Stats202A - HW 7

BHARGAV PARSI (804945591)

1) Error plot for Lasso for various Lambda values

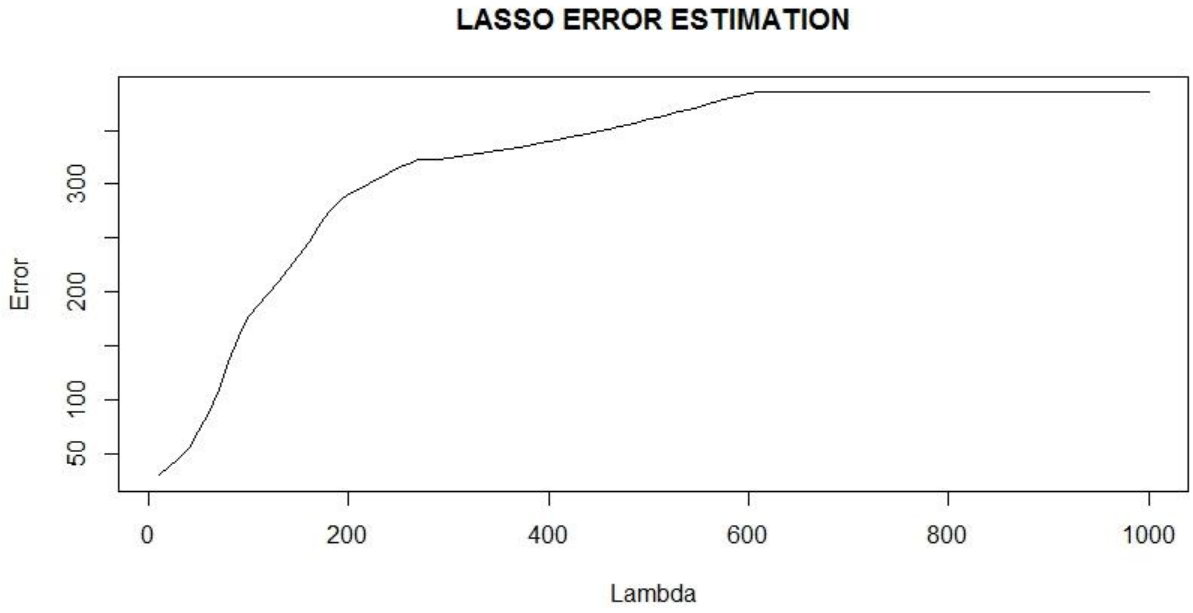

Fig 1: Error Vs Lambda

2) Analysis of Real World Data sets:
   I have used the Swiss Data set for Linear Regression and Ridge Regression.
   Below I have plotted a few variables vs the y value.

