Big Data

Linear Regression

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Sec** | **BN** | **ID** |
| Sara Gamal Gerges | 1 | 20 | 9210455 |
| Eman Ibrahim | 1 | 14 | 9210265 |

# Part 1

## Q1)

|  |  |
| --- | --- |
| **Std=2** | **Std=10** |
| **Std=20** | **Std=50** |

**Comment:**

**- smaller sd (as 2 &10) Data points are closely clustered around the line y=5+6xy, meaning the spread of the points is small.**

**-Higher sd (as 20 &50) introduces more randomness, making the plot appear noisier with points scattered further from the expected linear trend.**

## Q2)

**Gold: y=5+6x.**

|  |  |
| --- | --- |
| **Std=2**  **Y= 5.717+x\*5.862** | **Std=10**  **y= 3.823+x\* 5.999** |
| **Std=20**  **Y= 6.008 +x\*5.277** | **Std=50**  **Y=-6.828 +x\* 6.985** |

**Comment:**

**Higher sd makes more noisy data resulting in model’s coefficients being farer from the original coefficients.**

## Q3)

|  |  |
| --- | --- |
| **Std=2**  **OLS gave slope of 5.861764 and an R-sqr of 0.9881344** | **Std=10**  **OLS gave slope of 5.999485 and an R-sqr of 0.8157217** |
| **Std=20**  **OLS gave slope of 5.276517 and an R-sqr of 0.4150892** | **Std=50**  **OLS gave slope of 6.984824 and an R-sqr of 0.1443906** |

**Comment:**

**Higher sd results in getting worse R2 (far from 1) which means more error in the predictions.**

## Q4)

|  |  |
| --- | --- |
| **Std=2** | **Std=10** |
| **Std=20** | **Std=50** |

**Comment**

**- when sd is small, the residuals range is small(y-axis), indicating less error**

**- when sd is large, the residuals range is larger(y-axis), indicating more error**

**-points are randomly scattered (there is no pattern in the data) which indicate good model**

# Part 2

## Q5)

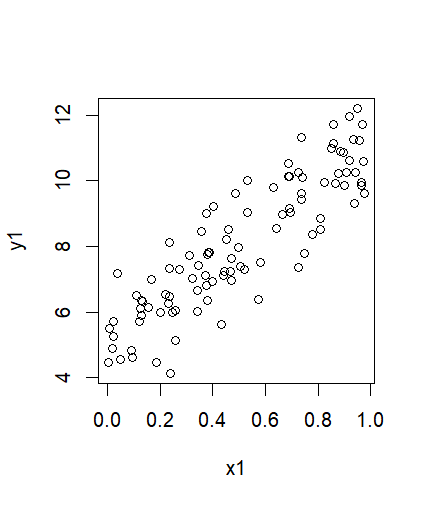
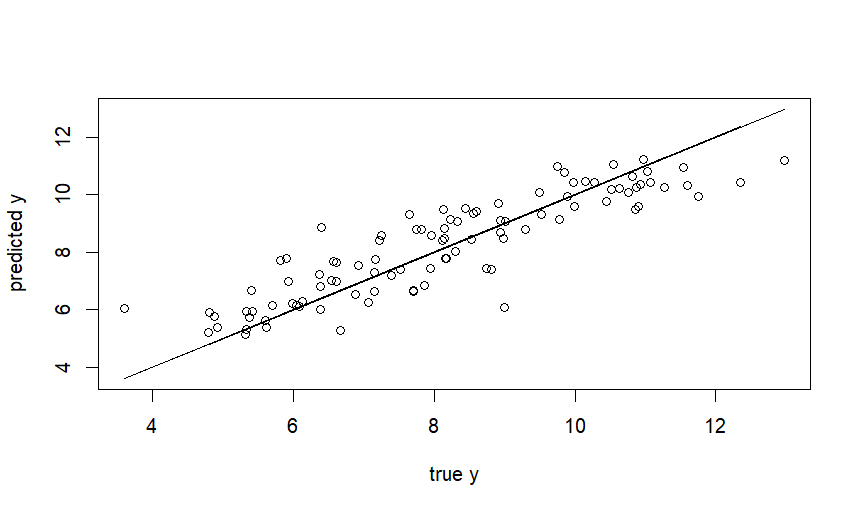
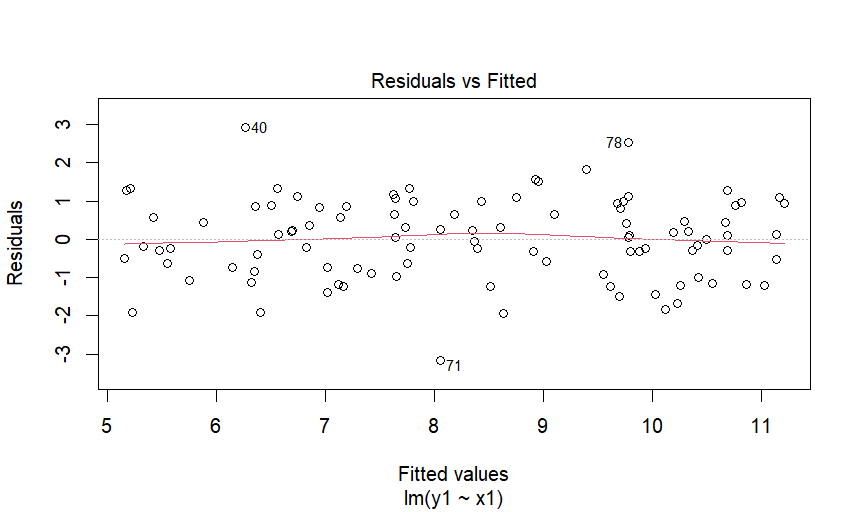


