Student name: SYED NOOR RAZI ALI

Student id : 2070326

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:

a) 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: $(Z^TZ)^{-1}Z^TY$
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?

```r
data("mtcars")
head(mtcars)
str(mtcars)
#A.

x<- mtcars[ ,c('cyl','disp','hp','wt','carb')]
x

#B.
y<- mtcars[ ,c('mpg')]
y

#C.
x<-cbind(count=1,x)

#D.
# convert dataframe to matrix
x1<- as.matrix(x)
x1
y1<- as.matrix(y)
y1

#E.
#Here's the formula: [T(x)X]-1*T(x)Y

# transpose the matrix
tx<-t(x1)
ty<-t(y1)

txx1<- tx %*% x1
tyy1 <- tx %*% y1
tyy1

library(MASS)
# for inverse,we use ginv command
inverse <- ginv(txx1)
inverse

beta <- inverse %*% tyy1

beta

#F.
model<- lm(mpg~ cyl+disp+hp+wt+carb, data= mtcars)
```

```
> str(y1)
 num [1:32, 1] 21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
> # transpose the matrix
> tx<-t(x1)
> ty<-t(y1)
> txx1<- tx %*% x1
> tyy1 <- tx %*% y1
> tyy1
           [,1]
count    642.900
cyl     3693.600
disp  128705.080
hp     84362.700
wt      1909.753
carb    1641.900
> library(MASS)
> # for inverse,we use ginv command
> inverse <- ginv(txx1)
> inverse
             [,1]          [,2]          [,3]          [,4]         [,5]         [,6]
[1,]  1.396389910 -0.2168589525  5.555427e-03 -2.383520e-03 -0.382871011  0.098254518
[2,] -0.216858953  0.0703757221 -8.390258e-04 -2.681453e-04  0.013715751 -0.010596503
[3,]  0.005555427 -0.0008390258  3.605971e-05 -2.799315e-05 -0.002163615  0.000847438
[4,] -0.002383520 -0.0002681453 -2.799315e-05  6.139347e-05  0.001673899 -0.001383005
[5,] -0.382871011  0.0137157515 -2.163615e-03  1.673899e-03  0.216799569 -0.051854297
[6,]  0.098254518 -0.0105965033  8.474380e-04 -1.383005e-03 -0.051854297  0.050306386
> beta <- inverse %*% tyy1
> beta
             [,1]
[1,] 40.815359236
[2,] -1.291898563
[3,]  0.011485584
[4,] -0.020352893
[5,] -3.846949031
[6,] -0.006746893
> str(x1)
```

```
Maserati Bora           1   8 301.0 335 3.570    8
Volvo 142E              1   4 121.0 109 2.780    2
> #F.
> model<- lm(mpg~ cyl+disp+hp+wt+carb, data= mtcars)
> model

Call:
lm(formula = mpg ~ cyl + disp + hp + wt + carb, data = mtcars)

Coefficients:
(Intercept)          cyl         disp           hp           wt         carb
  40.815359    -1.291899     0.011486    -0.020353    -3.846949    -0.006747

>
```

As we can see here we get the same beta co efficient values by using the formula and the “lm” function.

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:

a) 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 R^2 of the training set and the RMSE values of both the training and test sets. Is there evidence of overfitting?

