Linear Regression Report

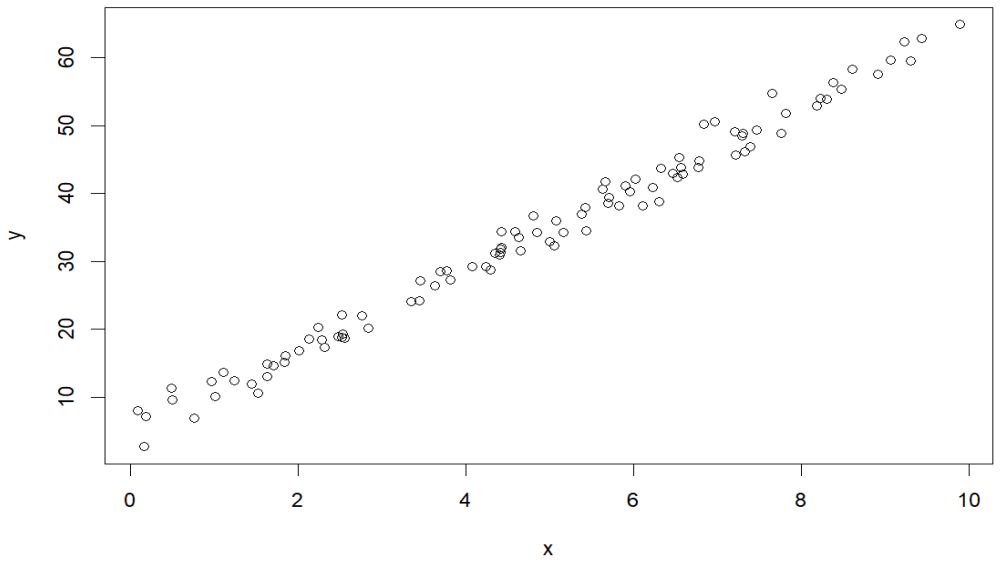
Name	Section	B.N	ID
Peter Atef	1	18	9202395
Beshoy Morad	1	19	9202405

Part (1):

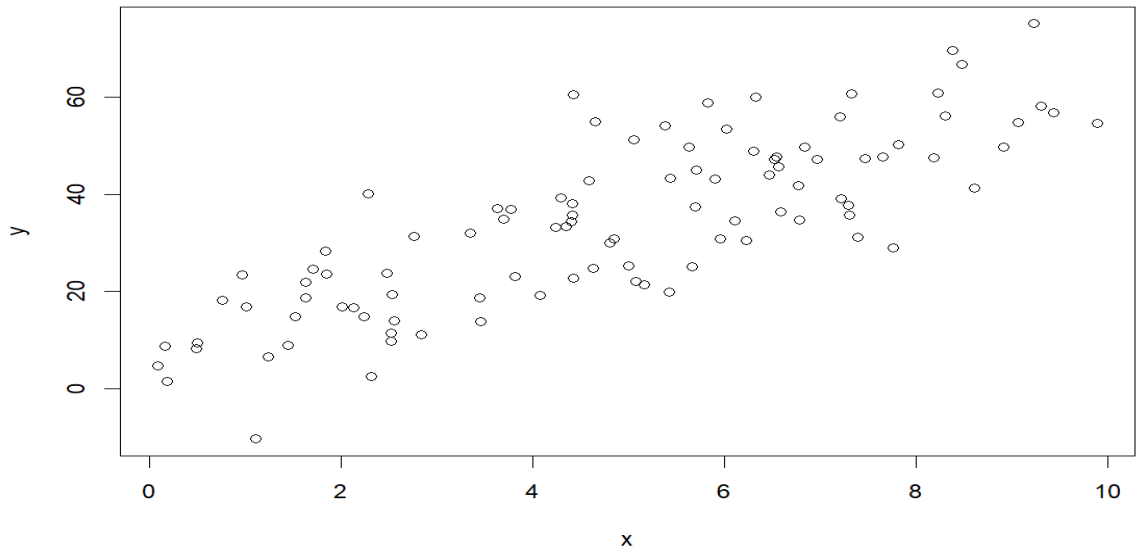
1. **Try changing the value of standard deviation (std). How do the data points change for different values of standard deviation?**

Ans:

