PROG8430 – Data Analysis, Modeling and Algorithms

Assignment 2

Multivariate Linear Regression

DUE BEFORE 10PM NOV 8

1. Submission Guidelines

All assignments must be submitted via the econestoga course website before the due date in to the assignment folder.
You may make multiple submissions, but only the most current submission will be graded.

SUBMISSIONS

In the Assignment 2 Folder submit:

1. Your R Code
2. Your report in Word, following the template from our MLR lecture and in the Assignment folder.

NOTE – If you use techniques or sources other than those shown in class, you should note these in your final report.

All variables in your code must abide by the naming convention [variable_name]_[intials]. For example, a variable I create for State would be State_DM.

You may only use the following 'R' packages:

1. **pastecs**
2. **corrgram**

THIS IS AN INDIVIDUAL ASSIGNMENT. UNAUTHORIZED COLLABORATION IS AN ACADEMIC OFFENSE. Please see the Conestoga College Academic Integrity Policy for details.

2. Grading

This assignment will be marked out of 50 and is worth 10% of your total grade in the course.
Late assignments will receive a 20% penalty.
Assignments received after start of class the day after due will receive a mark of 0.

3. Data

Your will be using the dataset diamond_val2.txt (a tab delimited text file) located in the Assignment folder.

The data dictionary is in Appendix One.

4. Background

The dataset contains information on the prices of diamonds.

Your task is to use multivariate linear regression to determine the factors that predict the price of a diamond.

Your work should follow the format of the sample report included in the Assignment folder.

5. Assignment Tasks

Nbr	Description	Marks
1	Data Transformation 1. As demonstrated in class, transform any variables that are required to conduct the regression analysis.	2
2	Descriptive Data Analysis 1. Create numeric and graphical summaries of the data (as demonstrated in class). 2. Comment on anything noteworthy or unusual. You are looking for distributions that seem reasonable and reflective of the data you are analysing.	2 2
3	Outliers 1. Create boxplots of all relevant variables to determine outliers. 2. Comment on any outliers you see.	1 1
4	Exploratory Analysis 1. Create QQNorm plots and numeric tests for normality of data and identify data that seems to be normal and not anything else that seems remarkable. If none do, state that. 2. Correlations: Create both numeric and graphical correlations (as demonstrated) and comment on noteworthy correlations you observe. Are these surprising? Do they make sense?	3 5
5	Model Development As demonstrated in class, create three models using three automatic variable selection techniques discussed in class (Full, Forward, Stepwise). For each model interpret and comment on the five main measures we discussed in class: 1. F-Stat 2. R-Squared value 3. Residuals 4. Significant variables 5. Variable Co-Efficients	

	1. Model with all variables included 2. Model with forward selection 3. Model with stepwise selection	5 5 5
	Model Evaluation – Verifying Assumptions 1. For all three models (notice, in the sample, it was only done for one model), as discussed and demonstrated in class, evaluate the independence of predictors, distribution of error terms (i.e. residuals) and homoscedasticity.	3 (for each model) = 9
	Final Recommendation 1. Based on your preceding analysis, recommend which of the three models should be used. **NOTE** – Even if none of the models meet *all* the assumptions of regression, choose the best of the three. In subsequent classes we will learn how to deal with these issues.	2
	Professionalism, Clarity and Proper Citations	8

APPENDIX ONE: diamond_val.txt DATA DICTIONARY

Variable	Description
Price	Price the diamond sold for
Carat	Size of diamond in carats
Clarity	A numerical measure of clarity associated with standard measures in diamonds
Color	A numerical measure of colour, also using standard diamond evaluations
Cut	A numeric measure of quality of cut (Excellent, Good, etc.
Source	The diamond manufacturing who mined, graded and cut the diamond.
Val	Insurance Value placed on the diamond
Year	The year the diamond was first cut.