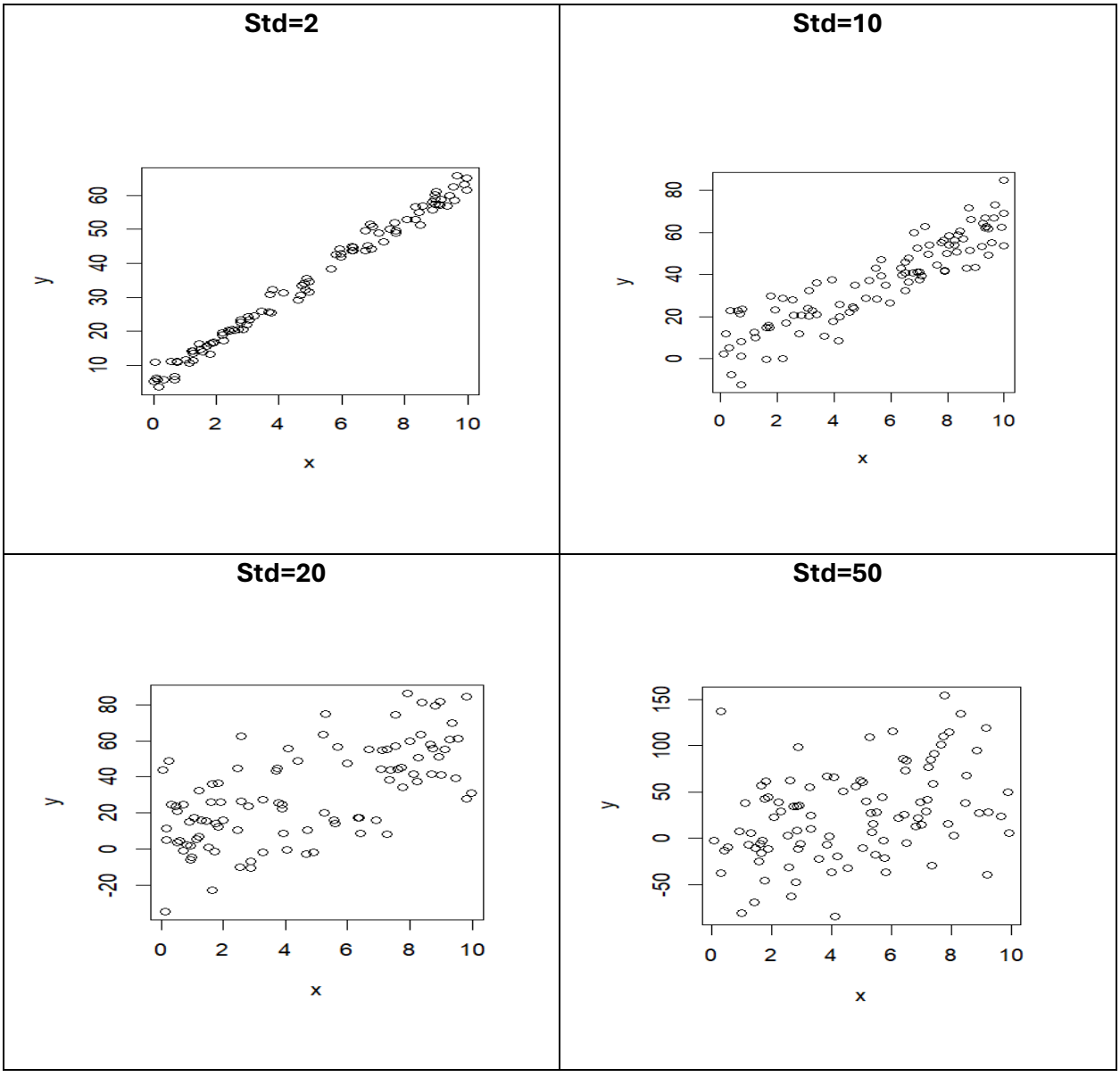
Big Data

Linear Regression

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

Part 1

Q1)



Comment:

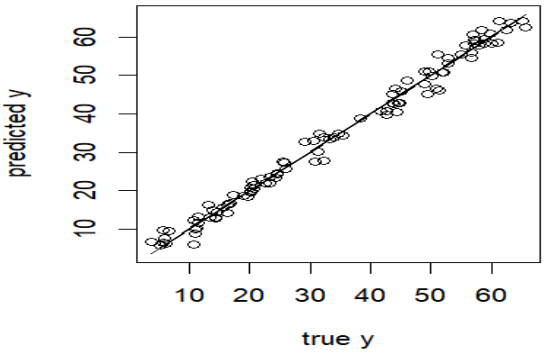
- smaller sd (as 2 &10) Data points are closely clustered around the line y=5+6xy, meaning the spread of the points is small.

