STAT448-25S1 Assignment 2

15% of your final grade

Instructions

Please adhere to the following guidelines for submitting your assignment:

1. Your assignment should be submitted on the Learn platform in one of the following formats: PDF or HTML R-markdown. This format should include written answers, R code, and results, all properly commented.
2. You have the option to complete the assignment in pairs. In such cases, please ensure that both names and student IDs are included on the assignment; BOTH students must submit the assignment.
3. It is crucial to carefully read and follow the provided instructions for the assignment. Make sure all questions are answered.
4. Answer all THREE questions.
5. Reminder: Ensure that your submission includes R code, output, and comments for the assigned tasks. These should be presented in a markdown HTML or PDF file. Clearly indicate what question or question part you are answering.

Hint: Summarising your results in tables makes comparison clearer and easier.

Question 1

(40 marks) In the folder for Assignment 2, question 1 there are 2 files: the RData file **Residen** and the excel file **Residential-Building-Data-Set.xlsx**. **Residen** is a copy of the dataset in the excel file where the variables names have been changed for use in R. The excel file also contains a description of the variables. More information on the dataset can be obtained on the UCI webpage:

https://archive.ics.uci.edu/ml/datasets/Residential+Building+Data+Set.

a). Explore with appropriate graphical tools the correlation between the variables.

b). Please fit a linear regression model to explain the "actual sales price" (V104) in terms of the of the other variables excluding the variable "actual construction costs" (V105). Explain the output provided by the **summary** function.

c). Fit a linear regression model to explain the "actual sales price" (V104) in terms of other variables (**excluding** the variable "actual construction costs" (V105)) using (i) **backwards selection** and (ii) **stepwise selection**. Compare these two models in terms of outputs, computational time, holdout mean square error and cross validation mean square error.

d). Fit a linear regression model to explain the "actual sales price" (V104) in terms of other variables (**excluding** the variable "actual construction costs" (V105)) using (i) Ridge regression (ii) LASSO regression with an appropriate λ determined by cross validation. Compare these two models in terms of outputs, computational time and cross validation mean square error.

e). Please give a possible explanation of why one method is better than the other in this case, considering also results from backwards and stepwise selection.

Question 2

(25 marks) Also included in the assignment folder is the file **parkinsons.csv**, this dataset [1] contains information on 42 patients with Parkinson's disease. The outcome of interest is **UPDRS**, which is the total unified Parkinson's disease rating scale. The first 96 features, **X1**–**X96**, have been derived from audio recordings of speech tests (i.e. there has already been a process of **feature extraction**) while feature **X97** is already known to be informative for **UPDRS**.

Read in the dataset and set up the model matrix **X**. Next, randomly split the data into a training set with 30 patients and a test set with 12 patients – remembering to state what your RNG seed was.

Standardize your training and test sets before your analysis: $X = scale(X)$ so that the absolute size of the estimated coefficients will give a measure of the relative importance of the features in the fitted model.

a). Confirm that a linear model can fit the training data exactly. Why is this model not going to be useful?

b). Now use the **LASSO** to fit the training data, using leave-one-out cross-validation to find the tuning parameter λ. (This means $nfolds = 30$. You will also find it convenient to set $grid = 10\hat{}seq(3, -1, 100)$ and $thresh = 1e - 10$). What is the optimal value of λ? and what is the resulting test error?

c). State your final model for the **UPDRS**. How many features have been selected? What conclusions can you draw?

d). Repeat your analysis with a different random split of the training and test sets. Have the same features been selected in your final model?

[1] This dataset has been derived from the Oxford Parkinson's Disease Telemonitoring Dataset available at http://archive.ics.uci.edu/ml/datasets/Parkinsons+Telemonitoring

Question 3

(35 marks) Another dataset included in the assignment folder is the file **Weather_Station_Data_v1.csv**, this dataset contains 2000 observations of weather station data, where the MEAN_ANNUAL_RAINFALL is given against thirteen attributes (features) of the weather stations.

There are no missing or undefined values in the dataset. The first row of the CSV contains the variable names.

Randomly split the data using a 80/20 split.

Train an **ElasticNet** model to predict **MEAN_ANNUAL_RAINFALL** using 10-fold cross-validation to optimize values for α and λ.

Plot the cross-validation results for your model.

Make predictions on your test set using both **lambda.min** and **lambda.1se**, show the coefficients of each model, report the MSE and/or RMSE of both, and the number of predictors used in each model.

Which model would you choose? Justify your answer.

The End