PILLA SATISH

STUDENT NO-210472095

Newcastle University

MAS8404 Project

Breast Cancer Data

Introduction

In the project we are analyzing the Breast cancer data which concerns characteristics of breast tissue samples collected from 699 women in Wisconsin using fine needle aspiration cytology (FNAC). This is a type of biopsy procedure in which a thin needle is inserted into an area of abnormal-appearing breast tissue.

Exploratory Data Analysis (EDA):

For the breast cancer data, we have omitted the null values using is.na function after omitting the null values we there are 683 rows and 11 columns. We also the variables are encoded as factors so we have converted the factors to quantitative variables.

	Id <dbl>	Cl.thickness <dbl>	Cell.size <dbl>	Cell.shape <dbl>	Marg.adhesion <dbl>	Epith.c.size <dbl>	Bare.nuclei <dbl>	Bl.cromatin <dbl>
1	1000025	5	1	1	1	2	1	3
2	1002945	5	4	4	5	7	10	3
3	1015425	3	1	1	1	2	2	3
4	1016277	6	8	8	1	3	4	3
5	1017023	4	1	1	3	2	1	3
6	1017122	8	10	10	8	7	10	9

Fig 1: Breast cancer data after omitting null values and converting variables factors to quantitative variables.

Numerical Summary:

By using cor function we can see how different variables are correlated to each other or we can check how they are co-related to our response variables in the breast cancer data.

```
                Cl.thickness Cell.size Cell.shape Marg.adhesion Epith.c.size Bare.nuclei Bl.cromatin
Cl.thickness       1.0000000 0.6424815  0.6534700     0.4878287    0.5235960   0.5930914   0.5537424
Cell.size          0.6424815 1.0000000  0.9072282     0.7069770    0.7535440   0.6917088   0.7555592
Cell.shape         0.6534700 0.9072282  1.0000000     0.6859481    0.7224624   0.7138775   0.7353435
Marg.adhesion      0.4878287 0.7069770  0.6859481     1.0000000    0.5945478   0.6706483   0.6685671
Epith.c.size       0.5235960 0.7535440  0.7224624     0.5945478    1.0000000   0.5857161   0.6181279
Bare.nuclei        0.5930914 0.6917088  0.7138775     0.6706483    0.5857161   1.0000000   0.6806149
Bl.cromatin        0.5537424 0.7555592  0.7353435     0.6685671    0.6181279   0.6806149   1.0000000
Normal.nucleoli    0.5340659 0.7193460  0.7179634     0.6031211    0.6289264   0.5842802   0.6656015
Mitoses            0.3545301 0.4654091  0.4468571     0.4249917    0.4811836   0.3490108   0.3536683
Class              0.7147899 0.8208014  0.8218909     0.7062941    0.6909582   0.8226959   0.7582276
                Normal.nucleoli   Mitoses     Class
Cl.thickness          0.5340659 0.3545301 0.7147899
Cell.size             0.7193460 0.4654091 0.8208014
Cell.shape            0.7179634 0.4468571 0.8218909
Marg.adhesion         0.6031211 0.4249917 0.7062941
Epith.c.size          0.6289264 0.4811836 0.6909582
Bare.nuclei           0.5842802 0.3490108 0.8226959
Bl.cromatin           0.6656015 0.3536683 0.7582276
Normal.nucleoli       1.0000000 0.4370424 0.7186772
Mitoses               0.4370424 1.0000000 0.4312971
Class                 0.7186772 0.4312971 1.0000000
```

Fig 2: Correlation matrix for breast cancer data

From above matrix we can see that cell.size, cell.shape, bare.nuclei are very strongly corelated to our response variables Class. Among predictor variables cell.size and cell.shape are strongly correlated to each other.

To find out how many women effected with benign and malignant in our dataset we can use table function. When we used tables function we found out that 444 out of 683 are suffering from benign and remaining 239 women are suffering from malignant.

```
table(breastcancer3$class)

  0   1
444 239
```

Fig 3: No of women effected with benign and malignant, 0 represent benign and 1 represent malignant.

