

CMP4011 Big Data and Cloud Computing

Lab 5 Linear Regression

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Sec** | **B.N** | **Code** |
| Basma Elhoseny | 1 | 16 | 9202381 |
| Sara El Zayat | 1 | 29 | 9202618 |

## 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?

Gold: y=5+6x. Y is a linear transformation of input feature x.

|  |  |  |
| --- | --- | --- |
| **std** | **2** | **50** |
| **figure** | A graph of a function  Description automatically generated with medium confidence | A graph of numbers and points  Description automatically generated |
| **Comment** | some noise is added to the points, so they are noisy around the straight line | Very high noise is added so data is completely random and no pattern |
| **std** | **0** | **100** |
| **figure** | A graph of a number of objects  Description automatically generated with medium confidence | A graph with numbers and dots  Description automatically generated |
| **Comment** | std = 0 then no noise we are just drawing the original function [St line] | Random as in case std=50 but the range of y here is larger range [more variant] |

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

y=b0+b1x Gold: y=5+6x.

|  |  |  |
| --- | --- | --- |
| **std** | **2** | **50** |
| **b0** | 5.351 | 2.582 |
| **b1** | 5.922 | 4.585 |
| **Comment** | • Values are near to the original coefficients | • Values deviate little from the gold ones |
|  | A graph with numbers and lines  Description automatically generated | A graph with a red line  Description automatically generated |
| **std** | **0** | **100** |
| **b0** | 5 | 12.681 |
| **b1** | 6 | 3.371 |
| **Comment** | •  The exact coefficients of the original equation.  •  Due to having noise of zero standard deviation | •  Very High Noise  •  We get large coefficients this is due to large randomness then no pattern in the data that the model can learn.  •  This high value indicates that the model is learning subset of the space and not able to learn rest |
|  | A graph of a graph  Description automatically generated with medium confidence | A graph with a red line  Description automatically generated |

**Note:** ML is all about learning from data so nothing could be leant from a random data

**Conclusion:** Higher std in noise make more noisy data so due learning this noisy data the model overfits because of trying to learn this noisy pattern this overfitting is obvious in the higher values for the coefficient as the data is more noisy

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

|  |  |  |
| --- | --- | --- |
| **std** | **2** | **50** |
| **R2** | 0.9899548 | 0.0582 |
| **Comment** | • Near 1 but not perfect 1 because data has some noise values so the no line can perfectly fit the data points [Data becomes not perfectly linear 😀] | • Very Small Value due to high value noise so fitting data point with line will get bad results [High Error] |
| **std** | **0** | **100** |
| **R2** | 1 | 0.008 |
| **Comment** | •  **1** Perfect 😊 because no noise is added and also data is generated from linear line so linear regression will perfectly learn the target function achieving R2 = 1 | •  Very Small even smaller than case of std 50 due to higher std for noise then we get more points more randomly deviated from the gold |

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

|  |  |  |
| --- | --- | --- |
| **std** | **2** | **50** |
| **Figure** |  |  |
| **Comment** | • Data Points are Randomly scattered which indicated good model :D | • Still Random but the problem if we saw the **range** of residuals (y-axis) it is **big comapred to case std=2** this means we have a lot of error the model isn’t able to capture [Linear Model and noisy data points 😐] |
| **std** | **0** | **100** |
| **Figure** |  |  |
| **Comment** | • Very Small Range of Residuals  • Constant Variance [Perfect Model Utopia] | •  Still Random but **huge range** for the residual values worse than case std = 50 this is normal because we have more noisy data so deviation from linear model is more so worse values ☹ |

## Part (2):

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

A graph of numbers and points

Description automatically generatedA line graph with numbers and points

Description automatically generated

**The Training & Testing Data are noisy linear [0.\*x2]**

A graph with numbers and dots

Description automatically generated

The residuals plot is random with no pattern exist which suggest it is good I think the only problem is bit large range for the residuals.

(Q6) Now, change the coefficient of the non-linear term in the original model for (A) training.

A graph of a graph

Description automatically generatedA graph of a number of points

Description automatically generated with medium confidence

**The Training & Testing Data are non-linear (Parabola like :D)**

A graph of a graph with numbers and a line

Description automatically generated with medium confidence

**Parabolic Pattern Exists in the Residual**

If a pattern exists in the residual plot, it suggests that the model's **predictions are not capturing all the information** present in the data, indicating potential **issues** with the model's specification or **assumption**. Which is clear than Linear Regression **assumes Linear Data** but the data we are using for test and train are from second order [Nonlinearly :D]

## Part (3):

### Q (7) Import the data set LungCapData.tsv. **What are the variables in this dataset**?

* LungCap [Continues]
* Age [Continues]
* Height [Continues]
* Smoke [Categorial]
* Gender [Categorial]
* Caesarean [Categorial] 🡪 [Target]

**Info About each col for better understanding for me:**

* LungCap🡺 This could represent lung capacity, which is the maximum amount of air that a person can inhale and exhale from their lungs.
* Caesarean 🡺 This variable likely represents whether the individuals were born via Caesarean section (C-section) or not.

