**Data Mining, Big Data and Analytics.**

Lab 6 – Linear Regression

|  |  |  |
| --- | --- | --- |
| Name | Section | BN |
| Donia Gameel | 1 | 24 |
| Heba Ashraf Raslan | 2 | 32 |

**Part (1):**

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

Higher standard deviations will result in more dispersed data points => data points are more spread out from the regression line.

|  |  |
| --- | --- |
| Sd=0.2 |  |
| Sd=2 |  |
| Sd=20 |  |

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

Increasing the standard deviation 🡺 the coefficients become further from the real coefficients of the model

|  |  |
| --- | --- |
| Sd=0.2 | Intercept=5.031 X=5.993 |
| Sd=2 | Intercept=4.677 X=6.070 |
| Sd=20 | Intercept=5.756 X=4.863 |

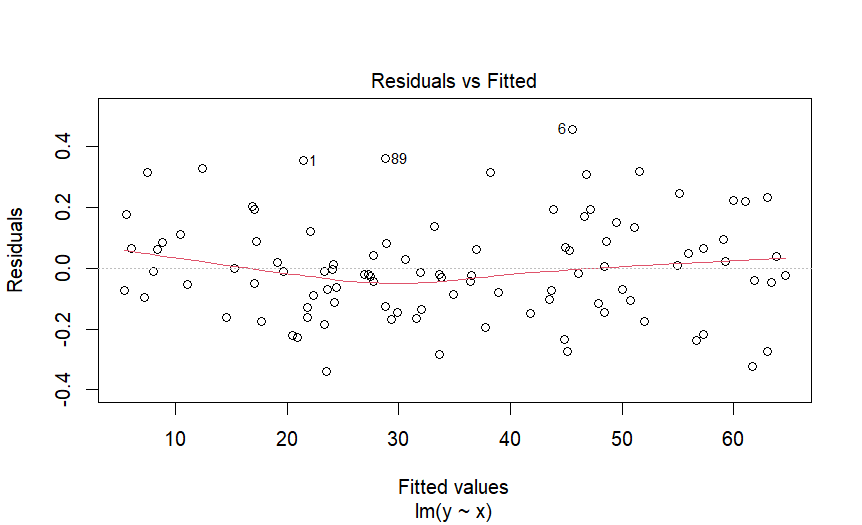
**3.** **How is the value of R-squared affected by changing the value of standard deviation in Q1?**

higher standard deviations leading to lower R-squared values, indicating a poorer fit.

|  |  |  |
| --- | --- | --- |
| Sd=0.2 |  | R-sqr= 0.9998653 |
| Sd=2 |  | R-sqr= 0.9887495 |
| Sd=20 |  | R-sqr= 0.3094882 |

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

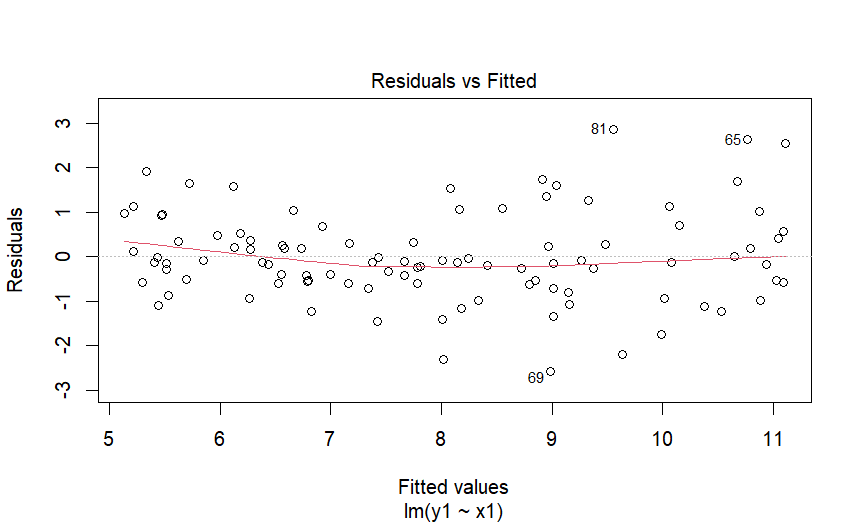
Yes it is. The good residual plot🡺 no pattern



**Part (2):**

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

Yes it is. The good residual plot🡺 no pattern

****

**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?**

Increasing the coefficient of the non-linear term leads to a more curved relationship between the variables, which reflected in the residual plot as a pattern of residuals that are not randomly scattered.

