Cairo University
Faculty of Engineering

Department of Computer
Engineering

Big Data – Lab 5 Requirement
Linear Regression

Team Info

Name	Code	Sec	BN
Abdelrahman Hamdy Ahmed	9202833	1	38
Abdelrahman Noaman Loqman	9202851	2	4

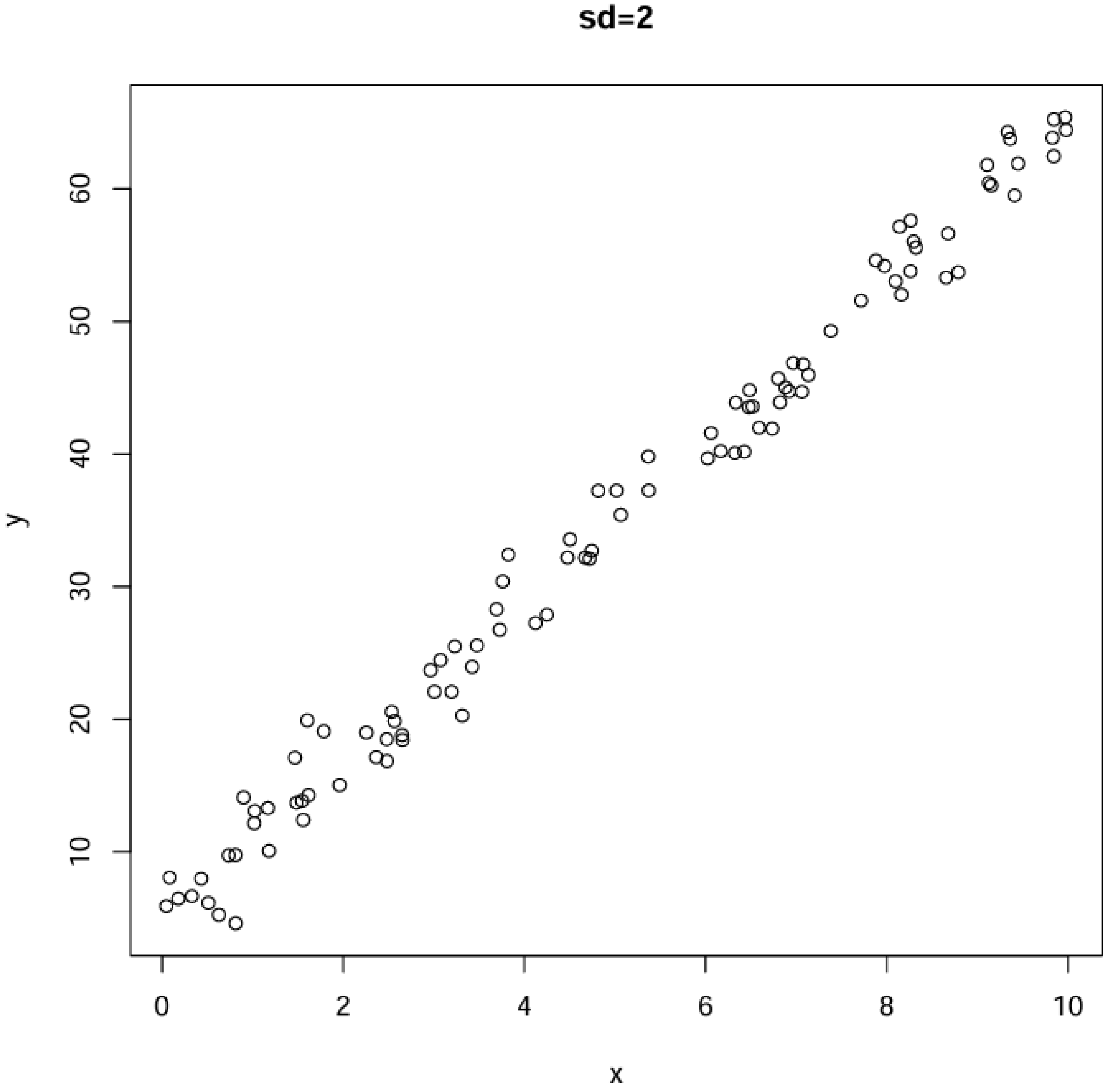
Date: 4/08/2024

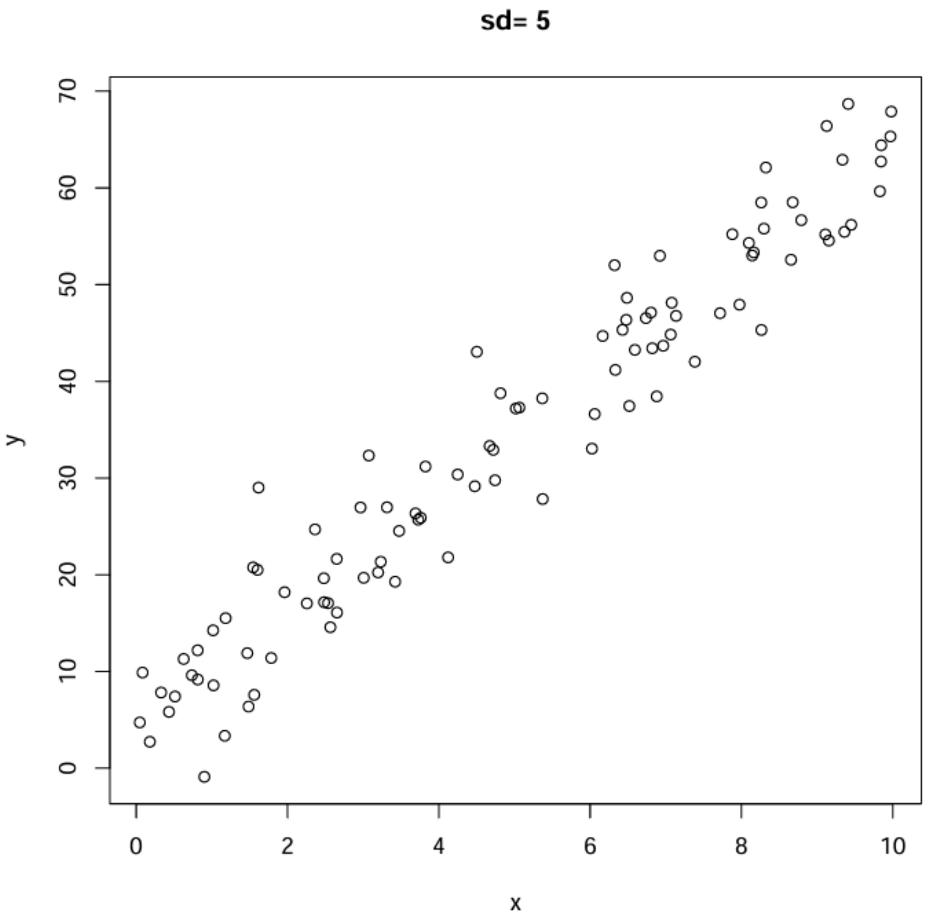
Presented to: Eng. Omar Samir

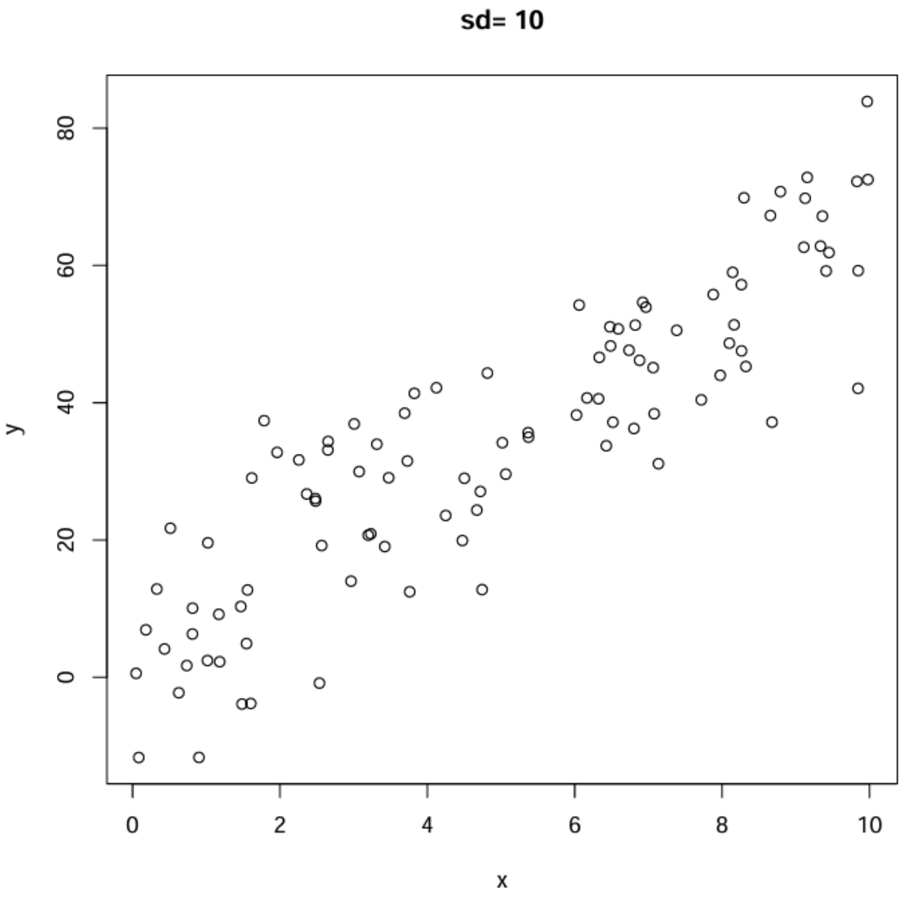
Answers to Theoretical Questions With Plots

Q1) **Try changing the value of standard deviation (SD). How do the data points change for different values of standard deviation?**

As the standard deviation increases, the data points become more dispersed, resulting in greater variability or "noise" within the dataset. This means that as sd grows, individual data points tend to deviate further from the average or mean, leading to a broader spread of values across the data set.




sd= 5


sd= 10

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

```
Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
      5.185        6.012

Call:
lm(formula = y2 ~ x)

Coefficients:
(Intercept)            x
      2.347        6.433
```

As we can clearly see when we change the value of standard deviation, the coefficients differ for both models. The first one is when sd was 2 & the second is sd = 10.

With higher noise, the model tries to adjust the coefficients where it can draw a line that fits the data with highest R-squared value and minimum residuals to ensure that the variance isn't too high.