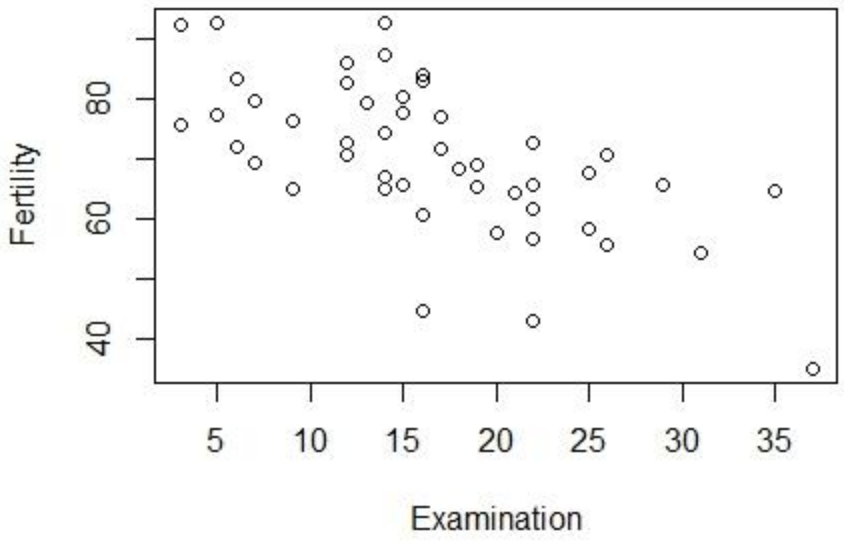

Fig 2 Fertility Vs Examination

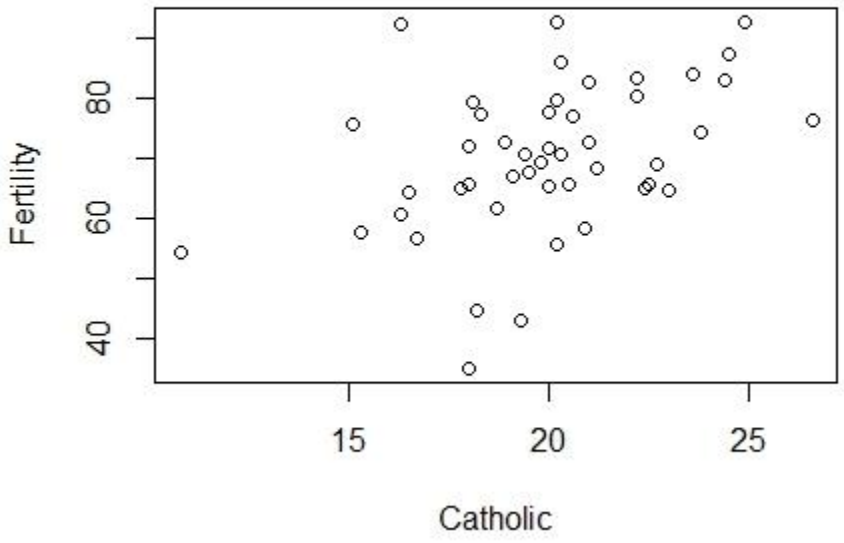

Fig 3 Fertility Vs Catholic

a) **Logistic Regression:**

I have used the binary dataset from the IDRE website("https://stats.idre.ucla.edu/stat/data/binary.csv"). This dataset has the independent variables as GRE, GPA and Rank. The dependent variable is whether that particular person got an Admit or not.

```
> myLogistic(x,y)
$coefficients
       gre        gpa       rank
 0.2233584  0.2510192 -0.4472078

$standard_error
      gre       gpa      rank
0.1147555 0.1140612 0.1082179

> print(glm(formula = y~x + 0, family = "binomial"))

Call:  glm(formula = y ~ x + 0, family = "binomial")

Coefficients:
   xgre     xgpa    xrank
 0.2217   0.2500  -0.4453

Degrees of Freedom: 400 Total (i.e. Null);  397 Residual
Null Deviance:      554.5
Residual Deviance: 519.9        AIC: 525.9
```

Fig 4: Logistic Regression

b) **Ridge Regression:**

Swiss dataset was used to compare results. It gives us the Standardized fertility measure and socio-economic indicators for each of 47 French-speaking provinces of Switzerland at about 1888.

```
> Ridge_P = myRidge(x[train,],y[train], lambda)
> Ridge_P
[1] 74.64436146 -0.27807670 -0.93900466 -0.35978119  0.06500147  1.37552338
> coef(Ridge_R)
6 x 1 sparse Matrix of class "dgCMatrix"
                            s0
(Intercept)      73.36350615
Agriculture      -0.26542433
Examination      -0.89519263
Education        -0.36435849
Catholic          0.06570399
Infant.Mortality  1.37394755
> |
```

Fig 5: Results of Ridge Regression

The myRidge result was compared with lambda = 0.1 with the glmnet library in R. We can see that the result is quite similar to the output of the R function.

c) PCA

Iris Dataset was used to analyse PCA. This famous (Fisher's or Anderson's) iris data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are *Iris setosa*, *versicolor*, and *virginica*.

```
> p = myPCA(t(a)%*%a)
> e = eigen(t(a)%*%a)
> p
$D
[1] 9208.30507  315.45432   11.97804    3.55257

$V
           [,1]       [,2]        [,3]       [,4]
[1,] -0.7511082 -0.2841749  0.50215472  0.3208143
[2,] -0.3800862 -0.5467445 -0.67524332 -0.3172561
[3,] -0.5130089  0.7086646 -0.05916621 -0.4807451
[4,] -0.1679075  0.3436708 -0.53701625  0.7518717

> e = eigen(t(a)%*%a)
> e
eigen() decomposition
$values
[1] 9208.30507  315.45432   11.97804    3.55257

$vectors
           [,1]       [,2]        [,3]       [,4]
[1,] -0.7511082  0.2841749 -0.50215472  0.3208143
[2,] -0.3800862  0.5467445  0.67524332 -0.3172561
[3,] -0.5130089 -0.7086646  0.05916621 -0.4807451
[4,] -0.1679075 -0.3436708  0.53701625  0.7518717
```

Fig 6: Analysis of PCA

d) Linear Regression

```
> myLM(x,y)
$coefficients
                  Agriculture      Examination        Education         Catholic
      66.9151817       -0.1721140       -0.2580082       -0.8709401        0.1041153
Infant.Mortality
       1.0770481

$standard_error
                  Agriculture      Examination        Education         Catholic
     10.70603759        0.07030392       0.25387820       0.18302860       0.03525785
Infant.Mortality
       0.38171965

> coef(lm(y~x))
      (Intercept)      xAgriculture      xExamination        xEducation         xCatholic
       66.9151817        -0.1721140        -0.2580082        -0.8709401         0.1041153
xInfant.Mortality
        1.0770481
```

Fig 7: Linear Regression

We can see the negative correlation between Examination and Fertility and positive correlation between Catholic and Fertility. This can be also observed from the coefficient values which are -ve and +ve.

e) Lasso Regression

I used the Swiss data set with Lambda = 10^seq(10,-2,length = 100). Below are the lasso solution paths for in built function(glmnet) and my function.

