B3T1102 Data Modelling & Analytics

Assignment 1: Linear Regression

Answer Form

Instructions

1. This is an individual assignment. You are to submit your attempts individually via Canvas.
2. You are required to submit two documents to Canvas, namely, this Answer Form, which you shall name as “<your student ID>\_AF1.docx” , and an Excel file, which you shall name as “<your student ID>\_NA1.xlsx” . Errors in your file name may potentially lead to a loss of marks, so please ensure that the file name is correct.
3. In this assignment, you will be required to write all numerical answers in the provided Excel answer file. The Excel answer file contains two columns. The first column lists the question numbers to which an answer is expected in the Excel answer file. Do not edit the contents of the first column. In the second column, you are to key in your answer to the corresponding question there.
4. For any numerical answers, when keying in to the Excel answer file, please round off your answers to 3 significant figures. If they are whole numbers, leave them as such.
5. For text-based answers, you can type your answers into this answer form. Where a text-based answer is expected of you, you will find a box. You may type you answers in the boxes provided.
6. This assignment counts towards 20% of the final grade. There is a maximum of 40 points awarded for this assignment. Any student who scores lower than 22 points for this assignment can opt for the improvement option. However, if a student gives a nil-return for this assignment, we will not grant the improvement option, as the student is assumed to have forfeited the assignment.
7. Late submissions will not be entertained, unless special approval has been obtained.

Deadline for handing in this assignment: Friday April 29, 2022, 23:59.

**Initialization**

On the Canvas assignment page, you will find an R script, called ‘Assignment 1.R’ Download it, as you will need to run the codes in there to complete this assignment. In the first line of the code, you will find:

>>> studentID <- <fill in your 6-digit student ID here>

Please do as instructed, by replacing the text right of the assignment operator “<-“ with your 6-digit student ID, **not** including the alphabets for your initials. If you do not do this correctly, your code might not run properly.

Once that is done, right below, you will see a list of libraries that would be required to load in order to run the code for this assignment. For whichever that you cannot load, please remember to install.packages .

Once you have verified that all the libraries are installed, please run the code from the first line, until the lines

>>> ###################

>>> ### END OF PART 1 ###

>>> ###################

At this point, running the code further might result in errors.

**Part 1**

You will find a data frame that is named as “df1\_stu”. You may work on this data frame. If at any point in time, you have messed up and need to recover this data frame, then use the following code:

>>> df1\_stu <- df1

Do not overwrite the data frame “df1”. If you do so, you may have to re-run the whole code from the start.

In this data frame, you will find 8 columns. The column ‘y’ is to be assumed as the outcome / dependent variable. The other columns labelled from ‘x1’ to ‘x7’ are to be assumed as the predictors / independent variables.

**Question 1**

**1 point**

Find the mean of y. Key your answer into the Excel answer file.

**Question 2**

**3 points**

Which predictor has the highest correlation with ‘y’? Key your answer into the Excel answer file. Your answer should be in the format: ‘x<a number>’.

**Question 3**

**3 points**

Build a linear model using all of the independent variables. Report the adjusted R-squared value in the Excel answer file.

**Question 4**

**1 point**

Report the p-value corresponding to the feature ‘x4’ in the Excel answer file.

**Question 5**

**1 point**

Depending on your answer for Question 4, conclude whether or not there is a relationship between the outcome variable ‘y’ and the feature ‘x4’, under significance level 2.5%. Key in ‘Yes’ or ‘No’ into the Excel answer file.

**Question 6**

**3 points**

The true relationship between the outcome variable and the predictors is in fact given by the following equation:

Is it always true that because the true relationship does involve all of the predictors, that we must therefore obtain a situation where the p-values are all significant? Explain.

No, this is because not every independent factor “contributes” equally to the independent variable Y. Independent variables x3 and x5 are multiplied with the smallest fractions when compared to the other independent variables. Therefore, changes in x3 and x5 are logically less significant in influencing Y. This can also be seen in the output of the linear model: x3 and x5 have the lowest p-values (although x5 is still just under SL 2.5%).

