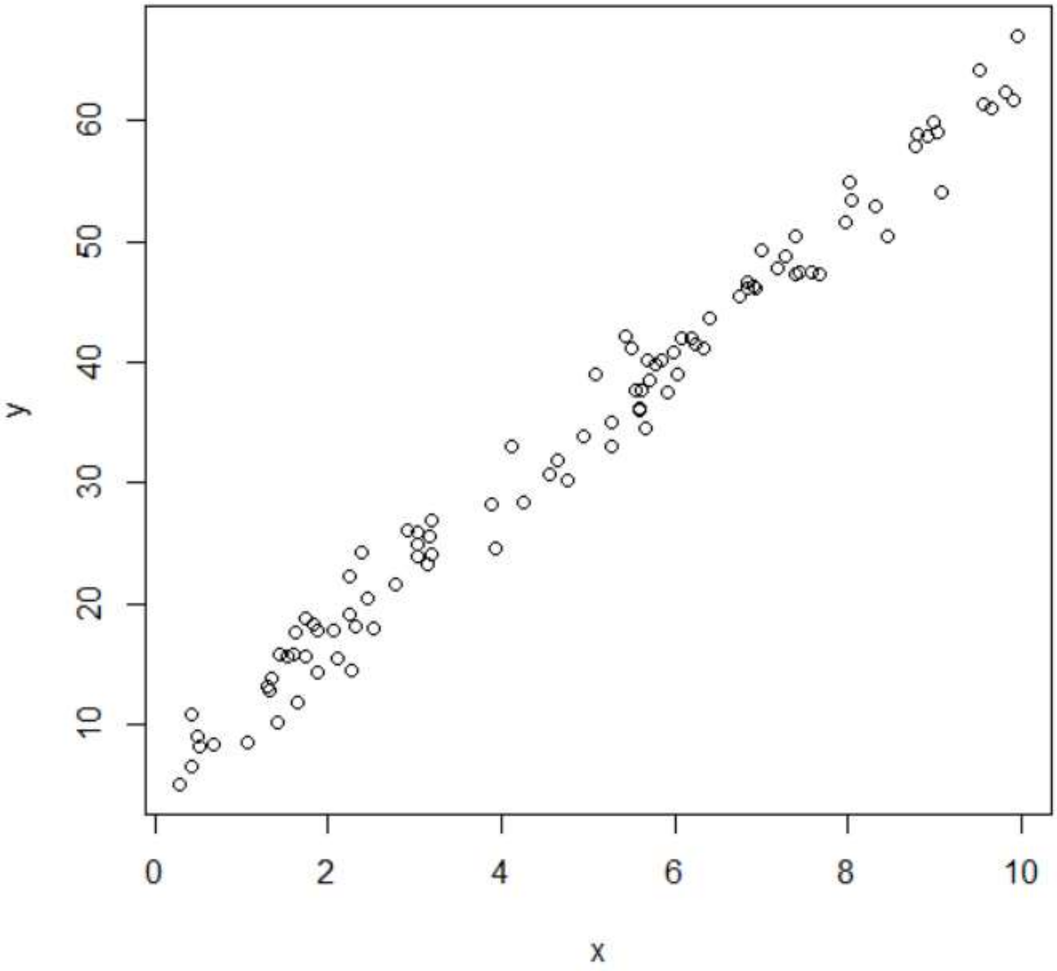
Lab5

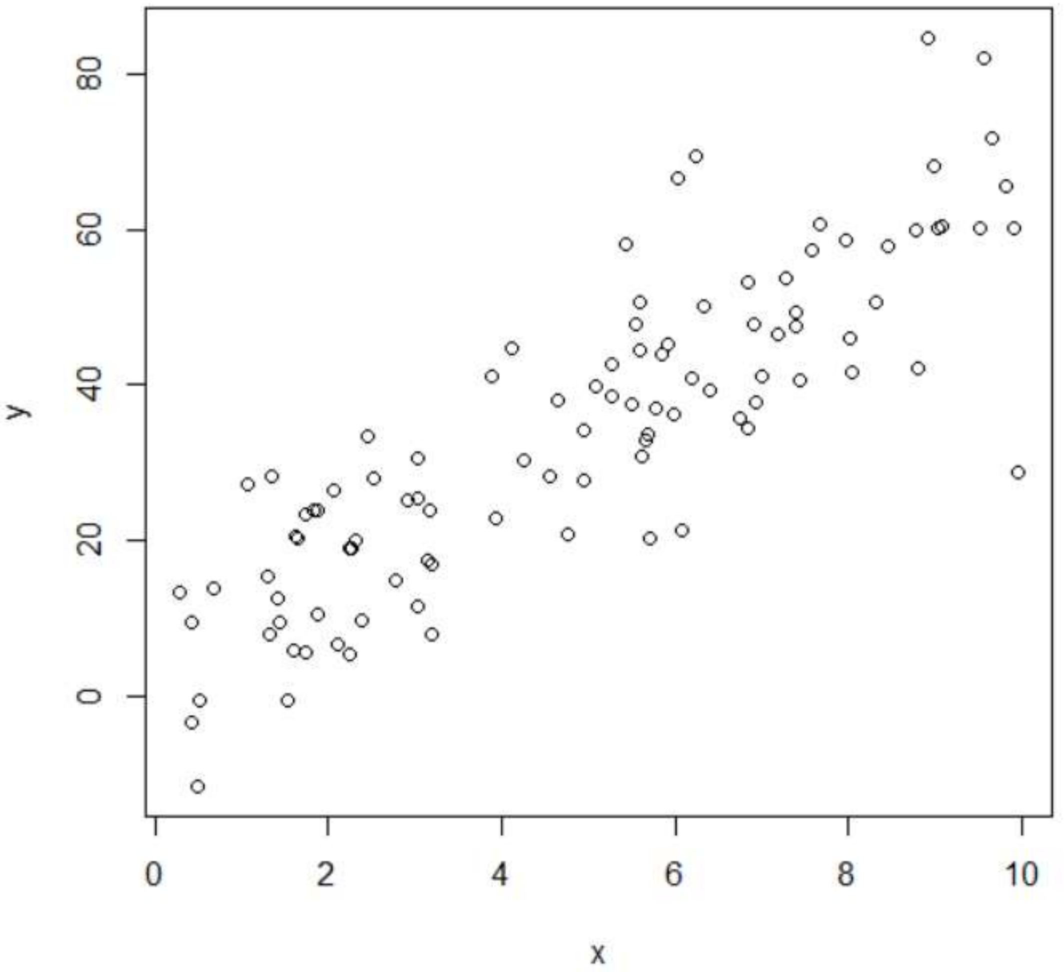
1.Try changing the value of standard deviation (sd). How do the data points change for different values of standard deviation?

As the value of sd increases the data points gets more scattered. “Increase the noise”

sd = 2



sd = 10



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

The coefficients of the linear model are changed by changing the value of standard deviation.

The model is trying to get a line that fits the data and minimizing the residuals.

sd = 2

```
Coefficients:
(Intercept)            x
      4.428        6.112
```

sd = 10

```
Coefficients:
(Intercept)            x
      3.271        6.310
```

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

As the value of sd increases the R-squared value decreases.

R-squared value measures how much the data points are scattered around the line.

100% -> data points perfectly on line.

0% -> data points aren’t correlated with the line. “linear regression isn’t suitable in this case”

