Lab 4

Linear Regression

Muhammad Ahmad Hesham Sec: 2 BN: 16

Muhammad Alaa Abdel-Khaleq Sec: 2 BN: 22

# Part 1

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

As shown in the following figures Figure 1, Figure 2, Figure 3, Figure 4 the change in the *sd* variable changes the variance of the noise in the input data so for large *sd* the points are more scattered from the line and for small *sd* the points are very close to a line i.e. the square error is minimal and will equal zero when *sd=0*

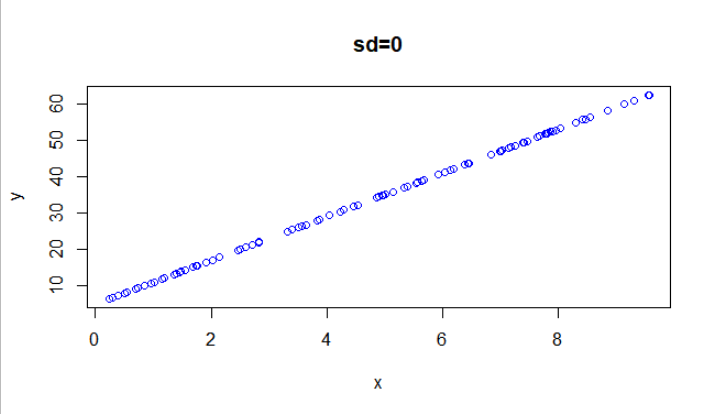


Figure 1

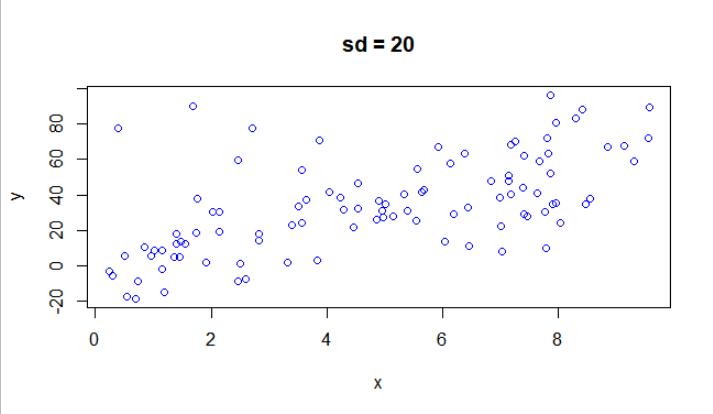


Figure 2

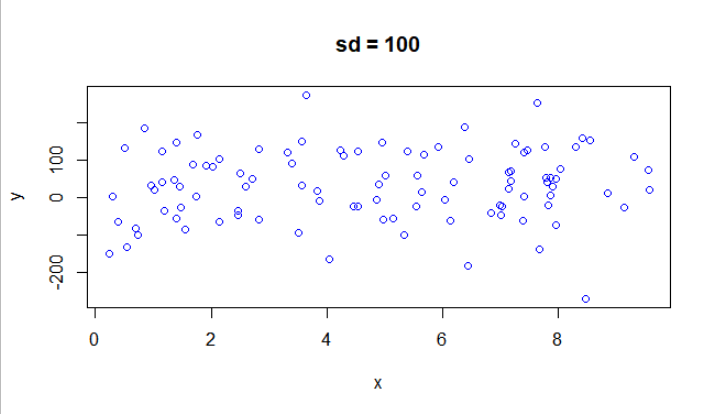


Figure 3

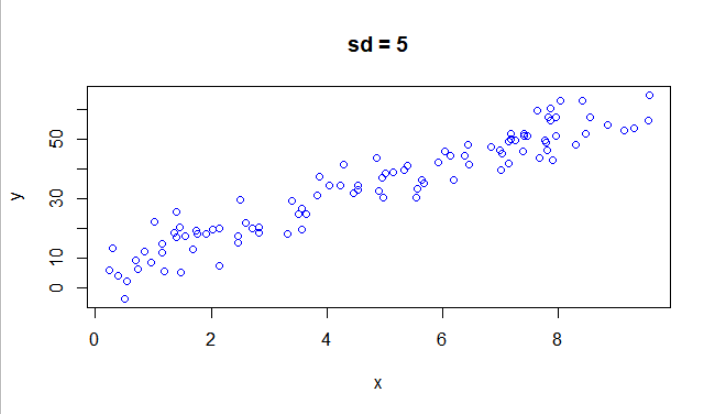


Figure 4

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

The higher the standard deviation the more deviated from the original line we get

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

When the standard deviation increases the value of the R-square decreases which has values between 0 and 1 where 1 means that all of the variation in the y can be explained by the variable x and the intercept. R-squared represents the scatter around the regression line the higher the scatter the lower the R-square value.

