**Student name: GOUTHAM SELVAKUMAR  
   
DSC 324/424 Assignment 2  
Due: Tuesday, April 12th at 11:59pm CST**

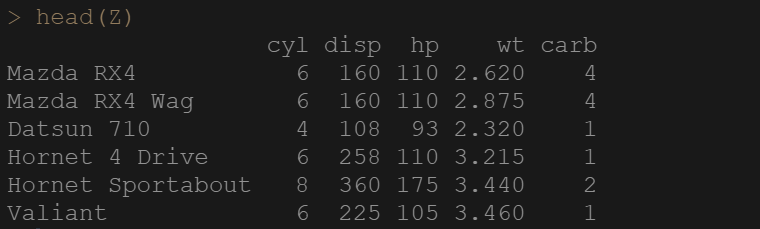
Complete the following problems and submit your answers in a single document.

All your answers should be supported with screenshots of any relevant outputs or graphs. When making conclusions, always reference specific numbers and your interpretations of them.

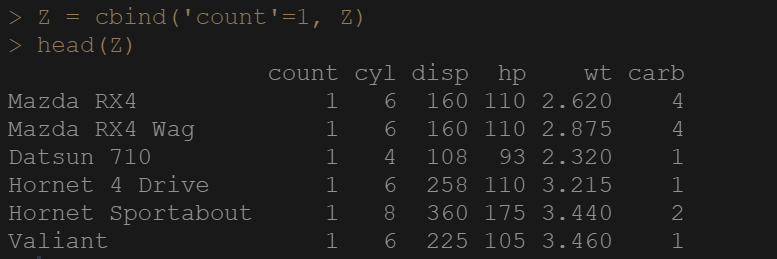
**1. Math of regression beta coefficients**: Use R to calculate the betas of the following dataset. Include your code and screenshot of the final output. Use the following steps as a guide:

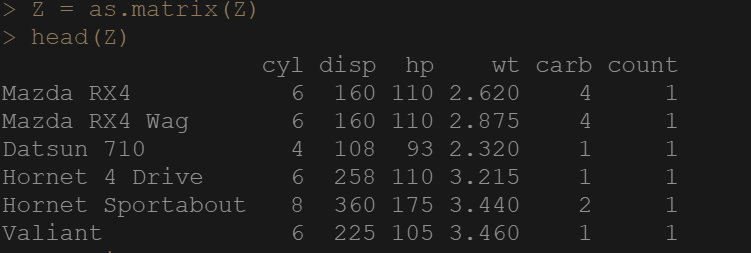
Using the dataset "mtcars" (built-in to RStudio), do the following:

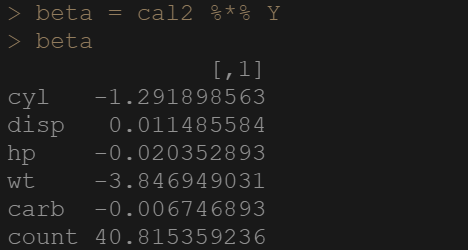
1. Copy the original dataset into a new variable Z with only these columns: cyl, disp, hp, wt, carb.

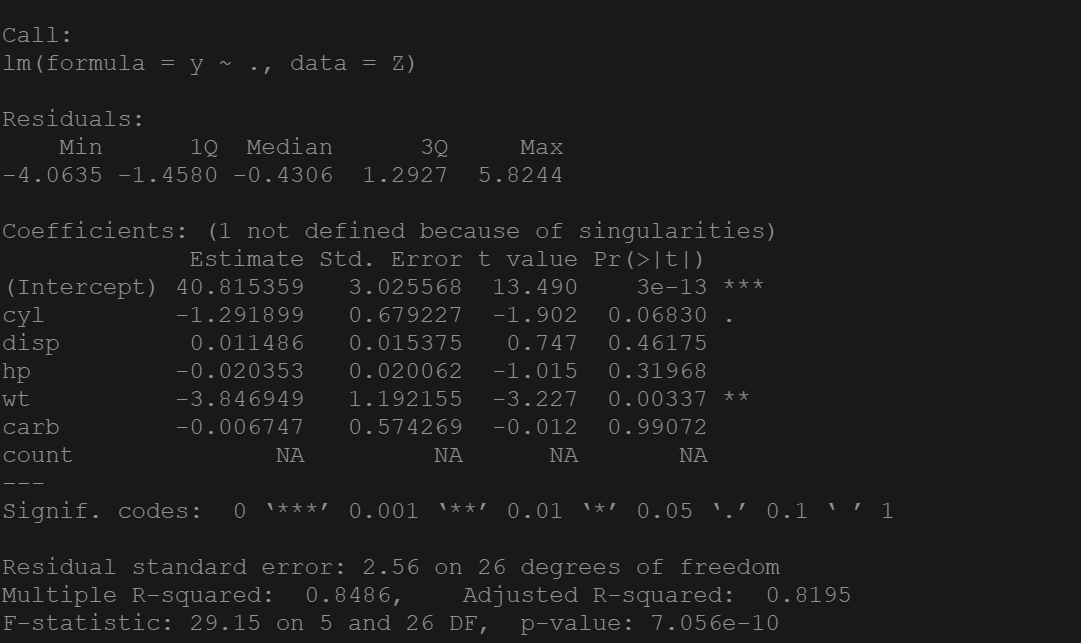
  
**b)** Copy the original dataset again into a new variable Y with only the mpg column.

  
**c)** Add a column of 1's to the front of your Z dataset. Name this variable "count".

  
**d)** Convert Z and Y into matrices.

  
**e)** Using the matrix operations in R, compute the beta coefficients for a regression predicting the mpg variable in vector Y using the variables in matrix Z. Include a screenshot of your results.  
Here's the formula: (ZTZ)-1ZTY