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

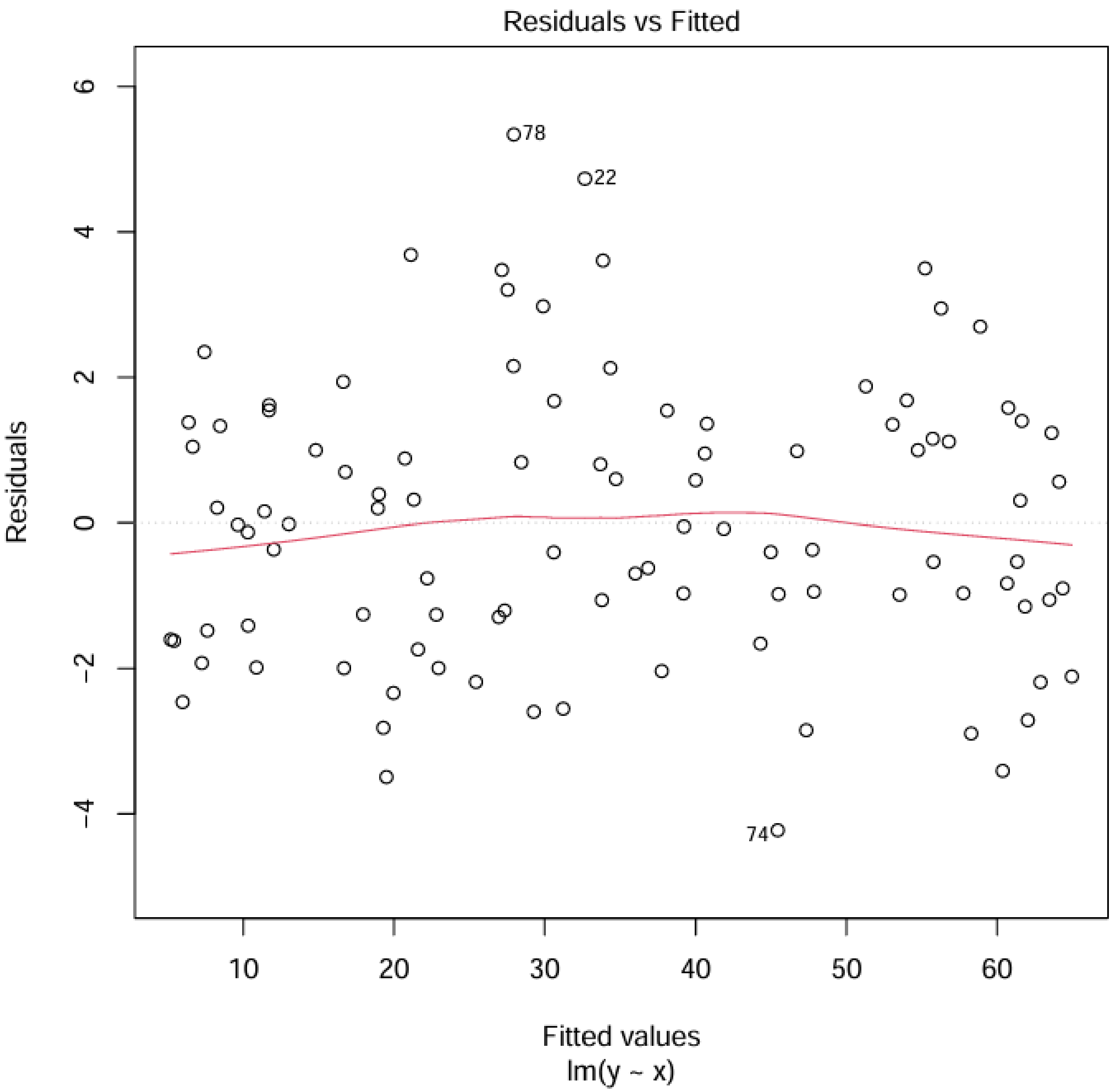
Model (sd = 2) OLS gave slope of 6.048241 and an R-sqr of 0.9860649
Model (sd = 10) OLS gave slope of 6.257811 and an R-sqr of 0.7108289

R-squared measures the proportion of the variance in the dependent variable that is explained by the independent variable(s) in a regression model.

When the standard deviation increases meaning that the data points are more scattered around the regression line, the R-squared value decreases because a larger portion of the total variance in the dependent is not accounted for by the model.

A very high R-squared percentage approaching 100% means that the data points are perfectly aligned on the line and predictions are most accurate. However, when coming close to 0% it means that the points are spread out everywhere away from the fitting line, this is when the model performance is poorest and can't accurately predict continuous values close to the true target.

