Lab 5

Predictive Analysis ||

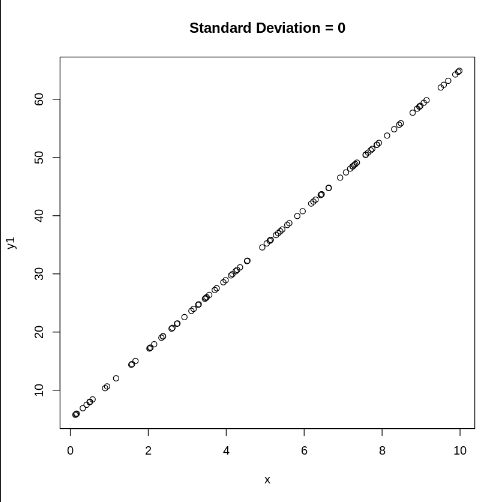
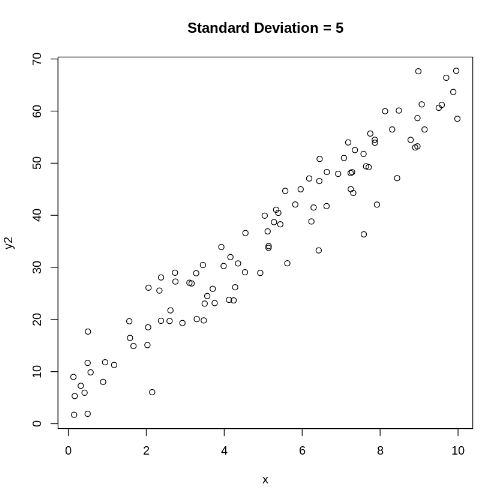
Karim Mahmoud Kamal Sec: 2 BN: 12

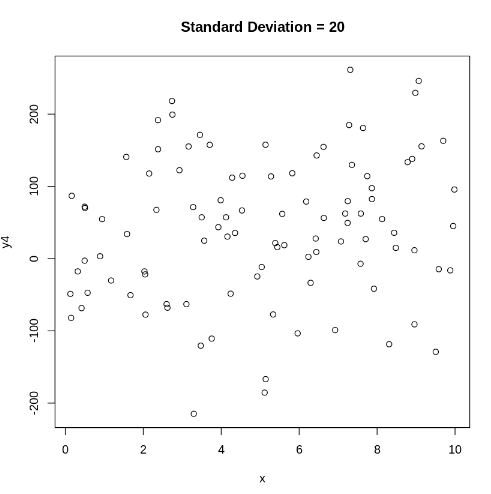
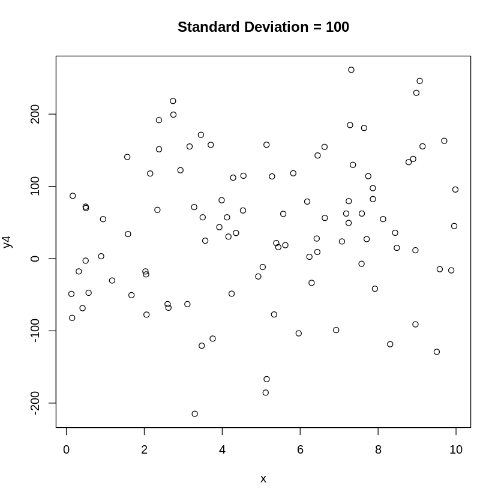
Mustafa Mahmoud Hamada Sec: 2 BN: 25

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.