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

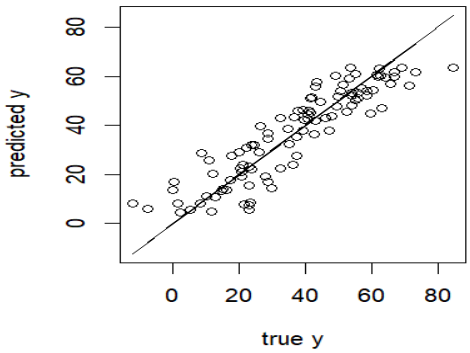
Q2)

Gold: y=5+6x.

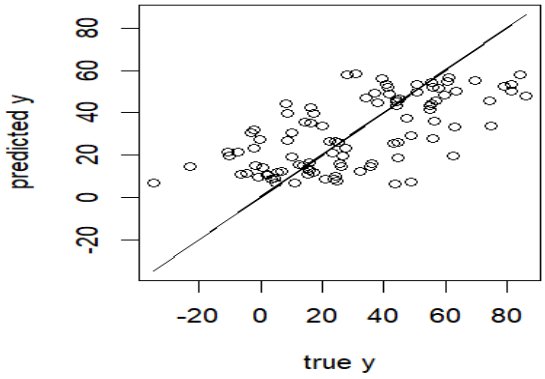
Std=2
Y= 5.717+x*5.862



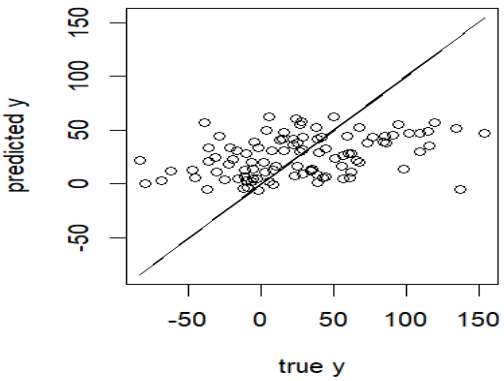
Std=10
y= 3.823+x* 5.999



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)

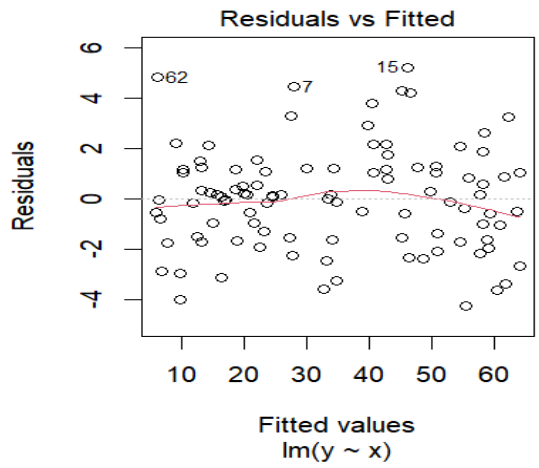
Std=2 **OLS gave slope of 5.861764 and an R-sqr of 0.9881344**	**Std=10** **OLS gave slope of 5.999485 and an R-sqr of 0.8157217**
Std=20 **OLS gave slope of 5.276517 and an R-sqr of 0.4150892**	**Std=50** **OLS gave slope of 6.984824 and an R-sqr of 0.1443906**

Comment:

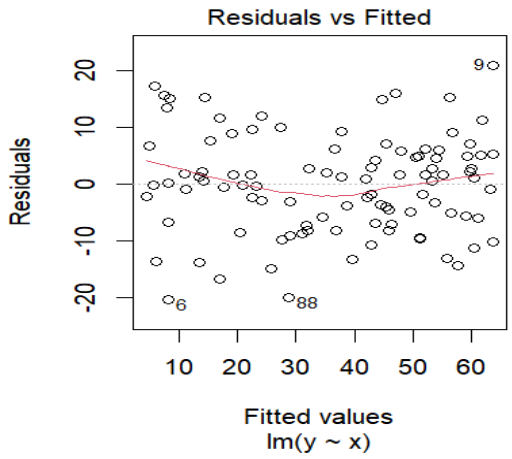
Higher sd results in getting worse R^2 (far from 1) which means more error in the predictions.