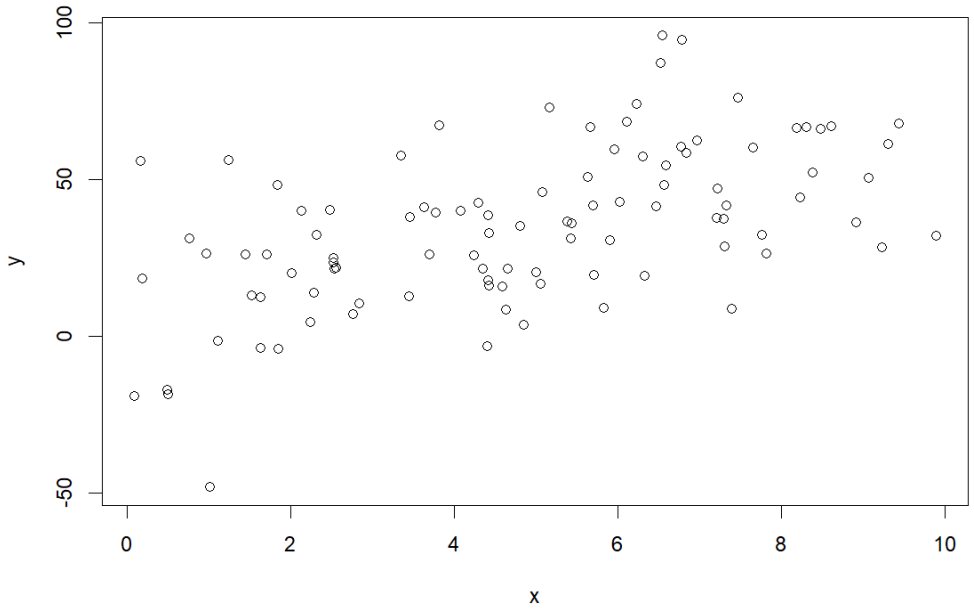
Std = 2



Std = 10



Std = 20



It’s noticed that by increasing the standard deviation, the data points are more scattered.

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

   Ans:

   For Std = 2:

   Y = 5.998 * X + 4.543

   For Std = 10:

   Y = 5.952 * X + 4.268

   For Std = 20:

   Y = 5.233 * X + 8.147

   Conclusion:

   Increasing the standard deviation, resulting in getting far away from the original coefficients of the equation.

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

   Ans:

   Std = 2 -> R^2 = 0.9827074

   Std = 10 -> R^2 = 0.7364146

   Std = 20 -> R^2 = 0.2583542

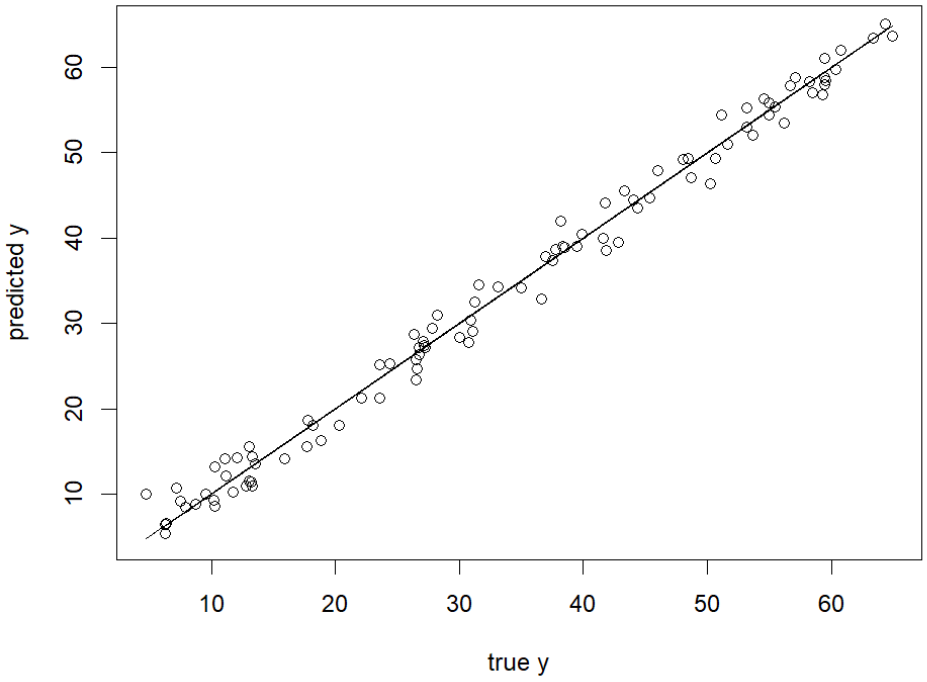
   Conclusion:

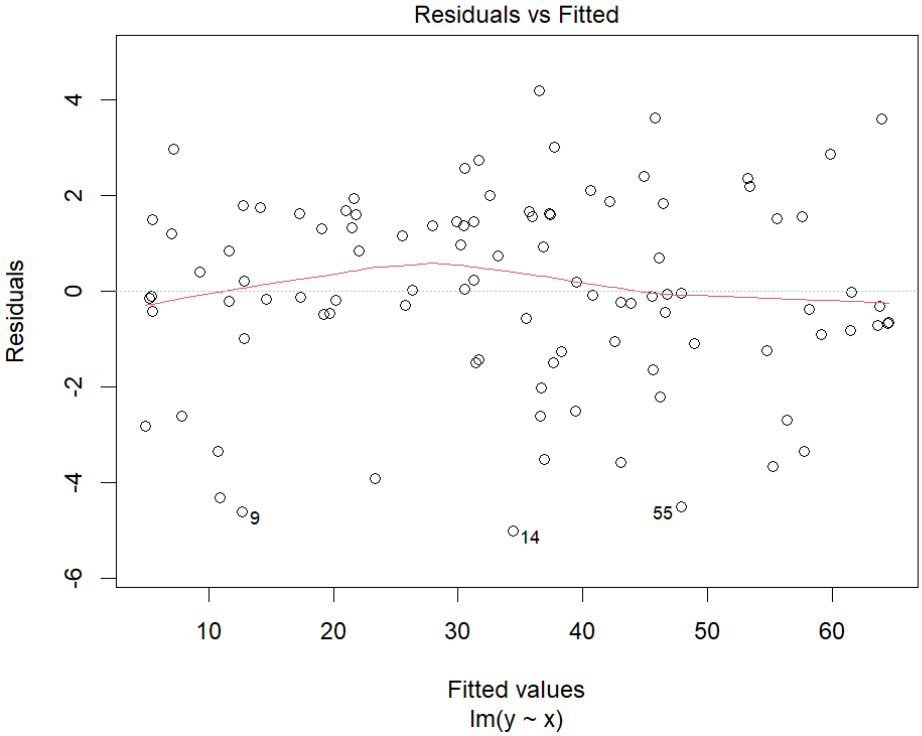
   Increasing the standard deviation, resulting in getting a worse R^2 is far from 1 (near zero). That means more errors in predictions.

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

   Ans:

   For std = 2:


predicted y
true y
10
20
30
40
50
60


Residuals vs Fitted
Residuals
Fitted values
lm(y ~ x)
9
14
55
-6
-4
-2
0
2
4
10
20
30
40
50
60

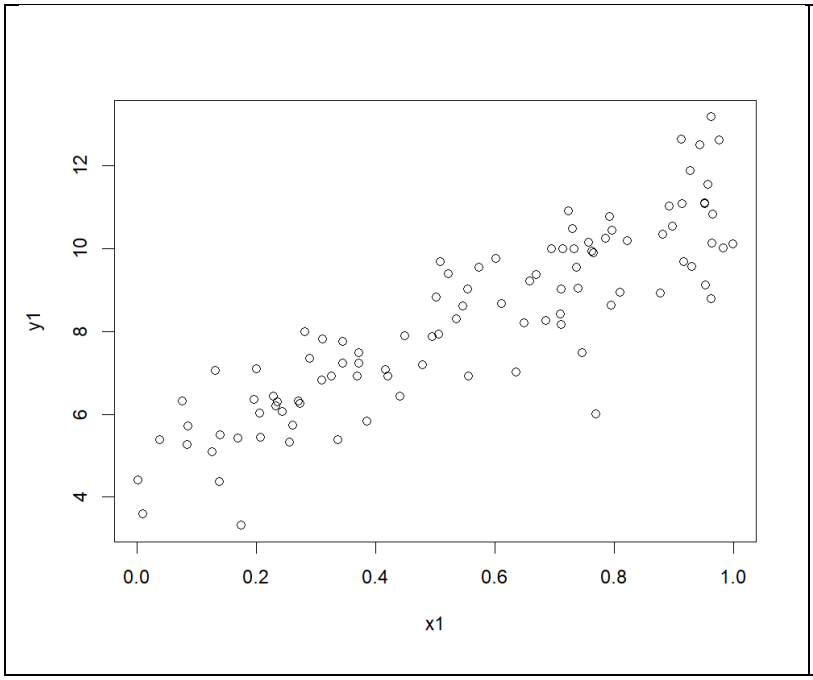
There is no pattern in the residual plot, so the linear regression assumption is still valid. So, it's a good residual plot and we can conclude that the linear regression model is appropriate as the points are randomly dispersed around the horizontal axis. Also, there are no patterns like U-shapes or curves (non-linear shapes)

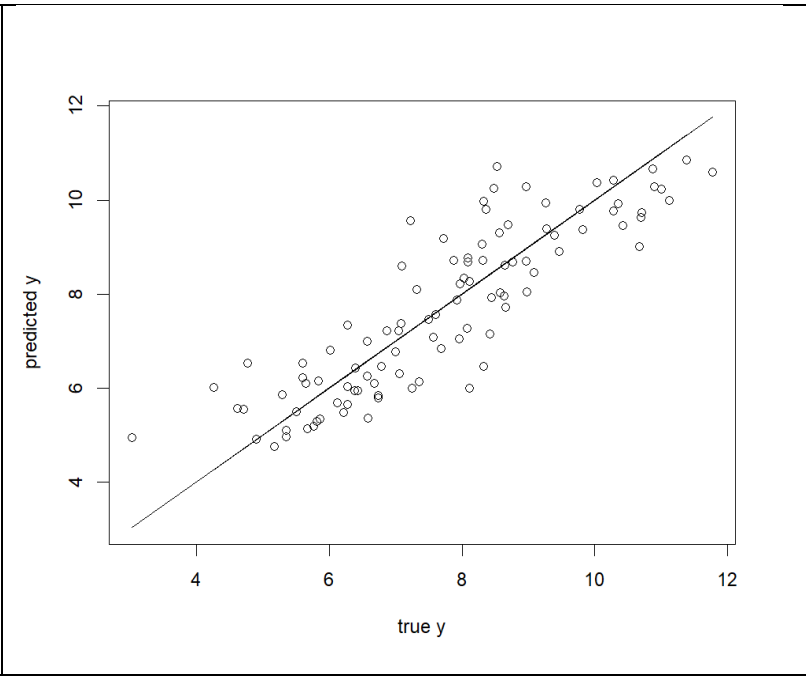
Part (2):