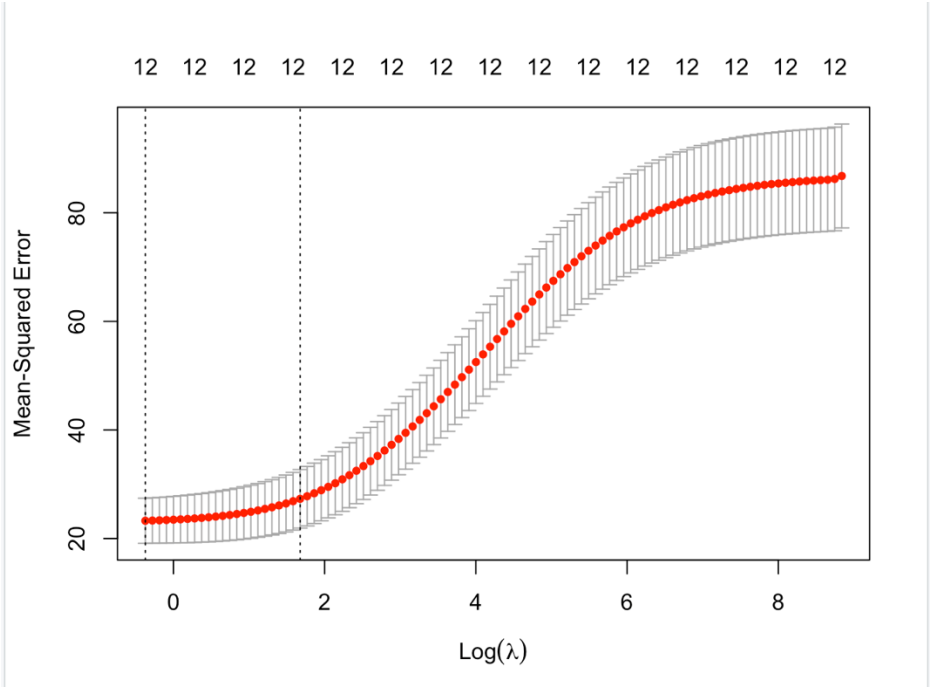
```
D:/DEPAUL MS DATA SCIENCE/DSC 424 Advance Data Analysis/HW2 Datasets/
> modeltrain <- lm(MEDV ~., data = housingtrain)
> summary(modeltrain)

Call:
lm(formula = MEDV ~ ., data = housingtrain)

Residuals:
     Min       1Q   Median       3Q      Max
-15.9605  -2.6653  -0.6272   1.7309  26.2670

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.965e+01  5.610e+00   7.069 7.94e-12 ***
CRIM        -1.299e-01  3.412e-02  -3.807 0.000165 ***
ZN           4.341e-02  1.570e-02   2.764 0.005994 **
INDUS        6.302e-03  6.958e-02   0.091 0.927884
CHAS         3.594e+00  9.454e-01   3.802 0.000168 ***
NOX         -2.197e+01  4.377e+00  -5.021 8.05e-07 ***
RM           4.229e+00  4.898e-01   8.634  < 2e-16 ***
AGE         -1.268e-04  1.511e-02  -0.008 0.993307
DIS         -1.529e+00  2.318e-01  -6.598 1.46e-10 ***
RAD          2.665e-01  7.341e-02   3.630 0.000324 ***
TAX         -1.134e-02  4.130e-03  -2.746 0.006338 **
PTRATIO     -9.828e-01  1.506e-01  -6.526 2.24e-10 ***
LSTAT       -4.665e-01  6.094e-02  -7.655 1.73e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.661 on 367 degrees of freedom
Multiple R-squared:  0.7571,    Adjusted R-squared:  0.7491
F-statistic: 95.31 on 12 and 367 DF,  p-value: < 2.2e-16

> rmse_hous_Train = sqrt(mean(modeltrain$residuals^2))
> rmse_hous_Train
[1] 4.580769
> # ols Predict on the test set
> olsPred = predict(modeltrain,housingTest )
> # Compute the RMSE of the predictions on the test set
> rmse_hous_Test = sqrt(mean((olsPred - housingTest$MEDV)^2))
> rmse_hous_Test #by ols
[1] 5.263608
>
```

With an adj R-square of 0.7491 and a residual error of 4.66 on 367 degrees of freedom, our model explains 74.91 percent of the variability in the dependent variable.

The train set's RMSE is 4.58, which is lower than the test set's RMSE (5.26). As a result of the smaller difference in RMSE. Independent variables are also behaving admirably well. As far as we can tell The model has no overfitting,

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.

```
Console  Terminal  Jobs
D:/DEPAUL MS DATA SCIENCE/DSC 424 Advance Data Analysis/HW2 Datasets/
> #2-b ridgeregression on train dataset
> set.seed(123)
> xTrain = as.matrix(housingtrain[, -13]) # take out MEDV
> yTrain = as.matrix(housingtrain[, 13]) #take MEDV dependent
> fitRidge = cv.glmnet(xTrain, yTrain, alpha=0, nfolds=7)
> fitRidge$lambda.min
[1] 0.6899987
> fitRidge$lambda.1se
[1] 6.434951
> plot(fitRidge)
>
```



The value of lambda vs Mean Square Error of Ridge Regression is shown in the graph.