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

The residual values provide us some measure of how well the line fits the data, that is, the **goodness of fit**. We can see that there is a linear pattern in the plot and the data was generated with a constant variance but we notice large values for the residual when the standard deviation is large i.e., the model doesn’t fit the data very well and the noise is high.

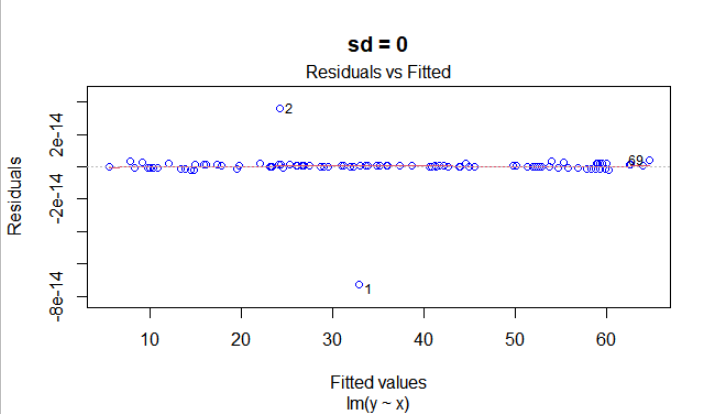


Figure 8

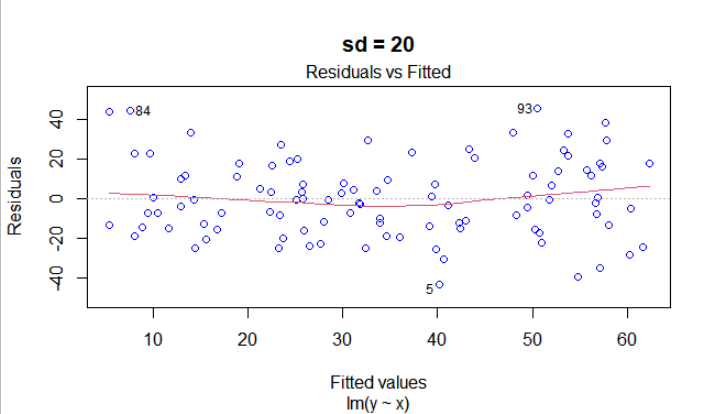


Figure 8

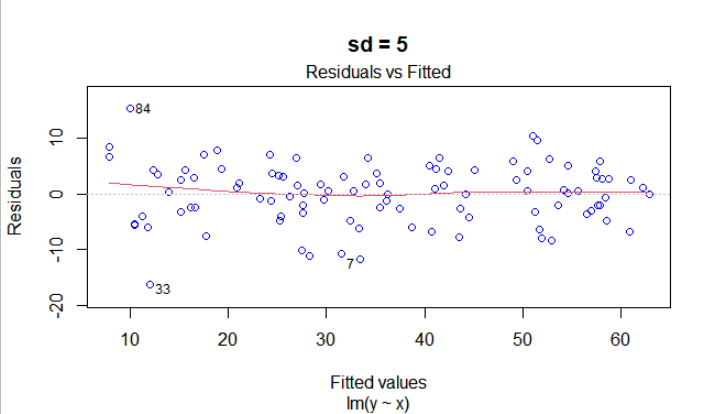


Figure 8

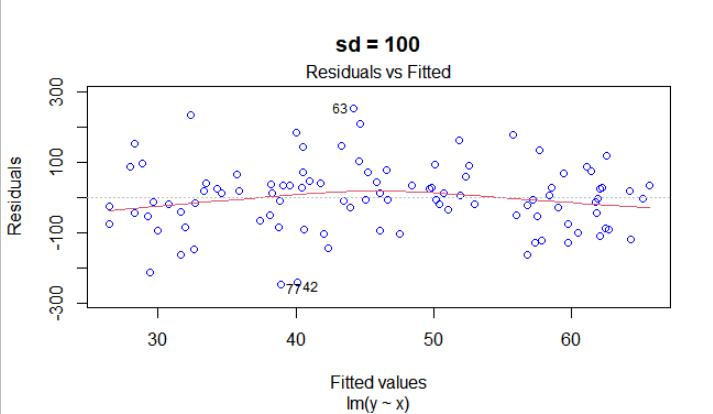


Figure 8

# Part 2

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

The residual values provide us some measure of how well the line fits the data, that is, the **goodness of fit**. We can see that there is a linear pattern in the plot as the non-linear term has a very small coefficient and the data was generated with a constant variance.

## Q6) What do you notice about the residual plot?

The residual values provide us some measure of how well the line fits the data, that is, the **goodness of fit**. We can see that there is a non-linear pattern in the plot as the coefficient of the non-linear part increases and the data was generated with a constant variance.