Q4)

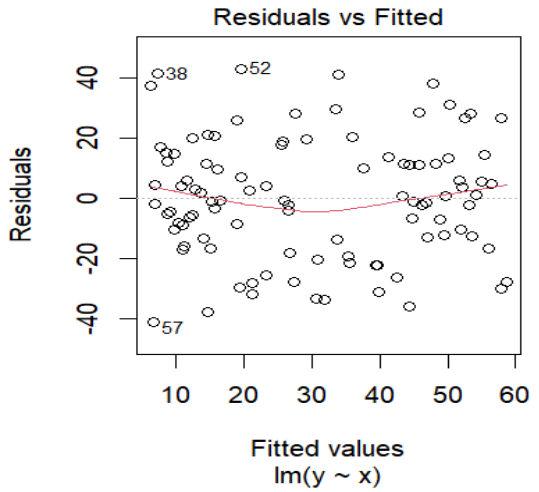
Std=2



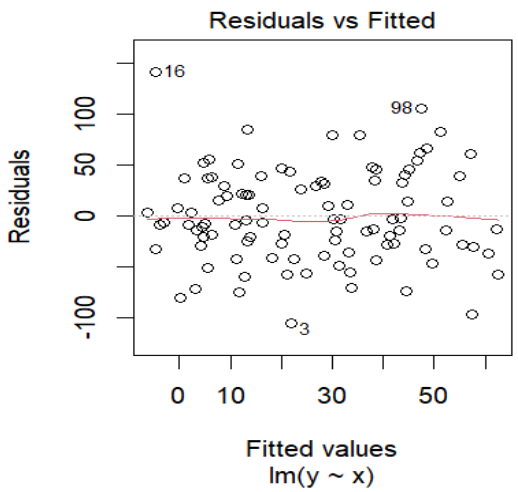
Std=10



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)

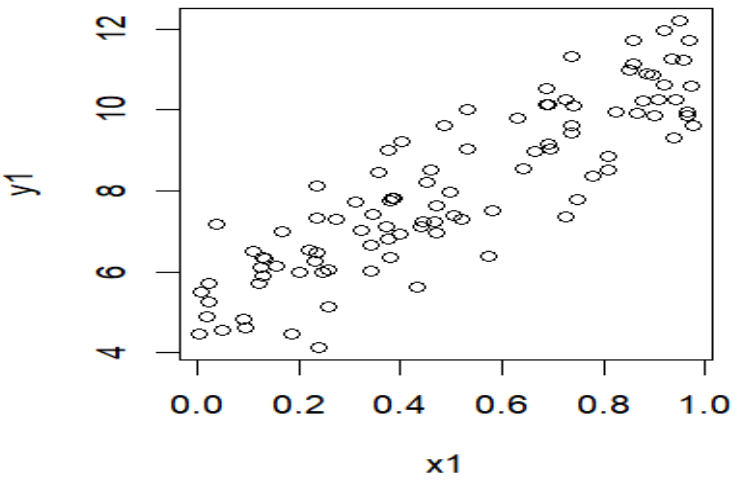


Figure 2 *training points*

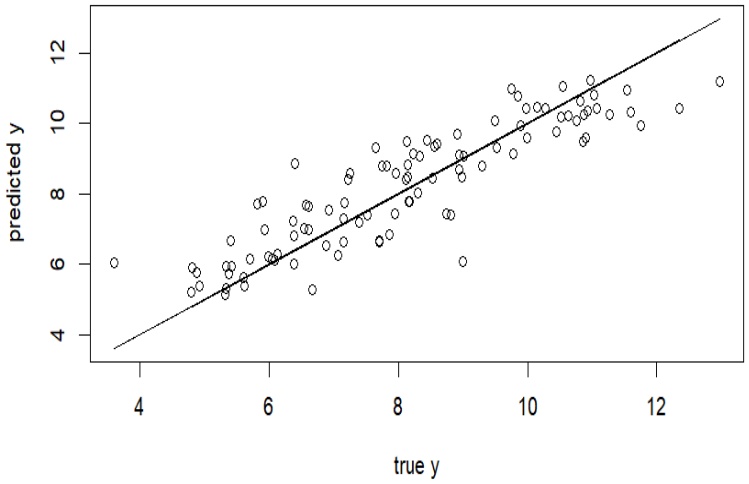
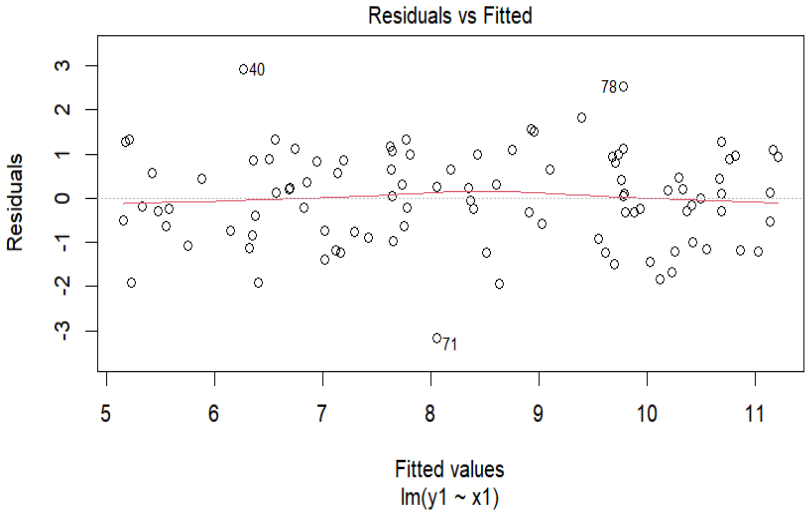


