## CS 498 AML – Assignment 6

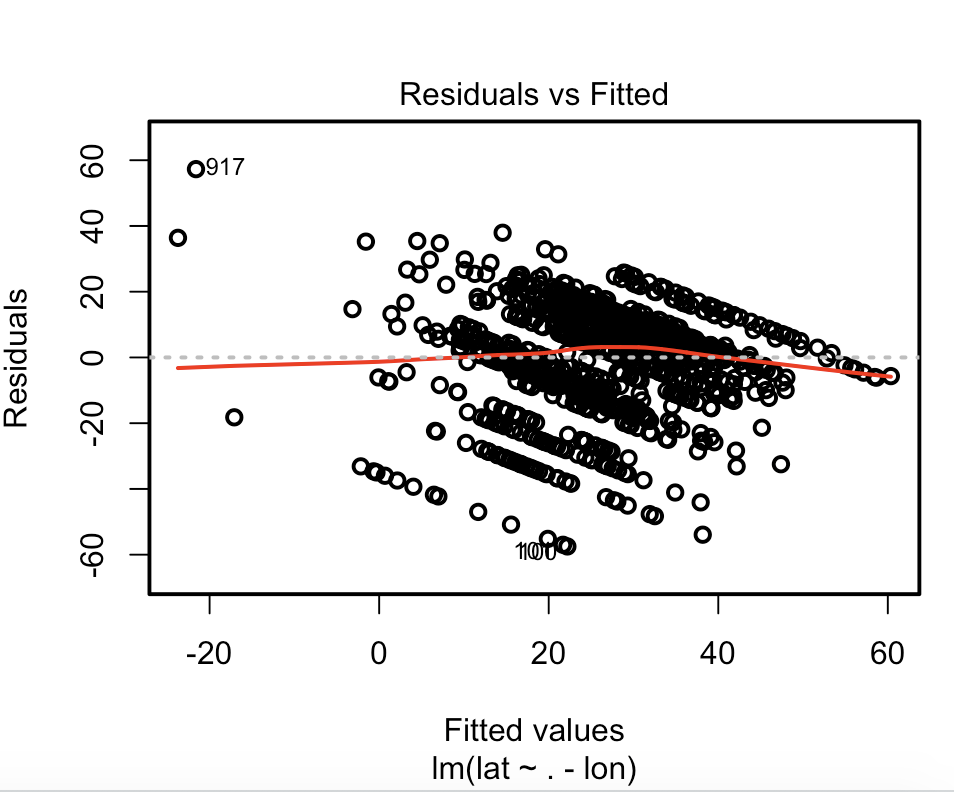
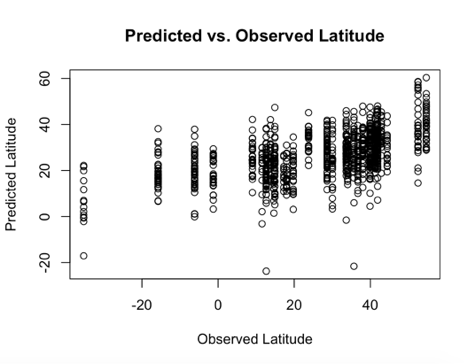
**Group:**

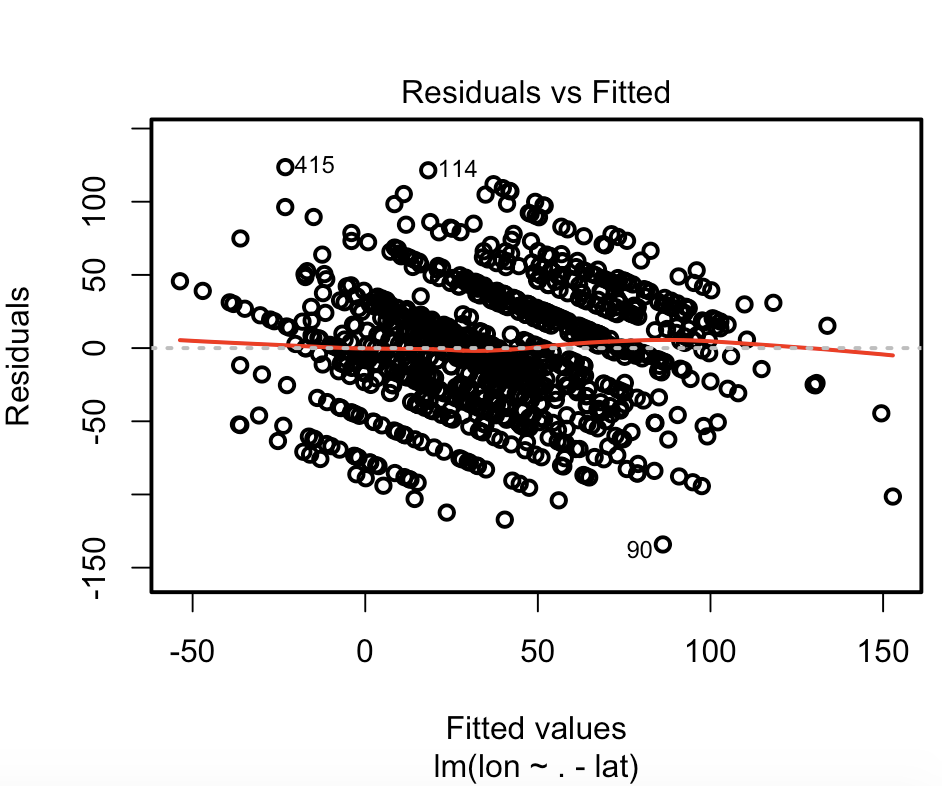
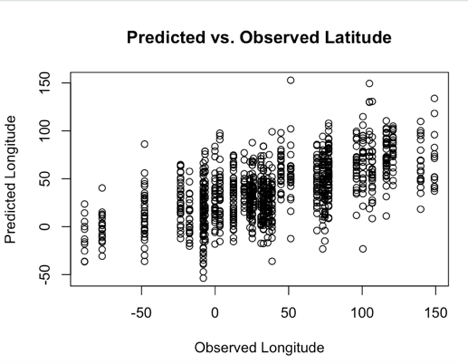
* **Behrooz Moradi Bajestani (bmora5)**
* **Mallory Brandt (mbrandt4)**
* **Yusuf Bhabhrawala (ynb2)**