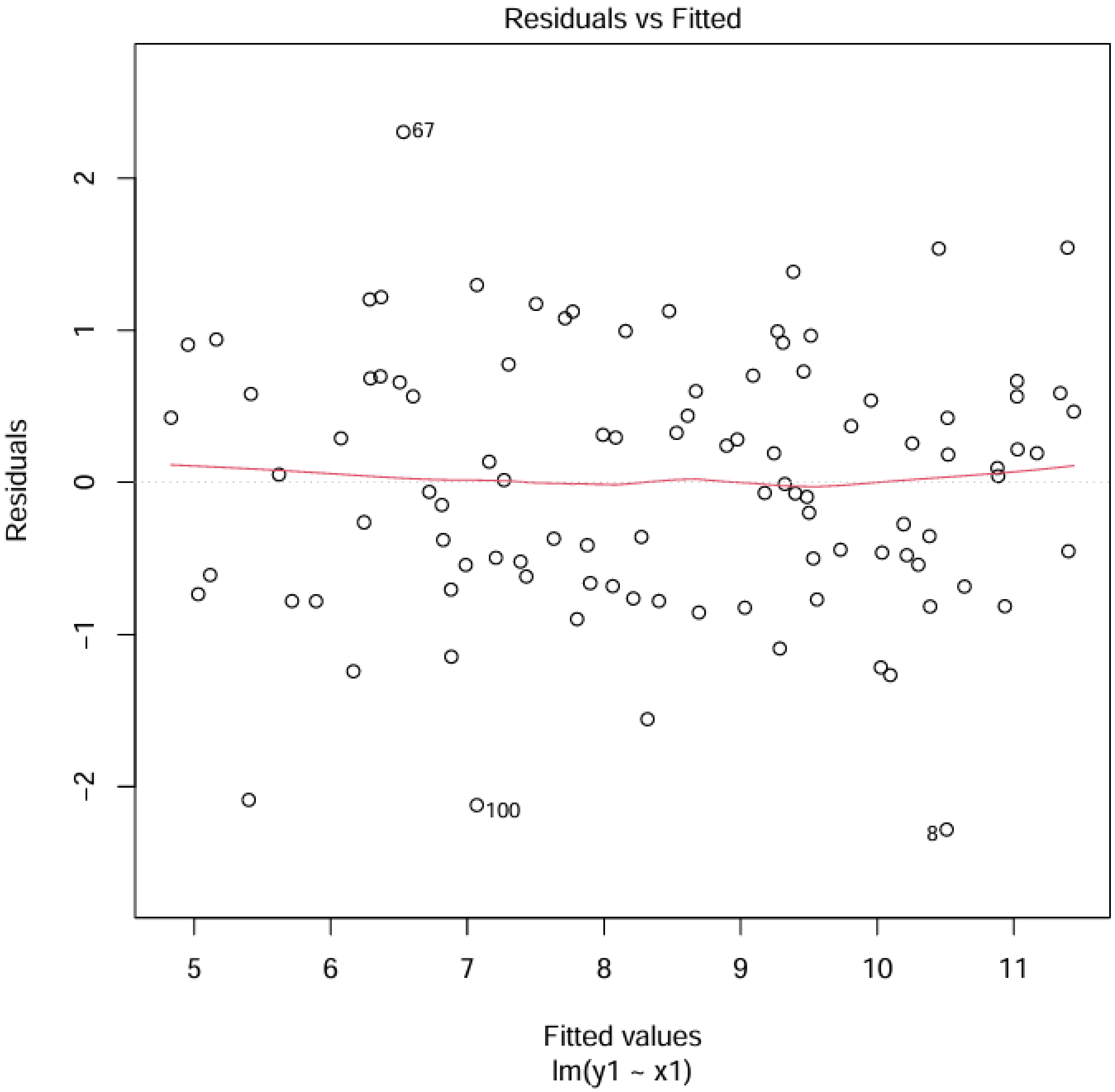
Q4) **What do you conclude about the residual plot? Is it a good residual plot?**



A residual plot is used to check the assumption of constant variance and to check model fit (is a line a good fit).
Good residual plot: no pattern

The plots are scattered randomly without following any specific pattern along the x-axis. This suggests that the residual plot is good because there is no discernible trend or structure in the way the points are distributed indicating that the model's errors or residuals are randomly distributed. Therefore, we can conclude that the linear model is suitable for this data because there is no apparent pattern in the plot.

Q5) **After adding slight non-linearity, what do you conclude about the residual plot? Is it a good residual plot?**