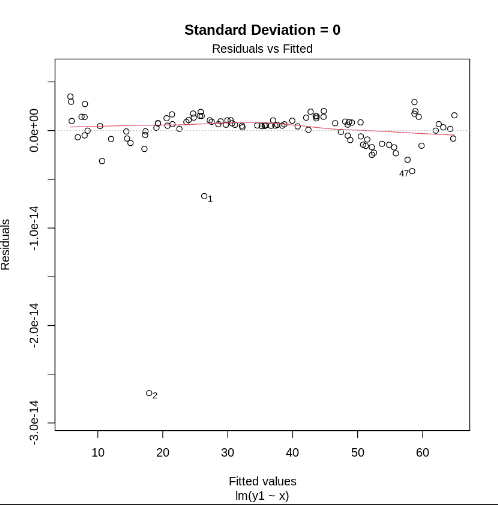
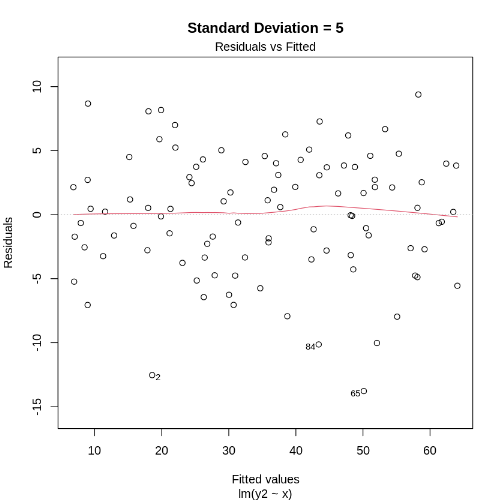
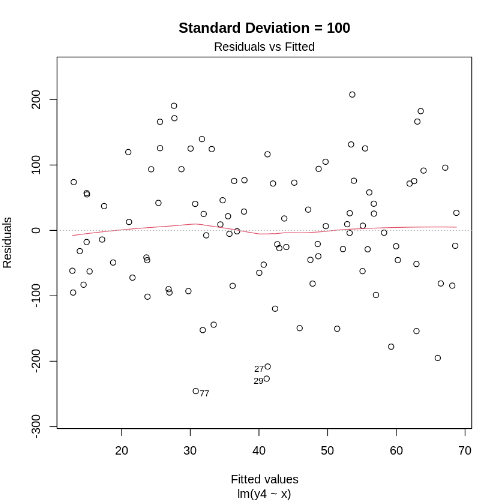
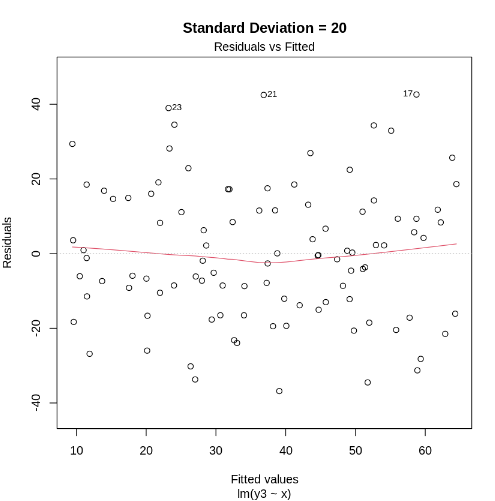
## Q2) How are the coefficients of the linear model affected by changing the value of standard deviation in Q1?

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.**

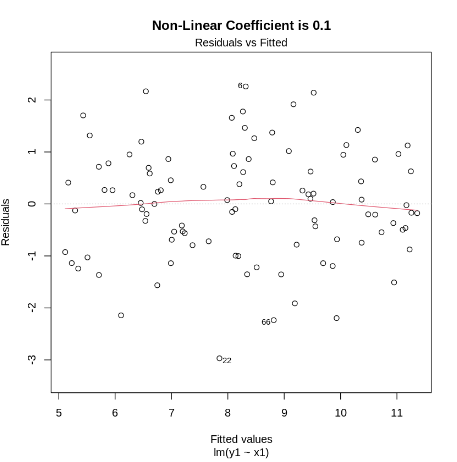
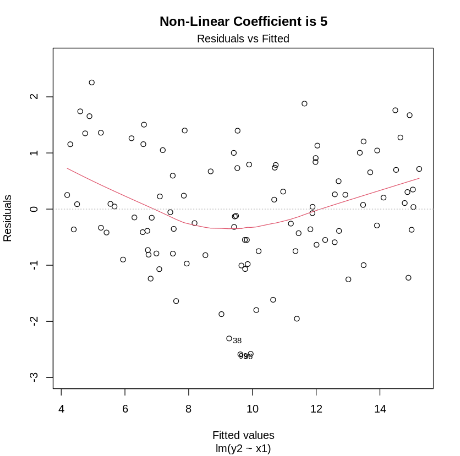
# Part 2