**1. Linear regression with various regularizers: The UCI Machine Learning dataset repository hosts a dataset giving features of music, and the latitude and longitude from which that music originates**[**here**](https://archive.ics.uci.edu/ml/datasets/Geographical+Original+of+Music)**. Investigate methods to predict latitude and longitude from these features, as below. There are actually two versions of this dataset. Either one is OK by me, but I think you'll find the one with more independent variables more interesting. You should ignore outliers (by this I mean you should ignore the whole question; do not try to deal with them). You should regard latitude and longitude as entirely independent.**

**First, build a straightforward linear regression of latitude (resp. longitude) against features. What is the R-squared? Plot a graph evaluating each regression.**

**Source code:** hw6-part1-lm-glmnet.R

* 





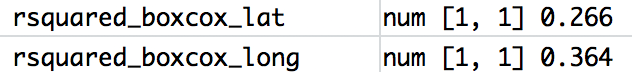
**Does a Box-Cox transformation improve the regressions? Notice that the dependent variable has some negative values, which Box-Cox doesn't like. why do you say so?**

**Source code:** hw6-part1-boxcox.R

UNTOUCHED R-SQUARED VALUES



BOXCOX R-SQUARED VALUES



After resetting the origins of latitude and longitude by adding 90 degrees and 180 degrees, respectively, performing our prediction and fitting, then resetting the origins, the boxcox transformation appears to be giving a slightly lower R-squared value than the original data. From this observation, the boxcox transformation does not appear to benefit the models.

**Use glmnet to produce:**

* **A regression regularized by L2 (equivalently, a ridge regression). You should estimate the regularization coefficient that produces the minimum error. Is the regularized regression better than the unregularized regression?**
* **A regression regularized by L1 (equivalently, a lasso regression). You should estimate the regularization coefficient that produces the minimum error. How many variables are used by this regression? Is the regularized regression better than the unregularized regression?**
* **A regression regularized by elastic net (equivalently, a regression regularized by a convex combination of L1 and L2). Try three values of alpha, the weight setting how big L1 and L2 are. You should estimate the regularization coefficient that produces the minimum error. How many variables are used by this regression? Is the regularized regression better than the unregularized regression?**

**Source code:** hw6-part1-lm-glmnet.R

Glment() function is used on raw data in order to to predict latitude and longitude from music features. According to the documentation of the cv.glmnet() function, the fuction does a 10-fold cross validiation by splitting the data into 10 sets, fitting the moder on 9 sets and evaluating on the one, repaeating this process 10 times, changing the sets each round. Therefore, the MSE values are obtained by fitted model being tested on one fold of data not used for training the model.

Two sets of linear regression models are fitted for latitude and longtitude vs. 166 features. A set of linear regression models fitted for each case include one cross validated unregurized linear model and five cross validated regulrized linear model: ridge (alpha = 0), lasso (alpha = 1), elastic net (alpha = 0.25) , elastic net (alpha = 0.50) , elastic net (alpha = 0.75). For being able to compare the results of the models with each other, all the models are fitted using the cv.glmnet() function and mean squared errors (MSE) reported by the function are obtained and analysed. In case of the cross validated unregurized linear model, the lambda and alpha coefficents are set to zero for cv.glmnet() function which according to the document of function, will cause function to perform as an unregularized linear model with no regularizer penalty. Table 1 and 2 show the MSE results obtained by fitting a set of linear models for latitude and longtitude labels, respectively.

