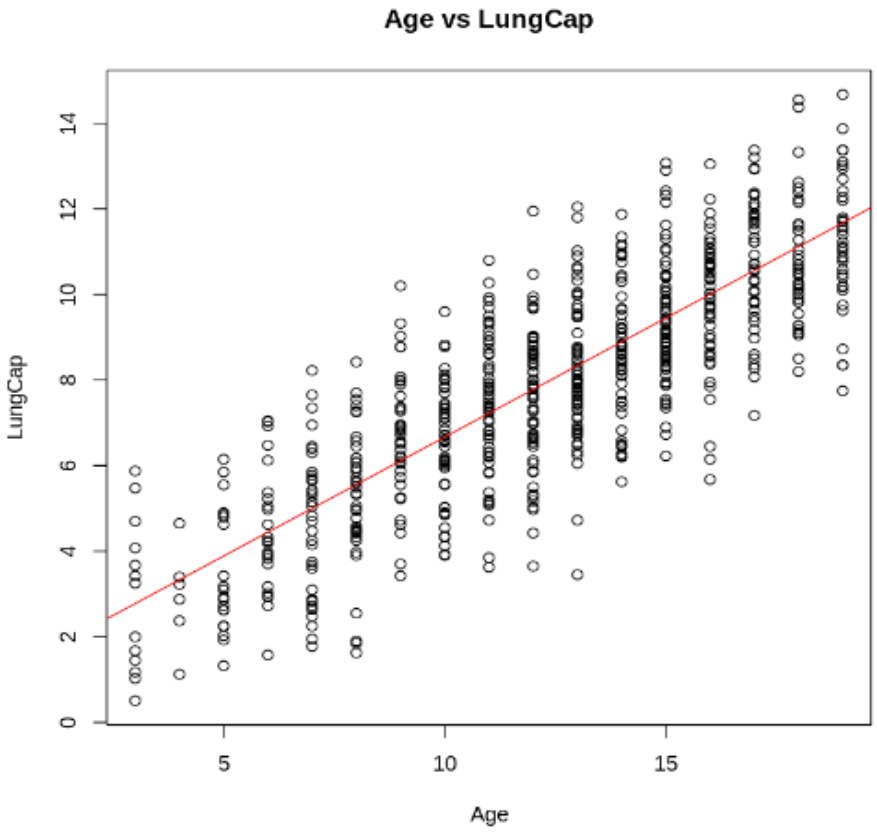
```
fit_lungCap <- lm(LungCap ~ Age + Smoke + Caesarean, data = dfm)
```

Q19 & Q20) Repeat Q16, Q17 for the new model. What happened?

In terms of the p-values, it may be worth considering the deletion of the caesarean variable from the model, given its large p-value compared to other variables.

A matrix: 4 × 4 of type dbl

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.1086723	0.18418699	6.019276	2.788391e-09
Age	0.5561667	0.01439403	38.638710	8.142669e-178
Smokeyes	-0.6431029	0.18680655	-3.442614	6.093803e-04
Caesareanyes	-0.1460278	0.13467911	-1.084265	2.786098e-01



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

2.280169