  
**f)** Now compute the same regression using the lm () function in R and include a screenshot of the coefficients. Are they the same as your manual calculations?

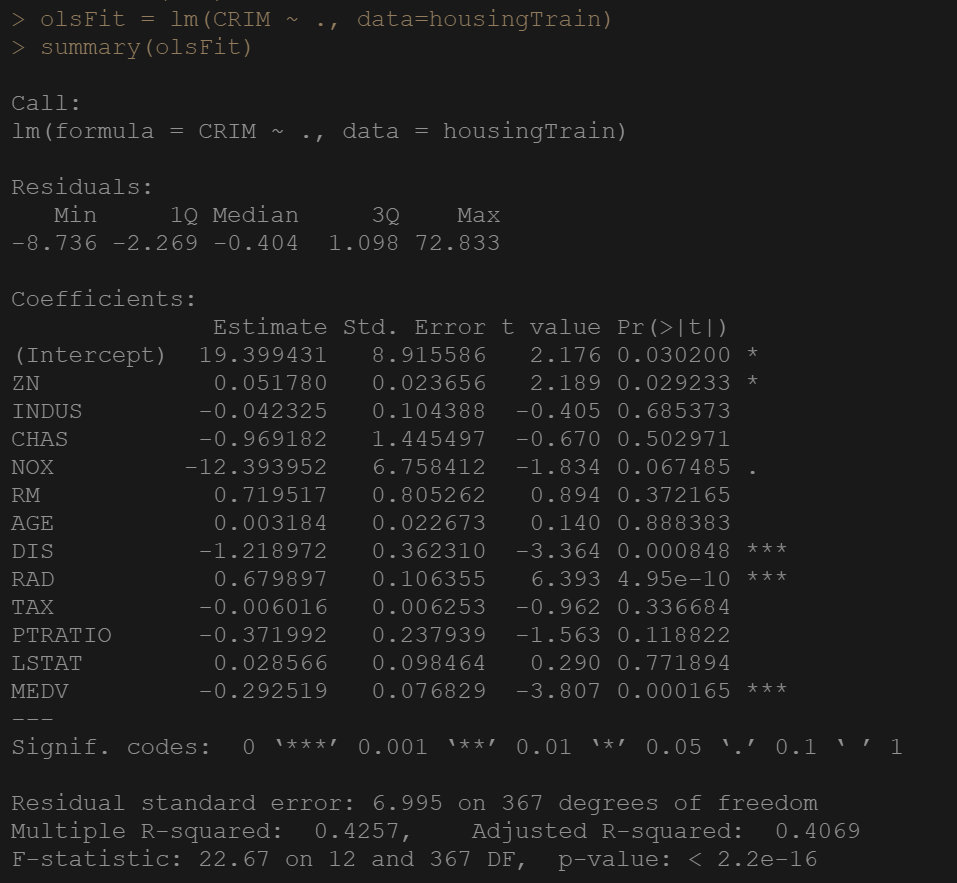


**From the significance code you can see that “wt” which is the weight of the car which is highly significant for ‘mpg’ mileage per gallon. Which is similar to the answer “D” that if weight increases by 1 unit then mileage of the car goes down by 3.846949 miles per gallon.**

