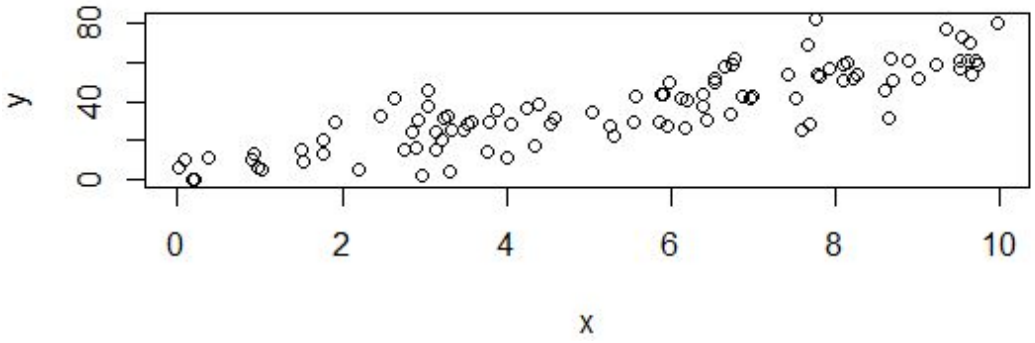
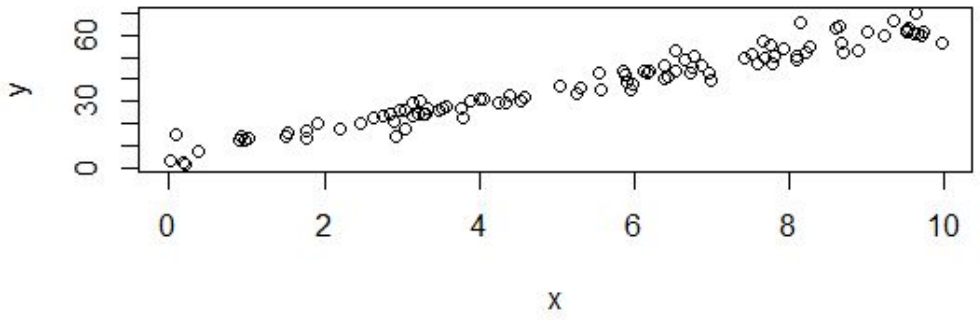
Part 1:

Q1) Try changing the value of standard deviation (sd) in the next command. How do the data points change for different values of standard deviation?

A1) Data become more scattered around the line y=5+6x as the sd increases. So higher sd → more (noise) scatter, less sd -> less (noise) scatter.
Sd = 10



Sd = 4



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

A2) Adding more noise i.e. increasing standard deviation would make it harder to estimate the correct coefficients so the standard error increases for the coefficients (coefficients take larger ranges) which means that the model's uncertainty increases. Less standard deviation(less noise) → easier → less std.Error.

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

A3) The value of R-squared increases as we decrease the standard deviation because the data become less scattered around the line and the opposite is true as we increase the standard deviation (more scatter) the R-squared value goes down.

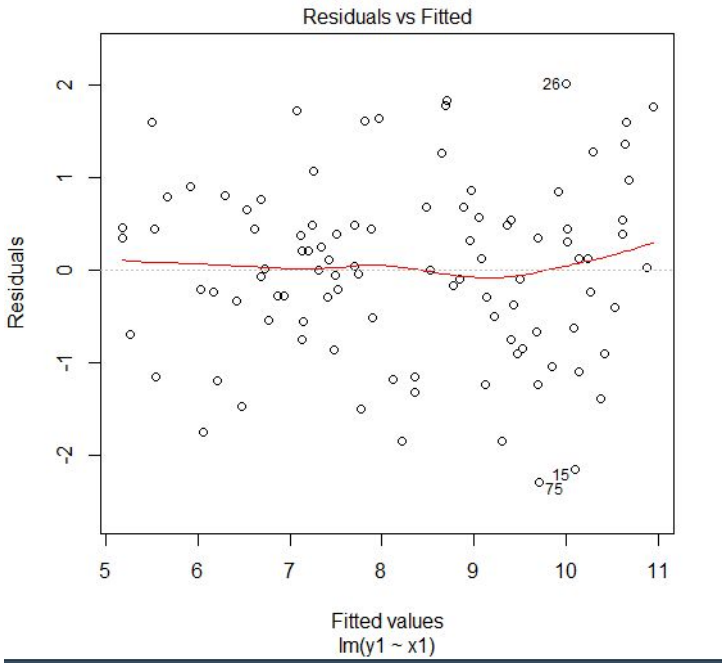
Q4) What do you conclude about the residual plot? Is it a good residual plot?