Percentage of women effected with benign and malignant stage:

Description: df [2 x 2]

Var1 <fctr>	Freq <dbl>
0	65.00732
1	34.99268

2 rows

Fig 4: Percentage of women effected with benign and malignant stage in the breast cancer data

From above we can clearly see that 65% of women are affected with benign and 35% are affected with malignant in Breast cancer data.

Graphical summary:

Frequency of cancer diagnosis:

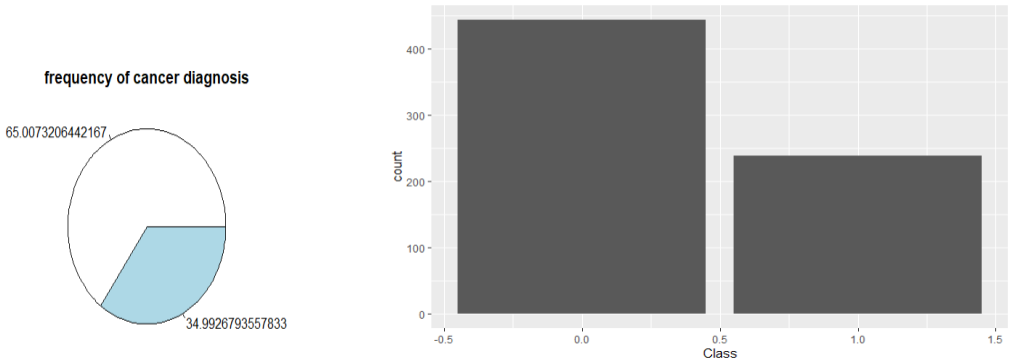

Fig 5: Frequency of cancer diagnosis and count of female affected with benign and malignant.

From above we can see that in Breast cancer data 65% women are suffering from benign and 35% are suffering for malignant.

To get an idea about relationships between variables we can plot a pairs plot of the predictors on Class (response variables), coloring the points according to whether the cancer is benign and malignant.

Fig 6: Scatter plot for the Breast Cancer data.

From the above plot we can clearly see that Cl.thickness and Cell.size are strongly correlated to each other. We can also say that as Cl.thickness, Cell.size, Bare.nuclei are increasing the cancer is being to malignant.

We can see it by plotting the bar graph between Class (response variables) and predictors variables (Cl.thickness, Cell.size, Bare.nuclei).

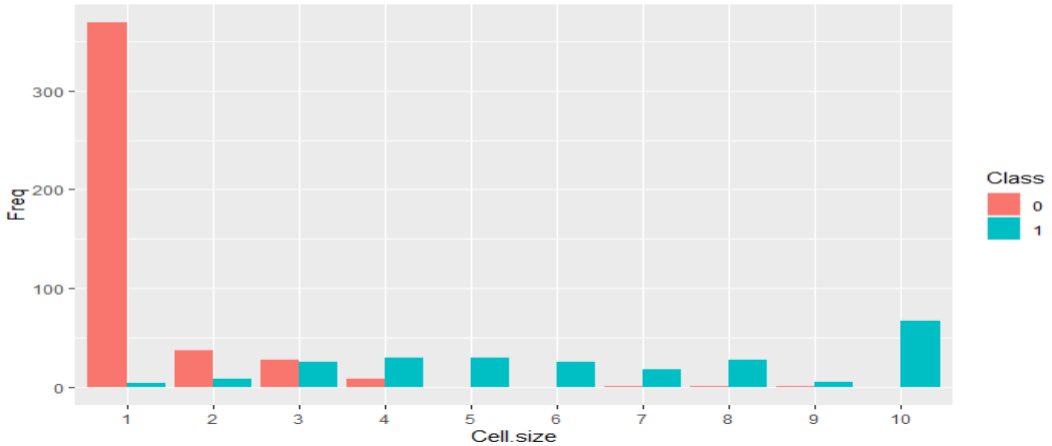

Fig 7: Bar plot to represent relationship between Cell.size and Class (response variables)