**2. Ridge and Lasso Regressions**

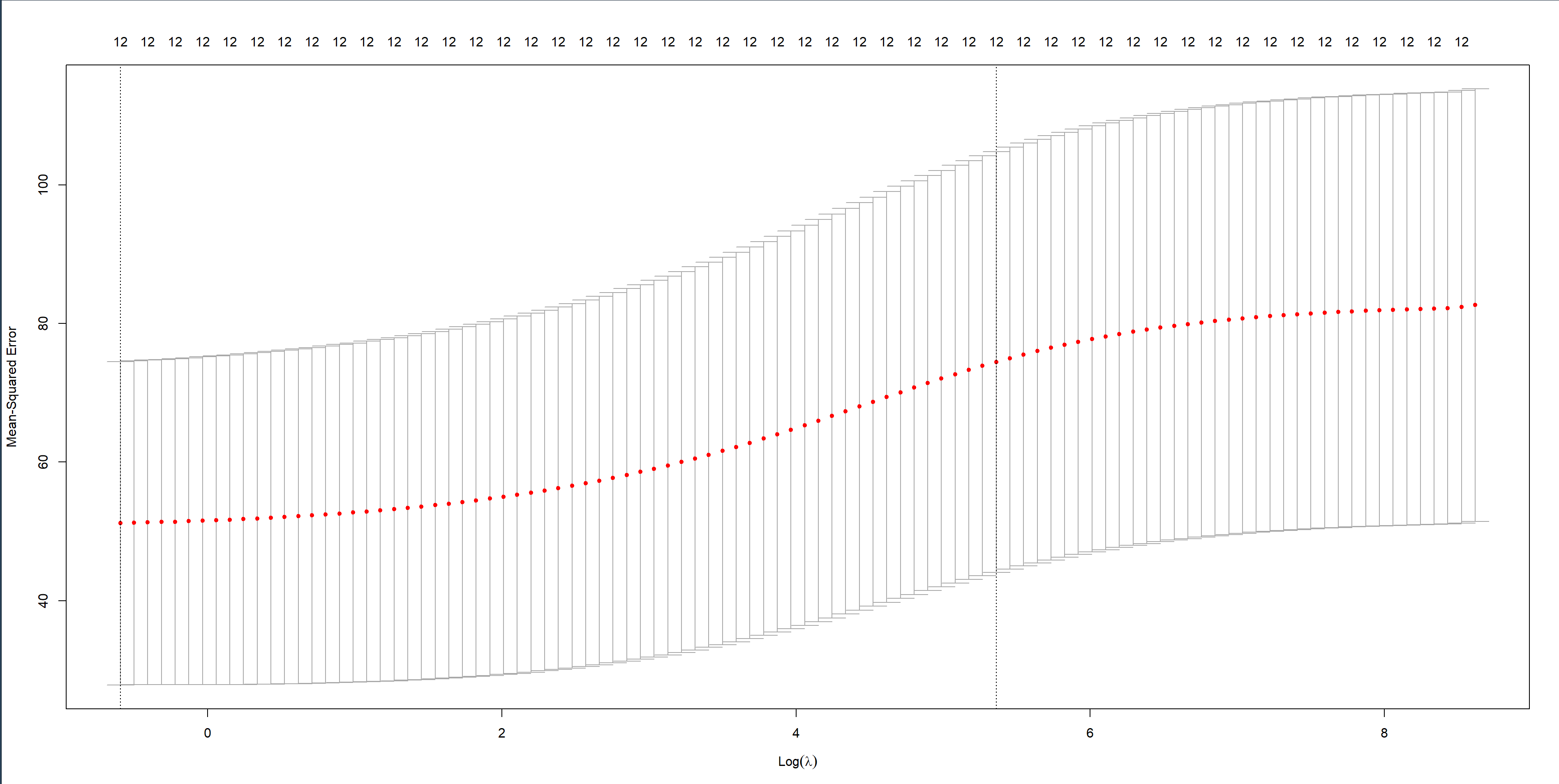
For this question, use "housingTrain.csv" for the training set and "housingTest.csv" for the test set. Investigate the effects of regularized regression with the following problems:

1. Run an OLS linear regression model on MEDV using the training set, then use your model to predict the values of the test set. Evaluate this model using the adjusted R2 of the training set and the RMSE values of both the training and test sets. Is there evidence of overfitting?



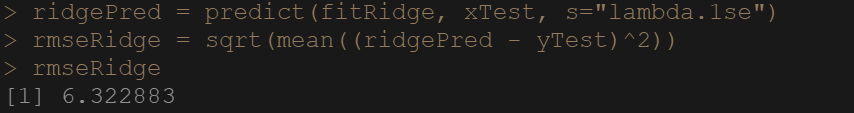
**From the O/P, we can say that there is notable overfitting. You can tell that the R-Square value of the full model and comparing the testing and training set.**

**b)** Use cross-validated ridge regression on the training set and plot the relationship between the cross-validated error and the log-lambda values. What can you discern from this graph? Include a screenshot.