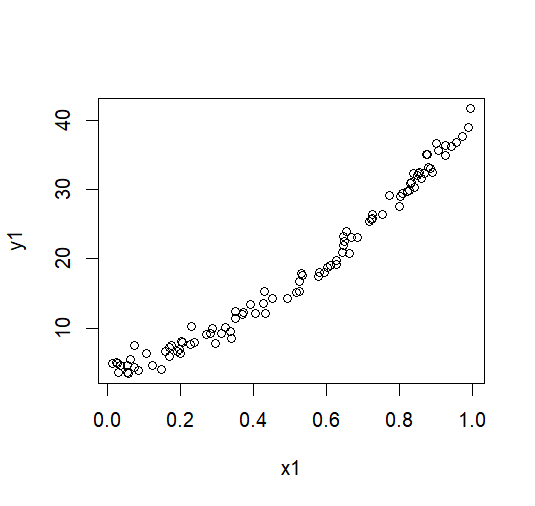
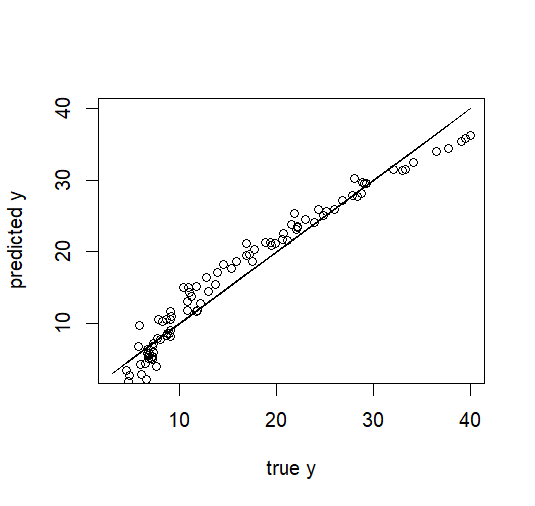
Figure 1 test points

Figure 2 training points



**Comment:** The residuals plot is random, there is no pattern, and the residual range is small which suggests good model and good plot

## Q6)

changing the nonlinear coefficient to 30. (y1 = 5 + 6\*x1 + 30\*x1\*x1 rnorm(100))