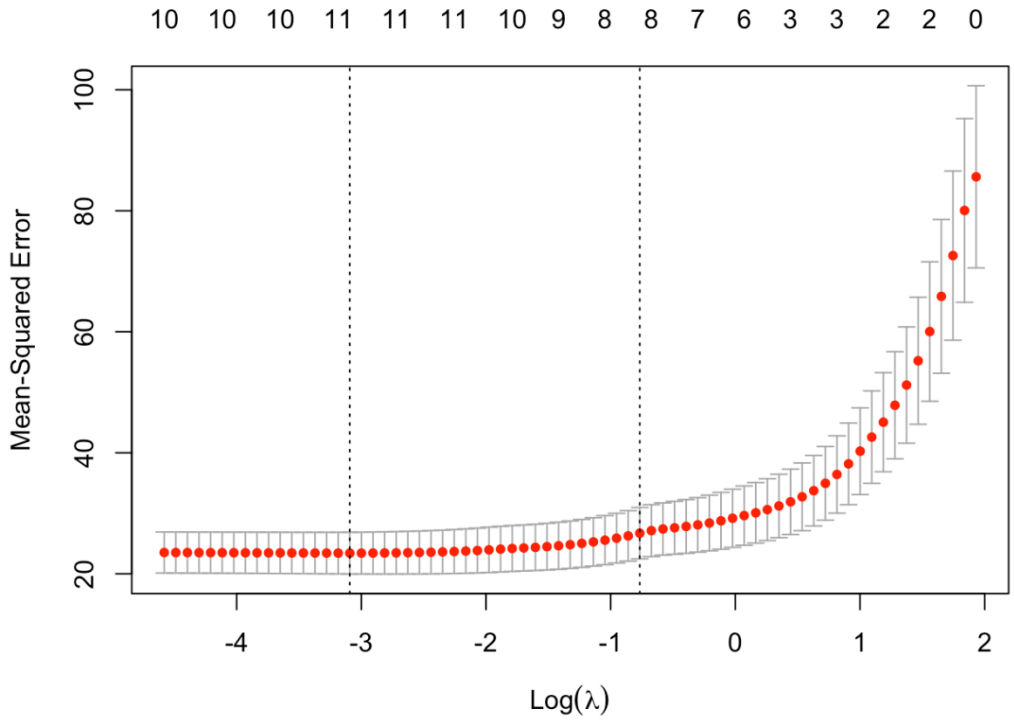
The graph shows that the lambda value of 0.68 has the smallest cross validation mean square error. The lambda1se error increases smoothly from left to right, however it appears that the independent variable is penalized more after that. Hence, between lambda min and lambda1se, we have a decent prediction window.

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.

```
Console  Terminal  Jobs
D:/DEPAUL MS DATA SCIENCE/DSC 424 Advance Data Analysis/HW2 Datasets/
> #To predict with this model, by passing it the test set and a lambda value.
> ridgePred = predict(fitRidge, xTest, s="lambda.1se")
> rmseRidge = sqrt(mean((ridgePred - yTest)^2))
> rmseRidge # The RMSE of the test set using Ridge regression.
[1] 5.756121
> rmse_hous_Test # The RMSE of the test set using OLS regression.
[1] 5.263608
>
```

The RMSE value of ridge regression is 5.75, whereas the OLS is 5.26 for the test set, which is nearly same, but suggests that the OLS outperforms ridge regression. However, rather than obtaining the only best fit, ridge regression provides us with essential bias to highlight.

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?



```
Console  Terminal  Jobs
D:/DEPAUL MS DATA SCIENCE/DSC 424 Advance Data Analysis/HW2 Datasets/
> fitLasso = cv.glmnet(xTrain, yTrain, alpha=1, nfolds=7)
> fitLasso$lambda.min
[1] 0.0215674
> fitLasso$lambda.1se
[1] 0.320269
> plot(fitLasso)
> lassoPred = predict(fitLasso, xTest, s="lambda.1se")
> rmseLasso = sqrt(mean((lassoPred - yTest)^2))
> rmseLasso
[1] 5.388097
> # comparision
> rmse_hous_Test # OLS
[1] 5.263608
> rmseRidge  #ridge
[1] 5.756121
> rmseLasso   #LASSO
[1] 5.388097
>
```

Comparing the RMSE values of test dataset of OLS, Ridge and Lasso regression.