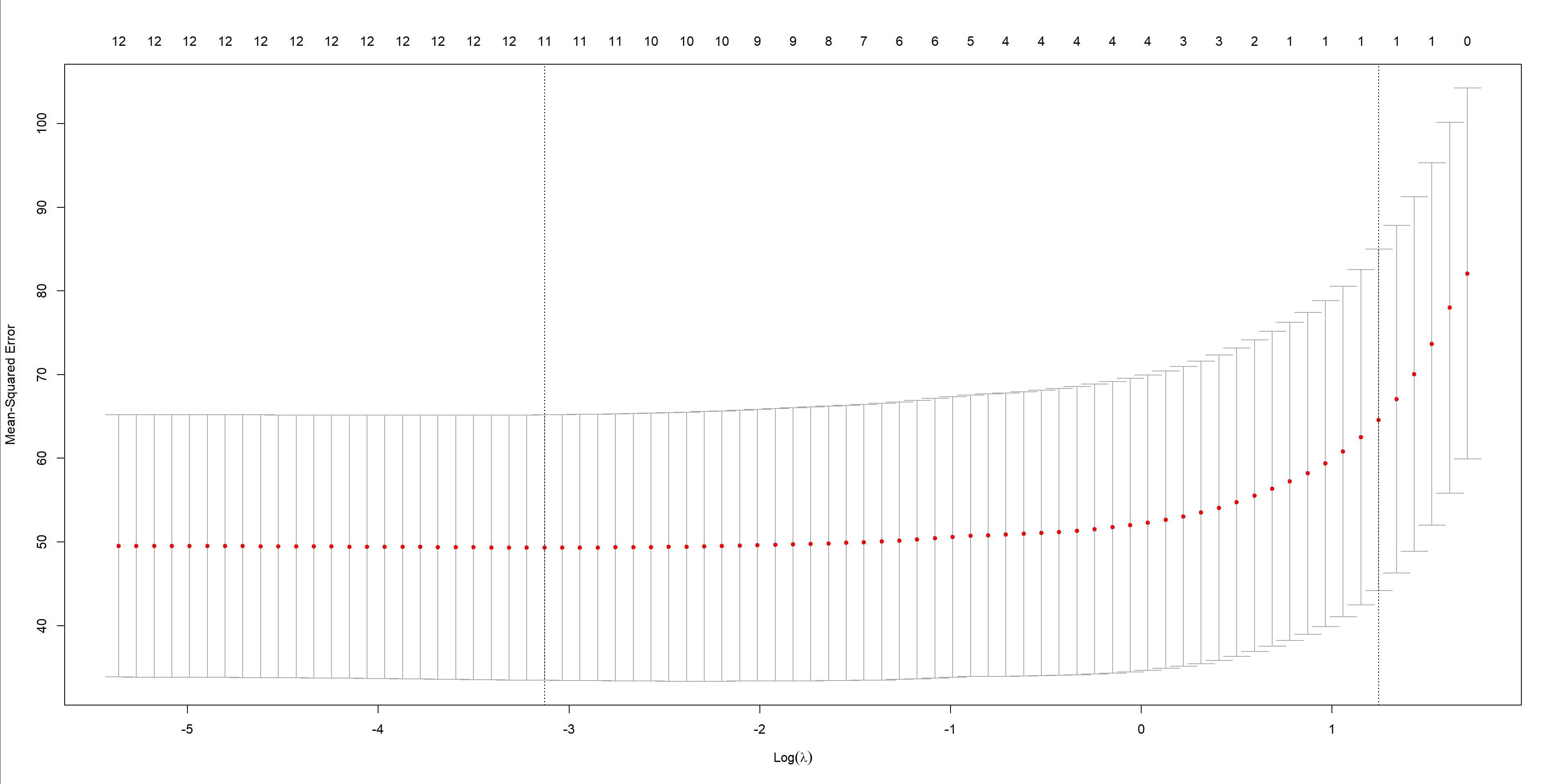
**From the results, you can say that the OLS model is not better than ridge regression as you can see the graph the RMSE value of the test data set. So, we can finalize that after analysis even though the model is slightly overfitted.**

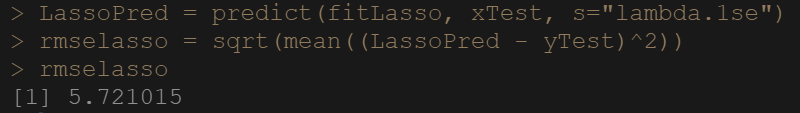
**c)** Use the model you built in b) to predict the values of the test set using the "lambda.1se" value. Evaluate this model using the same metrics as you did in a). How do these compare to the OLS regression model? How well is regularization working here? Be specific.

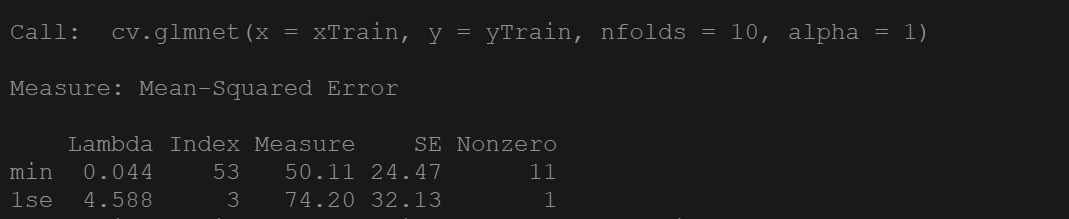


**RMSE Value for the ridge is 6.322883 which I have done for the test data set.**

**d)** Repeat parts b) and c) using a Lasso regression and compare your results. Is Lasso performing better or worse than the OLS and ridge regressions?





****

**We are following with the same steps but this time it is included with the lasso regression. From the graph we can see the value of the lambda in min is 0.044 and lambda 1se is 4.588. Lasso regression is predicted better as the new model is less overfitted and improving the regression.**

**e*)*** *Evaluate how the number of variables changes with the lambda values in your Lasso regression. How many variables are selected at lambda.1se? How much of the variance is captured at lambda.1se?*

As the lambda increases the number of variables decreases. At lambda.1se 5 variables were selected.