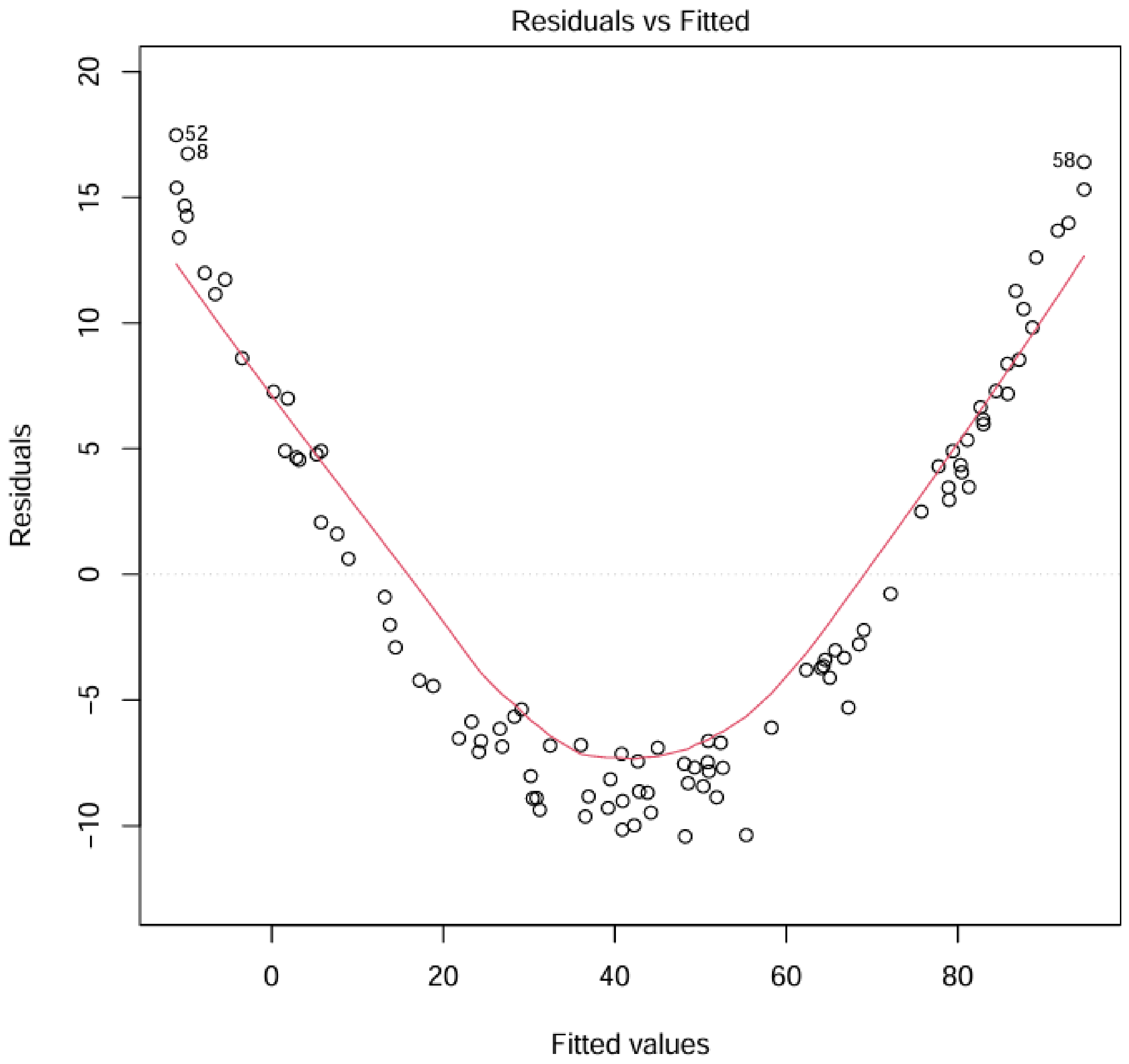
When slight nonlinearity is introduced into the model, it can alter the distribution of data points in the residual plot. Even though there may still be no discernible pattern, the points might not be as scattered as before
This may lead to a better fit between the observed data and the model predictions by capturing the underlying relationships between variables leading to better alignment with the output. As a result, the residuals, which are the differences between observed and predicted values, may be smaller in magnitude. With reduced residual

variability, the data points in the residual plot may appear less scattered. (Hence the reduced ranges compared to the previous plot)

This is still a good residual plot due to the same reasons we mentioned above (Absence of a pattern). Linear model is appropriate here for sure.

Q6) **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?**

When we increase the value of the non-linear term in the original model by a lot, the residual plot starts showing a curved pattern. In other words, by enhancing the influence of the non-linear term, the residual plot deviates from randomness. This means that the linear model doesn't fit the data well anymore. Instead, it suggests that a different model, like a quadratic one, might be a better fit for the data.



Q7) **Import the dataset LungCapData.tsv. What are the variables in this dataset?**

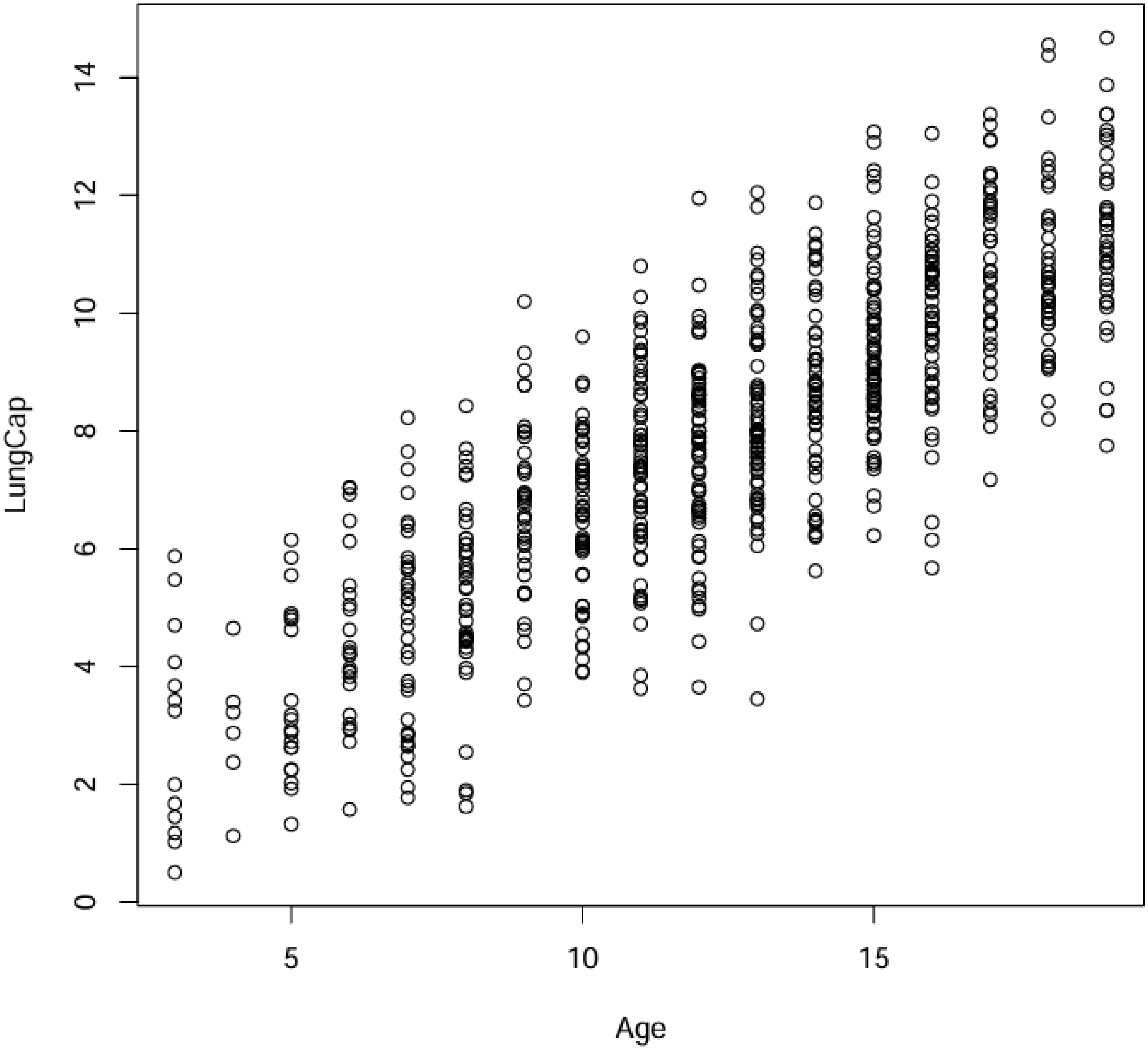
```
dataset <- read.csv("LungCapData.tsv", header=TRUE, sep="\t")
```