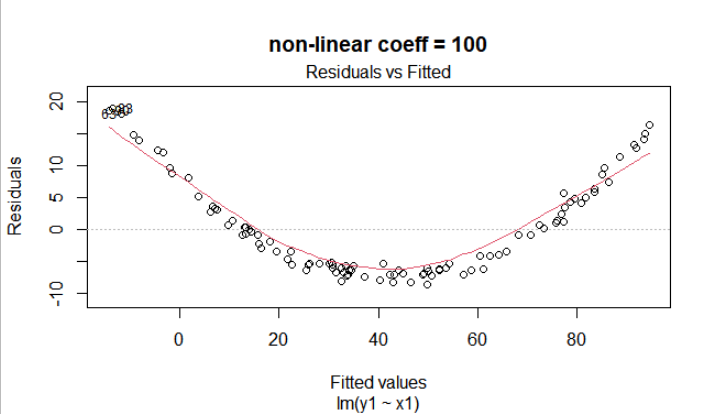


Figure 12

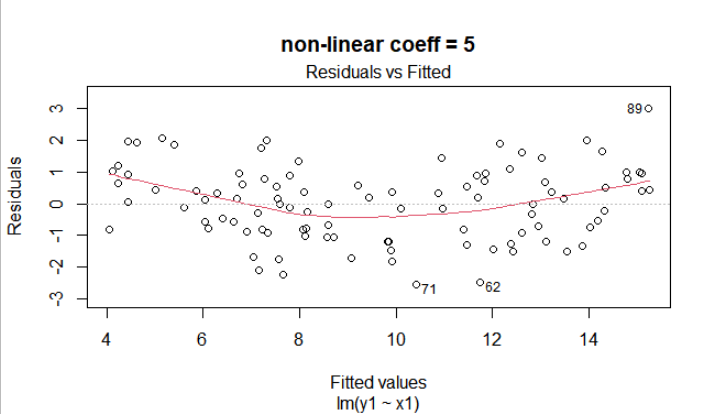


Figure 12

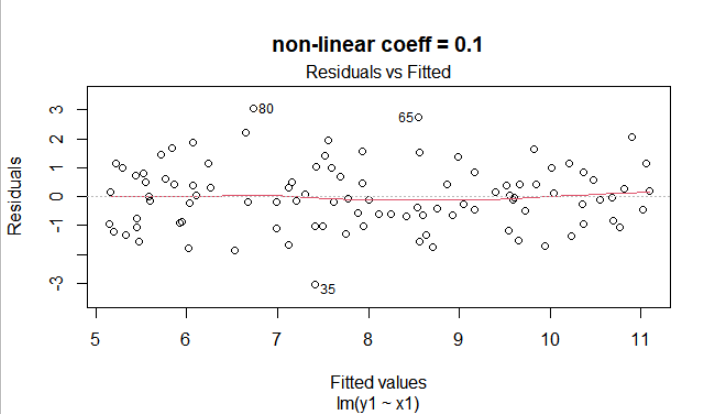


Figure 12

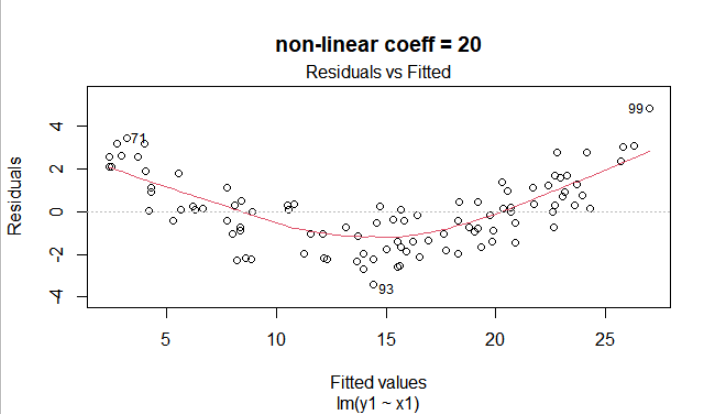


Figure 12

# 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).

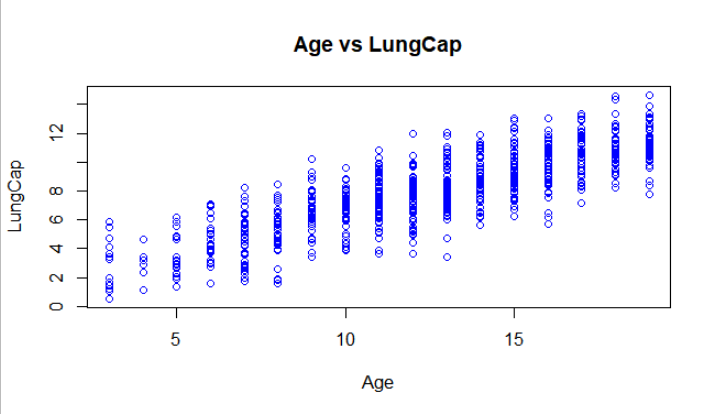


Figure 13

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

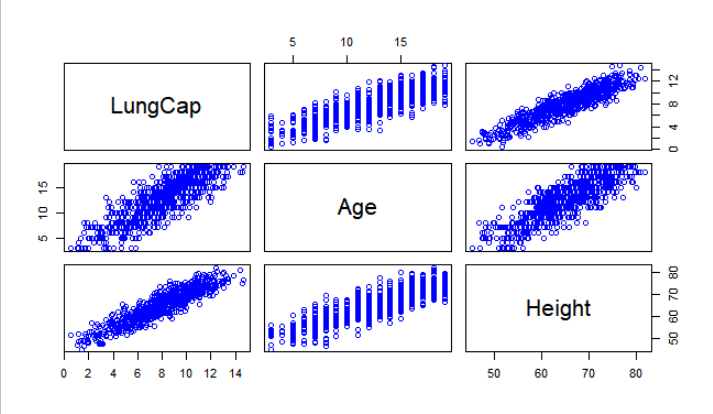


Figure 14

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

LungCap Age Height

LungCap 1.0000000 0.8196749 0.9121873

Age 0.8196749 1.0000000 0.8357368

Height 0.9121873 0.8357368 1.0000000

Figure 15

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

Height is more correlated to LungCap than Age is and there is a positive correlation. i.e., when one increases the other increases

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

Yes they are correlated because they have a large correlation coefficient, very close to 1. i.e., increase in height indicates an increase in lung capacity and vice versa. Also using cor.test with Pearson's product-moment correlation we verify this hypothesis.

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