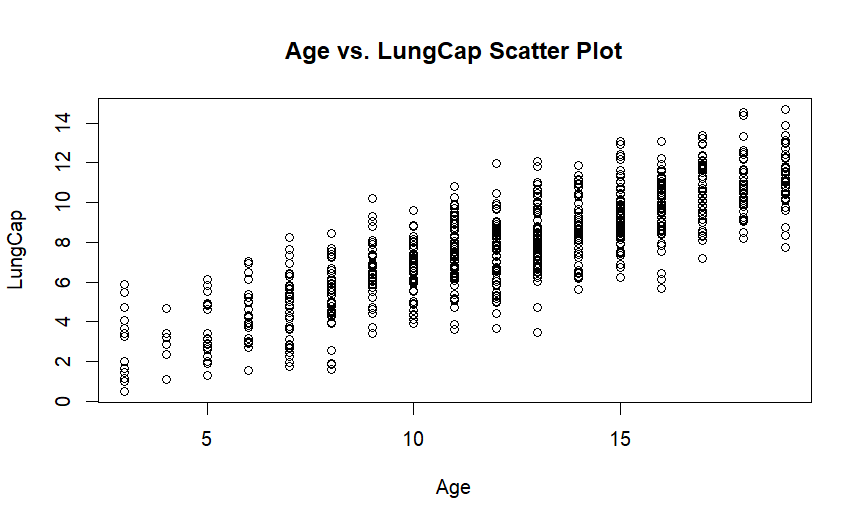
|  |  |
| --- | --- |
| Coefficient=0.1 |  |
| Coefficient=10 |  |
| Coefficient=100 |  |

**Part (3):**

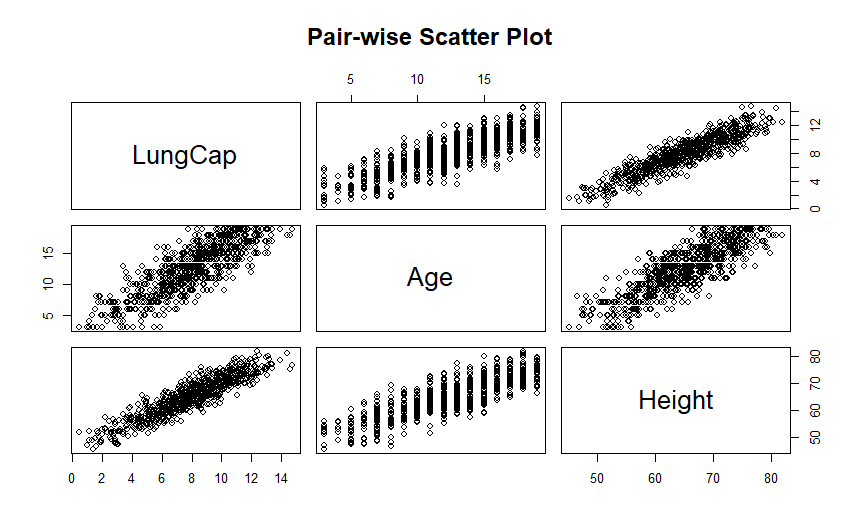
**7. Import the dataset LungCapData.tsv. What are the variables in this dataset?**

"LungCap" "Age" "Height" "Smoke" "Gender" "Caesarean"

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



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

****

**10. Calculate the correlation between Age and LungCap, and between Height and LungCap.**

Correlation between Age and LungCap: 0.8196749

Correlation between Height and LungCap: 0.9121873

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

Height is more strongly correlated with the dependent variable LungCap than Age.

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

Yes, based on the correlation coefficient of 0.9121873 between Height and LungCap, we can conclude that these two variables are strongly positively correlated. This means that as Height increases, Lung Capacity tends to increase as well.

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

model <- lm(LungCap ~ Age + Height + Smoke + Gender + Caesarean, data = lung\_data)

**14. Show a summary of this model.**

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

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

The R-squared value of this model is 0.8542.

It indicates the fraction of the variance in the dependent variable (LungCap) that is predictable from the independent variables (Age, Height, Smoke, Gender, Caesarean) in the model. In this case, an R-squared value of 0.8542 means that approximately 85.42% of the variance in LungCap can be explained by the independent variables in the model.