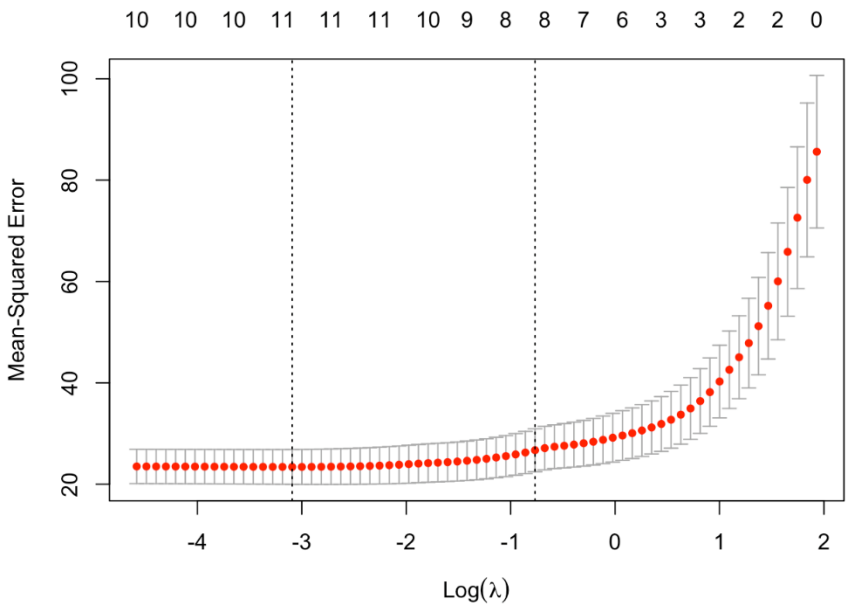
RMSE of OLS = 5.263608

RMSE of Ridge = 5.756121

RMSE of Lasso = 5.388097

RMSE of Lasso lies between the RMSE of OLS and Ridge, so we can say that it performs better than OLS, but not better than Ridge.

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?



From the graph we can observe that as the lambda value increases, the number of variables decreases.

We can select 8 variables.

At lambda1se 51% of the variance is explained.

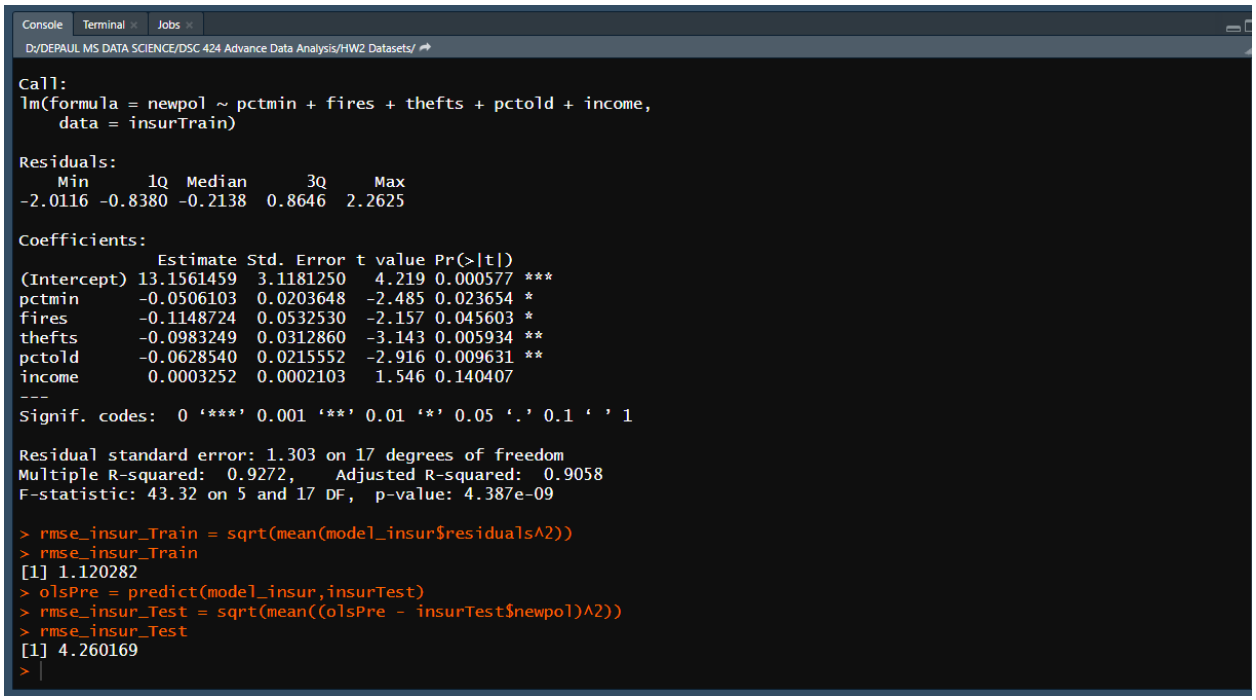
g) If you had to pick only one model to use from these three, explain which you would choose and why.