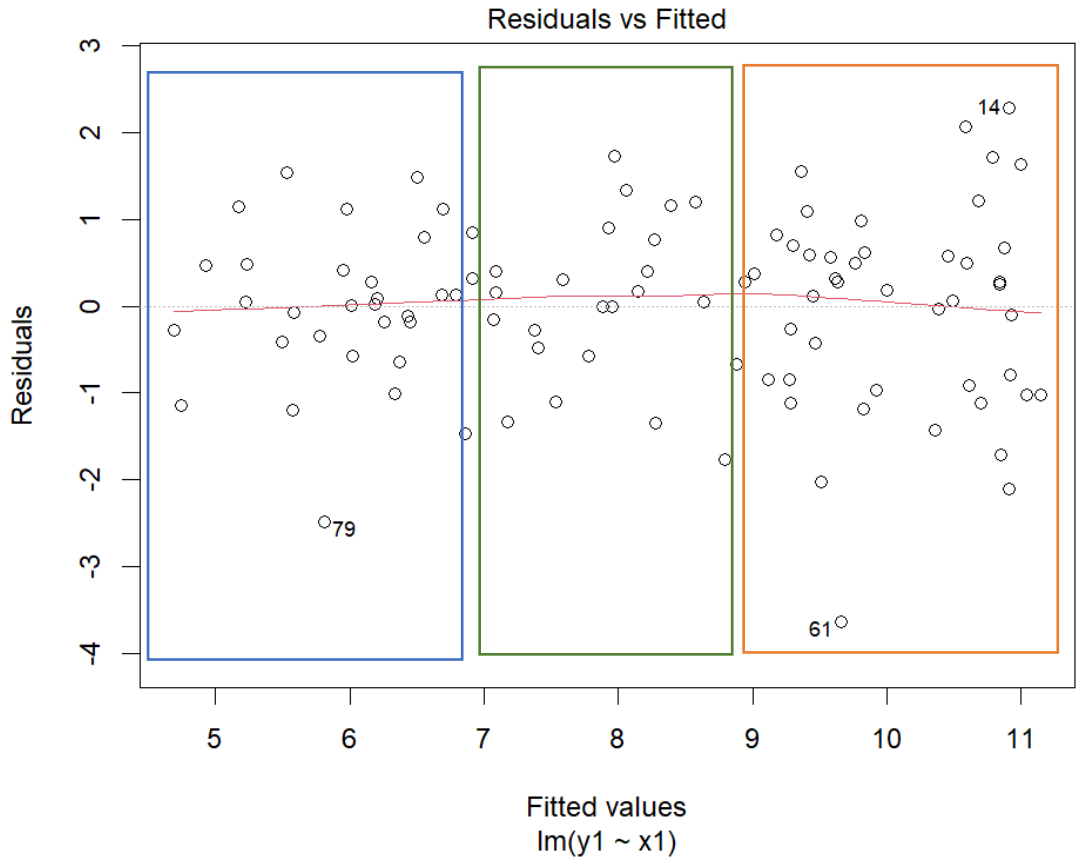
5. **What do you conclude about the residual plot? Is it a good residual plot?**

   Ans:

   That isn't a good residual plot because we can notice there is a pattern which is there are high residual values at the beginning of the x-axis and the residual values decrease then increase again so, it's non-linear. Also, there are a few data points with large positive or negative residuals which are considered outliers.