**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?**

Intercept: -11.32249

Age: 0.16053

Height: 0.26411

Smoke (Smokeyes): -0.60956

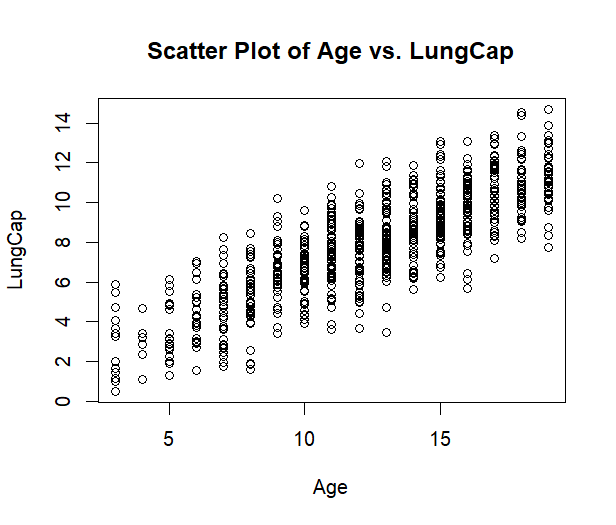
Gender (Gendermale): 0.38701

Caesarean (Caesareanyes): -0.21422

the Intercept's value is not reasonable, as it is high with a negative sign. Setting all other coefficients to small values could result in a negative lung capacity, which is not realistic. A large value of the intercept suggests strongly correlated inputs.

Based on the pairwise plot, we can observe that age and height are correlated. Therefore, including both as inputs to the model may not be necessary or beneficial; choosing one that provides more relevant information could be more appropriate.

**17. 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 line may not be displayed on the plot due to the high correlation between age and height. When two variables are highly correlated, the linear model may struggle to distinguish the individual effects of each variable, leading to difficulties in fitting a clear linear relationship. As a result, the linear model's line may not be visible or may not accurately represent the relationship between age and LungCap in the scatter plot.

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

model2 <- lm(LungCap ~ Age + Smoke + Caesarean, data = lung\_data)

Residuals:

Min 1Q Median 3Q Max

-4.7428 -1.0580 -0.0490 0.9985 4.2319

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.10867 0.18419 6.019 2.79e-09 \*\*\*

Age 0.55617 0.01439 38.639 < 2e-16 \*\*\*

Smokeyes -0.64310 0.18681 -3.443 0.000609 \*\*\*

Caesareanyes -0.14603 0.13468 -1.084 0.278610

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.514 on 721 degrees of freedom

Multiple R-squared: 0.6778, Adjusted R-squared: 0.6764

F-statistic: 505.5 on 3 and 721 DF, p-value: < 2.2e-16

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

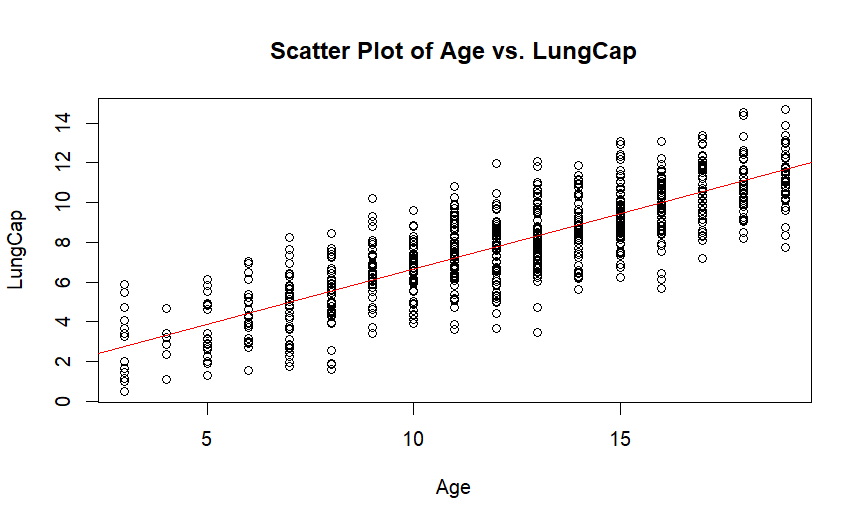
Intercept: 1.10867

Age: 0.55617

Smoke (Smokeyes): -0.64310

Caesarean (Caesareanyes): -0.14603

The red line appears because we removed one of each correlated input pair. .  
We removed height from the correlated pair (height, age) and gendermale from the pair ( gendermale, smokeyes)



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

The first 6 predicted values

1 2 3 4 5 6

4.445673 10.476571 9.861312 8.895007 3.889506 7.226506

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

MSE= 2.280169