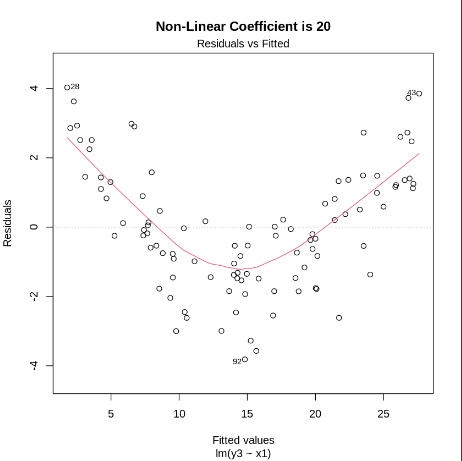
## Q5) What do you conclude about the residual plot? Is it a good residual plot?

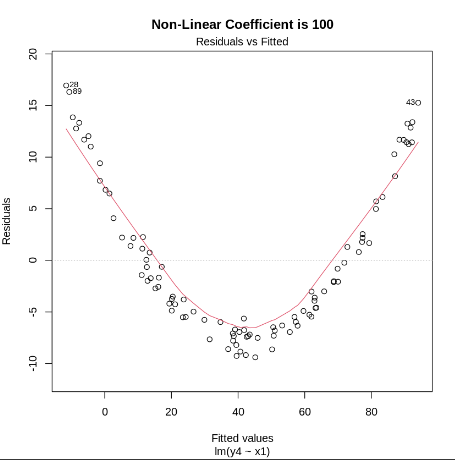
**The residual values serve as a measure of the goodness of fit**, indicating how well the line fits with the data. In this case, we observe **a linear pattern** in the plot, given the **negligible coefficient associated with the non-linear term**, and the data was generated with a constant variance, 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.





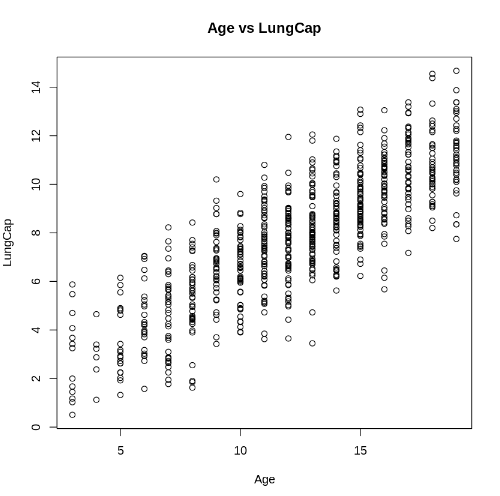


# Part 3

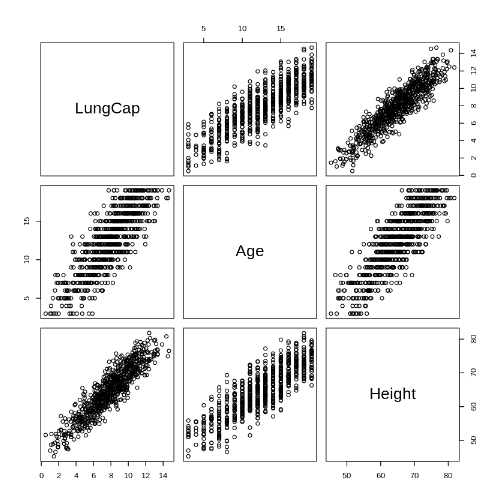
## Q7) What are the variables in this dataset?

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

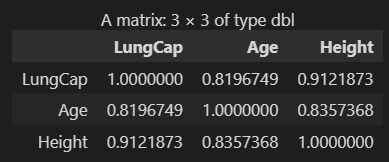
## Q8) Draw a scatter plot of Age (x-axis) vs. LungCap (y-axis).