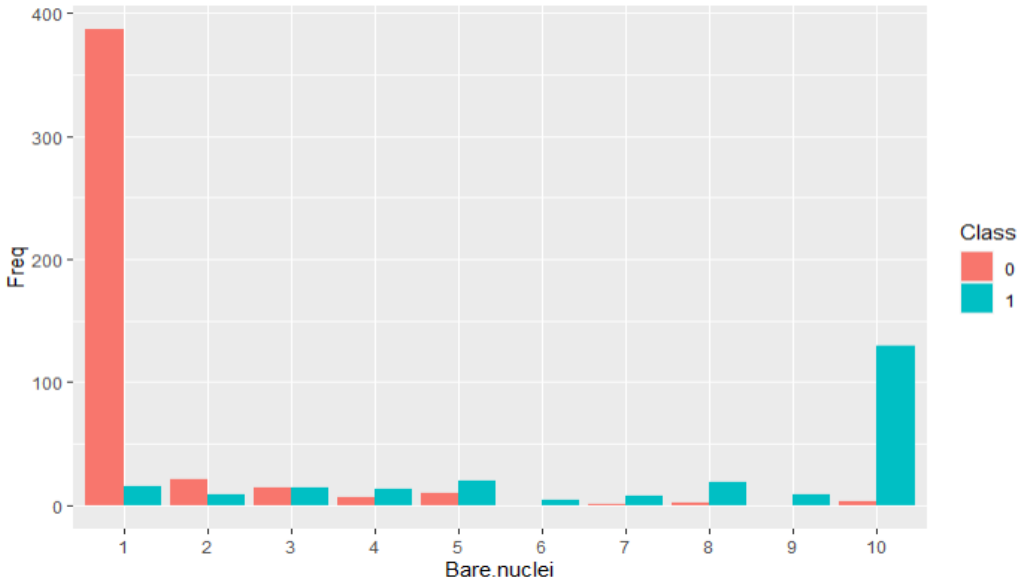

Fig 8: Bar plot to represent relationship between Bare.nuclie and Class (response variables)

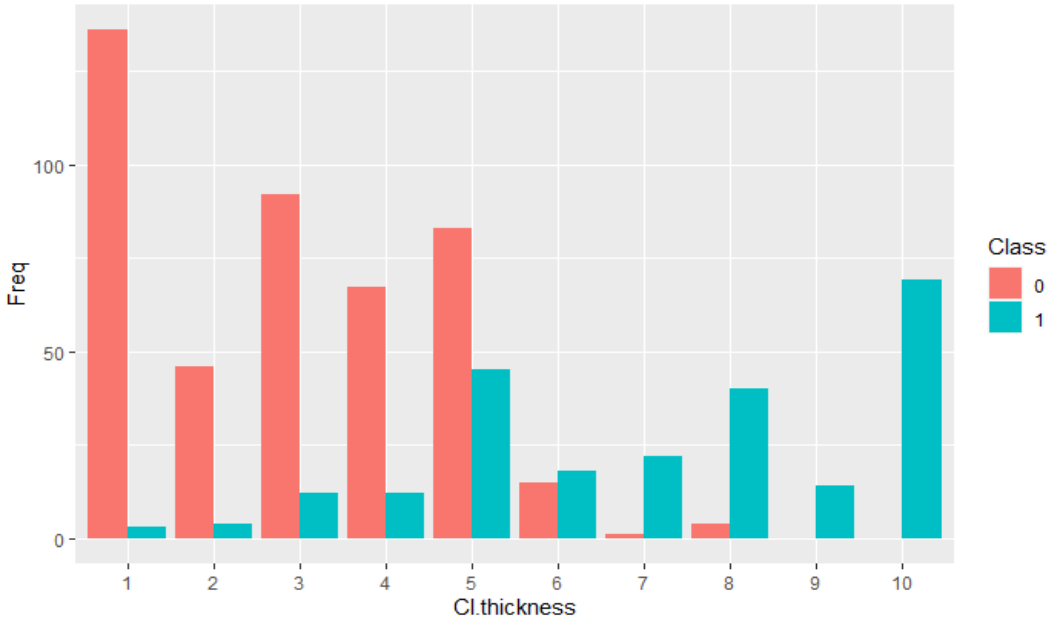

Fig 9: Bar plot to represent relationship between Cl.thickness and Class (response variables)

From the above plots we see there are some strong evidence of a higher incidence of cancer being malignant when the cell size and Bare.nuclei more than 3, similarly for Cl.thickness when the Cl.thickness is more than 5 there is higher incidence of cancer being malignant.