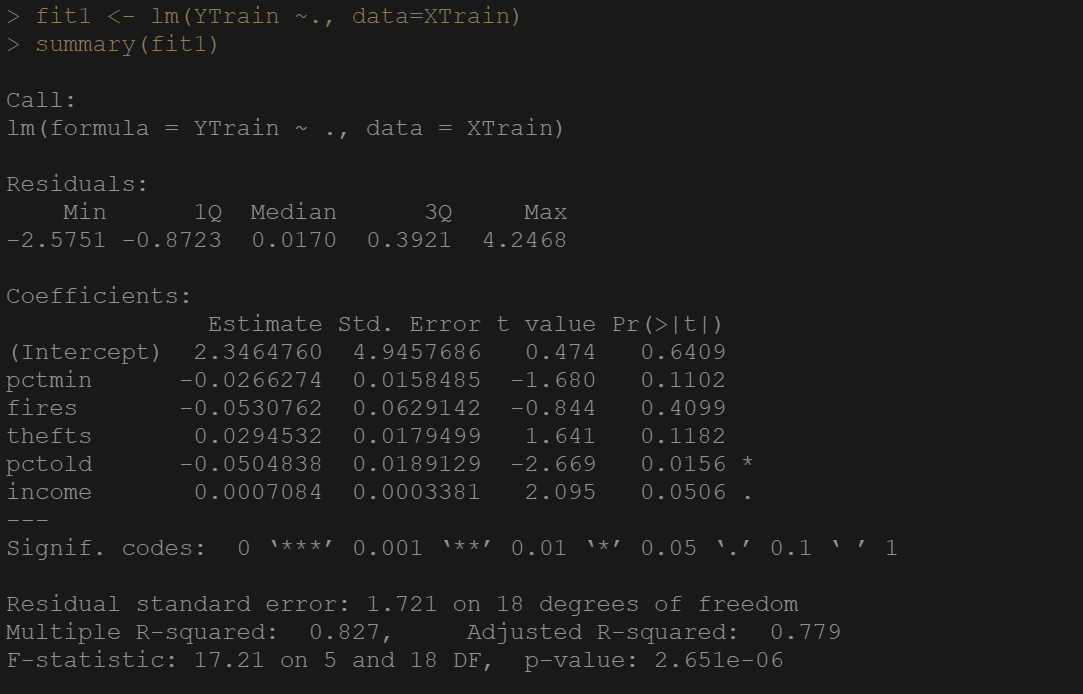
**g)** If you had to pick only one model to use from these three, explain which you would choose and why.

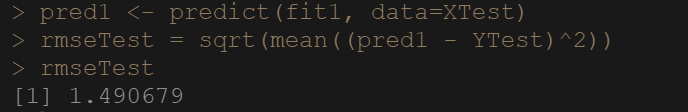
**I would choose the model-2 since the lambda value is reduced and the graph shows that the regression is reduced too.**

**3.** **Elastic Net Regression**

We'll now return to the insurance dataset from the Module 1 homework. This has also been split into training and test sets ("insurTrain.csv" and "insurTest.csv" respectively).

1. Run a multiple regression of NEWPOL using the following as independent variables: PCT-MINOR, FIRES, THEFTS, PCTOLD, and INCOME. Use the model to predict the test set. Can you find any evidence of overfitting?



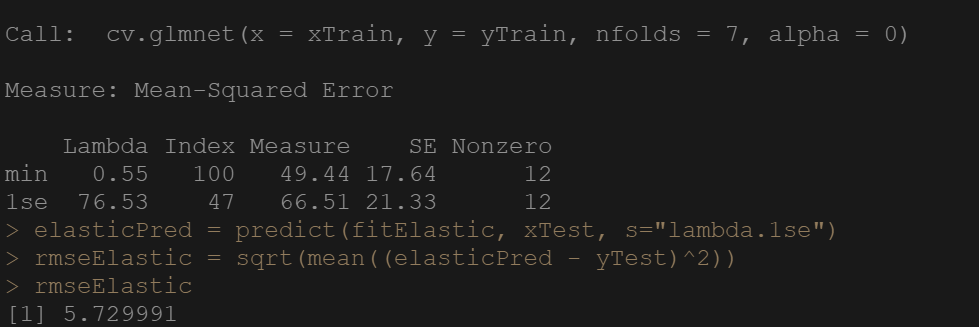


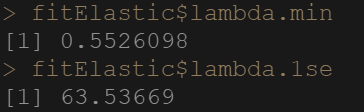
* **The RMSE value for the test set is 1.490. This indicates that model performed too well on test set than train set and it is a sign of multicollinearity.**
* **The R-Squared of the model is 0.827 which is pretty good. It can be inferred that 82.7% of variability in NEWPOL can be explained by the variables. The Pr value of pctmin, fires, and thefts are not significant. Both pctold, and income are significant under 5% confidence interval.**