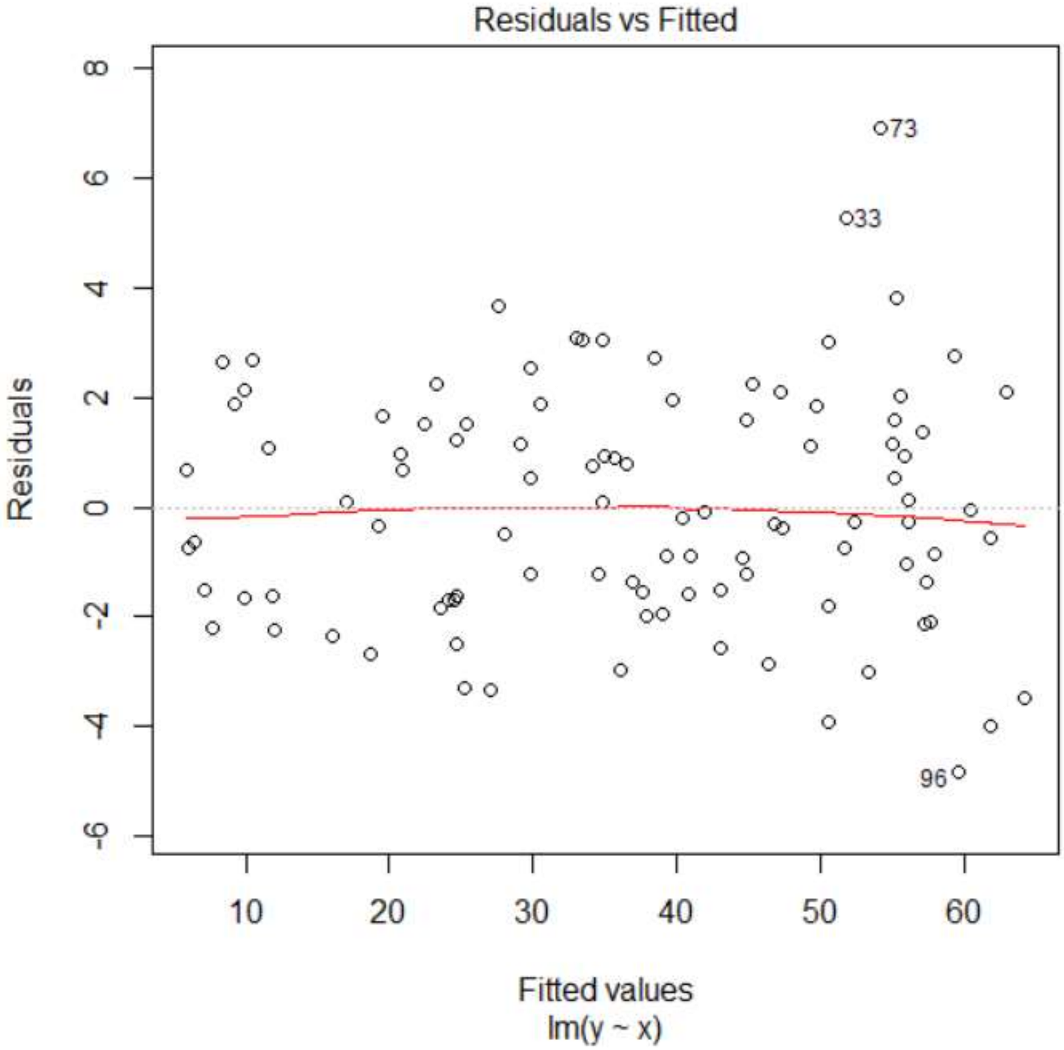
```
sd=2 : Adjusted R-squared:  0.9833

sd=10 : Adjusted R-squared:  0.7543
```

4. What do you conclude about the residual plot? Is it a good residual plot?

