CENTRAL UNIVERSITY

FAITH • INTEGRITY • EXCELLENCE

2018/2019 END OF SECOND SEMESTER EXAMINATIONS

FACULTY OF ARTS AND SOCIAL SCIENCES
DEPARTMENT OF ECONOMICS

ECON 206: APPLIED STATISTICS
(LEVEL 200)

Time Allowed: 3 Hours

Instructions:

1. ANSWER **ALL** QUESTIONS USING THE **DATA FILE** GIVEN.

2. ALL PROGRAM OUTPUTS SHOULD BE SAVED IN WORD DOCUMENT ON THE DESKTOP WITH **COURSE CODE & INDEX NUMBER** AS THE FILE NAME.

3. ALL COMMENTS ON PROGRAM OUTPUTS ARE TO BE DONE IN THE ANSWER BOOKLET.

Name of Examiner: Ernest Somua-Wiafe

ATTEMPT ALL QUESTIONS

1) Refer to the ***CarValue*** data, which report information on the value of cars which is based upon five-year owner costs, overall road-test scores, and predicted-reliability ratings.

Using a national average of 12,000 miles per year, an average cost per mile driven is used as the measure of five-year owner costs. Road-test scores are the results of more than 50 tests and evaluations and are based on a 100-point scale, with higher scores indicating better performance, comfort, convenience, and fuel economy.
Predicted-reliability ratings (1 = Poor, 2 = Fair, 3 = Good, 4 = Very Good, and 5 = Excellent) are based upon data from Consumer Reports' Annual Auto Survey.
A car with a value score of 1.0 is considered to be an "average-value" car. A car with a value score of 2.0 is considered to be twice as good a value as a car with a value score of 1.0; a car with a value score of 0.5 is considered half as good as average; and so on.
The data for three sizes of cars (13 small sedans, 20 family sedans, and 21 upscale sedans), including the price ($) of each car tested, are contained in the file named ***CarValues***.

(a) Descriptive Statistics

(i) Consider the following all the variables in the dataset given;
Which of the variables are *qualitative* and which are *quantitative discrete* and *quantitative continuous*?
Determine the *level of measurement* for each of the variables.

(ii) For the variable price, select an appropriate class interval and organize the prices into a distribution table displaying the frequency, percentage relative frequency and percentage cumulative relative frequency. Write a brief report summarizing your findings.

(iii) What is the mean and median of price and briefly comment on the distribution based on these measures?

(iv) What is the range of sales prices? What is the standard deviation? What peculiar advantage does the standard deviation have over the range in measuring variation? In addition, what peculiar advantage does the range have over the standard deviation in measuring variation?

(v) Create a multiple box plot graphs for the variables price, cost per mile, road-test score and value score in separate panels of the same graph on different scales of measurement. Briefly comment on the distribution of the variables as displayed by the graphs.

(vi) Obtain a multiple scatter plot for value score against price, cost per mile, and road-test score by placing all the three scatter plots in separate panels on the same graph at different scales on the axes. Comment on the linear relationship between the value score as compared with the other variables in your plot.

(b) Confidence interval and hypothesis testing

(i) Develop a 95% confidence interval for cost/mile of cars. Comment on your output.

(ii) Obtain a cross tabulation for the grouped prices and the size of car. Conduct a hypothesis test to determine whether there exist an association between prices and the size of car at a 0.05 level of significance.

(iii) Conduct a hypothesis test to determine whether there exist an association between size of car and the predicted reliability? Use a 0.05 level of significance.

(iv) At $\alpha = 0.05$, is there a significant difference in the mean price of small sedan cars, family sedan cars and upscale sedan cars?

(v) At $\alpha = 0.05$, can you conclude that there is a significant difference in the mean price of cars in relation to their predicted reliability?

(c) Simple linear regression

(i) Let price be the dependent variable and cost per mile the independent variable. Determine the simple linear regression model.

(ii) Obtain a fitted line plot for the linear regression model formulated displaying a 95% confidence interval and a 95% prediction interval of the model. Comment on how well the model predict attendance in terms of salary.

(iii) In a tabular form, estimate the price of car given the cost/mile of 0.90, 0.91, 0.92, 0.93, 0.94, 0.95, 0.96, 0.97, 0.98, 0.99 giving the 95% confidence interval and the 95% prediction interval for you're the prices with the given values of cost/mile.

(d) Multiple linear regression

(i) Consider the value score of the car as the response variable, develop a correlation matrix for the dataset. Which predictor variable has the strongest correlation with the dependent variable? Does it appear there will be any problems of multicollinearity?

(ii) Determine your final regression model considering the variables that are significant in the model using a 0.05 significance level. Indicate briefly how you selected the variables to include in the equation?

(iii) Write out the regression equation and interpret its practical application. Report and interpret the adjusted R-square.

(iv) Conduct a global test of hypothesis to determine whether any of the regression coefficients differ from zero. Use the .05 significance level.

(v) Determine the residuals for the regression equation. Use a histogram to verify whether the distribution of the residuals is approximately normal or not.

(vi) Develop a normality plot and perform the Anderson-Darling test of normality of the residuals from the final regression model formulated. Is it reasonable to conclude that the normality assumption has been met? Comment on the plot and the p-value realized.

(vii) Plot the residuals against the fitted values from the final regression model developed. Plot the residuals on the vertical axis and the fitted values on the horizontal axis. Can you conclude from your plot that the assumption of constant variation has been satisfied? Explain.

(viii) Comment on whether the residuals are randomly and independently distributed using the residuals versus order observation plot for the final regression model.