Hw 2 Report

Gouthamrajan Nadarajan

Problem 1

Using the cereal dataset and fitting a linear model for nutritional rating, I found that the mean square error for the training and testing data was 7.14E-14 and 1.37E-13 respectively. We would normally expect that the testing error would be slightly higher than the training which is what we see in this case. Overall, we can see that a linear model very accurately fits the data in this case.

Performing forward subset selection, I plot in Figure 1 the curve below where the Y axis is the error and the x is the number of variables.

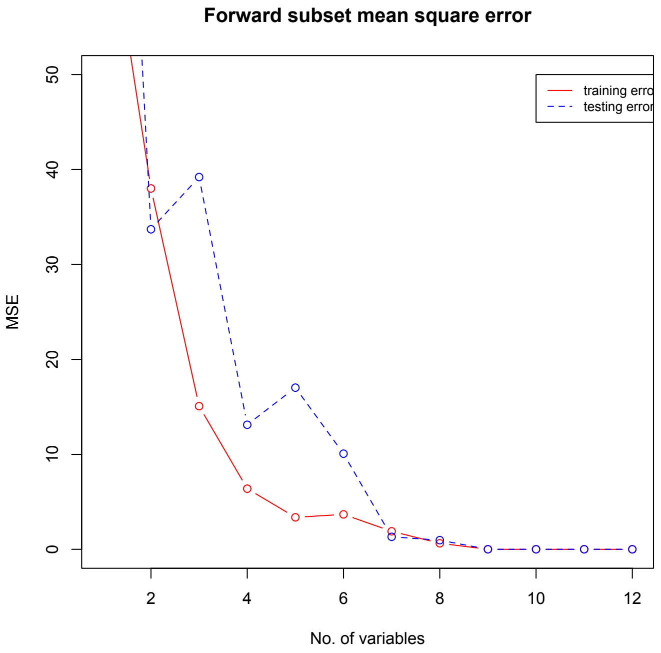


Figure 1: Forward subset selection mean square error

Performing exhaustive subset error on the same dataset, I plot a similar plot of MSE error in Figure 2.

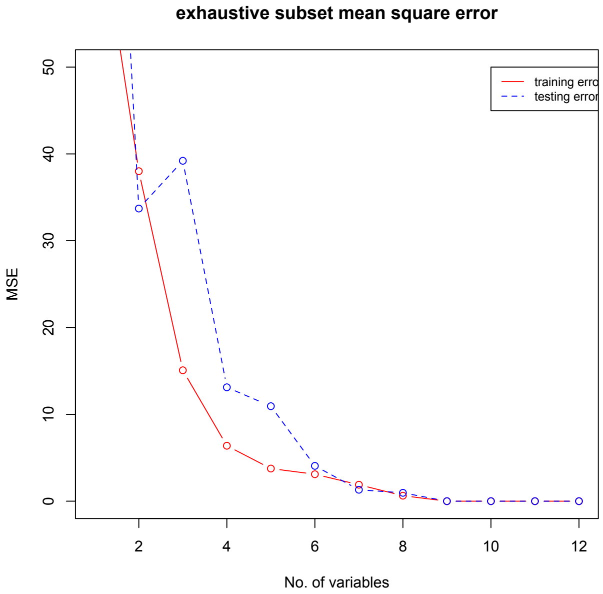


Figure 2: Exhaustive subset selection mean square error

When comparing the testing error between exhaustive subset selection vs. forward subset selection, we can see that exhaustive subset selection has the lower error which is to be expected as exhaustive subset selection looks at all possible combinations of variables. Nonetheless, the differences in the MSE is only noticeable from 4-7 variables and converges to about the same after 8 variables.

Comparing the training vs. testing error for both graphs, you can see it follows very closely for both graphs indicating there is no significant overfitting of the data. For picking the best model to use for predictive purposes, I would pick the exhaustive model using 9 variables. This is because even though the error does decrease slightly more for more variables, having a simpler model will be more generalizable for future use cases.

Problem 2

In this problem, we use the zip code data and compare the classification performance of KNN and linear regression approaches. For preprocessing, I first set the 4's to 0 and set the 7's to 1 for the input data. Then for the model prediction, because linear regression is designed for a continuous response variable, I used a threshold where values greater than 0.5 were given a value of 1 and values less than 0.5 were given a value of 0.