Plots are randomly placed with no pattern around the x-axis.

Yes, it is a good residual plot. There is no pattern in the ploting.

So, we conclude that linear model is appropriate here.



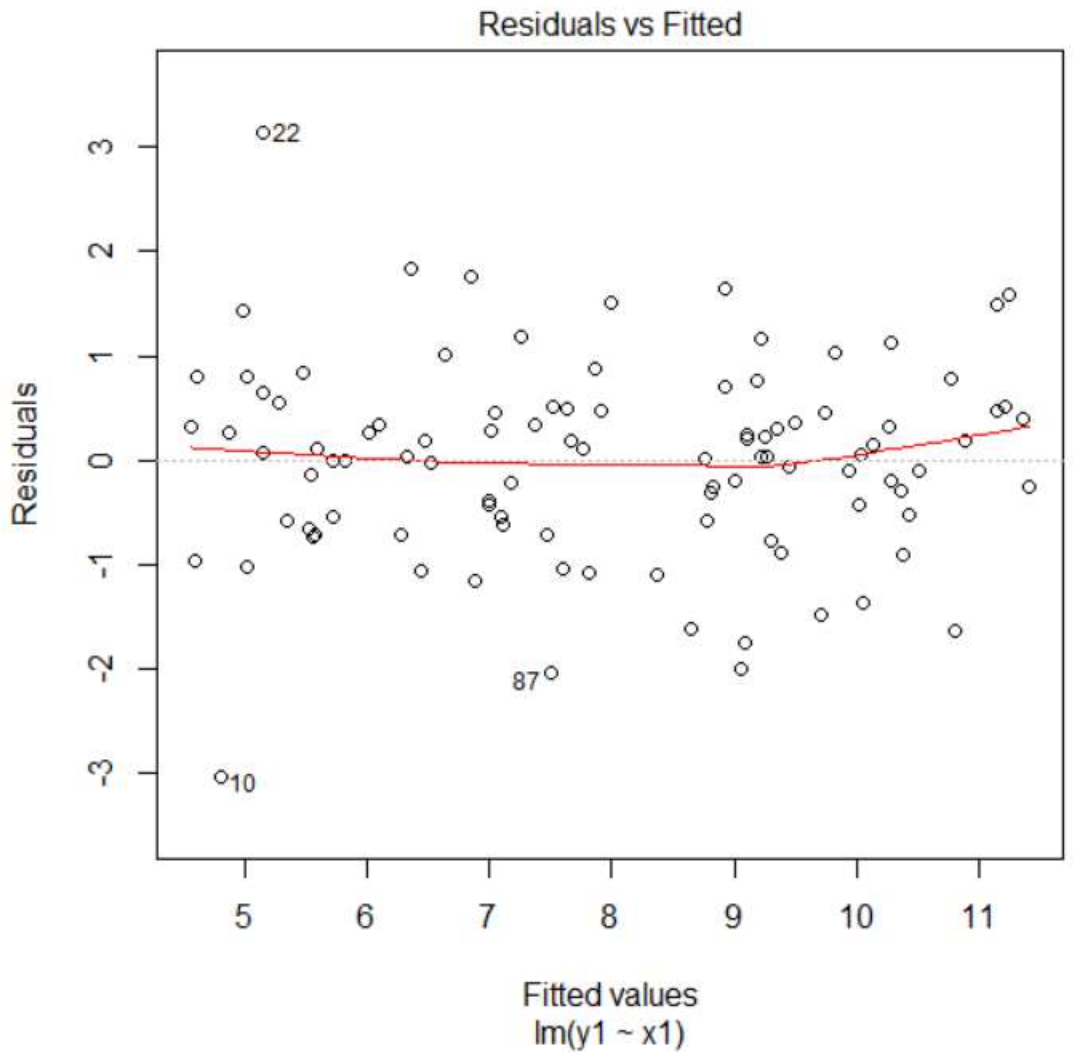
Part (2):