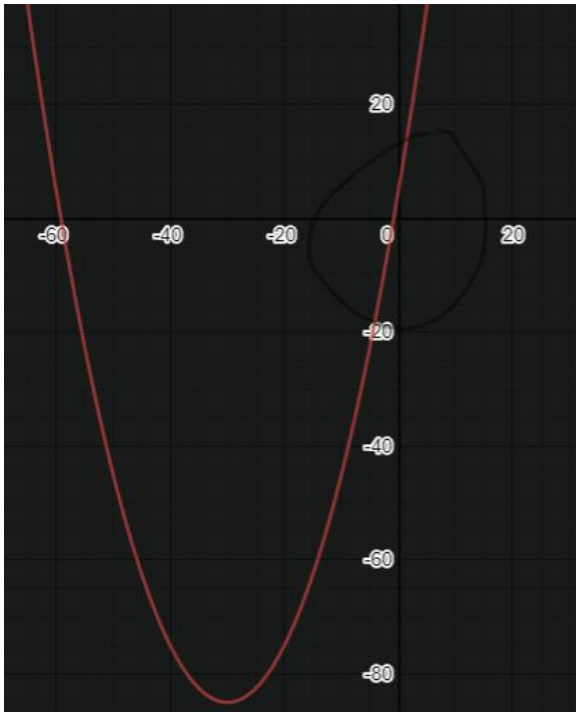
A4) I think it's good since i see no pattern in the residual plot and the variance in the plot is quite similar (almost constant variance).

Part 2:

Q5) What do you conclude about the residual plot? Is it a good residual plot?

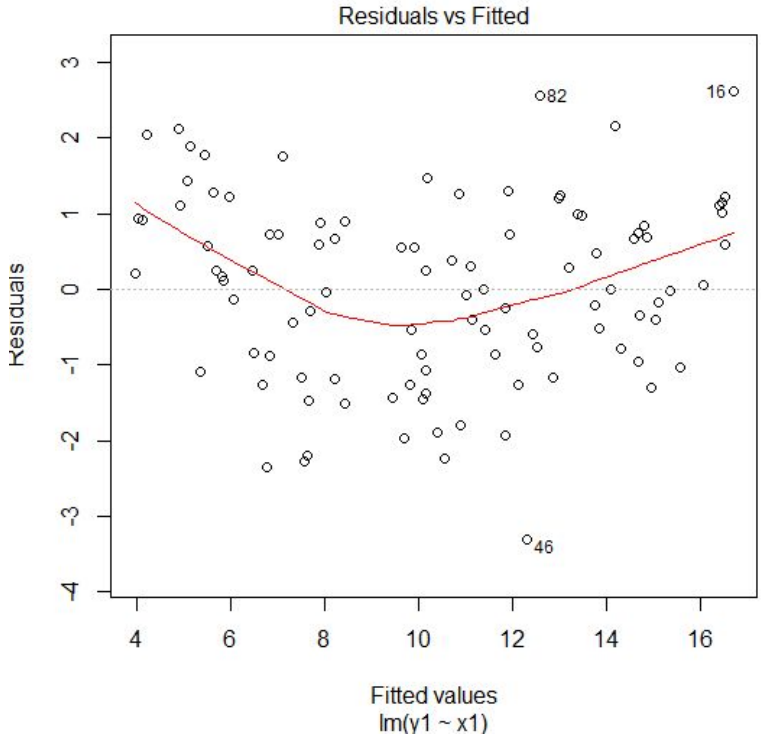
A5) I think it's good since i see no pattern in the residual plot although the data is not linear but this is because the coefficient of the non-linear term is so small and the function is almost linear In the region where the values of x exist.