Table 1. MSE results obtained by fitting linear models for latitude vs. 116-feature matrix

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Latitude | | | | | |
| Unregularized | Regularized - ridge | Regularized - lasso | Regularized - elastic net | Regularized - elastic net | Regularized - elastic net |
| Best Lambda | 0 | 5.78 | 0.458 | 1.83 | 0.916 | 0.611 |
| Alpha | 0 | 0 | 1 | 0.25 | 0.5 | 0.75 |
| MSE | 285 | 280 | 279 | 280 | 280 | 277 |
| Non-zero Variables | 116 | 116 | 21 | 34 | 22 | 21 |

Table 2. MSE results obtained by fitting linear models for longtitude vs. 116-feature matrix

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Longtitude | | | | | |
| Unregularized | Regularized - ridge | Regularized - lasso | Regularized - elastic net | Regularized - elastic net | Regularized - elastic net |
| Best Lambda | 0 | 4.5 | 0.27 | 1.43 | 0.539 | 0.327 |
| Alpha | 0 | 0 | 1 | 0.25 | 0.5 | 0.75 |
| MSE | 1929 | 1873 | 1873 | 1887 | 1882 | 1851 |
| Non-zero Variables | 116 | 116 | 58 | 97 | 88 | 86 |

In case of linear regression regularized by L2 (equivalently, a ridge regression), the best regularization coefficient Lambda for latitude and longtitude values are 5.78 and 4.5 producing a minimum MSE values of 280 and 1873 (see Table 1 and 2). Comparing the regularized ridge regression with unregularized linear model, it can be concluded that ridge regularization is improving the accuracy of the model.

For linear regression regularized by L1 (equivalently, a lasso regression), the best accuracies are obtained with regularization coefficients of Lambda 0.458 and 0.27 along with alpha of 1 for latitude and longtitude, respectively. Obtained minimum MSE values corresponding to these regularization coefficients are 279 and 1873 for latitude and longtitude, respectively (see Table 1 and 2). Comparing the regularized lasso regression with unregularized linear model, it can be concluded that lasso regularized regression is better than the unregularized ones. In case of lasso regularized regression, 21 and 58 non-zero variables are used for latitude and longtitude cases, respectively.

Finally, in regard to linear regression regularized by elastic net (equivalently, a regression regularized by a convex combination of L1 and L2) three alpha values of 0.25, 0.5 and 0.75 are used. The regularization coefficient that produces the minimum error are provided in Table 1 and 2 for three various values of alpha along with number of variables used by each regression model. Comparing the MSE values, it can be conluded that the best regression model is obtained by using an elastic net regularization technique and this methods definitely is better that the unregularized regression model.

**2. Logistic regression: The UCI Machine Learning dataset repository hosts a dataset giving whether a Taiwanese credit card user defaults against a variety of features** [**here**](http://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients)**. Use logistic regression to predict whether the user defaults. You should ignore outliers, but you should try the various regularization schemes we have discussed.**

**Source code:** hw6-part2.R

Raw data for Taiwanese credit card user defaults database are downloaded from the webpage and the headres are modified in the downloaded xls file to remove header rows (more than one). The data file is then converted to csv format (also provided in the submitted homework). The .csv raw data file is read into the R and the database is created. As preparation of the databse, the feature database is generated from the 23 feature columns and labels are converted to factors with two classes of 0 and 1. The label and factor datasets are then used for cv.glmnet() function of R with parameters of family="binomial", type.measure = 'class' which performs similar to a binary logistic regression analysis. Performance of the binary logistic regression is evaluated using the misclassification error reported by the cv.glmnet() function.

According to the documentation of the cv.glmnet() function, the function does a 10-fold cross validiation by splitting the data into 10 sets, fitting the model on 9 sets and evaluating on the 1, repeating this process 10 times, changing the sets each round.

First an unregularized linear model is fitted by assigning lambda and alpha coefficients of zero (0) to cv.glmnet() function. According to the Eq. 1. mentioned below, by feeding zero values for lambda and alpha coefficients to the cv.glmnet(), it will behave as if an unregulrized linear model is fitted.

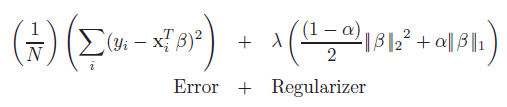


Table 1 below shows results for the unregularized linear model fitted for Taiwanese credit card user defaults database.

Table 1. Results for unregularized linear model fitted on Taiwanese credit card user defaults database.