5. What do you conclude about the residual plot? Is it a good residual plot?

Plots are still randomly placed with no pattern around the x-axis. ”actually with a slightly pattern in it but I think it isn’t significant so we can neglect it”

Yes, it is a good residual plot. There is no pattern in the plotting.

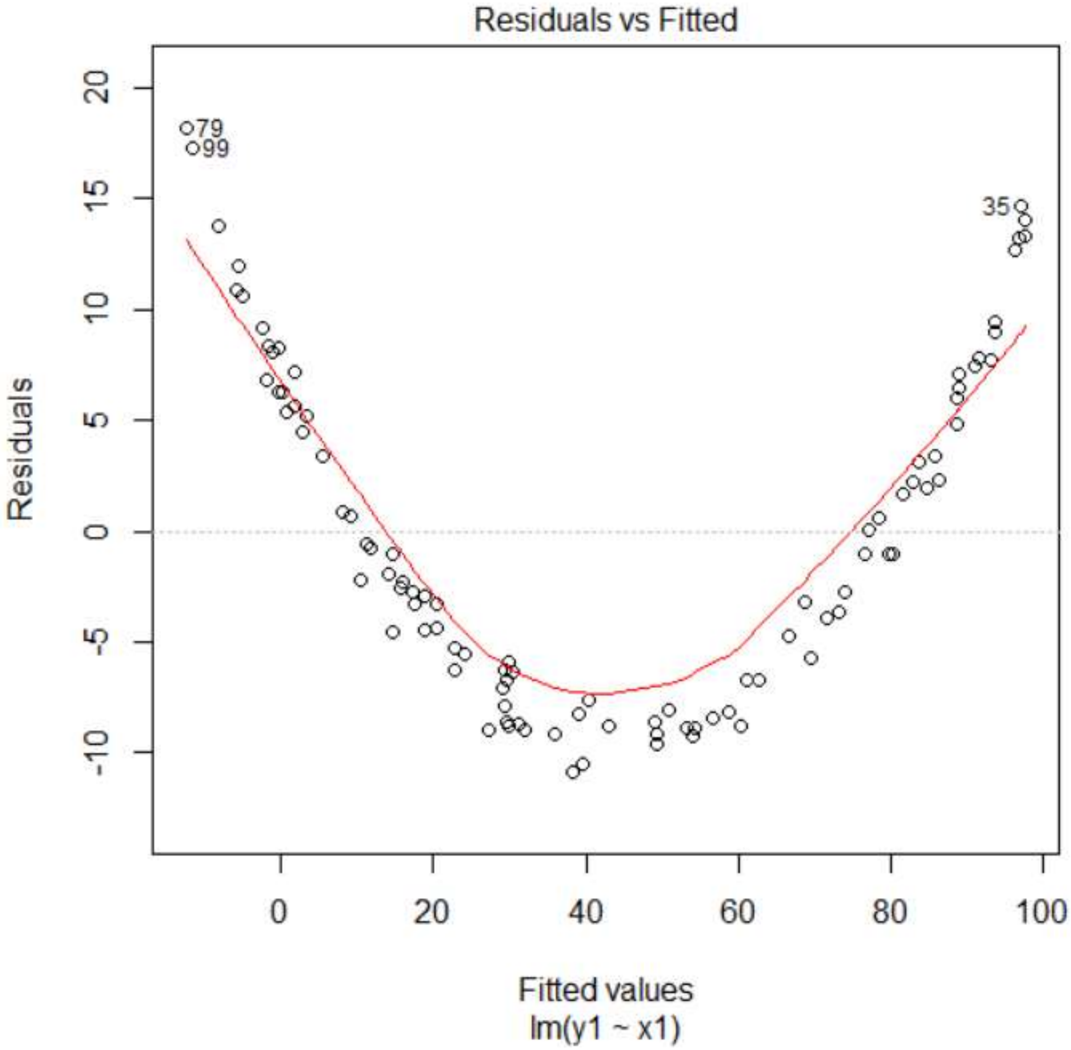
So, we conclude that linear model is still appropriate here.