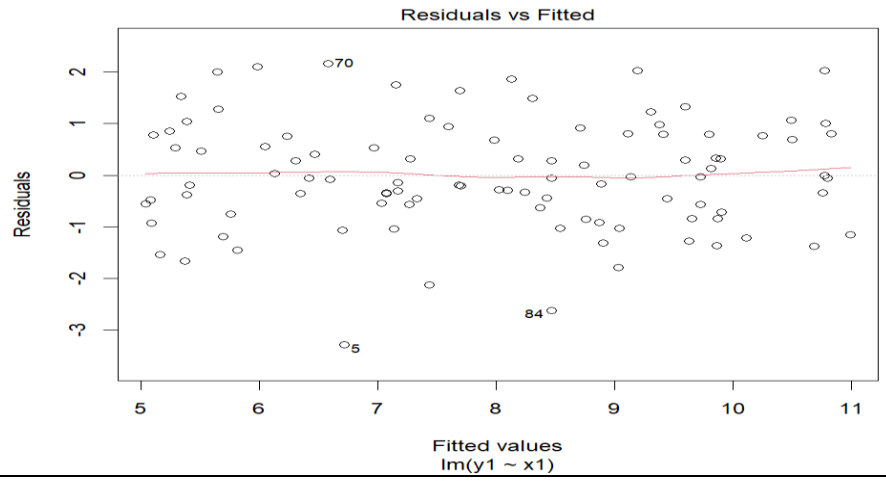




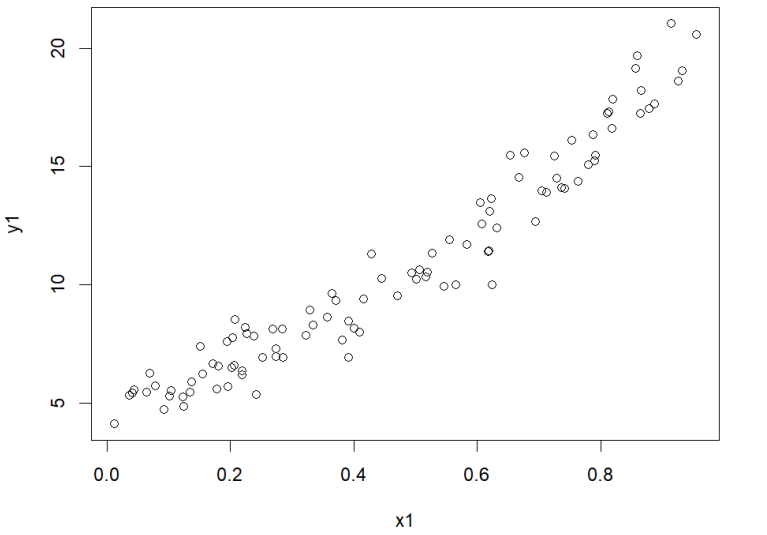


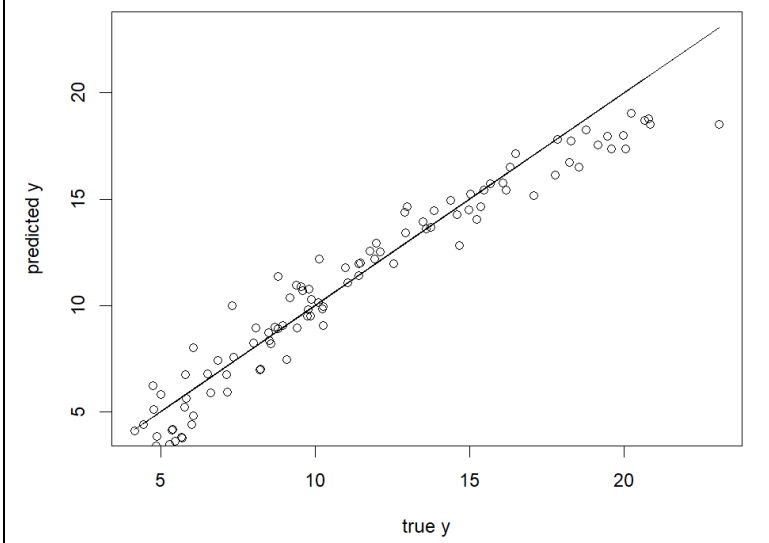
6. **Now, change the coefficient of the non-linear term in the original model for (A) training and (B) testing to a large value instead. What do you notice about the residual plot?**

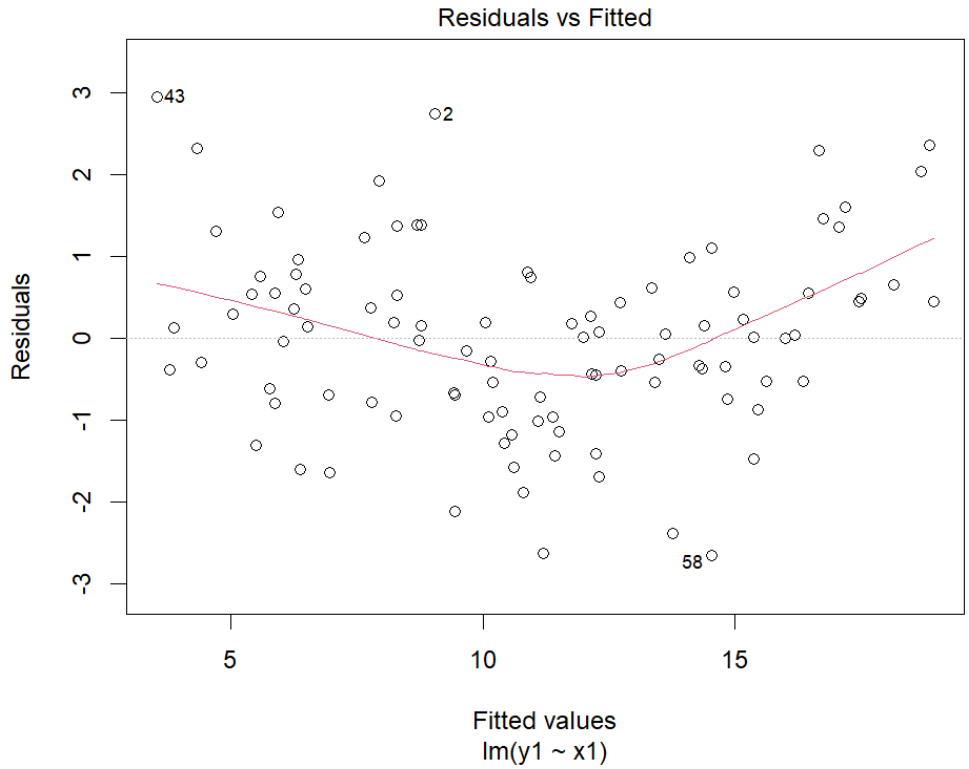
   Ans:

For non-linear term's coefficient = 0.1	

For non-linear term’s coefficient = 10







It's noticed:

- Non-Constant Variance: The spread of residuals appears to change as we move along the x-axis. Initially, the residuals are more tightly clustered around zero, but as the x-values increase, the spread becomes wider.
- More non-linearity as the quadratic term becomes dominant.

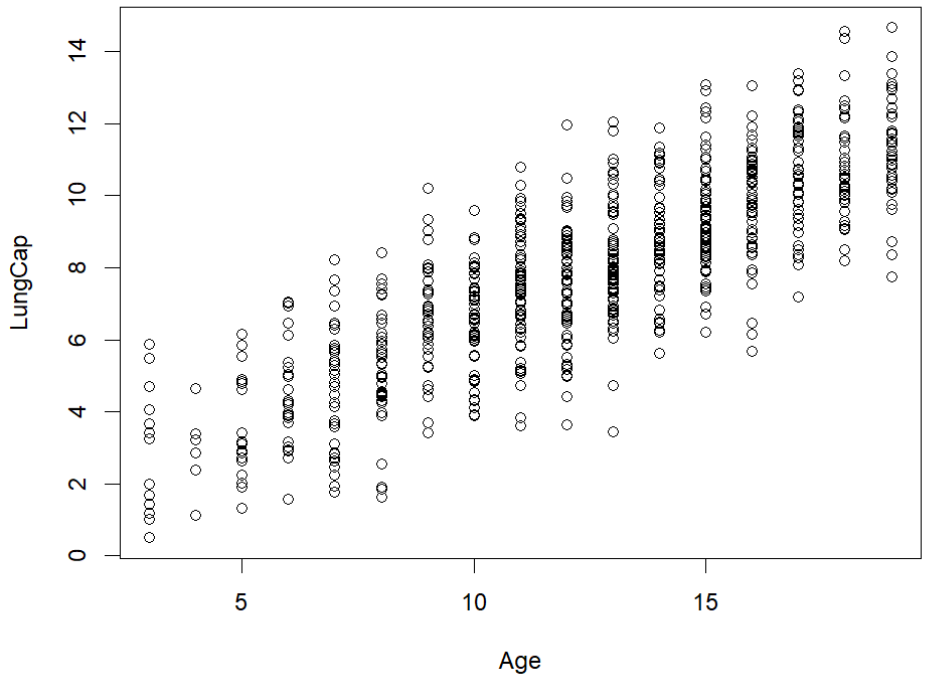
Part (3):

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

```
LungCapData <- read.table("LungCapData.tsv", header = TRUE, sep = "\t")
```