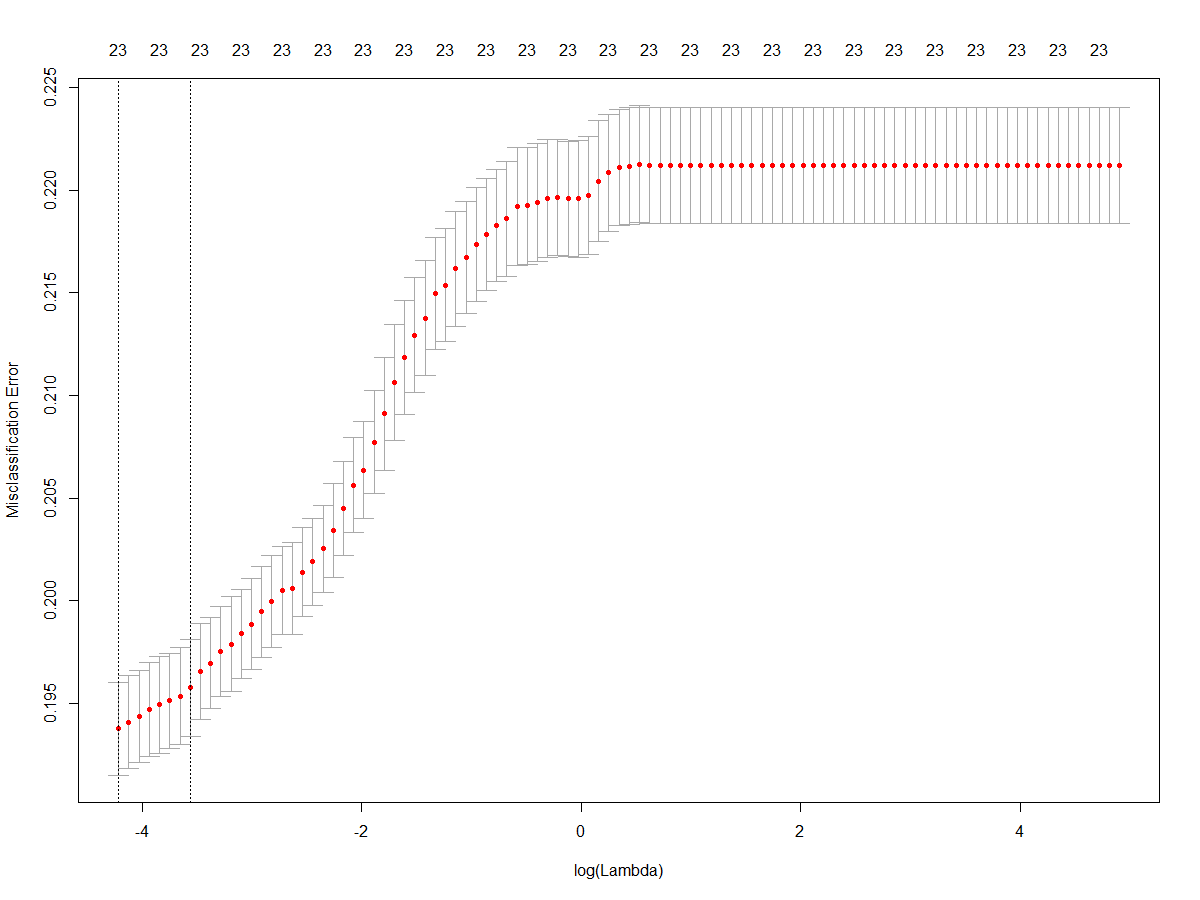
|  |  |
| --- | --- |
|  | Unregulrized Linear Model |
| Alpha | 0 |
| Lambda | 0 |
| Misclassification Error | 0.1895333 |
| Accuracy | 0.8104667 |