6. Now, change the coefficient of the non-linear term in the original model for (A) training and (B) testing to a large value instead. What do you notice about the residual plot?

By increasing the value of the non-linear term the residual plot became a problematic (not good) plot.

There is a quadratic pattern in the plot which gives an indication that the linear model isn’t appropriate here and we should choose another quadratic model.



Part (3)

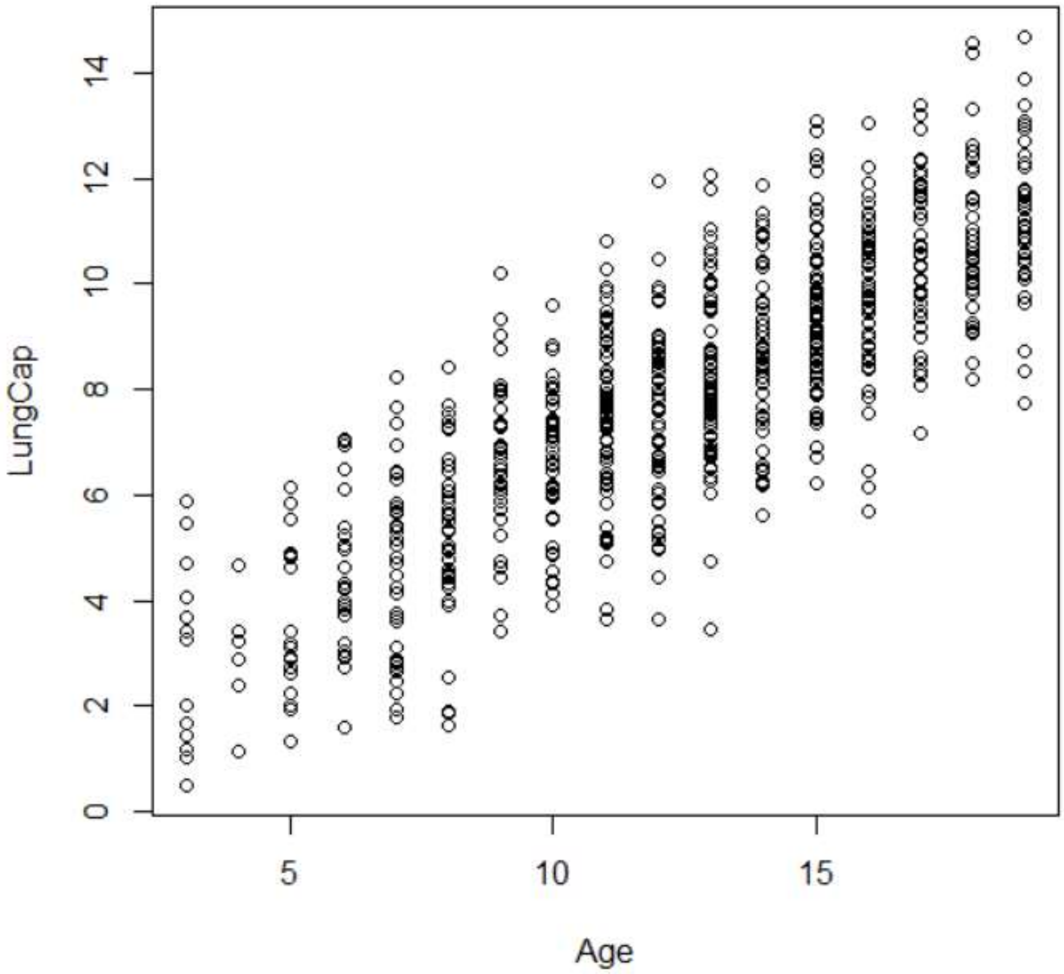
7. Import the dataset LungCapData.tsv. What are the variables in this dataset?

```
dfm <- read.csv("LungCapData.tsv", sep = "\t")
```