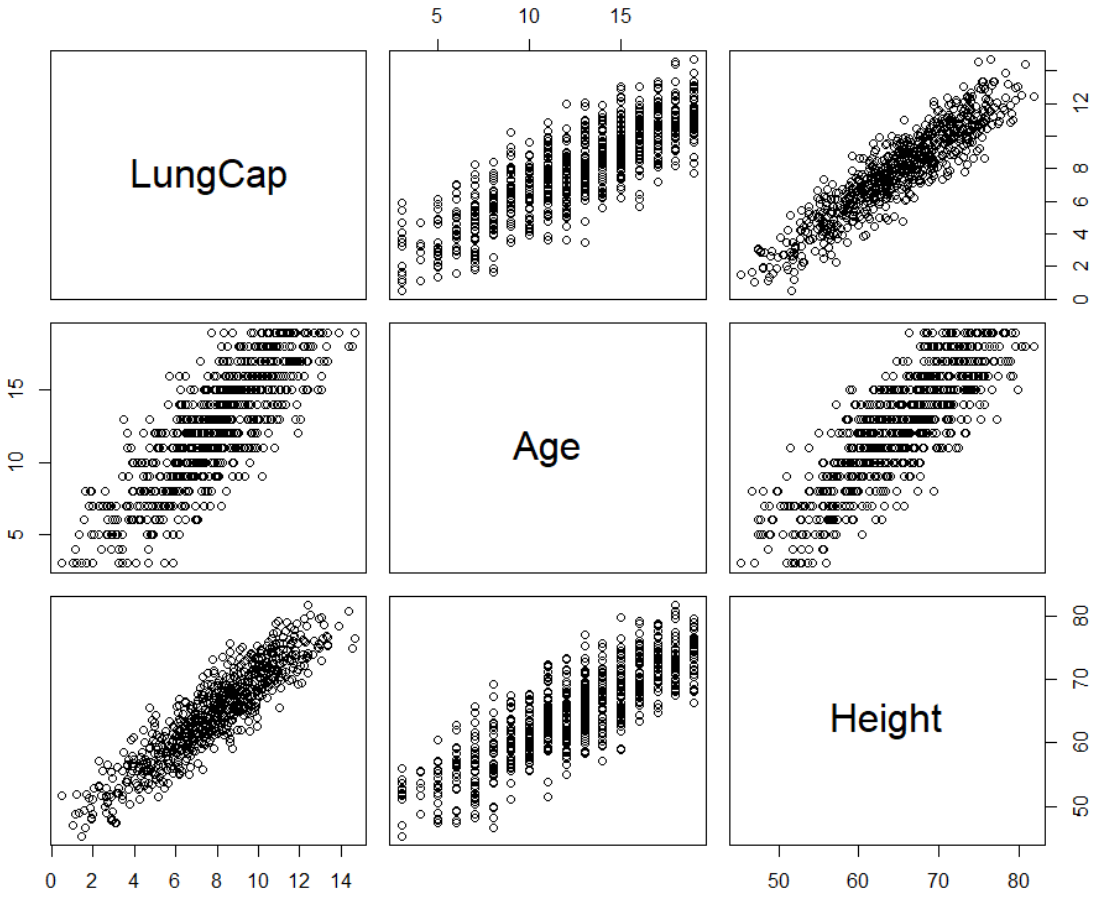
8. Draw a scatter plot of Age (x-axis) vs. LungCap (y-axis). Label x-axis "Age" and y-axis "LungCap"

```
plot(LungCapData$Age, LungCapData$LungCap, 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(LungCapData[, c("LungCap", "Age", "Height")])
```



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

```
age_lung <- cor(LungCapData$Age, LungCapData$LungCap)
# 0.8196749
```

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

```
height_lung <- cor(LungCapData$Height, LungCapData$LungCap)
# 0.9121873
Height is more correlated to LungCap
```

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

```
# The correlation between Height and LungCap is 0.9121873 which is a
high correlation.
# This means that the two variables are correlated.
```

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

```
model <- lm(LungCap ~ ., data = LungCapData)
```

14. Show a summary of this model.

```
d2 <- summary(model)
d2
```

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

```
cat("R-sqr = ", d2$r.squared, "\n")
# R-sqr is 0.8542478 which is a high value. This means that the model
is a good fit.
```

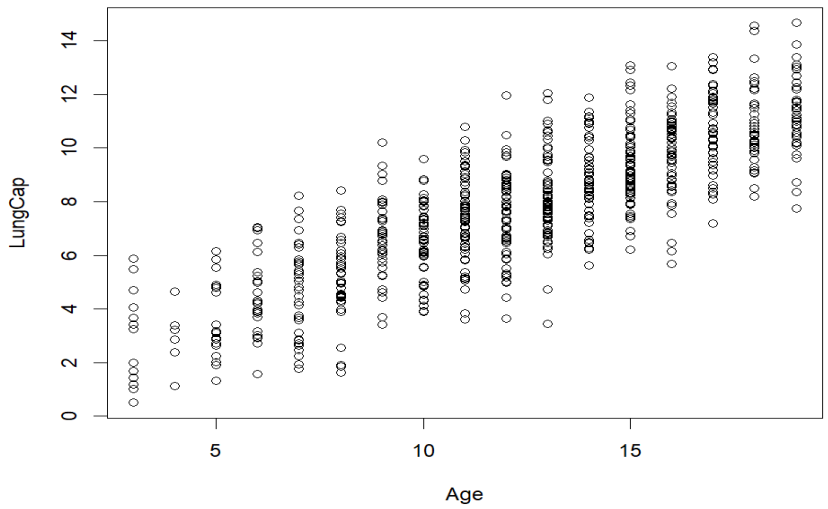
16. Show the coefficients of the linear model. Do they make sense? If not, which variables don't make sense to you? What should you do?

```
• cat("OLS gave slope of ", d2$coefficients)
• # Age:
• # The coefficient for age is approximately 0.16.
• # It suggests that, on average, for each one-unit increase in age,
  the response variable increases by 0.16 (assuming other predictors
  are constant).
• # This makes sense, as older individuals might have different
  outcomes compared to younger ones.
• # Height:
• # The coefficient for height is approximately 0.26.
• # It implies that, on average, for each one-unit increase in height,
  the response variable increases by 0.26 (assuming other predictors
  are constant).
• # I think it's larger than the coefficient for age because taller
  people might be younger and stronger and have higher lung capacity.
• # Smokeyes:
• # The coefficient for smokeyes is approximately -0.61.
• # It suggests that individuals with “smokeyes” (presumably related
  to smoking) have a lower response value.
```

```
# Negative coefficients make sense if we assume that smoking negatively impacts the outcome.
# Gendermale:
# The coefficient for gendermale is approximately 0.39.
# It indicates that being male is associated with a higher response value so they have higher lung capacity.
# This aligns with Caesareanyes because being a woman is associated with having Caesarean delivery and vice versa.
# Caesareanyes:
# The coefficient for caesareanyes is approximately -0.21.
# It suggests that individuals who had a Caesarean section have a lower response value.
# This makes sense if we consider the impact of Caesarean delivery on health outcomes so they have lower lung capacity.
```