Figure 1 *test points*



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))

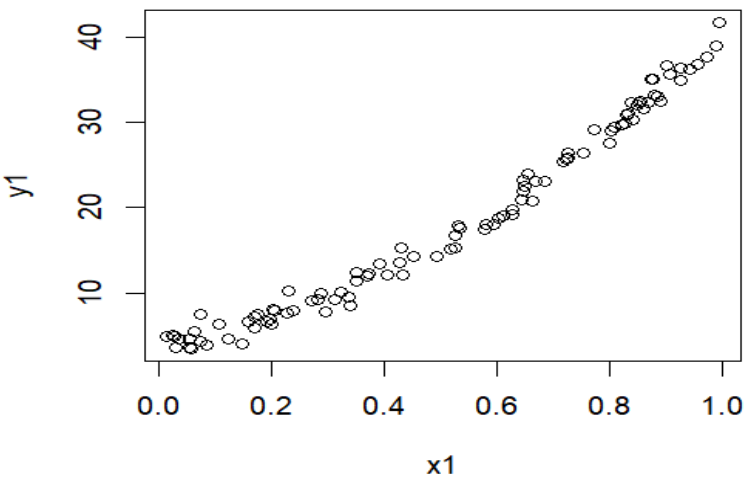


Figure 4 *training points*

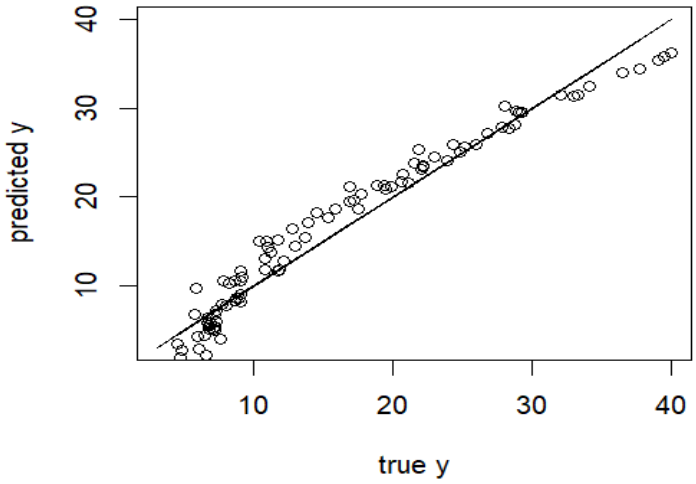
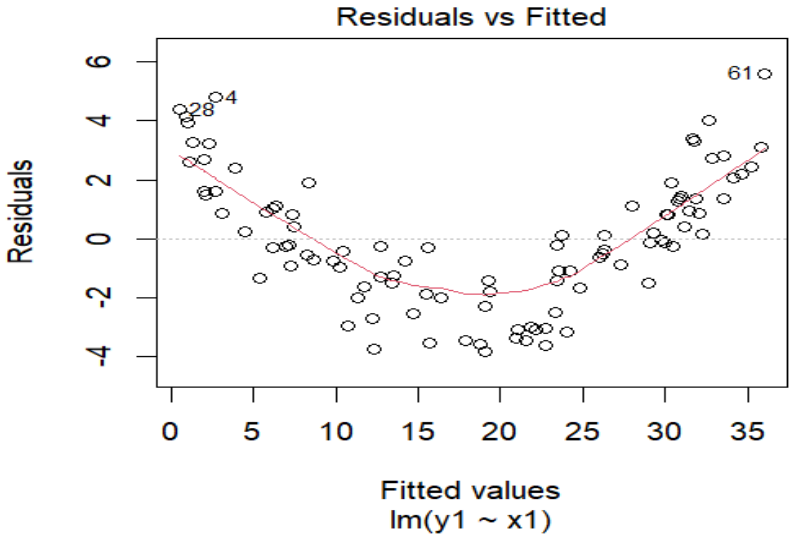