Variables: LungCap Age Height Smoke Gender Caesarean

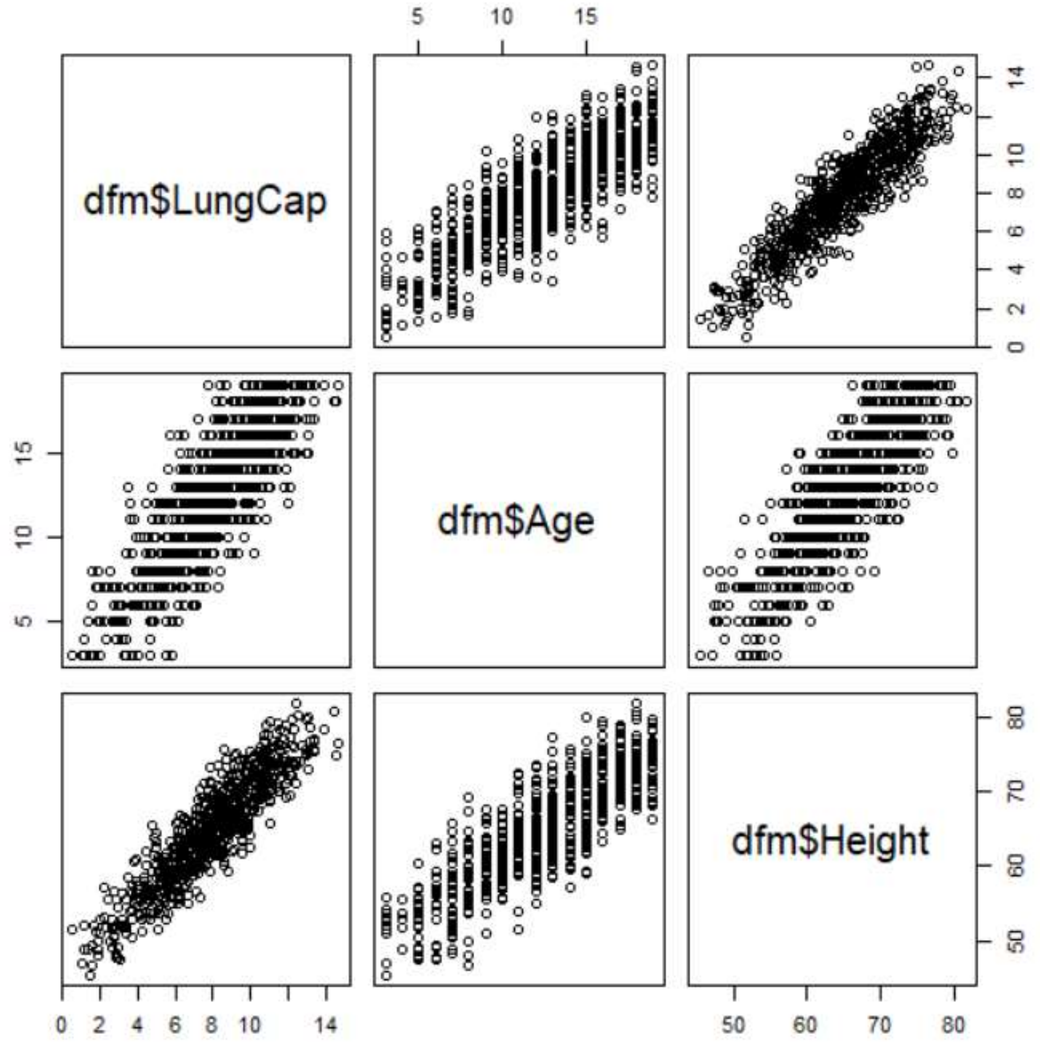
8. Draw a scatter plot of Age (x-axis) vs. LungCap (y-axis). Label x-axis "Age" and y-axis "LungCap"

```
par(mfrow=c(1,1)) # parameters for the next plot
plot(dfm$Age, dfm$LungCap, type="p", xlab="Age", ylab="LungCap")
```



9.Draw a pair-wise scatter plot between Lung Capacity, Age and Height.
Hint: Check the tutorial slides for how to plot a pair-wise scatterplot

```
pairs(dfm$LungCap~dfm$Age+dfm$Height)
```



10. Calculate the correlation between Age and LungCap, and between Height and LungCap. Hint: You can use the function cor.

```
cor(dfm$Age, dfm$LungCap)
0.8196749

cor(dfm$Height, dfm$LungCap)
0.9121873
```

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

correlation between Height and lungCap gave larger value. SO, Height is more correlated to LungCap

12. Do you think the two variables Height and LungCap are correlated? Why?

Yes

It gave the value of correlation is 0.912 which is a high value.