0.8542478. the R-squared represents the scatter around the regression line the higher the scatter the lower the R-square value. The value in our case is very close to 1 meaning that a line can fit the data well

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

Yes they make sense regarding the signs of the coefficients, the intercept is large because the zero lies outside the observed data of course there is no negative capacity but a normal new born can’t be shorter than 45 cm which will get a positive value for the lung capacity.  
regarding the p values we might consider deleting the caesarean variable as it has a very large p value compared to other variables.

Figure 16

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

Intercept is at -11 and the slope is 0.16 so the change in age can’t capture the line for example -11+20\*0.16 is -7.8 which is outside the range of this plot.

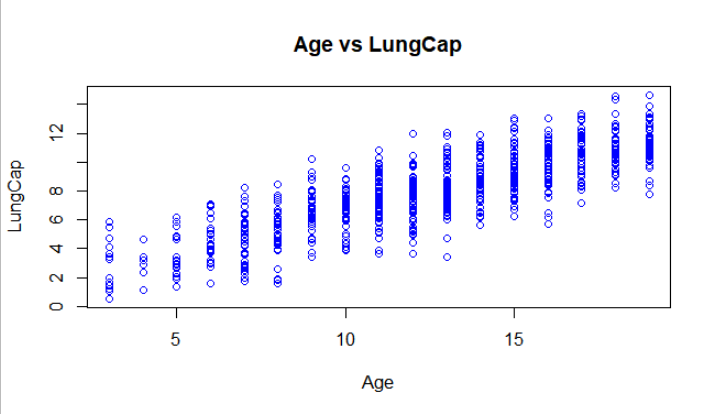


Figure 17

## Q19) Repeat Q13 but with these variables Age, Smoke and Cesarean

regarding the p values we might consider deleting the caesarean variable as it has a very large p value compared to other variables.

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

Figure 18

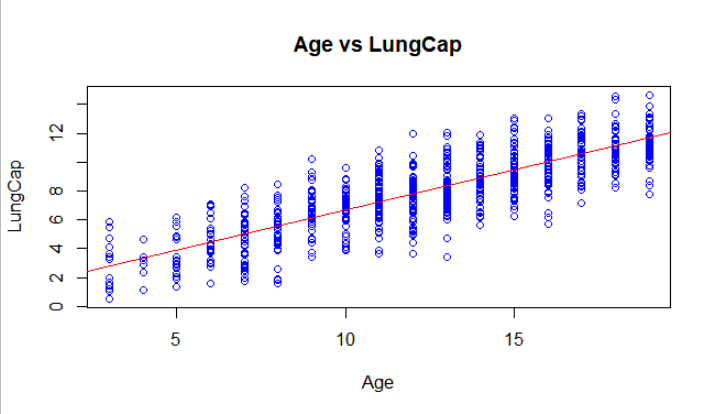


Figure 19

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

2.280169