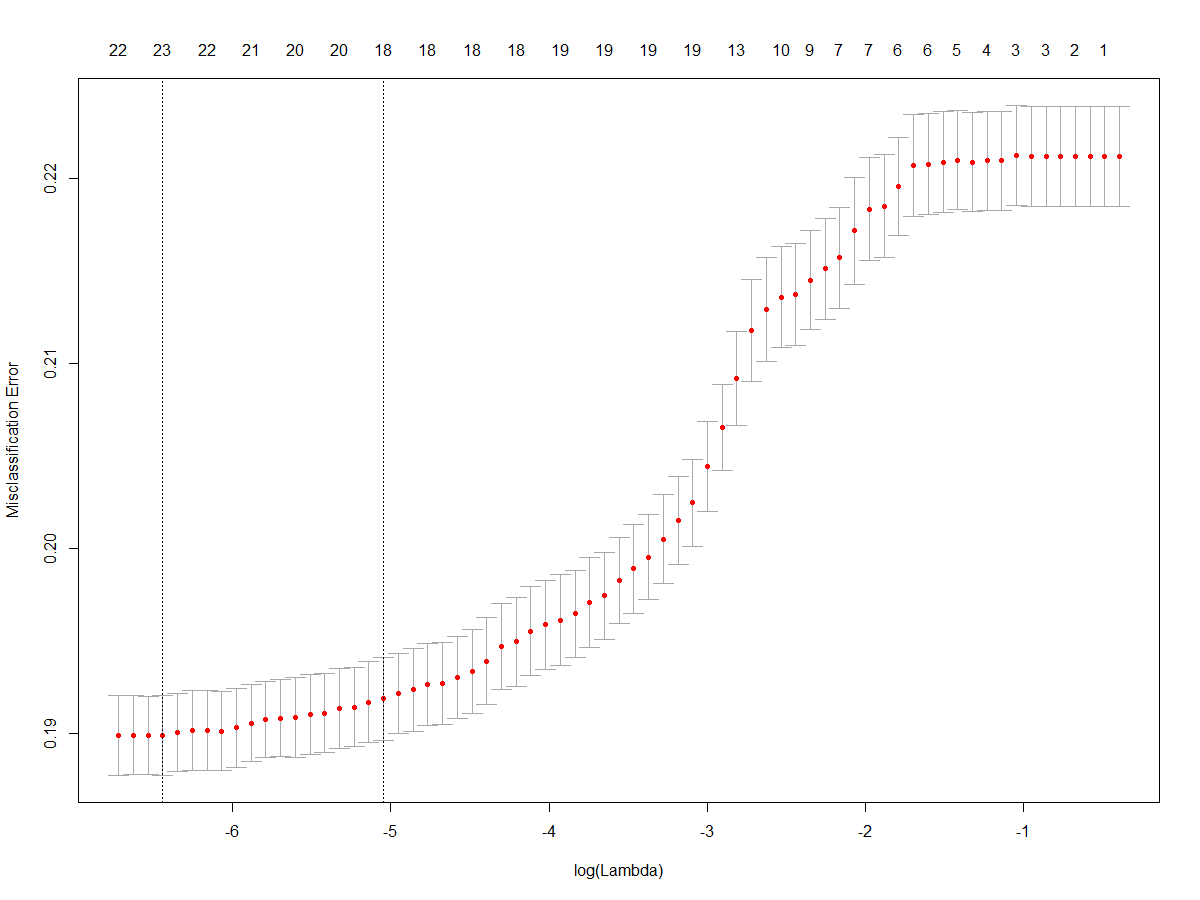
To study performance of the regulrized linear models, cv.glmnet() function is used with various values of lambda and alpha. For ridge regression alpha of zero, for lasso regression an alpha of 1 and for elastic net regression alpha values between 0 and 1 are used. In these cases, the lambda values are not determined for the function and sc.glmnet searches for the best lambda value itself. It is worth mentioning that for each case of specific lambda and alpha values, a 10-fold cross validation is performed by the function. The results of the regularized linear models are provided in Table 2. A selection of the cross validated misclassification error results are presented in Figures 1 to 6 below.

According to the obtained results, the accuracies does not vary significantly ranging only from 78% to 81%. The maximum accuracy (resp. least misclassification error) is 81.07 % obtained from 4th scenario of the elastic net regularized regression in which an alpha of 0.3 is used. Also it can be seen that the regression accuracy is slightly deacrased by using regularization and unreglrized regression models stands in the second place with a misclassification error of 81.05%.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Scenario 1  (ridge reg) | Scenario 2  (Elastic net) | Scenario 3  (Elastic net) | Scenario 4  (Elastic net) | Scenario 5  (Elastic net) | Scenario 6  (Elastic net) | Scenario 7  (Elastic net) | Scenario 8  (Elastic net) | Scenario 9  (Elastic net) | Scenario 10  (Elastic net) | Scenario 11  (lasso reg) |
| Alpha | 0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1 |
| Lambda for Minimum MCE | 0.0147951 | 0.0013798 | 0.0013232 | 0.0010625 | 0.0008746 | 0.0005293 | 0.0006399 | 0.0011545 | 0.0006963 | 0.00061892 | 0.00055703 |
| Misclassification Error | 0.1934333 | 0.1900333 | 0.1897333 | 0.1893333 | 0.1898667 | 0.1898 | 0.1898667 | 0.1898667 | 0.1897333 | 0.18963333 | 0.1898 |
| Accuracy | 0.8065667 | 0.8099667 | 0.8102667 | 0.8106667 | 0.8101333 | 0.8102 | 0.8101333 | 0.8101333 | 0.8102667 | 0.81036667 | 0.8102 |

Table 2. Results for regulrized linear model fitted on Taiwanese credit card user defaults database

Figure 1. Cross validated misclassification errors for various Lamdas and Alpha of 0 (ridge regression)

Figure 2. Cross validated misclassification errors for various Lamdas and Alpha of 0.2 (elastic net regression)

