UNIVERSITI TUNKU ABDUL RAHMAN
ACADEMIC YEAR 2019 / 2020
Assignment

Instructions:

* This is an individual assignment.
* Refer to the dataset and answer section A and B questions. All the chart, measurements or outputs must be produced by statistical software.
* Compile the report with a cover page. The cover page consists of student's name and id no, unit code and name, program name.
* The report must be prepared by Microsoft Word, the font size is 12 and the font style is Times New Roman.
* Submit soft copies of your reports to limhuait@utar.edu.my. You should have the three following documents submitted:
  - Word Document
  - Pdf
  - Related document/Output from statistical software (e.g. Section B- Excel, SPSS, R programming, Python, and etc.; Section C- article)
* Report submission duration: **9 April 2020 till 14 April 2020**

Section A:

The cigarettes dataset cigarettes.xlsx contains $n = 25$.

Variables Name	Count
Brand name (Brand)	25
Tar amount in milligrams (Tar)	25
Nicotine amount in milligrams (Nicotine)	25
Weight in grams (Weight)	25
Carbon monoxide amount in milligrams (CO)	25

The United Kingdom Trade & Investment department investigate various domestic cigarettes annually according to their tar, nicotine, and carbon monoxide amount as each of these substances is hazardous to a smoker's health. Past researches have indicated that increases in the tar and nicotine amount of a cigarette will result in an increase in the carbon monoxide emitted from the cigarette smoke. The amount of Carbon monoxide (in milligrams) will be stated as the dependent variable of the case.

Using any statistical software with $\alpha = 0.05$ answer the following questions:

(a) Plot the scatter plot of CO against Nicotine, where CO is the dependent variable and Nicotine is the independent variable. Determine whether there are any unusual data points based on

the plot. Discuss the reason you chose these point(s) as being unusual if any. What problems might these point(s) cause in a simple linear regression analysis?

(b) Find the correlation between CO and Nicotine? Is there a significant correlation between CO and Nicotine?

(c) Fit a simple linear regression (SLR) model of CO against Nicotine, where CO is the dependent variable and Nicotine is the independent variable. Construct a plot of the residuals against the fitted values, a normal Q-Q plot of the residuals, and a bar plot of Cook's distances for each observation. Comments on the model assumptions and on any unusual data points based on these plots.

(d) Remove ONE (but NOT more than 1) of the unusual observations from the data and re-fit the SLR model in part (c). Repeat the plots you constructed in part (c) for this new model. Based on these plots, discuss whether or not the new model has solved any problems that you may have identified in part (c) with the model assumptions?

Section B:

Answer the following questions by considering the cigarettes dataset and solutions in section A.

(a) Find the ANOVA (Analysis of Variance) table for the new linear regression model in section A part (d) and interpret the results of the F test.

(b) Find and interpret the coefficient of determination for this model.

(c) Using CO as the dependent variable and Tar, Nicotine and Weight as explanatory variables, fit a multiple regression model. Investigate *the order* in which you fit the 3 explanatory variables in the model and examine the ANOVA table for each model. The experiment is to understand "Do the models change depending on the order in which you include the explanatory variables?"

Section C:

Write an essay about applying *categorical variables* in building regression model. In your easy, provide an example of the application in any field or area, and then make a prediction of the model built. (300 ~ 500 words)