Figure 3 *test points*



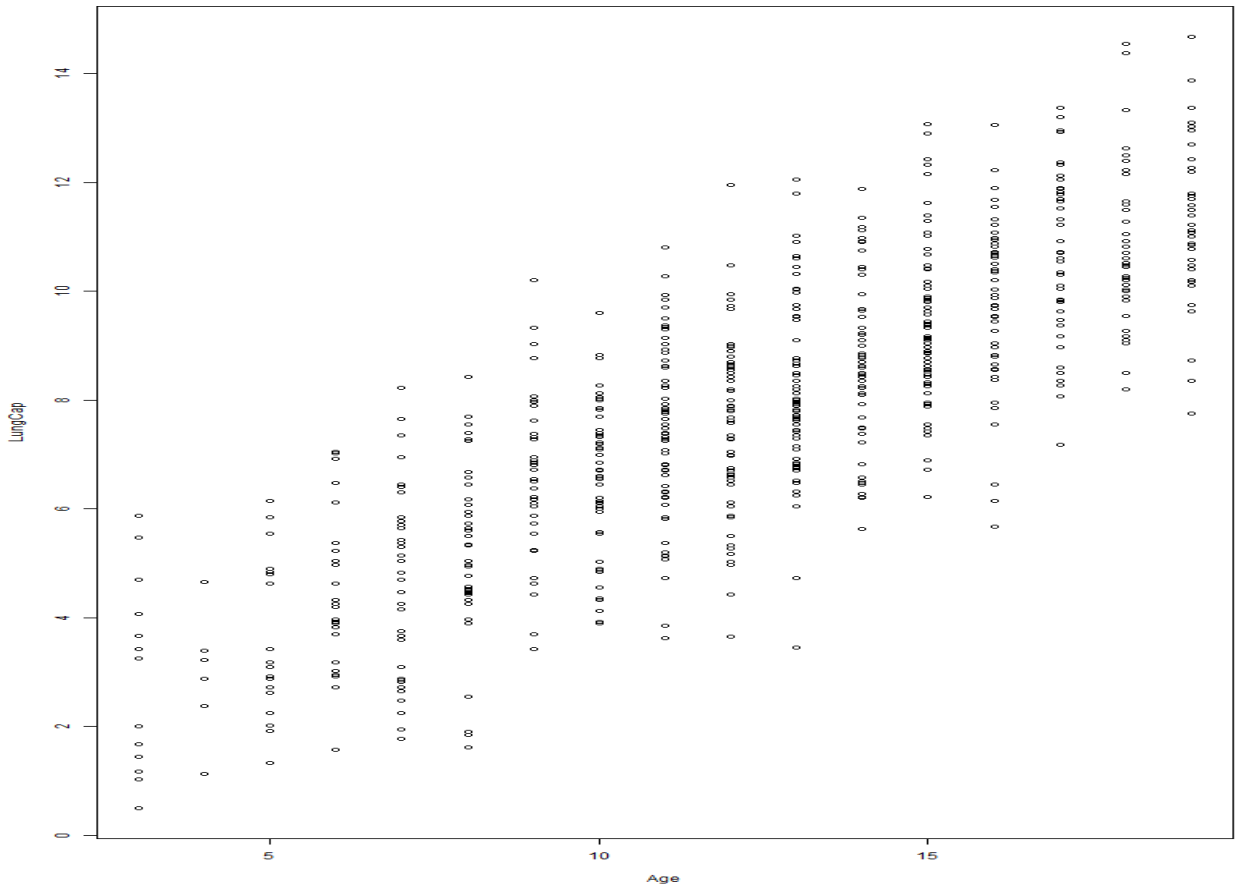
Comment: The residuals plot is not random; a pattern exists which indicates that the model is not capturing all the information in the data which is clear because the model makes wrong assumption that the data is linear when in reality the data is not linear.