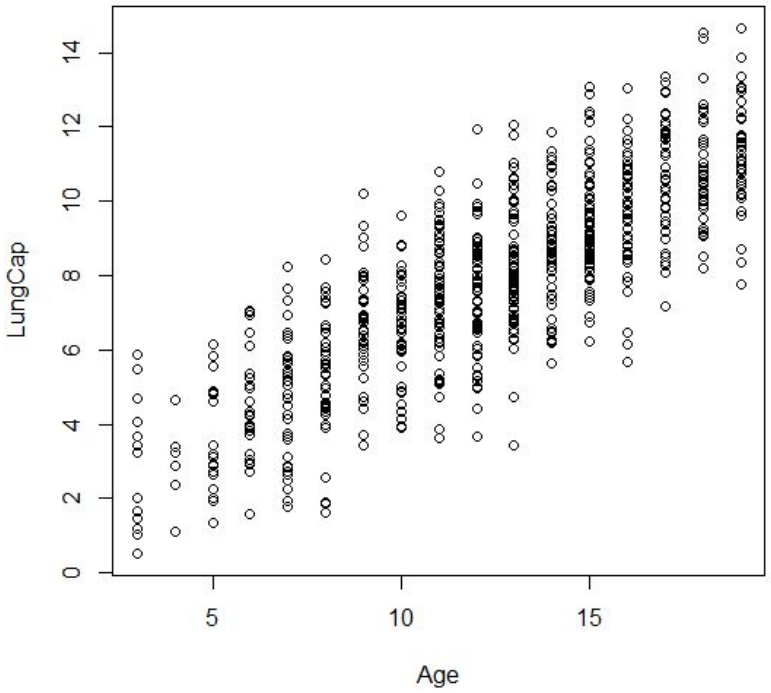
Q6) Now, change the coefficient of the non-linear term in the original model for (A) training and (B) testing to a large value instead. What do you notice about the residual plot?

A6) non-linear coefficient = 7. We start to notice the problem in the residual plot. We start to see the non-linear (quadratic) pattern in the plot and that the data is quadratic not linear.

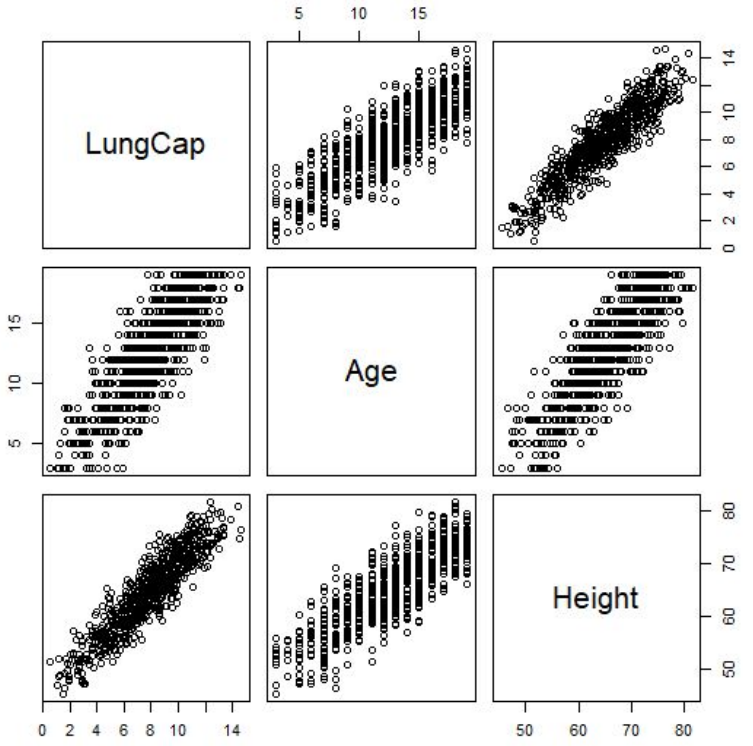


Part 3:

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



Q9) Draw a pairwise scatter plot between Lung Capacity, Age and Height.