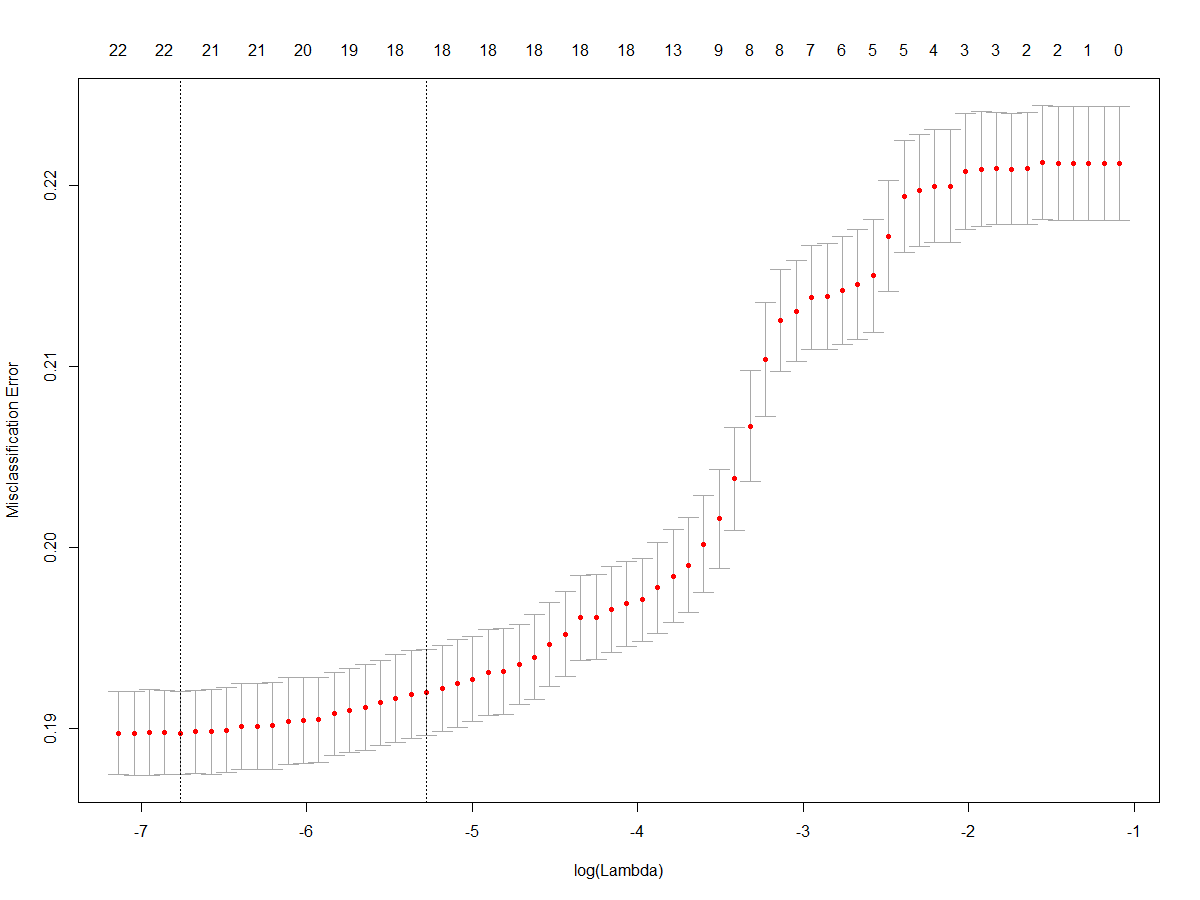


Figure 3. Cross validated misclassification erros for various Lamdas and Alpha of 0.4 (elastic net regression)

