SAI Assignment 5

Work should be submitted on Blackboard by 23:59 on Sunday, December 11th

Regression Modelling of Air pollution

This assignments is related to a data collected on a random sample of 60 customers from a large clothing retailer. The manager of the store is interested in predicting how much a customer will spend on his or her next purchase based on one or more of the available explanatory variables.

The variables of interest are described below:

`Amount` Net dollar amount spent by customers in their latest purchase from this retailer;

`Recency` Number of months since the last purchase;

`Freq12` Number of purchases in the last 12 months;

`Dollar12` Dollar amount of purchases in the last 12 months;

`Freq24` Number of purchases in the last 24 months;

`Dollar24` Dollar amount of purchases in the last 24 months;

`Card` 1 for customers who have a private-label credit card with the retailer, 0 if not,

where the `Amount` of money spent by a customer is the response variable, while the rest of variables are explanatory variables.

Questions of Interest

1. Using a subset selection procedure on the variables above, construct a model that **best** predict the `Amount` of money spent by a customer. Summarise the key features of the model fit and performance, and check the underlying assumptions.
2. Run a bootstrap simulation with a bootstrap sample size of 10,000 to confirm the estimate of the parameters in your chosen model above. Provide a 95% bootstrap confidence interval for your bootstrap estimates of parameters and compare your results with your findings in (1). Use the command `set.seed("your student id")` to initialise the random number generator, before you start the bootstrap sampling.
3. The first person in the dataset is a customer with the following information:

```
##   Amount Recency Freq12 Dollar12 Freq24 Dollar24 Card
## 1     30       6      3      140      4      225    0
```

What does your final model from part (1) say about the amount of money spend by this person in the store. Provide an appropriate interval estimation at $\alpha = 0.05$ significance level for the amount of money spend by this person in the store

Data Management

The CSV format datafile containing these data is provided on Blackboard. The following code will read in the datafile, wrangle it prior to your analysis and produce some summaries of the data:

```
# load the required packages
library(tidyverse)

# read the data in R
# you need to make sure the data file and rmarkdown file are both in the same folder.
clothing <- read.csv("clothing.csv")

# the first few rows of the data
head(clothing)

# column names
names(clothing)

# summary of each column
summary(clothing)

# Data wrangling:
# A careful examination of Table 3.5 reveals that the first three values for
# Amount are zero because some customers purchased items and then returned them.
# We are not interested in modeling returns, so these observations
# will be removed before proceeding.
# The last row of the data indicates that one customer spent $1,506,000 in the store.
# A quick consultation with the manager reveals that this observation is a
# data entry error, so this customer will also be removed from our analysis.
# We can now proceed with the cleaned data on 56 customers.

clothing <- clothing %>%
  filter(! Amount %in% c(0,1506000))
```

Hints:

- You need to produce suitable graphs and statistics to explore the relationships between the features and target variable, including suitable matrix plot(s).
- Train a maximal multiple regression model, using all the features, and identify any signs of multicollinearity/overfitting. You may want to consider exploring the quadratic effect of predictors or create new explanatory variables using the current predictors.
- Apply all the feature subset selection techniques covered. You should consider best subset regression, all possible subsets regression, all four stepwise selection methods and lasso regression. Compare and contrast the results from all these subset selection approaches in justifying the subset of features that should be included in the final multiple regression model.
- Give an interpretation of the estimated effects of each included feature in the final trained model (including the straight line trend) on the respone of interest.
- Provide some regression diagnostic plots to determine if the assumptions underlying the multiple linear regression model are valid.
- How well does the model perform? Is it good at prediction?

Submission Presentation

You should summarise the above analysis in a short **6-8 page report**. You are unlikely to be able to include all of the output/results in your report and you should not try to do so. The report should give an overview of the results from your analysis, you will need to be concise and include only the key results/graphs/table that are needed to justify the conclusions drawn.

You do not need to include the `R` code in the page limit for the report. But you should include all of your `R` code, **to reproduce the full analysis**, as an Appendix or separate script file. Alternatively, if you wish to use `RMarkdown` to present your report then you should submit both the PDF and RMarkdown file in your submission, so that your full `R` code can be assessed. Your code should include comments to explain what is being achieved at each step.

Do not simply copy and paste the output from `R`, as this is often extremely inefficient. For example, the output from the `lm()` and `step()` provide lots of result you do not need. Concisely present the results as though for a professional report.

Graphs are expected to be legible, have relevant axis labels and titles, and possibly legends if they have many types of items displayed.

A key focus of the assessment is of your understanding of the models/methods/algorithms, when it is appropriate to use them, how to apply them and how to interpret the results. Your presentation and `R` code will also be considered.