To see what variables are in this dataset, we can simply print it.

The variables are: LungCap Age Height Smoke Gender Caesarean

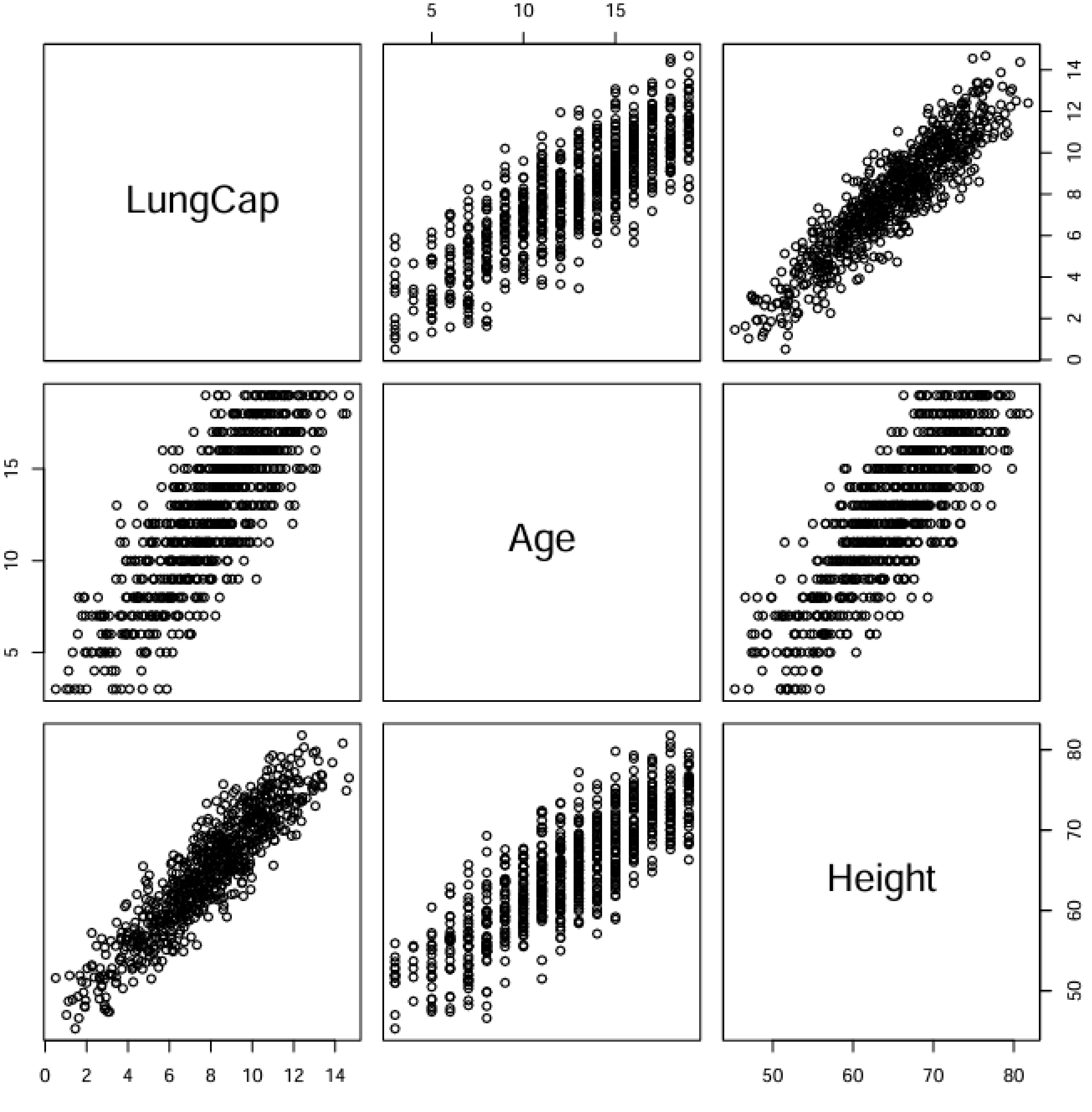
Q8) **Draw a scatter plot of Age (x-axis) vs. LungCap (y-axis). Label x-axis "Age" and y-axis "LungCap"**

```
par(mfrow=c(1,1))
plot(dataset$Age, dataset$LungCap, xlab="Age", ylab="LungCap")
```