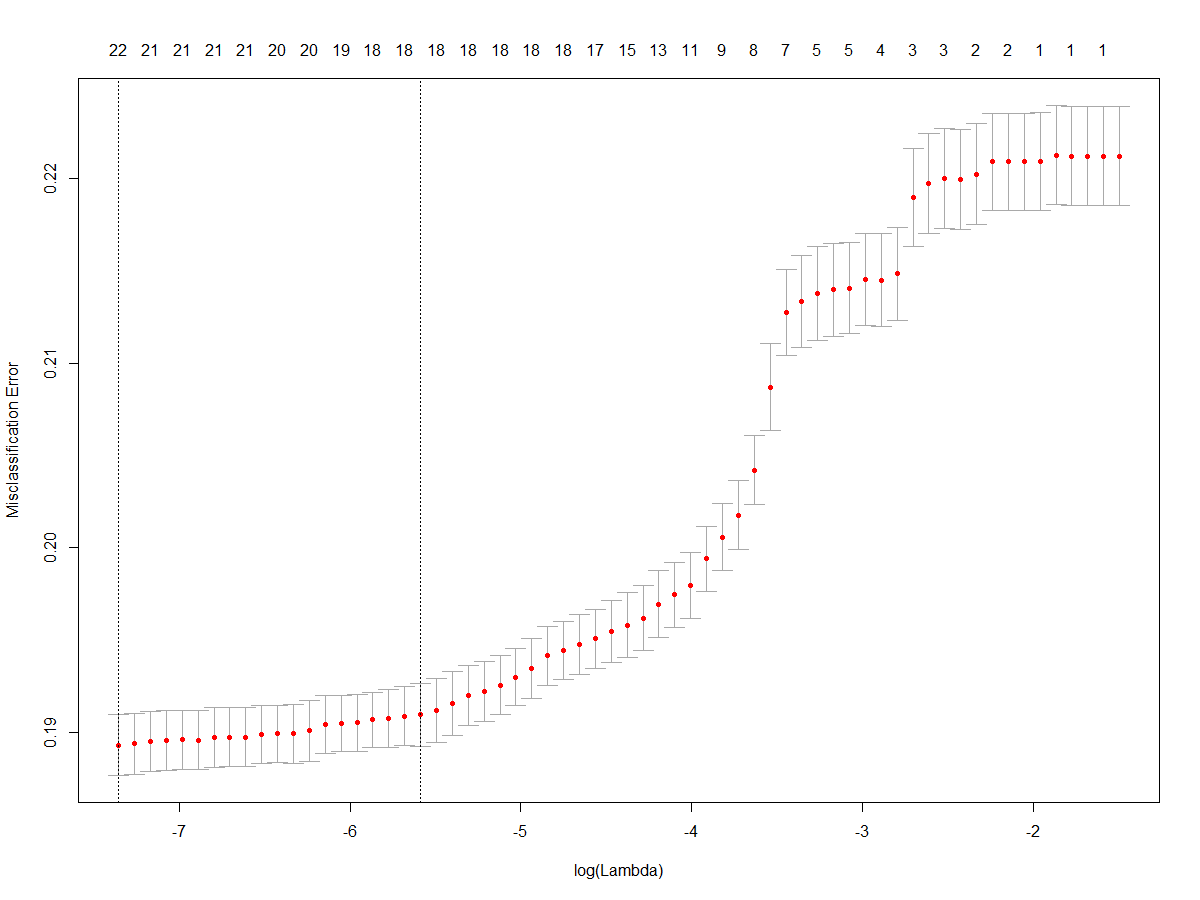


Figure 4. Cross validated misclassification errors for various Lamdas and Alpha of 0.6 (elastic net regression)