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

```
plot(LungCapData$Age, LungCapData$LungCap, xlab = "Age", ylab = "LungCap")
abline(model, col = "red")
```



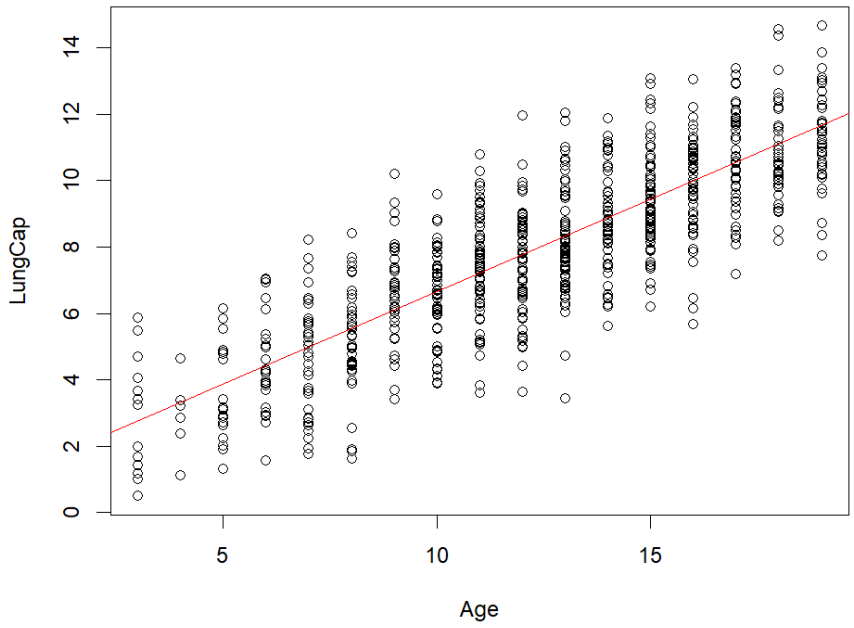
18. Repeat Q13 but with these variables Age, Smoke, and Cesarean as the only independent variables.

```
model <- lm(LungCap ~ Age + Smoke + Caesarean, data = LungCapData)
```

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

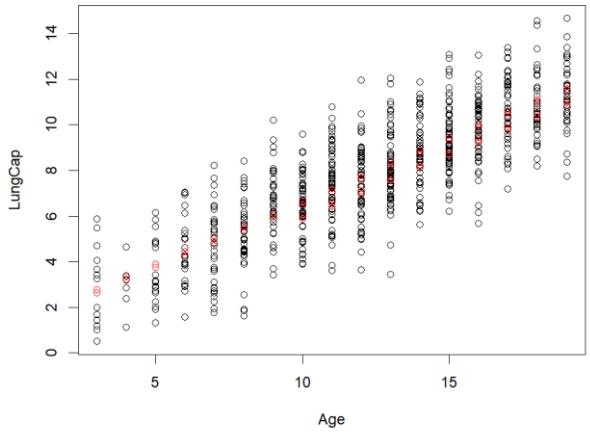
```
d3 <- summary(model)
d3

cat("R-sqr = ", d3$r.squared, "\n")
# It's noticed that we got a lower R-squared value than the previous model.
cat("OLS gave slope of ", d3$coefficients)
# Age:
# The coefficient for age is approximately 0.56.
# It suggests that, on average, for each one-unit increase in age, the response variable increases by 0.56 (assuming other predictors are constant).
# This makes sense, as older individuals might have different outcomes compared to younger ones.
# Smokeyes:
# The coefficient for smokeyes is approximately -0.64.
# It suggests that individuals with “smokeyes” (presumably related to smoking) have a lower response value.
# Negative coefficients make sense as smoking negatively impacts the outcome.
# Caesareanyes:
# The coefficient for caesareanyes is approximately -0.15.
# It suggests that individuals who had a Caesarean section have a lower response value.
# This makes sense if we consider the impact of Caesarean delivery on health outcomes.
# Also, the absolute value of Caesareanyes is less than the absolute value of Smokeyes which means that the impact of smoking is more than the impact of Caesarean delivery on lung capacity.
plot(LungCapData$Age, LungCapData$LungCap, xlab = "Age", ylab = "LungCap")
# now we can see the line because we have only two coefficients
abline(model, col = "red")
```



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

```
ypred <- predict(model)
# plot the predicted values
plot(LungCapData$Age, LungCapData$LungCap, xlab = "Age", ylab = "LungCap")
points(LungCapData$Age, ypred, col = "red")
```



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

```
mse <- mean((LungCapData$LungCap - ypred)^2)
mse #2.280169
```