For linear regression**, I found the error rate (# samples misclassified/total samples) for training and testing as 0.7324595 and 0.1148805 respectively.**

For KNN classification, we found the following error rates for the appropriate values of K in table 1.

Value of K	KNN error rate on training data	KNN error rate on testing data
1	0	0.01310717
3	0.0046260601387818	0.0154202
5	0.0053970701619121	0.01773323
7	0.00771010023130301	0.01696222
9	0.00693909020817271	0.01850424
11	0.00925212027756361	0.01850424
13	0.00848111025443331	0.02004626
15	0.00925212027756361	0.02158828

Table 1: KNN error rates

Here we can see that the KNN model performs best when using K= 1 neighbor and the error generally increases for higher values of K. Interestingly, we see that the KNN model outperforms linear regression even when using the highest value of K. Explanations as to why linear regression performs poorly when compared to KNN could be that this dataset has a highly non-linear relationship between the predictors and the response variables that KNN is able to elucidate well but linear regression is not.

Problem 3

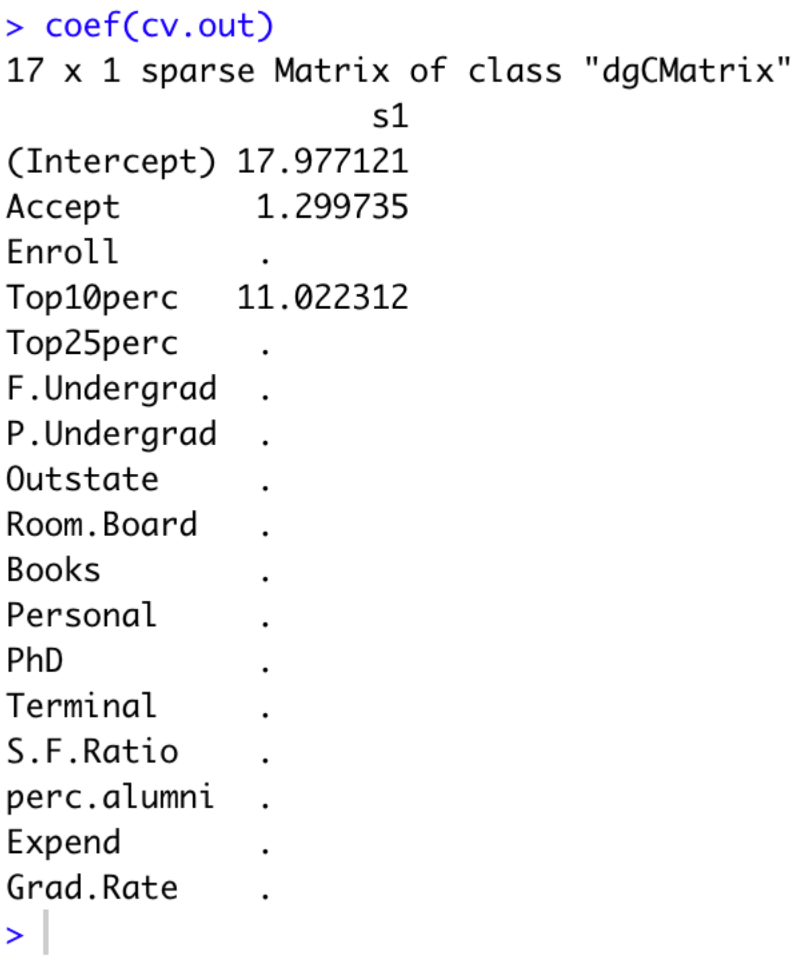
I had to do some preprocessing before working with the College dataset. This included normalizing the variables Acceptance(by the total number of applications) and Enrollment (by the total number of acceptances). I also randomly divided the overall dataset into an 66:33 % split for training and testing.

After fitting a linear model, I found that the testing set mean square error was 5892178.

For ridge regression, we find the best lambda using cross validation of 423.8751. Using this lambda value, I found the mean square error on the testing data to be 1650781. A graph of MSE vs. Lambda is shown in Figure 3

Comparing this to a model trained using lasso regression, I similarly determined lambda using cross validation and obtained a lambda value of 106.2669. I then found the mean square error on the testing data to be 811364.3. A graph of MSE vs. Lambda is shown in Figure 4.