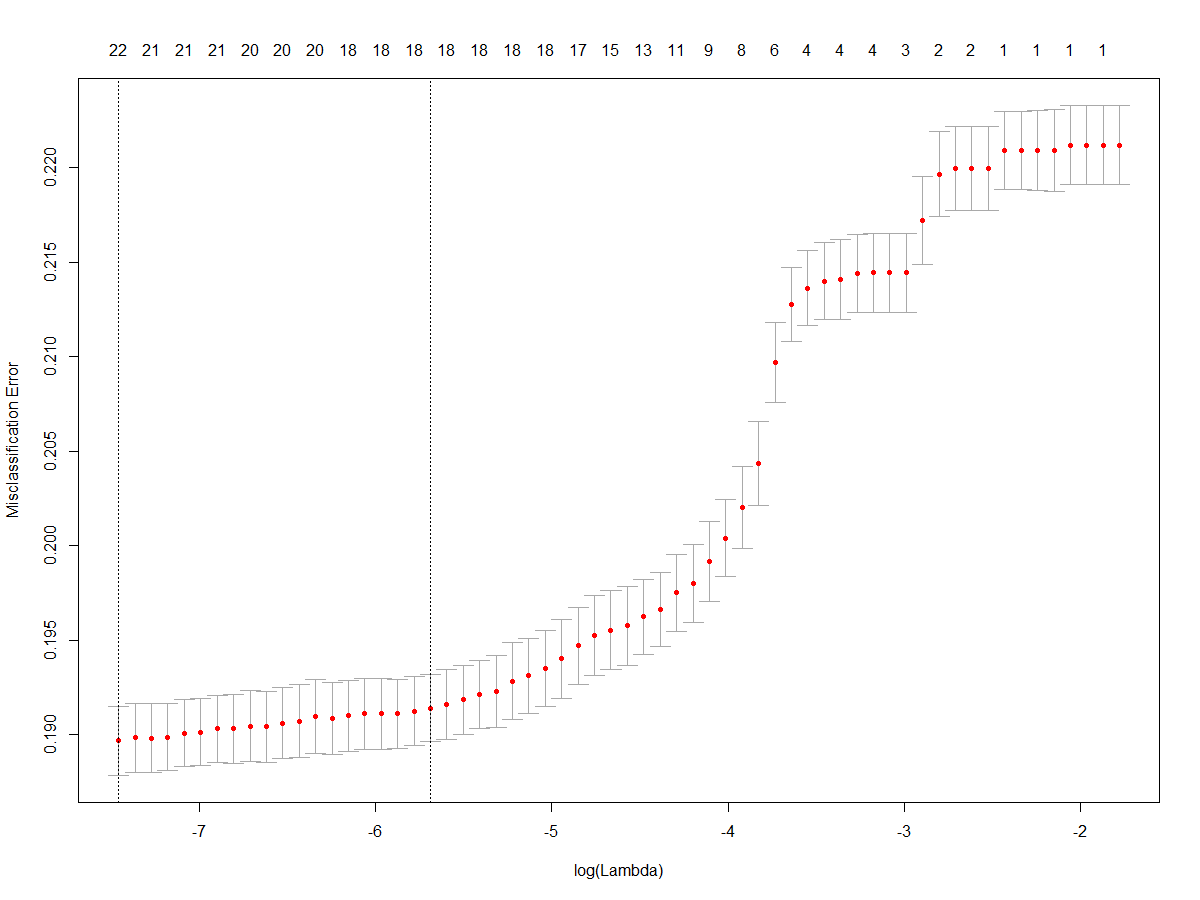


Figure 5. Cross validated misclassification errors for various Lamdas and Alpha of 0.8 (elastic net regression)

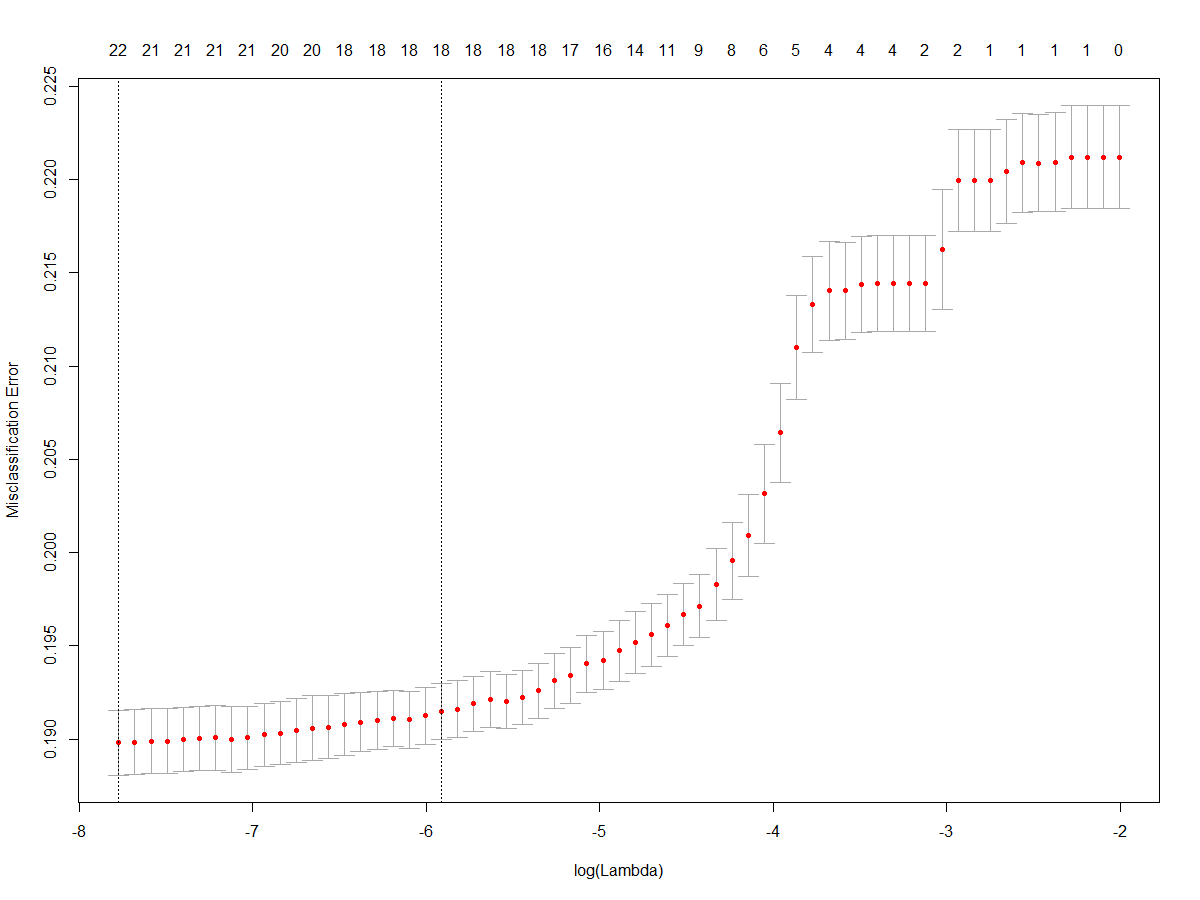


Figure 6. Cross validated misclassification errors for various Lamdas and Alpha of 1.0 (lasso regression)