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

A graph of a scatter plot

Description automatically generated

**In general, there is increasing lung Cap as Age increases.**

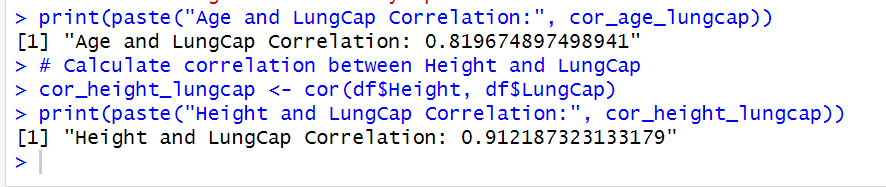
### (Q9) Draw a pair-wise scatter plot between Lung Capacity, Age and Height.

A graph of different sizes and shapes

Description automatically generated with medium confidence

**Height and Lung Capacity are more correlated than Age and Lung Capacity**

### (Q10) Calculate correlation between Age and LungCap, and between Height and LungCap



**The correlation between Height & Lung Capacity is more than that between Age and Lung Capacity which agrees with results in Q 9 😀**

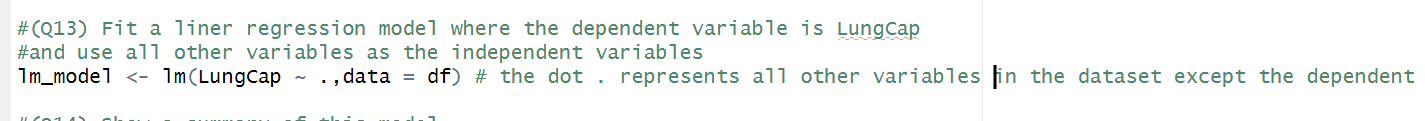
### (Q11) Which of the two input variables (Age, Height) are more correlated to the dependent variable (LungCap)?

**Height**

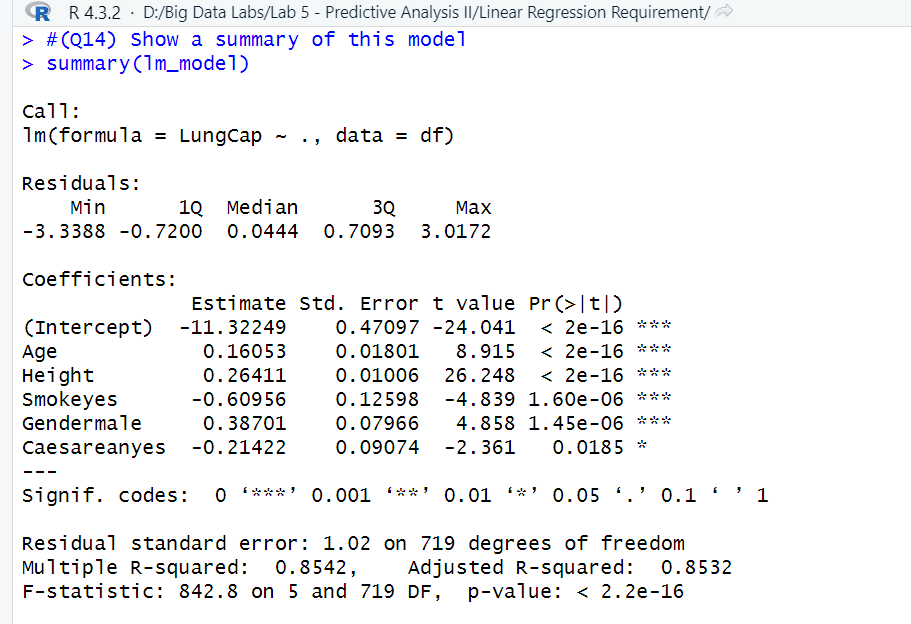
### (Q12) Do you think the two variables (Height and LungCap) are correlated? why?

Yes, they are. In practical terms, this could imply that taller individuals tend to have larger lung capacities compared to shorter individuals. However, correlation does not imply causation, so further analysis would be needed to determine the exact relationship between these variables.