**b)** Run five Elastic Net regressions at the following alpha values: 0, 0.25, 0.5, 0.75, 1

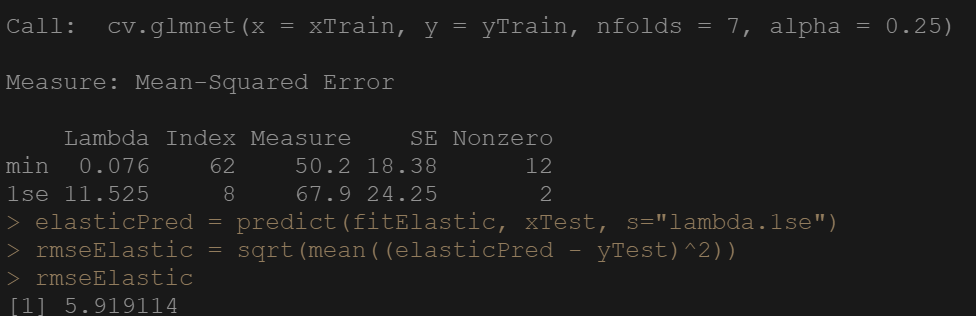
**The following code runs the specific alpha values respectively,**

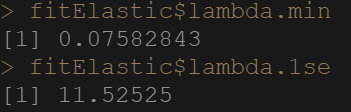
**For alpha value 0,**



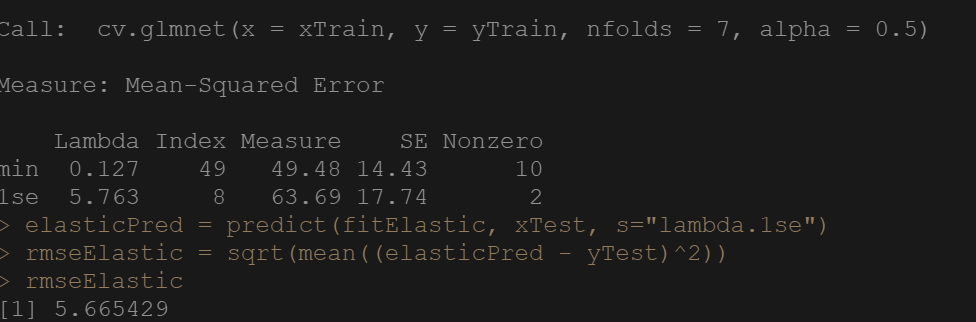
****

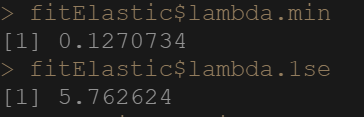
**For alpha value 0.25,**

****

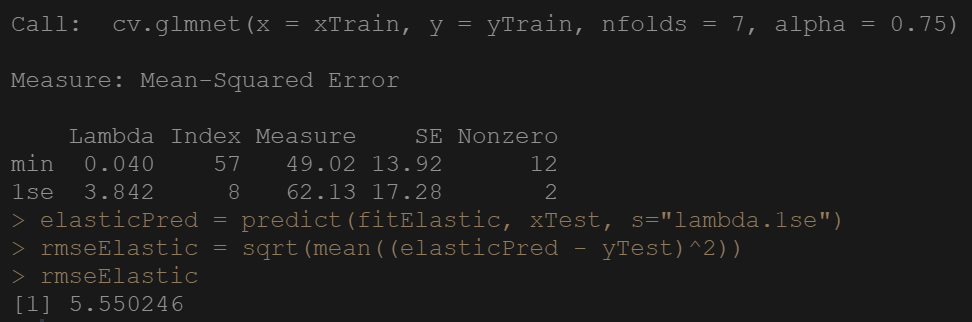
****

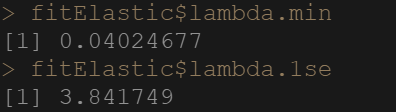
**For alpha value 0.5,**

****

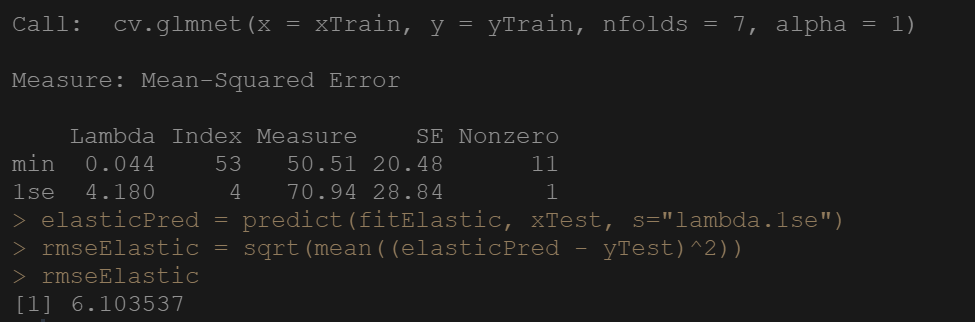
****

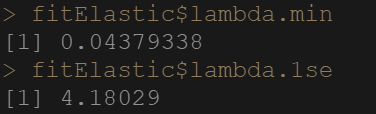
**For alpha value 0.75,**

****

****

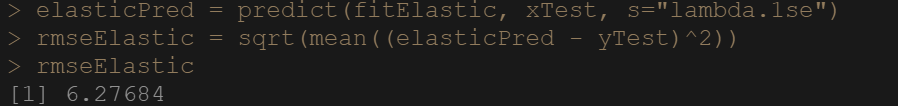
**For alpha value 1,**

****

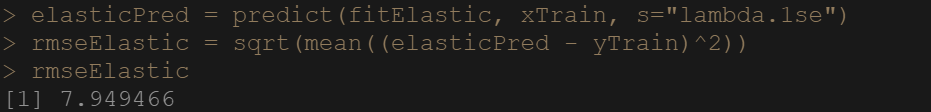
