PROG8430 – Data Analysis, Modeling and Algorithms

Assignment 4

Multivariate Linear Regression

DUE BEFORE MARCH 30, 2023; 10PM

1. Submission Guidelines

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

SUBMISSIONS

In the Assignment 4 Folder submit:

1. Your *.Rmd file. This file must have all output already run *and* your comments and answers to the questions. If you do not include your output, I will *not* be running your code to generate it. I may, however run the code to verify the results.
2. The *.pdf or *.doc file that is produced from your code.

DO NOT PUT THE DOCUMENTS IN TO A ZIP FILE!

PLEASE NOTE: The marks on the assignment are generally awarded 50% for the actual R code and calculations and 50% for interpretation and demonstration that you understand what you have done.

EXAMPLES: The example output provided is simply to demonstrate what a typical submission might look like. You can use it as a basis, but your submission must be in your own words. Submissions that simply "cut and paste" my example commentary will be marked 0.

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

THIS IS AN INDIVIDUAL ASSIGNMENT. UNAUTHORIZED COLLABORATION IS AN ACADEMIC OFFENSE AS IS DIRECT 'CUTTING AND PASTING' FROM OTHER SOURCES. Please see the Conestoga College Academic Integrity Policy for details.

Remember the discussion forums on eConestoga are a great place to ask questions.

2. Grading

This assignment will be marked out of 30 and is worth 12.5% of your total grade in the course.

Assignments submitted after 10pm will be reduced 20%. Assignments received after 8:00am the morning after the due date will receive a mark of 0%.

Assignments which do not follow the submission instructions may have marks deducted.

3. Data

Each student will be using the study dataset:

STUDY DATASET:

PROG8430-Assign04-23W.txt

Appendix one contains a data dictionary for the study file.

4. Background

A major mail-order company tracks the time (in days) it takes for customers to receive their orders (each row in the dataset represents one order).

Your task will be to use multiple linear regression to determine the factors which contribute to, and help predict, the delivery time (variable: DL). As always, imagine this analysis will be generating a report that will be presented to a client.

All of the tasks have been demonstrated in class. A careful review of your notes from the lectures should give you everything you need to complete these tasks.

5. Assignment Tasks

NOTE – For each step/task, include sufficient commentary in your submission to demonstrate understanding of the steps you have taken.

Nbr	Description	Marks
1	Preliminary and Exploratory	8
	1. Rename all variables with your initials appended (just as was done in assignment 1,2 and 3)	
	2. Examine the data using the exploratory techniques we have learned in class. Does the data look reasonable? Are there any outliers? If so, deal with them appropriately.	
	3. Using an appropriate technique from class, determine if there is any evidence if one Carrier has faster delivery times than the other. Make sure you explain the approach you took and your conclusions.	
	4. As demonstrated in class, split the dataframe into a training and a test file. This should be a 80/20 split. For the set.seed(), use the last four digits of your student number. The training set will be used to build the following models and the test set will be used to validate them.	
2	Simple Linear Regression	8
	1. Correlations: Create both numeric and graphical correlations (as demonstrated in class) and comment on noteworthy correlations you observe. Are these surprising? Do they make sense?	
	2. Create a simple linear regression model using time for delivery as the dependent variable and weight of the shipment as the independent. Create a scatter plot of the two variables and overlay the regression line.	
	3. Create a simple linear regression model using time for delivery as the dependent variable and distance the shipment needs to travel as the independent. Create a scatter plot of the two variables and overlay the regression line.	
	4. As demonstrated in class, compare the models (F-Stat, R^2, RMSE for train and test, etc.) Which model is superior? Why?	
3	Model Development – Multivariate	6
	As demonstrated in class, create two models, one using *all* the variables and the other using backward selection. This should be built using the train set created in Step 2. For each model interpret and comment on the main measures we discussed in class (including RMSE for train and test). (Your commentary should be yours, not simply copied from my example).	
4	Model Evaluation – Verifying Assumptions - Multivariate	4
	For both models created in Step 4, evaluate the main assumptions of regression (for example, Error terms mean of zero, constant variance and normally distributed, etc.)	
5	Final Recommendation - Multivariate	1
	Based on your preceding analysis, recommend which of the two models from step 4 should be used and why.	
	Professionalism, Clarity and Proper Citations	1

APPENDIX ONE: STUDY FILE DATA DICTIONARY

Variable	Description
DL	Time for delivery (in days, rounded to nearest 10th)
VN	Vintage of product (i.e. how long it has been in the warehouse).
PG	How many packages of product have been ordered
CS	How many orders the customer has made in the past
ML	Distance the order needs to be delivered (in km)
DM	Indicator for if the product is manufactured in Canada (C) or elsewhere (I)
HZ	Indicator for if the product is designated as Hazardous (H) or not (N).
CR	Indicator for which Carrier delivered the item (Def Post, or Sup Del)
WT	Weight of the shipment (in decagrams)