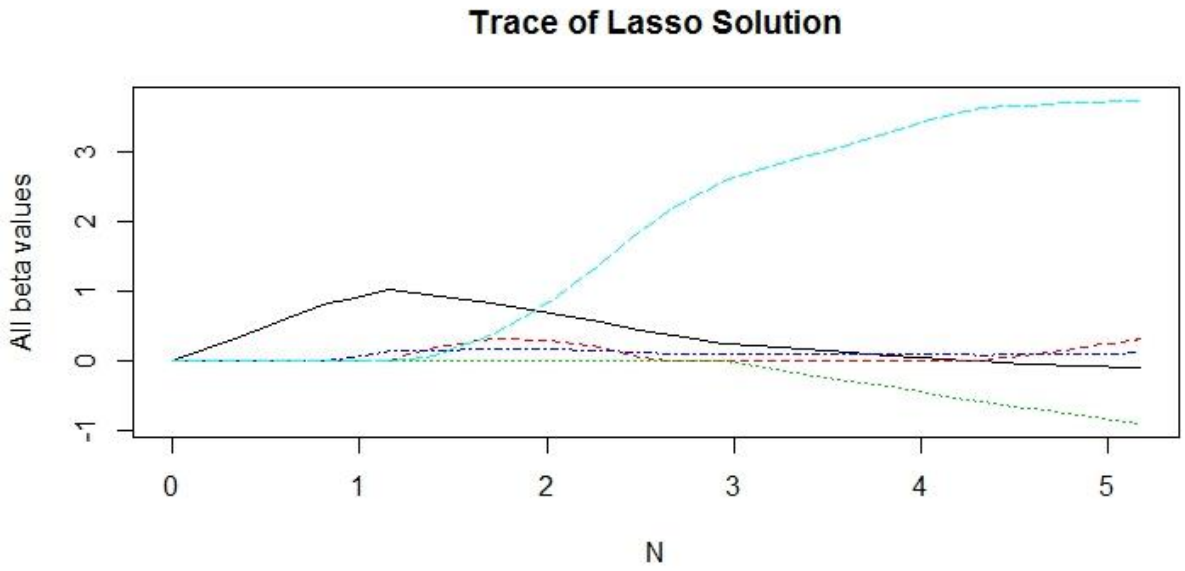

Fig 8: Lasso Solution for myLasso Function

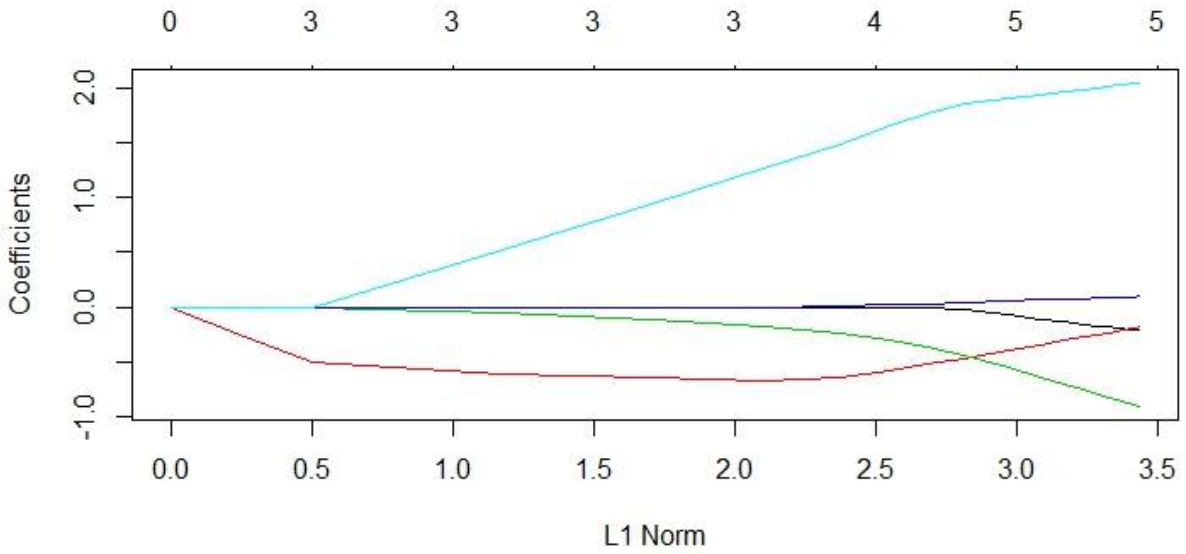

Fig 9: Lasso Solution for Inbuilt Lasso Function