## 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.

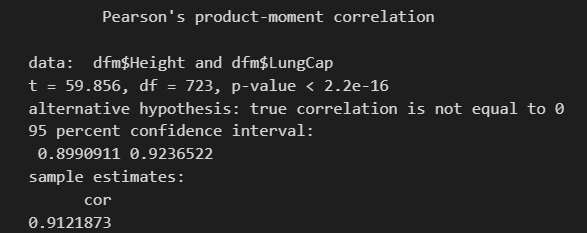


## Q11) Which of the two input variables Age and Height are more correlated to the dependent variable LungCap?

**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.



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



## Q14) Show a summary of this model.

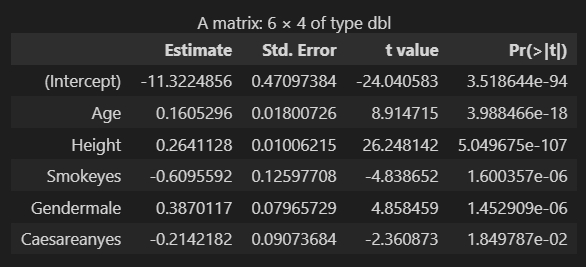
## Q15) What is the R-squared value here ? What does R-squared indicate?

**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?

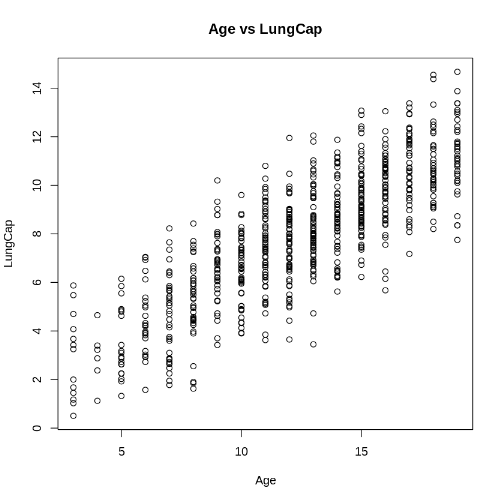
**Yes**, the signs of the coefficients make sense. The large intercept is logical because zero lies outside the observed data range. It's expected that lung capacity cannot be negative, and a normal newborn typically can't be shorter than 45 cm, resulting in a positive value for lung capacity.

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.



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

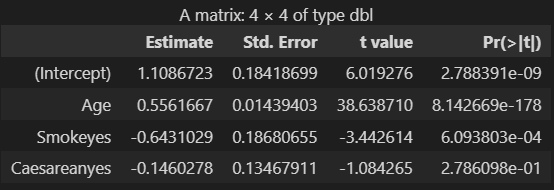
**The intercept is at -11**, and **the slope is 0.16**, so **the change in age alone cannot accurately capture the line.** For example, when we compute -11 + 10 \* 0.16, the result is -9.4, which lies outside the range of this plot.

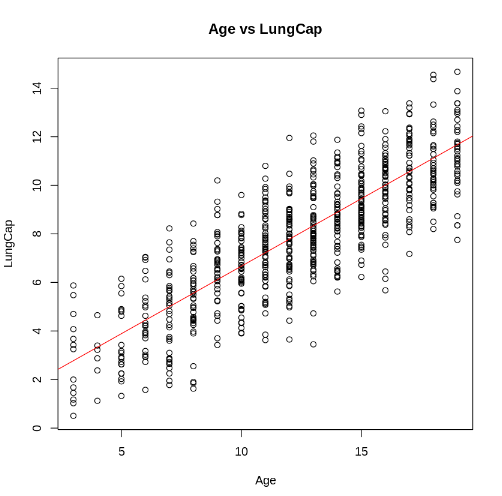


## Q18) Repeat Q13 but with these variables Age, Smoke and Cesarean as the only independent variables



## 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.



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

**2.280169**