**Question 7**

**2 points**

Make a Q-Q plot corresponding to the linear model that you built in Question 3. Put a screenshot of it in the box below. Based on this Q-Q plot, what can you conclude about the assumed noise term in the linear model?

Chart, line chart

Description automatically generated

The Q-Q plot indicates that the residuals and the data itself follow a normal distribution, with a few outliers at the tails. Based on this Q-Q plot, it can be determined that the assumption for the normally distributed noise (term) holds.

**Part 2**

Run the code in ‘Assignment 1.R’ from where you stopped earlier in Part 1, until the lines

>>> ###################

>>> ### END OF PART 2 ###

>>> ###################

At this point, running the code further might result in errors.

You will find a data frame that is named as “df2\_stu”. You may work on this data frame. If at any point in time, you have messed up and need to recover this data frame, then use the following code:

>>> df2\_stu <- df2

Do not overwrite the data frame “df2”. If you do so, you may have to re-run the whole code from the start.

In this data frame, you will find 6 columns. The column ‘y’ is to be assumed as the outcome / dependent variable. The other columns labelled from ‘x1’ to ‘x5’ are to be assumed as the predictors / independent variables.

**Question 8**

**2 points**

Build a linear model using all of the independent variables. Report the value of the smallest p-value observed for any of the predictors in the Excel answer file.

**Question 9**

**2 points**

Examine if there is multicollinearity amongst the predictors by computing the appropriate statistic for the predictors. Report the largest value of this statistic in the Excel answer file.

**Question 10**

**3 points**

If you discovered that there was multicollinearity in Question 9, then remove predictors until you obtain a model that no longer has multicollinearity, by removing the predictors with the largest statistic computed in Question 9 first. Report the p-value in the new model corresponding to the originally multicollinear predictor in the Excel answer file. If there is no multicollinearity in Question 9, report “NULL” in the Excel answer file.

**Question 11**

**1 point**

What can you conclude from the value you obtained in Question 10?

After removing x2 (which had the highest VIF value), there is now a statistically significant (at SL 2.5%) relationship between x1 (the originally multicollinear predictor) and the dependent variable Y.

**Question 12**

**4 points**

You are told that there is interaction between two of the variables in your new model. In this question and the next, key into the Excel answer file, the names of these 2 variables in the format ‘x<a number>’. Key in the one with the smaller number in the name here, and the larger number in Question 13.

**Question 13**

See Question 12.

**Question 14**

**5 points**

As a result of Questions 12 and 13, build a new model that correctly factors in the interaction variables, and excludes non-significant variables. State the formula for your model below. With this new model, plot the residuals against all of the predictors that were not involved in Questions 10-13 in the box below. What can you conclude from this plot?

model\_q14 <- lm(y ~ x1 + x3 + x4 \* x5, data = df2\_stu)

The points are relatively evenly distributed around the zero horizontal line. This could be an indication that the variable might not violate the linearity assumption, although that cannot be concluded for sure based on this graph. Therefore, it is smart to test whether adding a non-linear (squared) x3 can produce a better model.

**Question 15**

**4 points**

In light of your conclusion in Question 14, build a new model that corrects for the issue identified. Perform a statistical test to verify if this new model is better than the previous model in Question 14. Report the p-value of this test in the Excel answer file. If the p-value is very small, e.g. reflected as “< 2.2e-16” in the outputs of R, then simply report this as 0.

**Question 16**

**4 points**

Create a plot that might identify if there is heteroskedasticity in your model, and place it in the box below. What is the conclusion that you may form from your plot?

Chart, scatter chart

Description automatically generated

From the graph, it can be concluded that the variance of the error term seems to be (relatively) constant over the range of measured independent variable values.

A BP Test confirms this at a significance level of 2.5% (as previously defined).

**Question 17**

**1 point**

Name another check that has yet to be performed on the model to verify that it fulfils the assumptions of linear regression.

An assumption that has not been tested is autocorrelation. A check should be performed to verify that this assumption holds.