From the plotting, as the height increases the Lungcap increases.

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

```
model2 <-
lm(dfm$LungCap~dfm$Age+dfm$Height+dfm$Smoke+dfm$Gender+dfm$Caesarean)
```

14. Show a summary of this model.

```
summary(model2)
```

15. What is the R-squared value of this model? What does R-squared indicate?

0.842984

R-squared value measures how much the data points are scattered around the line.

100% -> data points perfectly on line.

0% -> data points aren’t correlated with the line. “linear regression isn’t suitable in this case”

So, 0.842984 indicates that linear regression is suitable here.

16. Show the coefficients of the linear model. Do they make sense? If not, which variables don't make sense? What should you do?

```
cat("coefficients ", d2$coefficients[,1],  "and an R-sqr of ", d2$r.squared, "\n")
coefficients -11.32249 0.1605296 0.2641128 -0.6095592 0.3870117 -0.2142182
and an R-sqr of  0.8542478
```

No, it doesn’t make sense

the coefficents are so small while the correlation between Age/Height and LungCap is large.

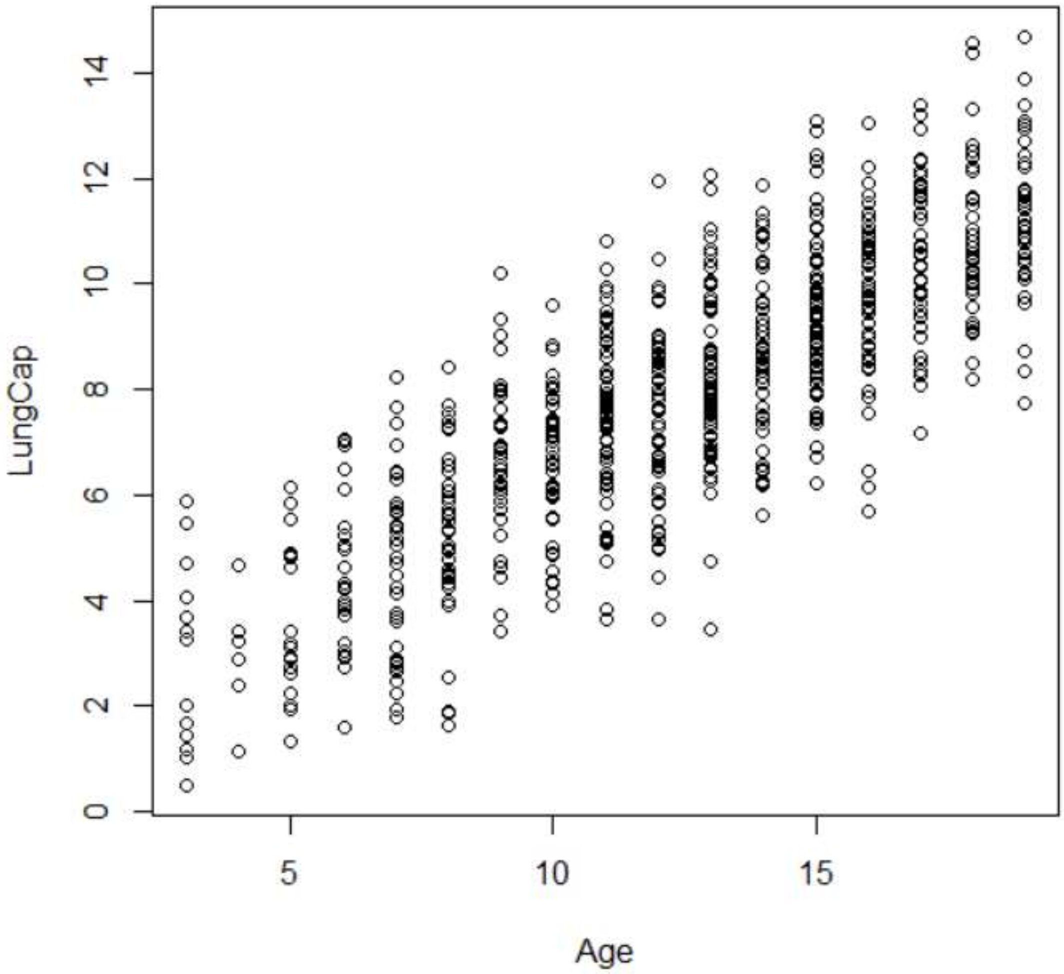
we may eliminate/reselect variables because the correlation between Height and Age is large.

cor(dfmAge, dfmHeight) = 0.8357368

17. Redraw a scatter plot between Age and LungCap. Display/Overlay the linear model (a line) over it. Hint: Use the function abline(model, col="red"). Note (1): A warning will be displayed that this function will display only the first two coefficients in the model. It's OK. Note (2): If you are working correctly, the line will not be displayed on the plot. Why?