Part3

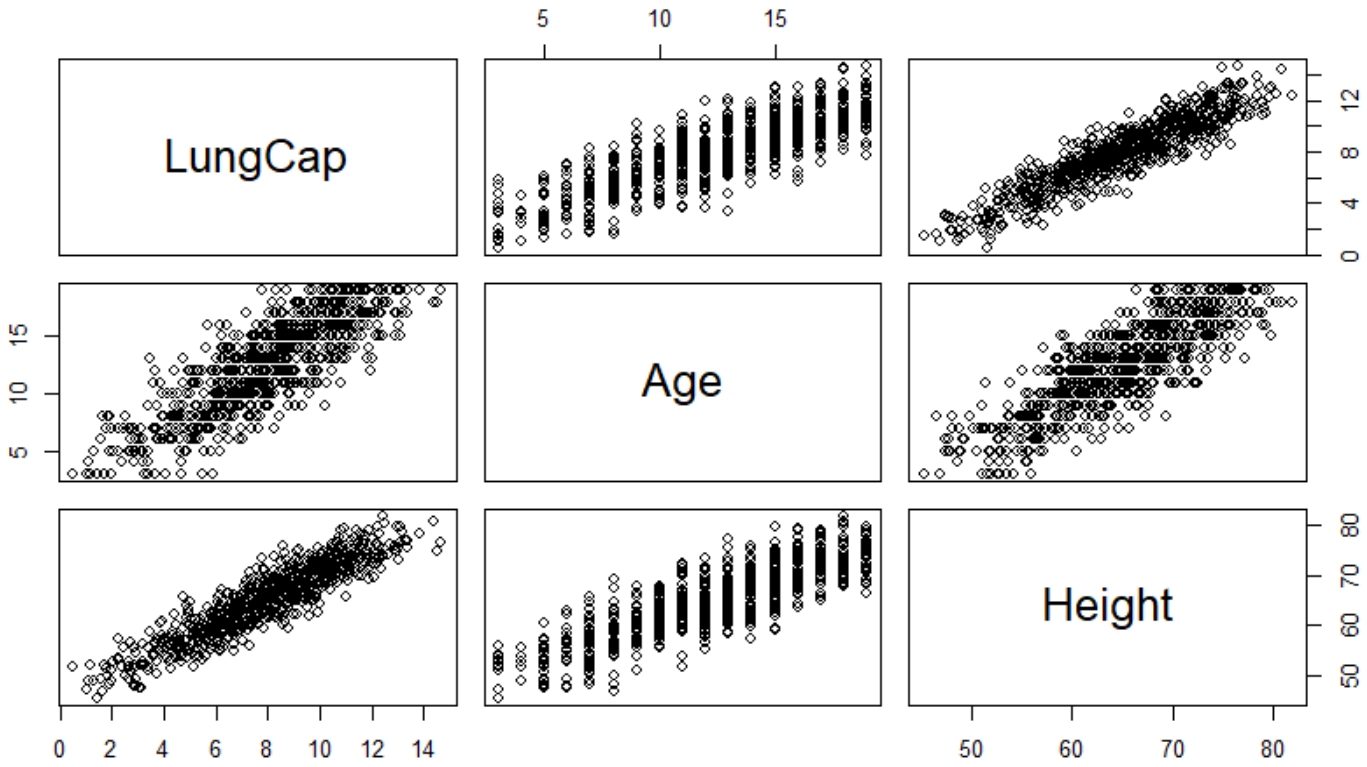
Q7) the variables in this dataset are:

- 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)

```
> cor(LungCap$Age, LungCap$LungCap)
[1] 0.8196749
> cor(LungCap$Height, LungCap$LungCap)
[1] 0.9121873
>
```

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)

```
#(Q13) Fit a liner regression model where the dependent variable is LungCap
#and use all other variables as the independent variables
model <- lm(LungCap ~ ., data=LungCap)
```

Q14)

```
> summary(model)

Call:
lm(formula = LungCap ~ ., data = LungCap)

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)

```
> #(Q15) What is the R-squared value here ? What does R-squared indicate?
> summary(model)$r.squared
[1] 0.8542478
>
```

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**:

- **R^2 = 0** means the independent variables explain none of the variance in the dependent variable. (bad fit)
- **R^2 = 1** means the independent variables explain all the variance in the dependent variable. (good fit)

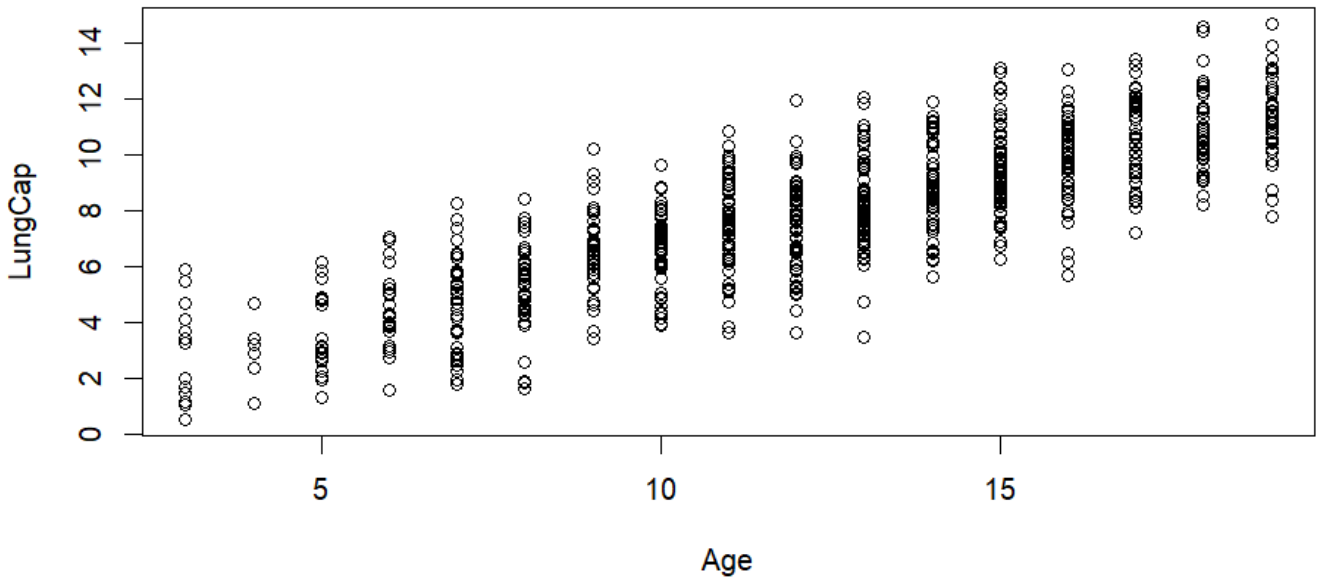
Q16)