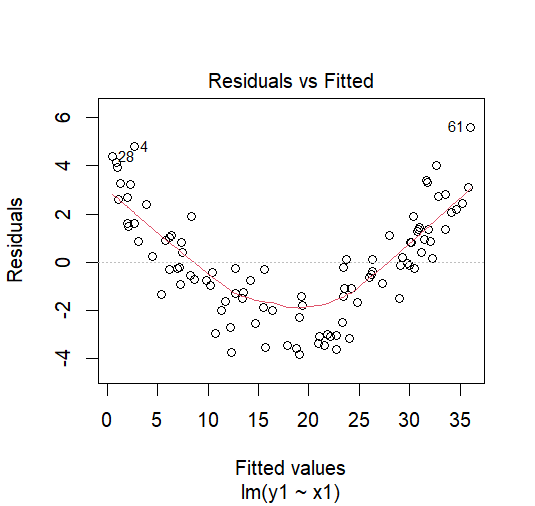


Figure 1 test points

Figure 2 training points

**Comment:** The residuals plot is not random; a pattern exists which indicates that the model is not capturing all the information in the data which is clear because the model makes wrong assumption that the data is linear when in reality the data is not linear.

# Part3

## Q7) the variables in this dataset are:

* LungCap
* Age
* Height
* Smoke
* Gender
* Caesarean

## Q8) It’s clear that as the age increases the lungCap increases.

## Q9) It’s clear that lungcap is more correlated with height

## Q10)

A close-up of a text

AI-generated content may be incorrect.

## Q11)

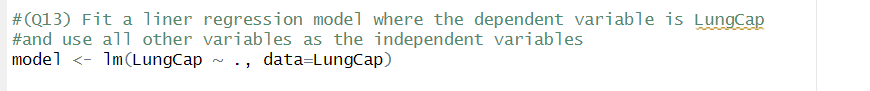
**Height**

## Q12)

Yes, height and lung capacity are correlated (the correlation between them =0.912 which means there is correlation between them)

This implies that taller people have larger lung capacity.

## Q13)

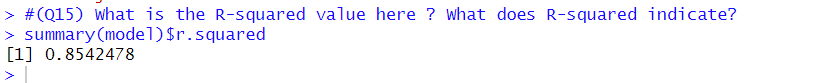


## Q14)

A screenshot of a computer

AI-generated content may be incorrect.

## Q15)



R-squared value= 0.8542478

R-squared is a measure that indicates how well the independent variables explain the variance in the dependent variable in a regression model. It ranges from **0 to 1**:

* **R² = 0** means the independent variables explain none of the variance in the dependent variable. (bad fit)
* **R² = 1** means the independent variables explain all the variance in the dependent variable. (good fit)

## Q16)

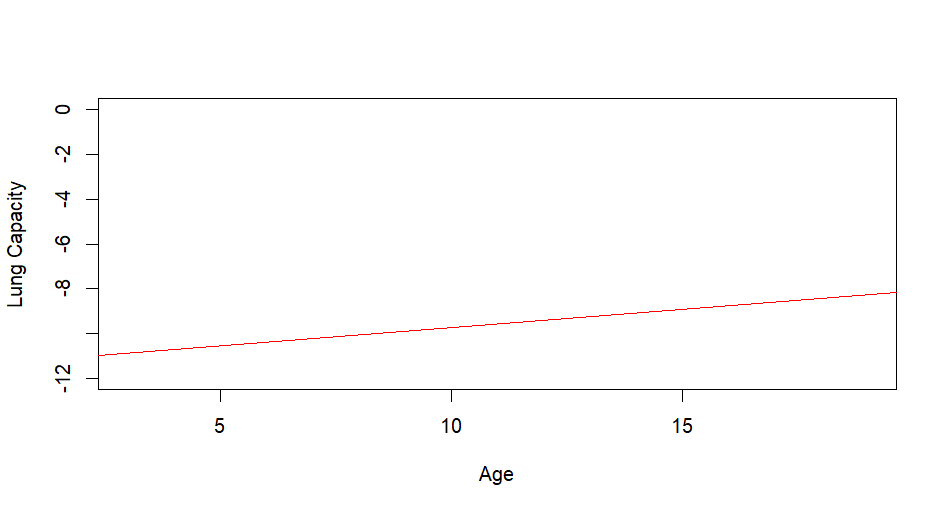
A screenshot of a computer

AI-generated content may be incorrect.