### (Q13) Fit a liner regression model where the dependent variable is LungCap and use all other variables as the independent variables.



### (Q14) Show a summary of this model.



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

A screenshot of a computer

Description automatically generated

It indicates that 85.42% of the variance in the dependent variable LungCap is explained by the independent variables (Used by the model). In other words, if we draw the linear Model plane in space and the training points, we will see the points are highly around the plane.

**Note: Difference between multiple vs adjusted R-squared:**

|  |  |
| --- | --- |
| **Multiple R-squared (R2)** | **Adjusted R-squared(R2adj)** |
| • Measures the proportion of the variance in the dependent variable that is explained by the independent variables in the model.  • Increases whenever a new predictor is added to the model, even if the predictor is not relevant.  • Does not penalize for overfitting or the inclusion of irrelevant predictors. | • Adjusts the R2 value to account for the number of predictors in the model.  • Penalizes the inclusion of irrelevant predictors by decreasing if unnecessary predictors are added to the model.  • Provides a more accurate measure of the goodness of fit when comparing models with different numbers of predictors. |

Generally, **R2adj is lower than the R2 if unnecessary predictors are included in the model**. The numbers above are very close to the third digit after decimal point so we will see if there are unnecessary predictors?! I Think no

### (Q16) Show the coefficients of the linear model. Do they make sense?

A close-up of a computer screen

Description automatically generated

Yes, they make Sense. 😉😉

* Age and Height have positive coefficients, which is logic becoming older means your body is growing up. Same for height taller means your larger capacity😁
* Smoking has a negative coefficient, yes smoking causes lung problems including cause of smaller lung capacity.
* GenderMale coefficient is 0.38701. Since this coefficient is positive, it means that there is a positive association between being male and lung capacity, holding other variables constant.

If an individual is female, then the coefficient for Gendermale does not directly apply because Gendermale would be 0 for females. Which means Males has more lung capacity which may be biologically correct [**I googled it**]

A screenshot of a computer

Description automatically generated

* CaesareanYes coefficient is negative this means that individuals with a history of Caesarean birth, on average, have a lung capacity that is lower by 0.21422 units compared to individuals without a history of Caesarean birth, holding all other variables constant. Which may seem logic 😊

A screenshot of a computer

Description automatically generated

### (Q17) Redraw a scatter plot between Age and LungCap. Display/Overlay the linear model (a line) over it.

A graph of a scatter plot

Description automatically generated

The Linear Model Line isn’t shown 🤔🤔?

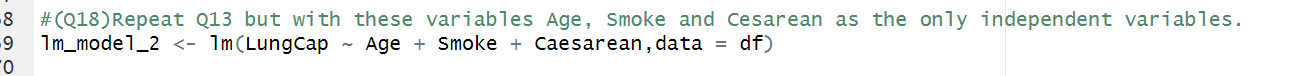
The line of the model to be drawn will have the first coefficients only which are the y-intercept and the coefficient of the Age. LungCap=0.1605296 Age - 11.3224856

A graph with a green line

Description automatically generated

**The line of separator is below the graph values we have its y-intercept -11.32 and x-intercept=70 so we cannot see it here 😳**

### (Q18) Repeat Q13 but with these variables Age, Smoke and Cesarean as the only independent variables.



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

#### Show the coefficients of the linear model. Do they make sense?

A close-up of a computer screen

Description automatically generated

I see the most significant difference is that y-intercept is positive and the Age coefficient is 0.55 instead of 0.16 which I think is logic we have remove other predictors so instead model positively more rely on Age

#### Redraw a scatter plot between Age and LungCap. Display/Overlay the linear model (a line) over it.

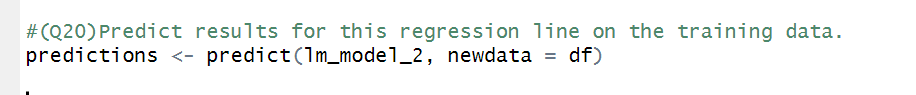
The line of the model to be drawn will have the first coefficients only which are the y-intercept and the coefficient of the Age. LungCap=0.556 Age + 1.1086

A graph with a red line

Description automatically generated

**The line of separator is below the graph values we have its y-intercept -11.32 and x-intercept=70 so we cannot see it here 😳**

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

A computer error message

Description automatically generated

**I Think it will be lower for Model (1)**

A computer code with blue text

Description automatically generated

For Model 1 we get less MSE [Better Model] This sounds logic because in Model 2 we have dropped regraters (Height, Gender) which are significantly important. This AGREES with [Q (15)](#_(Q15)_What_is) where R-squared is high 😎