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

pairs(dataset[,c("LungCap", "Age", "Height")])



Q10) **Calculate correlation between Age and LungCap, and between Height and LungCap.**

cor(dataset$Age, dataset$LungCap) → 0.8196749
cor(dataset$Height, dataset$LungCap) → 0.9121873

Q11) **Which of the two input variables (Age, Height) are more correlated to the dependent variable (LungCap)?**

The correlation between Height and LungCap is higher than the correlation between Age and LungCap, therefore we can deduce that Height is more correlated to LungCap that Age is. Also it’s visible from the plots above that height is more strongly correlated with LungCap.

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

The correlation between them is 0.9121873 which is fairly high so we can comfortably say that they’re correlated. From the graph, we can see that as height increases, the LungCap increases as well by almost the same scale.

Q13) **Fit a linear regression model where the dependent variable is LungCap and use all other variables as the independent variables**

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

Q14) **Show a summary of this model**

modelSummary <- summary(model)

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

print(modelSummary$r.squared) → 0.8542478

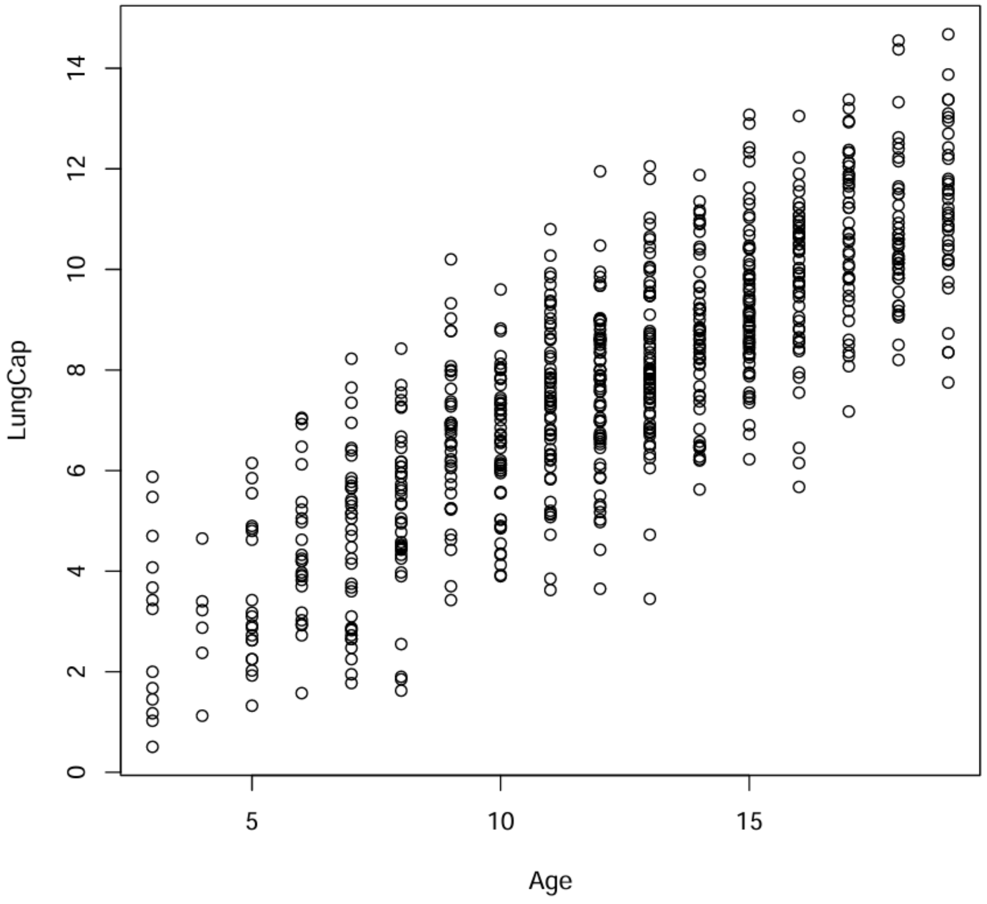
This means that about 85.4% of the variation in the dependent variable (the one we're trying to predict) is explained by the independent variable(s) in the model. This indicates how well the model fits the data: the closer the R-squared value is to 1, the better the model fits the data.

Q16) **Show the coefficients of the linear model. Do they make sense? If not, which variables don't make sense? What should you do?**

Coefficients → -11.32249 0.1605296 0.2641128 -0.6095592 0.3870117 -0.2142182

First 2 coefficients belong to the Age and Height variables. The numbers don’t make sense since we already showed above that the correlation between LungCap and both Age and Height is high since they spread similarly in the plots & in the same direction. Also, when we printed the correlation it turned out to be 0.8196749 between Age & LungCap and 0.9121873 for Height & LungCap. The coefficients set don't line up with this analysis so we know it doesn't make sense. This happened because there is correlation between the "independent variables" Age and Height = 0.8357368, meaning that the effect between Age and LungCap & Age and Height contradicted with each. In other words, the un-explained magnitude reduction is because the Age is supposed to be independent but is instead affected by the Height so it deviated away from correlating strongly with LungCap. Solution: Apply dimensionality reduction by removing some correlated features that could be marked as redundant information causing the decrease of the accuracy of our linear regression model and we'll see proof in the next question how reducing the number of features and ensuring that they're all independent would impact the predictions.