Out of the 3 models OLS, Ridge, and LASSO, I would select Lasso because it predicts better than OLS and Ridge and It is simpler.

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).

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



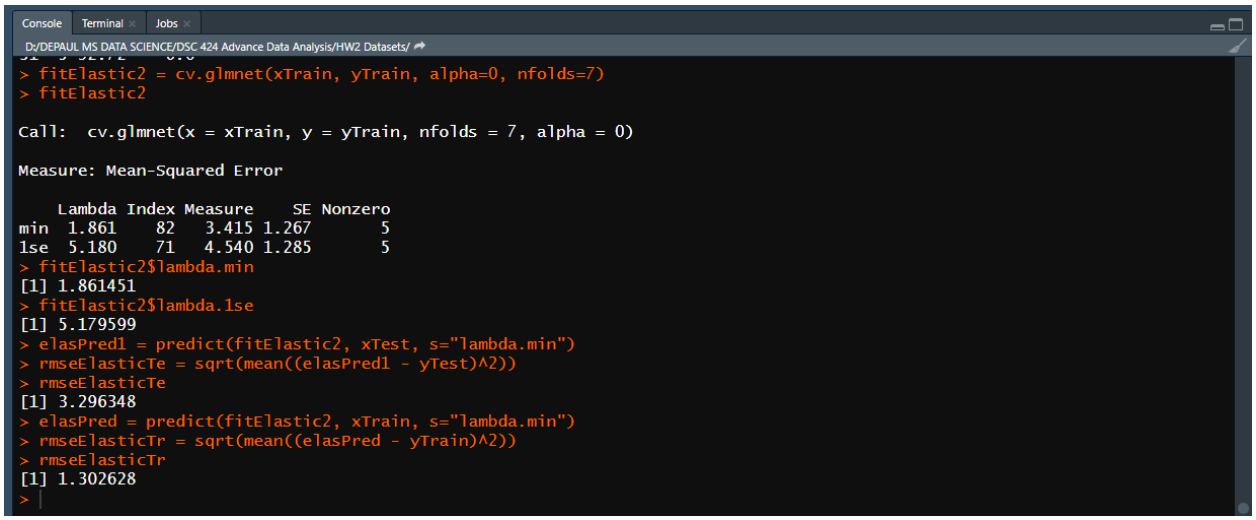
The adjusted R-squared of the model is 0.9058.

The RMSE of training dataset is 1.120282, and the RMSE of testing dataset is 4.260169.

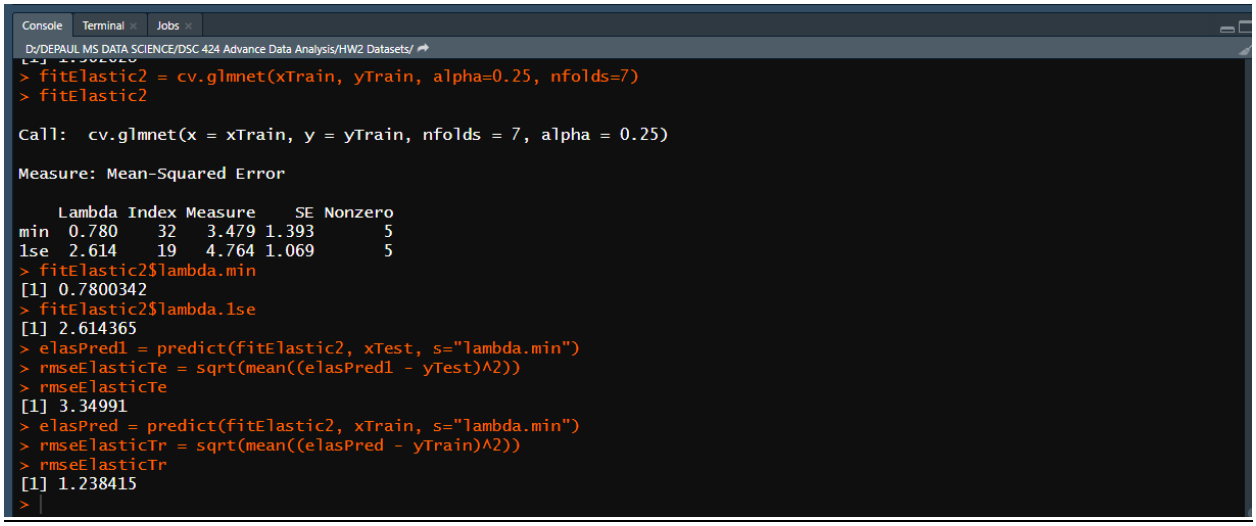
There is evidence of overfitting because the adjusted R-squared is very high and The difference between the RMSE of test and training data is high.