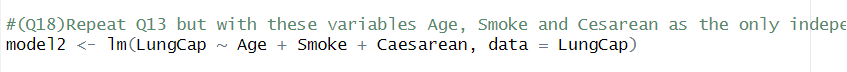
Yes, they make sense

* **Age (0.16):** which implies that older individuals tend to have slightly higher lung capacity.
* **Height (0.26):** which implies that taller individuals have higher lung capacity, which makes sense.
* **Smokes (Yes) (-0.61):** Smoking negatively affects lung capacity, which aligns with medical findings.
* **Gender (Male) (0.39):** Males tend to have slightly higher lung capacity.
* **Caesarean (Yes) (-0.21):** which implies that caesarean birth slightly lowers lung capacity.

## Q17)

The line is not displayed because it’s below the y-range shown in the previous image

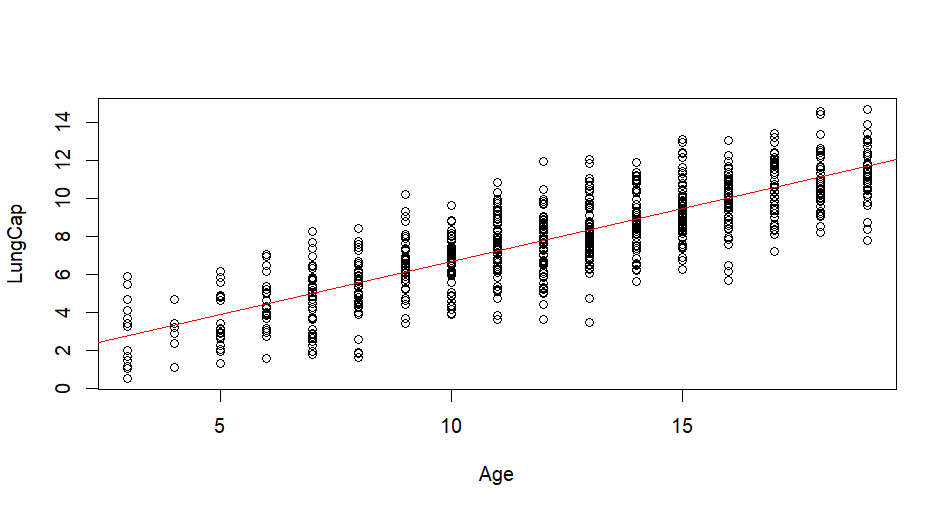
## Q18)



## Q19)

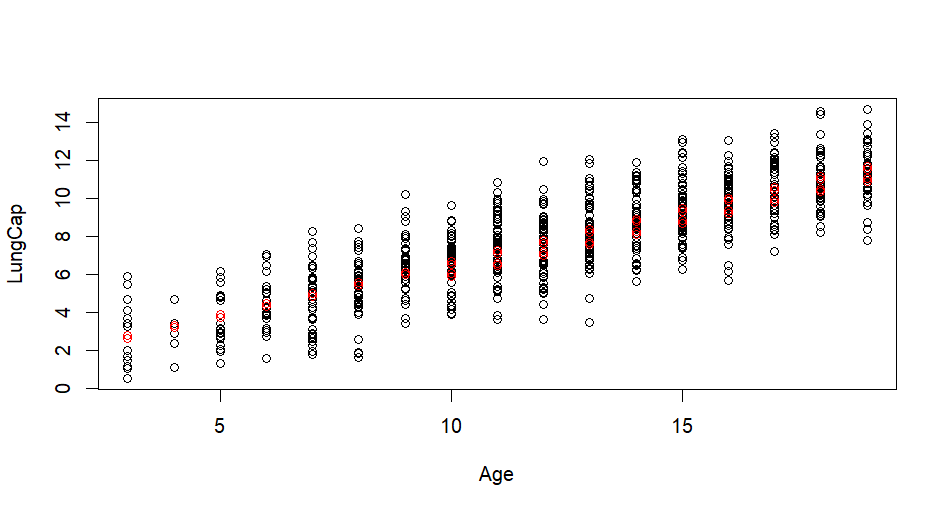
A computer code with numbers and symbols

AI-generated content may be incorrect.

Yes, the coefficients make sense for the same reasons as Q16

The most significant change is the age coefficient increased from 0.16 to 0.55(which mean is model heavily relies on the age coefficient compared to the previous model )

## Q20)

A computer screen shot of a computer code

AI-generated content may be incorrect.

## Q21)

For second model

A computer error message

AI-generated content may be incorrect.

For first model

A close-up of a computer code

AI-generated content may be incorrect.