What we notice is that the algorithms with regularization (lasso and ridge) both have a lower error than the baseline linear regression method. Comparing the methods with regularization, lasso regression performed an order of magnitude better. This could be due to the fact that there were many variables that had no effect on the classification and thus setting their coefficients to zero helped classification performance. This is shown below where I have shown the coefficients calculated using lasso.



```
> coef(cv.out)
17 x 1 sparse Matrix of class "dgCMatrix"
                   s1
(Intercept) 17.977121
Accept       1.299735
Enroll       .
Top10perc   11.022312
Top25perc    .
F.Undergrad  .
P.Undergrad  .
Outstate     .
Room.Board   .
Books        .
Personal     .
PhD          .
Terminal     .
S.F.Ratio    .
perc.alumni  .
Expend       .
Grad.Rate    .
> 
```

Interestingly, Of the models that were predicted to be most inaccurate, I found that those colleges were mostly public universities (with the exception of a few). This could be explained by these public universities having predictors that do not map to the response variable in a similar fashion as do private universities. Below I show the universities that had the highest mean square error using the lasso method.

```
> MSE_test_each[order(MSE_test_each, decreasing=TRUE)[1:10],,drop=FALSE]
                                              s1
Brown University                        34153211
Purdue University at West Lafayette     25286802
SUNY at Binghamton                      17475740
Pennsylvania State Univ. Main Campus    16093349
James Madison University                12243791
SUNY College at Cortland                11794096
Yale University                         10862155
University of Massachusetts at Amherst   9097148
Bloomsburg Univ. of Pennsylvania         7599391
University of Central Florida            7013551
```