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

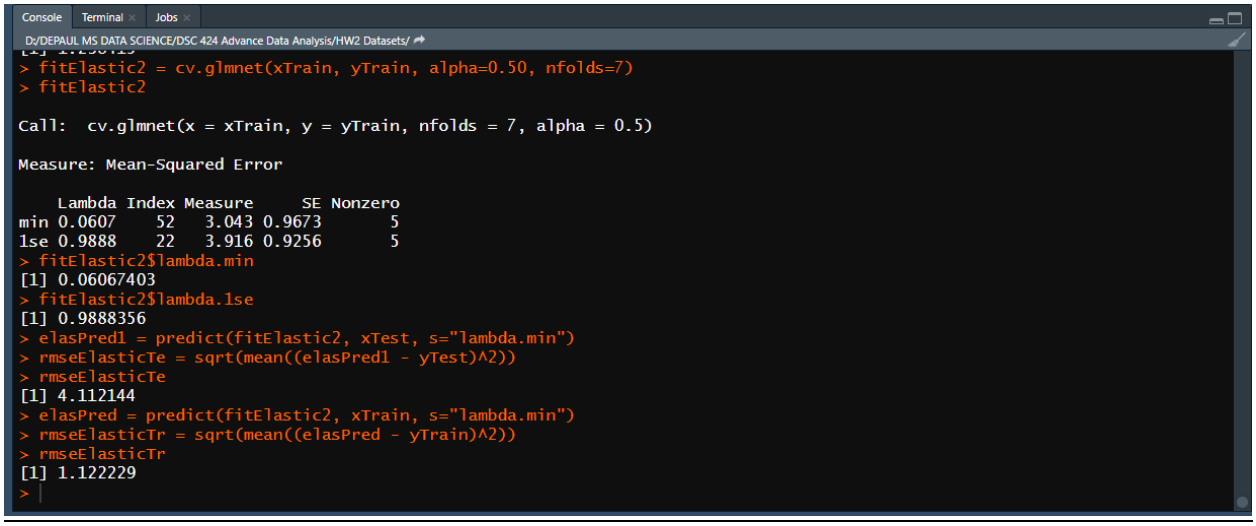
Alpha 0



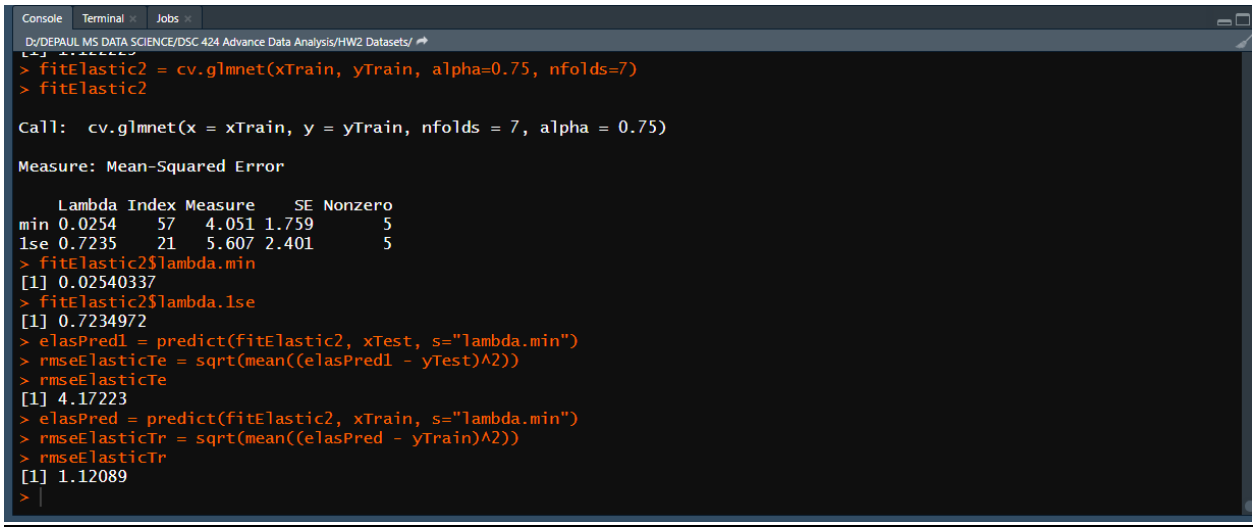
Alpha 0.25



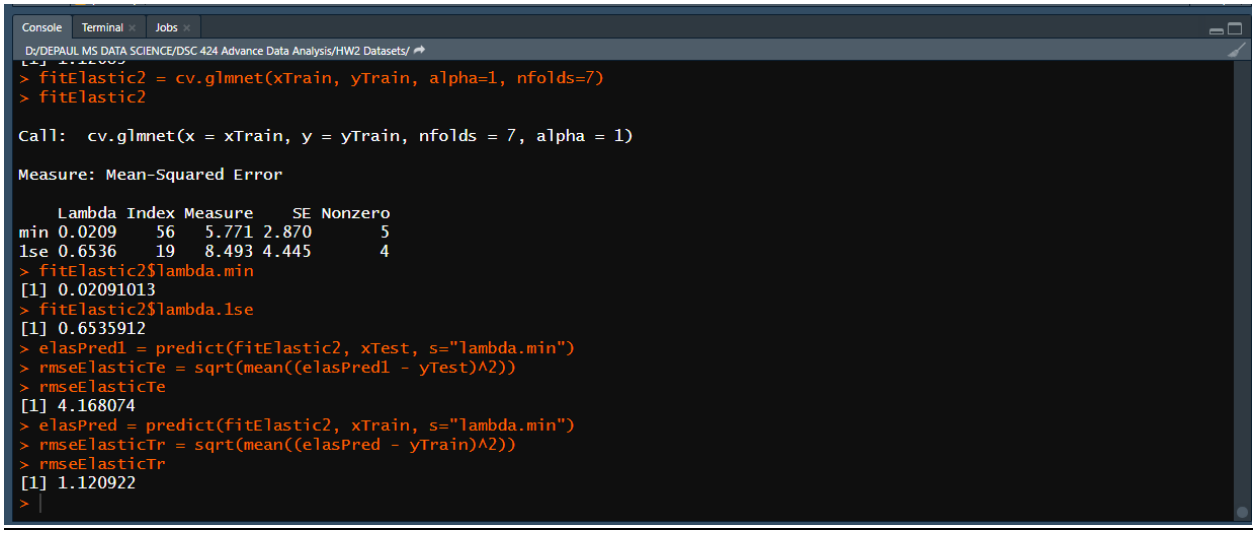
Alpha 0.50



Alpha 0.75



Alpha 1



c) Report the R^2 and training/testing RMSE values for each alpha.

Alpha value = 0

R-squared = 0.52564

Training RMSE = 1.30

Testing RMSE = 3.29

Alpha value = 0.25

R-squared = 0.5334128

Training RMSE = 1.23

Testing RMSE = 3.34

Alpha value = 0.5

R-squared = 0.5312085

Training RMSE = 1.12

Testing RMSE = 4.11

Alpha value = 0.75

R-squared = 0.5287431

Training RMSE = 1.12

Testing RMSE = 4.17

Alpha value = 1

R-squared = 0.5324399

Training RMSE = 1.120

Testing RMSE = 4.1668

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?

The model with alpha value = 0 is performing well. As the RMSE of both the training and testing are minimum, and the difference between the two RMSE is less when compared to other models.

The model chosen by me is influenced by Ridge, as the alpha value = 0.

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:

a) What is the size of the dataset relative to the number of independent variables?

There are 395 rows and 39 columns. Hence, the size of the dataset relative to the number of independent variables is very huge.

b) Is there evidence of overfitting in their dataset?

Yes, As the size of the dataset is large with too many variables and therefore there might be possibility that the variance in the data can be explained with some less variables.

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

They used the MSPE – mean squared prediction error to evaluate the performance.

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