Q17) **Redraw a scatter plot between Age and LungCap. Display/Overlay the linear model (a line) over it. If you are working correctly, the line will not be displayed on the plot. Why?**



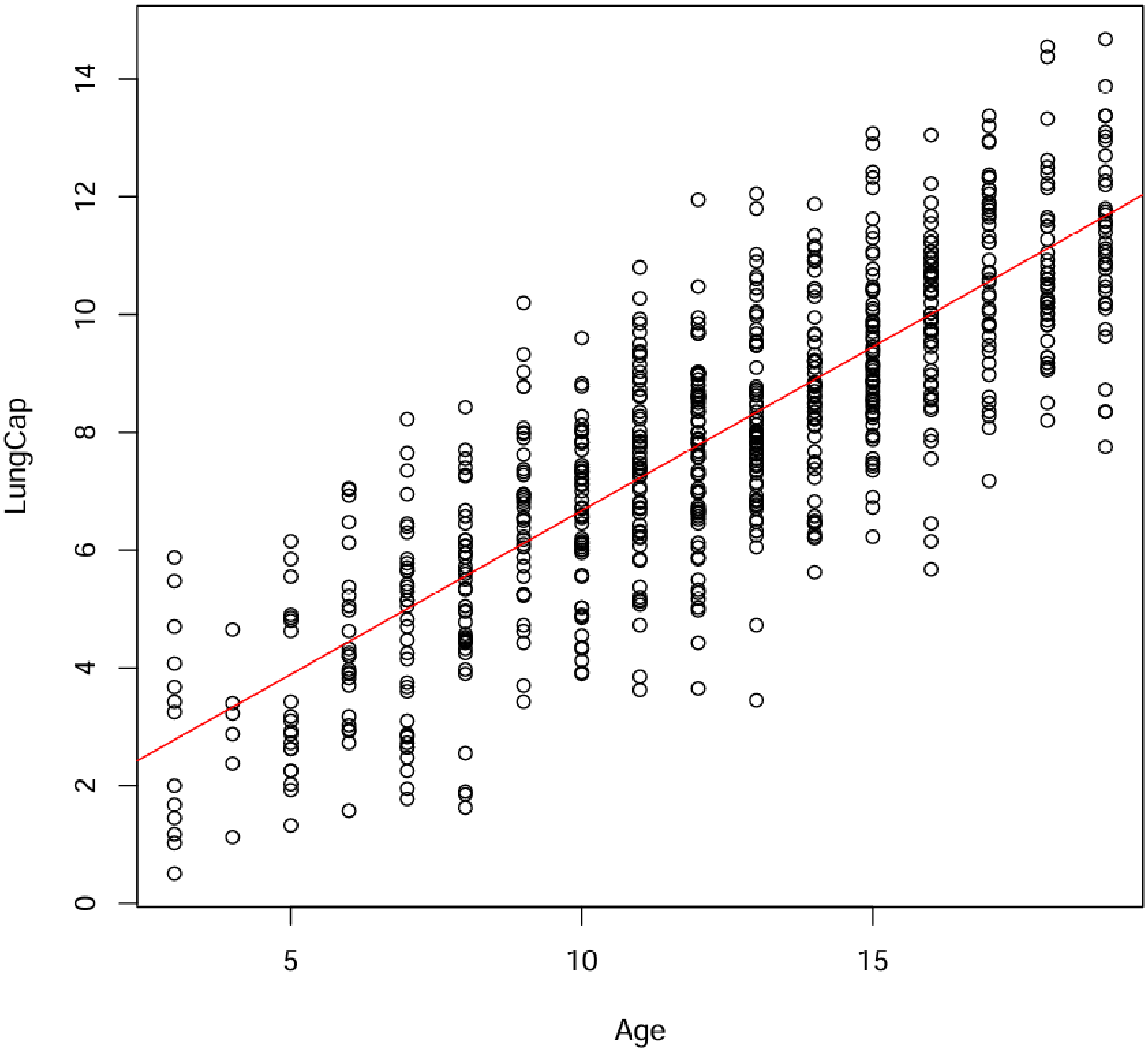
The model coefficients above show that the b0 (y-intercept) is out of the graph scale (0 – 14). Its value is -11.32249 which is below the graph, therefore it wasn't visible when plotted. This also shows that the model fits the data terribly since it doesn't align with it, not even slightly.

Q18) **Repeat Q13 but with these variables Age, Smoke and Caesarean as the only independent variables.**

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

Q19) **Repeat Q16, Q17 for the new model. What happened?**

Coefficients of the model are: 1.108672 0.5561667 -0.6431029 -0.1460278



After limiting the independent variables to only Age, Smoke and Caesarean, we have made sure that there is no correlation or dependency between them. This way the coefficient of Age increased noticeably indicating that correlation with LungCap is positive and strong. We now see the red line because the y-intercept is in the range of the graph y-axis. Also the line fits the data well running through the middle of all scattered data points leading to better R-squared value.

Q20) **Predict results for this regression line on the training data.**

```
ypred <- predict(model)
```

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

```
mse <- mean((dataset$LungCap - ypred)^2)
Mean Squared Error is: 2.280169
```