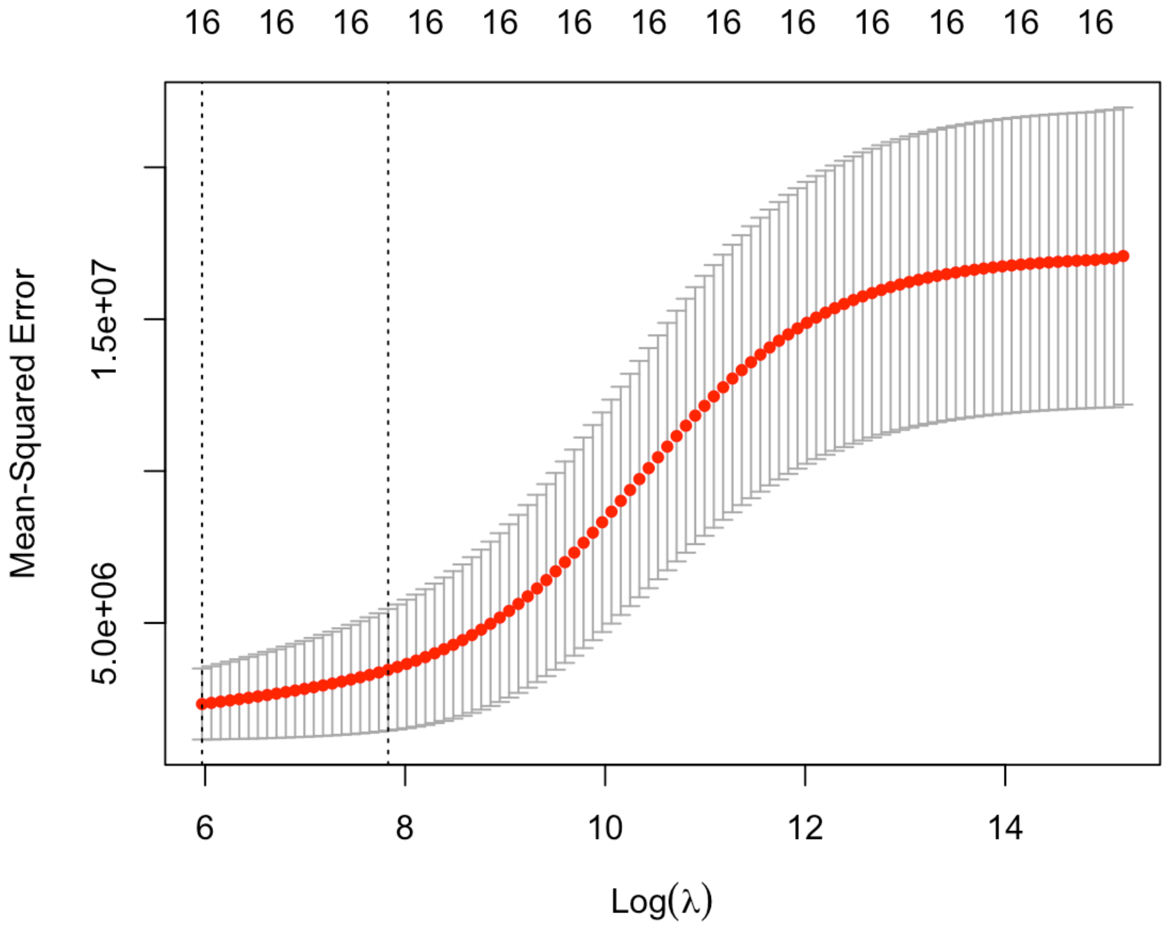


Figure 3: Lasso Regression MSE vs. Lambda. Here as lambda increases, we first see a drastic increase in MSE which then levels off.

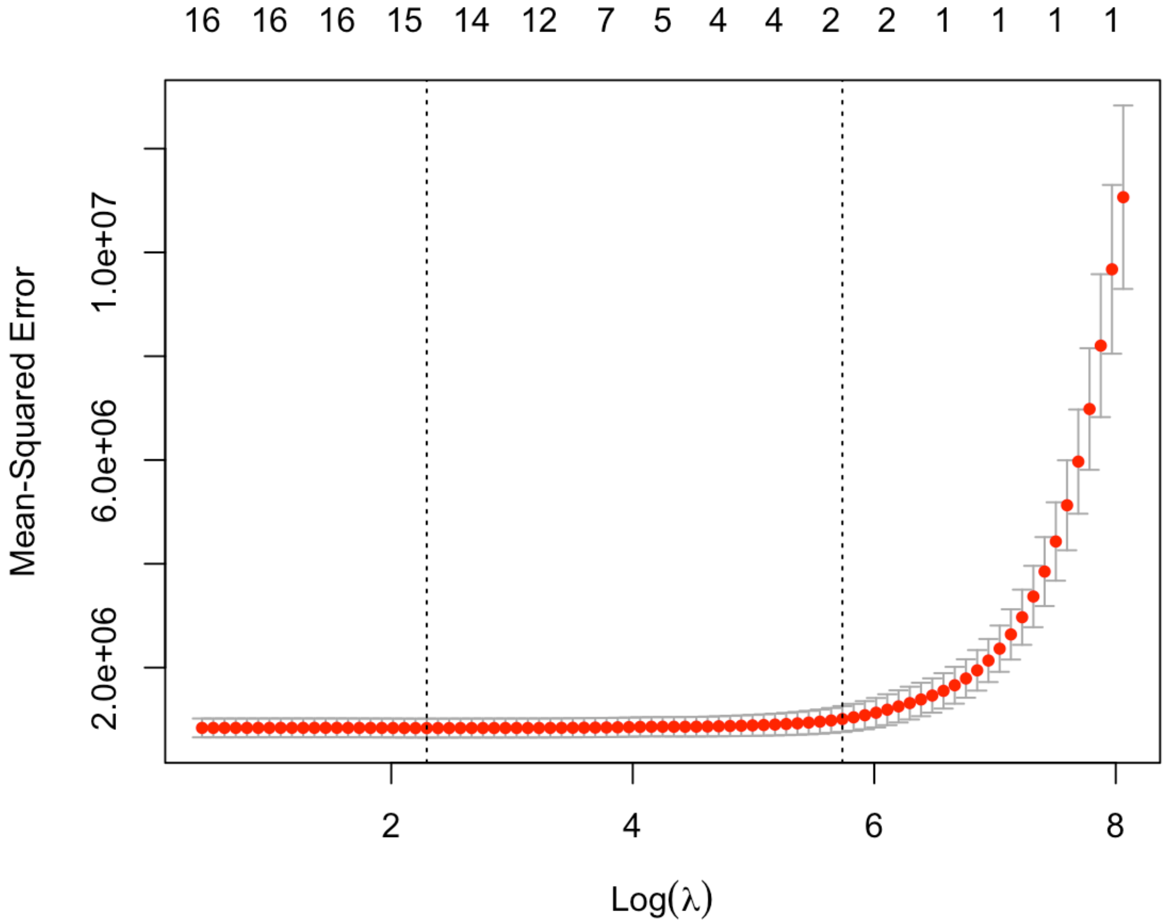


Figure 4: Lasso Regression MSE vs. Lambda. Here we see only small changes as lambda increases until about lambda=6 where we see a drastic shift.