PROG8430 – Data Analysis, Modeling and Algorithms

Assignment 4

Multivariate Linear Regression

DUE BEFORE NOV 16, 2021; 10PM

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 3 Folder submit:

1. Your R Code
2. Your report in Word, following the template from our MLR lecture and in the Assignment folder. NOTE: This is a rough guide to show the general layout of your submission. Your submission should follow the assignment tasks.

PLEASE NOTE: Your Word document forms the basis of your evaluation and will be what is marked. The R Code is there for me to verify your results and therefore your commentary and output *must* be included in your Word document to be evaluated.

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.

DO NOT PUT THE DOCUMENTS IN TO A ZIP FILE!

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 includes all variables that are provided in the dataset. You must rename these as they are read in as well.

You may only use the two 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 is worth 12.5% of your total grade in the course and you can expect it to take five to eight hours. It is out of 25 marks overall.

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_Assign_MLR_21F.Rdata

Appendix one contains a data dictionary for the study file.

4. Background

A survey of 2033 residents of Canada was conducted to determine the key factors associated with political engagement. A variety of variables were measured and recorded including some tests they were asked to complete. Appendix 1 contains the data dictionary for the data set. One group of respondents ("Treat") were given additional education on political matters while the other ("Control") were not.

Your task will be to used multiple linear regression to determine the factors which contribution to Political Awareness (variable: *Pol*). 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 (note – "in class" includes the pre-recorded material). 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	Reduce Dimensionality 1. Apply the Missing Value Filter to remove appropriate columns of data. 2. Apply the Low Variance Filter to remove appropriate columns of data. 3. Apply the High Correlation Filter to remove appropriate columns of data.	3
2	Data Transformation As demonstrated in class, transform any variables that are required to conduct the regression analysis (e.g. categorical variables to dummies).	2

	NOTE: This was/will be demonstrated in the lecture on Multivariate Regression. If you have not yet had this lecture, you can skip this step for now.	
3	Outliers 1. Create boxplots of all relevant variables (i.e. numeric, non-binary) to determine outliers. 2. Comment on any outliers you see and deal with them appropriately.	2
4	Exploratory Analysis 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
5	Simple Linear Regression 1. Create a simple linear regression model using PoI as the dependent variable and score as the independent. Create a scatter plot of the two variables and overlay the regression line. 2. Create a simple linear regression model using PoI as the dependent variable and scr as the independent. Create a scatter plot of the two variables and overlay the regression line. 3. Compare the models. Which model is superior? Why?	4
6	Model Development - Multivariate As demonstrated in class, create two models using two automatic variable selection techniques discussed in class (Full, Backward). For each model interpret and comment on the five main measures we discussed in class. (Your commentary should be yours, not simply copied from my example): 1. F-Stat 2. R-Squared value 3. Residuals 4. Significant variables 5. Variable Co-Efficients	4
7	Model Evaluation – Verifying Assumptions - Multivariate 1. For both models (as discussed and demonstrated in class) evaluate the main assumptions of regression: Error terms mean of zero, constant variance and normally distributed.	4
	Final Recommendation - Multivariate 1. Based on your preceding analysis, recommend which of the models should be used. **NOTE** – Even if none of the models meet *all* the assumptions of regression, choose the best of the two. In subsequent classes we will learn how to deal with these issues.	1
	Professionalism, Clarity and Proper Citations	3

APPENDIX ONE: STUDY FILE DATA DICTIONARY

Variable	Description
id	UserID (unique to each respondent)
group	Treatment or Control group
hs.grad	Graduated High School (Y or N)
nation	Nationality (Region)
gender	Male, female, undisclosed
age	Age in Years
m.status	Marital Status
political:	Political Affiliation
n.child	Number of Children
income	Annual Household Income
food	Pct of Income to Food
housing	Pct of Income to Housing
other	Pct of Income to Other Expenses
score	Score on Political Awareness Test
scr	Standardized Score Test
time1	Pct of Time Taken on Test
time2	Time Taken on Section 1 (Standardized)
time3	Time Taken on Section 2 (Standardized)
Pol	Measure of Political Involvement