****

***c)*** *Report the R2 and training/testing RMSE values for each alpha.*

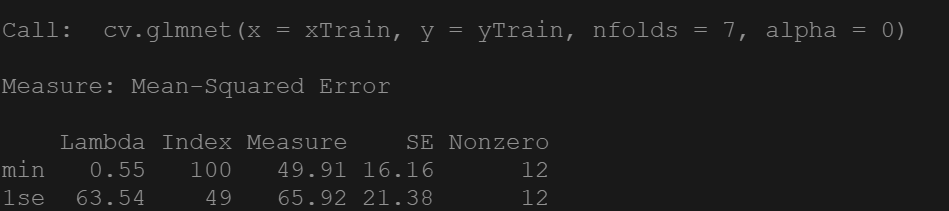
RMSE value for alpha 0 is 6.27684 (Testing Set)



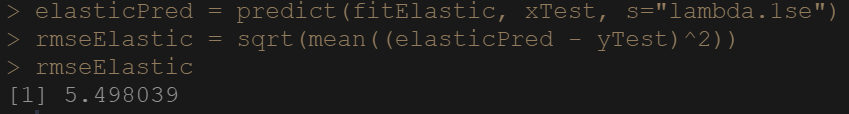
RMSE value for alpha 0 is 7.949466 (Training Set)



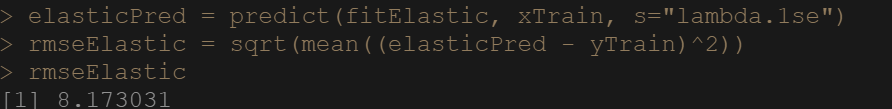
R-Squared for alpha 0



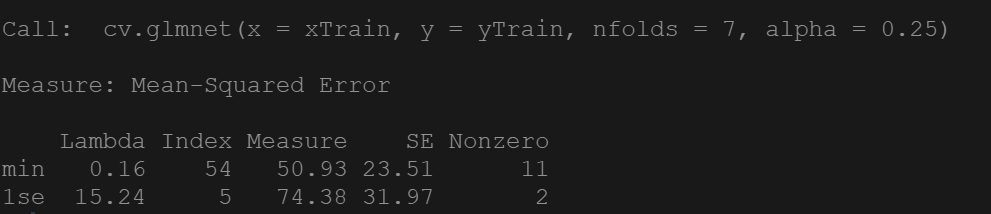
RMSE value for alpha 0.25 is 5.498039 (Testing Set)



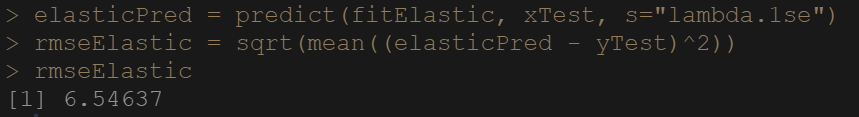
RMSE value for alpha 0.25 is 8.173031 (Training Set)

****

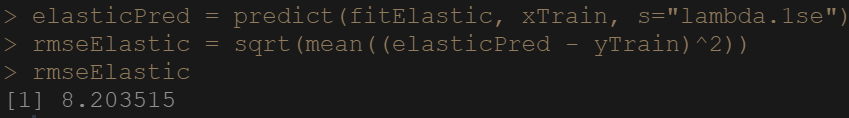
R-Squared for alpha 0.25

****

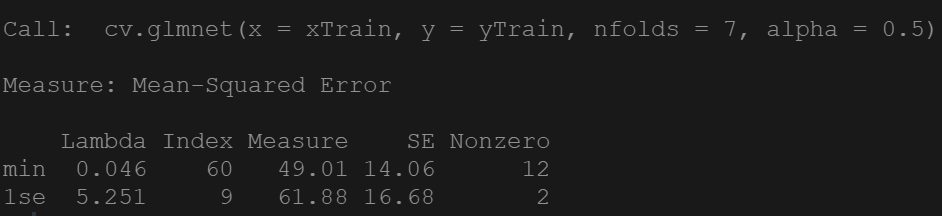
RMSE value for alpha 0.5 is6.54637 (Testing Set)



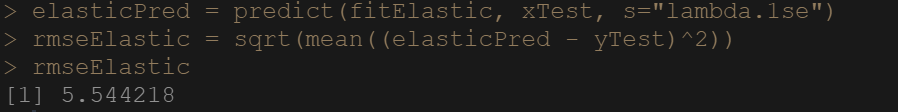
RMSE value for alpha 0.5 is 8.203515 (Training Set)