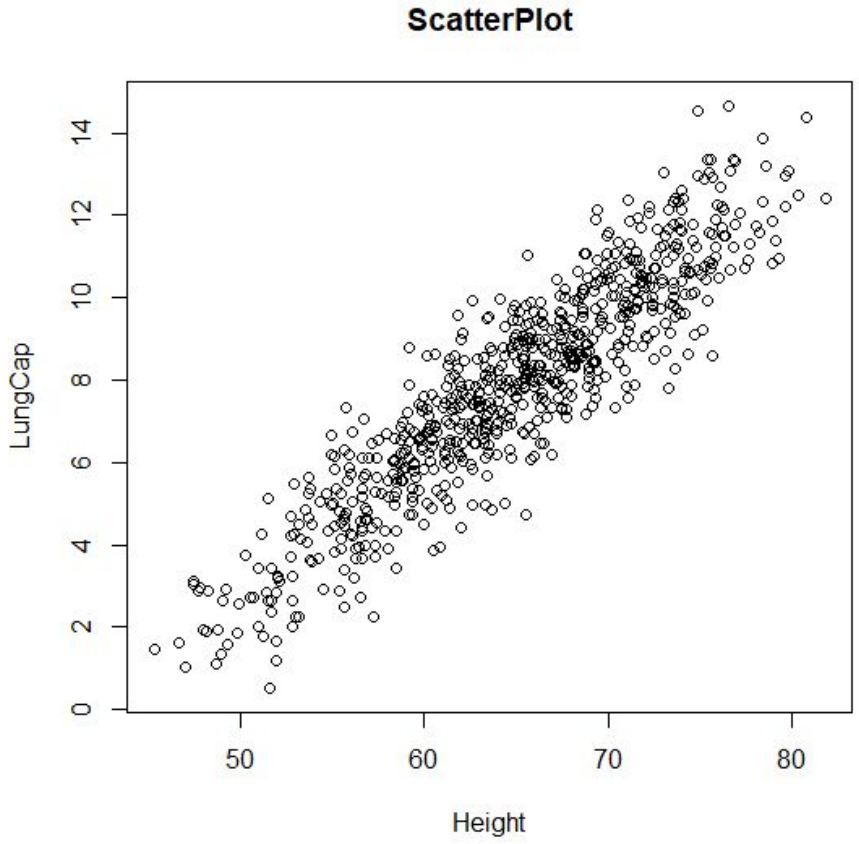


Q11) Which of the two input variables (Age, Height) are more correlated to the dependent variable (LungCap)?

A11) Height is more correlated to the dependent variable (cor = 0.9121873).

Q12) Do you think the two variables (Height and LungCap) are correlated ? why ?

A12) yes, because when the height goes up the LungCap goes up and vice versa Linear relation according to the plot.



Q15) What is the R-squared value here ? What does R-squared indicate?
A15) R-squared value = 0.8532 -> means that there is not much scatter around the line.

Q16) Show the coefficients of the linear model. Do they make sense? If not, which variables don't make sense? What should you do?
A16) Yes, I think they make sense. But Caesareanyes can be removed since it has low significance.

```
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 *
---
```

Q17) Why the line isn't displayed?
A17) Because the Intercept = -11.747065 so the line is out of the plot's boundaries i guess.

Q19) Repeat Q16, Q17 for the new model. What happened?
A17) yes, the coefficients make sense and the line is displayed now on the scatter plot (Intercept = 1.10867)