```
> #(Q16) Show the coefficients of the linear model. Do they make sense?
> #If not, which variables don't make sense? What should you do?
> coef(model)
  (Intercept)           Age        Height     Smokeyes    Gendermale Caesareanyes
  -11.3224856     0.1605296     0.2641128    -0.6095592     0.3870117    -0.2142182
>
```

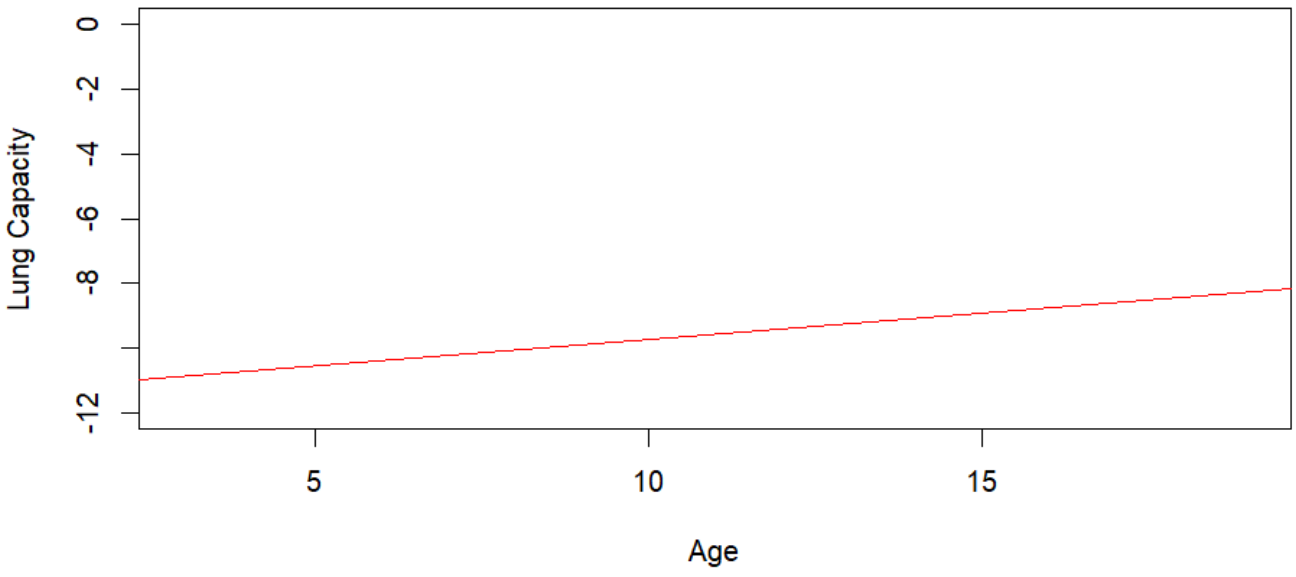
Yes, they make sense

- **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)

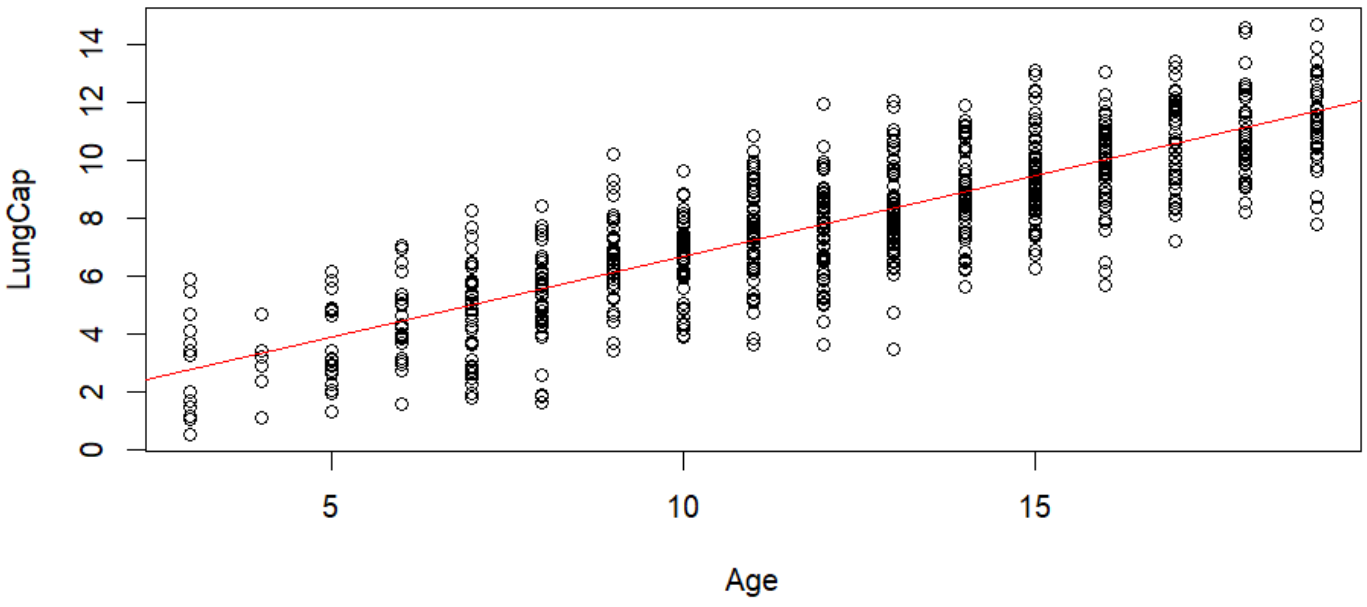
```
#(Q18)Repeat Q13 but with these variables Age, Smoke and Cesarean as the only indepe
model2 <- lm(LungCap ~ Age + Smoke + Caesarean, data = LungCap)
```

Q19)

