**Homework 8 Applied Task**

**Linear Model Selection and Regularization**

**Due March 23rd 11:59 PM (15 points)**

IST 5535 – Spring 2020, Chen

**Name**: Sayantan Majumdar

The data file "HW8Data.csv" contains a dataset of 150 observations. The dataset contains 6 predictors x1, x2, x3, x4, x5, and x6. The y variable is the response.

Complete the following tasks:

1. Univariate feature selection is to statistically test the relationship between each independent variable and the response variable. A simple approach of univariate feature selection is to calculate Pearson correlation coefficient. If a feature is statistically associated with the response, then this feature will be kept. Or else, it will be removed from the modeling. Implement the simple univariate feature selection on the whole dataset and list the variables that should be included in the final linear regression model.

|  |
| --- |
| Q1: What variables are in the final model identified by the simple univariate feature selection?  Your Answer: From the correlation matrix and feature importance plots, it is clear that x4, x5, and x6 should be removed before constructing the linear model. This is because x1, x2, and x3 have high correlation with the response variable, y. So, in the final model, x1, x2, and x3 are present. The final model is:  wherein BIC = 625.5419. Also is not statistically significant. |

1. Use the regsubsets() method in the leaps package to conduct the best subset selection on the whole dataset. Draw a plot to show the dynamics of RSS, Adjusted R2, Cp, and BIC across different number of variables.

|  |
| --- |
| Q2: If we use BIC as the criterion of model selection, what is the final model? Is the model consistent to the model identified at step 1 using the simple univariate feature selection?  Your Answer: If we use BIC (=-188.0302305) as the criterion of model selection, the final model should include 3 predictors. These include x1, x2, and x4. Hence, the final model is:  Clearly, by comparing the BIC values, we see that the best subset selection helps to develop a better model.  Therefore, the best subset selection method is not consistent with the univariate feature selection approach because it considers x4 to be of higher significance than x3 (which is correct in this case). |

1. Use the regsubsets() method to conduct the forward stepwise selection on the whole dataset. Draw a plot to show the dynamics of RSS, Adjusted R2, Cp, and BIC across different number of variables.

|  |
| --- |
| Q3: If we use BIC as the criterion of model selection, what is the final model? Is the model consistent to the model identified at step 1 using the simple univariate feature selection?  Your Answer: The forward stepwise selection results in the same model as the best subset selection method. So, it is not consistent with the univariate feature selection. |

1. Use the regsubsets() method to conduct the backward stepwise selection on the whole dataset. Draw a plot to show the dynamics of RSS, Adjusted R2, Cp, and BIC across different number of variables.

|  |
| --- |
| Q4: If we use BIC as the criterion of model selection, what is the final model? Is the model consistent to the model identified at step 1 using the simple univariate feature selection?  Your Answer: The backward stepwise selection results in the same model as the best subset selection method. So, it is not consistent with the univariate feature selection. |

1. Fit a lasso model to fit the whole dataset. Use 10-fold cross-validation to select the optimal value of λ. Create plots of the cross-validation error as a function of λ. Use the optimal value of λ to build the final lasso model.

|  |
| --- |
| Q5: What is the final linear model if the optimal λ is used? Does the final model do variable selection?  Your Answer: The final linear model (λ = 0.01321941) is:  It doesn’t do any variable selection as none of the coefficients are very close to zero. |

Submission:

1. Submit this Word document with your answers to the conceptual problems.
2. For the programming task, submit the R Markdown and the HTML report generated from the R Markdown. The HTML report should be neatly organized with a good structure and format.