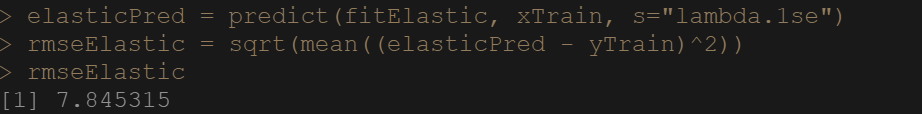
R-Squared for alpha 0.5



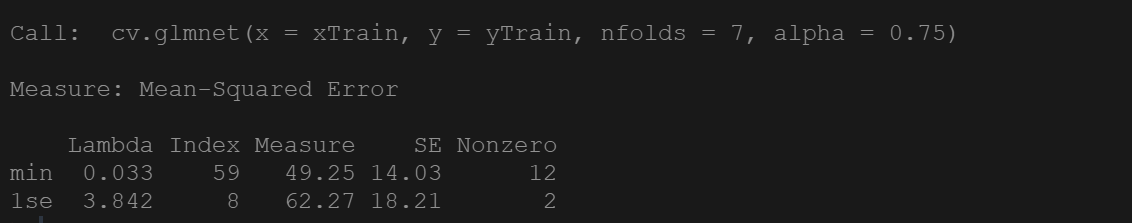
RMSE value for alpha 0.75 is 5.544218 (Testing Set)



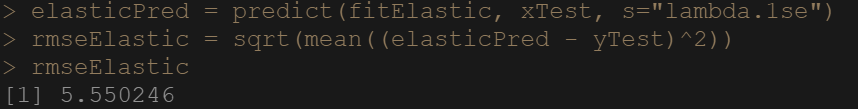
RMSE value for alpha 0.75 is 7.845315 (Training Set)



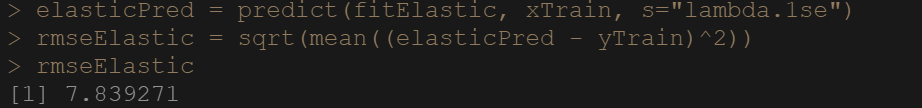
R-Squared for alpha 0.75

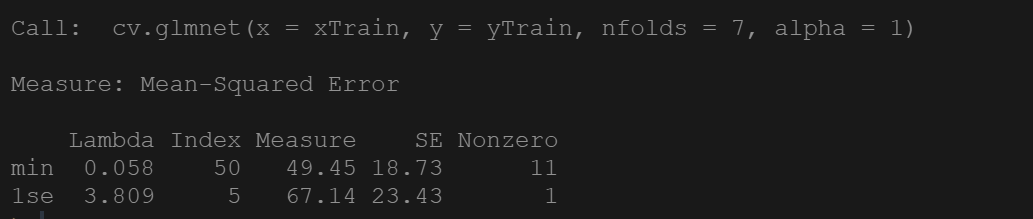


RMSE value for alpha 1 is 5.550246 (Testing Set)



RMSE value for alpha 1 is 7.839271 (Training Set)



R-Squared for alpha 1

**d)** Which of these models would you say is performing the best and why? Based on the alpha, is your chosen model influenced more by Lasso regression or Ridge regression?

**Based on the following alpha values of 0, 0.25, 0.5, 0.75, 1 from the table above we can assume that it falls under the Ridge Regression with the given values that corresponds to the respective alpha values specifically.**

**4. Grad only; Undergrad extra credit: Paper Review**

Read the posted paper "Adding bias to reduce variance in psychological results." Answer the following questions in detail. You should be able to write at least two or three sentences for each:

1. What is the size of the dataset relative to the number of independent variables?

* **There are 395 rows and 39 columns. After pre-processing in actual data it was 395 rows and 30 columns.**
* **Data was collected by Paul Cortez at the University of Minho in Portugal in 2008.**
* **It contains the math score of Portugal student.**

**b)** Is there evidence of overfitting in their dataset?

* **First, we need to understand what is overfitting. When your model is too good to be true then that means you may have overfitting in the model.**
* **They haven’t particularly mentioned that there is overfitting but with the use of new methods they have overcome overfitting that is the evidence.**
* **So, there was overfitting but with the use of penalized methods they fixed it.**

**c)** How do they evaluate the performance of each of the regularized regression techniques?

* **They evaluate the performance of each of the regularized regression techniques as below, to evaluate the performance of the regularized regression techniques they have calculated the mean square prediction error.**
* **OLS method have high mean value which is 9.41. In the overall research paper, OLS model is not that good as compared to the regularized regression techniques such as ridge regression and lasso regression.**

**d)** Are there any issues that you can identify with the way they are evaluating performance?

* **In my knowledge, they have done a really good job with the penalized regression method to show comparison with the OLS method.**
* **My insight into the field isn't that incredible which might oblige me to find any issue with the model since I was perusing that exploration to comprehend albeit, this research to limit the MSPE shows some punished relapse strategies to beat the exemplary OLS and stepwise regression coefficients.**