```
> #(Q19)Repeat Q16, Q17 for the new model. What happened?
> coef(model2)
 (Intercept)           Age      Smokeyes Caesareanyes
   1.1086723     0.5561667    -0.6431029    -0.1460278
> plot(LungCap$Age, LungCap$LungCap, xlab="Age", ylab="LungCap")
> abline(model2, col="red")
Warning message:
In abline(model2, col = "red") :
  only using the first two of 4 regression coefficients
>
```

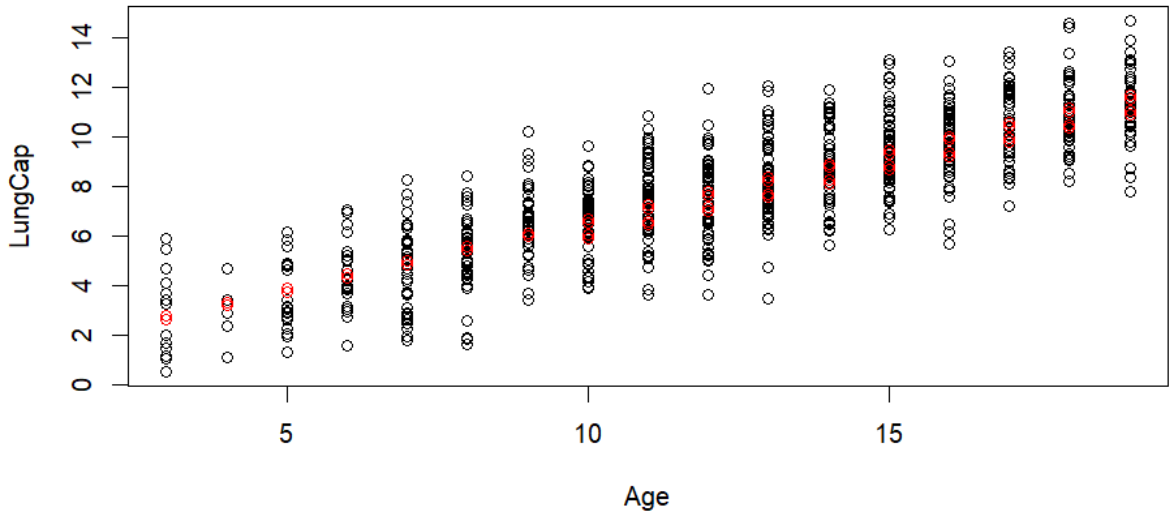
Yes, the coefficients make sense for the same reasons as Q16

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)

```
#(Q20)Predict results for this regression line on the training data.
ypred <- predict(model2)
ypred
#show the predicted values
plot(LungCap$Age, LungCap$LungCap, xlab="Age", ylab="LungCap")
points(LungCap$Age, ypred, col="red")
```



Q21)

For second model

```
> #(Q21)Calculate the mean squared error (MSE)of the training data.
> MSE = mean((LungCap$LungCap - ypred)^2)
> MSE
[1] 2.280169
>
```

For first model

```
> ypred <- predict(model)
> MSE = mean((LungCap$LungCap - ypred)^2)
> MSE
[1] 1.031418
>
```