There is no line

Because the coefficient of Age = 0.1605296 which is so small and the intercept is -11.32249 so the line intersect X axis at a value greater than the scale of the graph.



18. Repeat Q13 but with these variables Age, Smoke and Cesarean as the only independent variables.

```
model3 <- lm(dfm$LungCap~dfm$Age+dfm$Smoke+dfm$Caesarean)
d3 <- summary(model3)
```

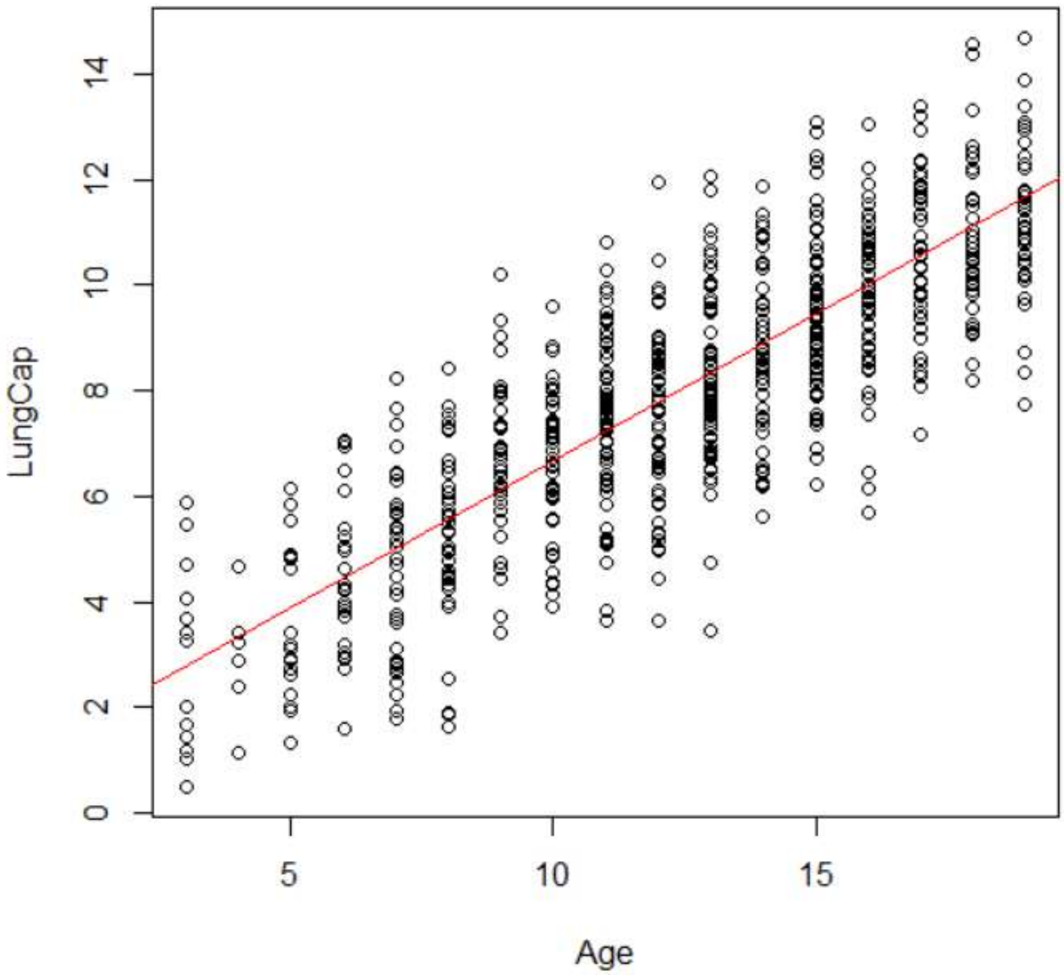
19. Repeat Q16, Q17 for the new model. What happened?

coefficients: 1.108672 0.5561667 -0.6431029 -0.1460278

and an R-sqr of 0.6777835

The coeffiecients make more sense now.

The line is shown.



20. Predict results for this regression line on the training data.

```
pred <- predict(model3)
```

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

```
MSE <- mean(model3$residuals^2)
MSE
```

```
2.280169
```