While applying correlation function we see that there is strong relationship between Cl.thinkness and Cell.size ,Plotting the scatter plot to see the relationship with response variable(Class).

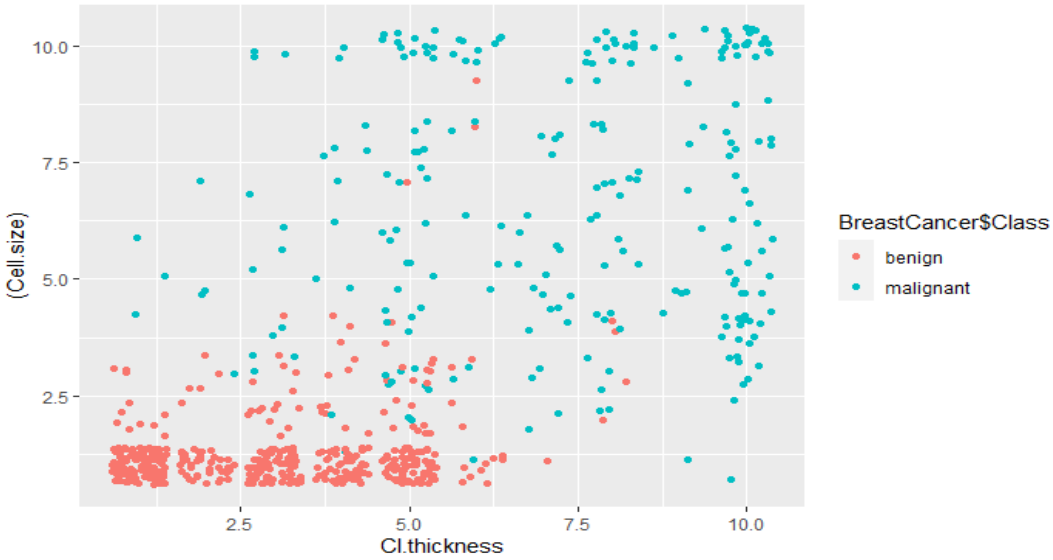

Fig 10: Scatter plot to show the relationship between response variables (Class) and predicted variables (Cl.thickness, Cell.size).

From above we can again clearly see that when Cell.size is less than 2.5 and Cl,thickness is less then 5 majority of the data(Breast cancer data) is at benign state. Which tell us as Cell.size and Cl.thickness increases there is higher incidence the cancer being malignant.

Building classifiers:

Applying logistic regression:

Fitting a logistic regression model for y(BreastCancer3$Class) in terms of the predictors: Cl.thickness up to normal.nuclei and mitoses.

We can then summarize the fit model using summary function,

```
Call:
glm(formula = BreastCancer3$Class ~ ., family = "binomial", data = BreastCancer3)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.4855  -0.1152  -0.0619   0.0222   2.4702

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)   -10.110096   1.173774  -8.613  < 2e-16 ***
Cl.thickness    0.535256   0.141938   3.771 0.000163 ***
Cell.size      -0.005943   0.209158  -0.028 0.977332
Cell.shape      0.322136   0.230644   1.397 0.162510
Marg.adhesion   0.330694   0.123462   2.679 0.007395 **
Epith.c.size    0.096797   0.156568   0.618 0.536415
Bare.nuclei     0.383015   0.093865   4.080 4.49e-05 ***
Bl.cromatin     0.447401   0.171392   2.610 0.009044 **
Normal.nucleoli 0.213074   0.112894   1.887 0.059109 .
Mitoses         0.538551   0.325615   1.654 0.098138 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 884.35  on 682  degrees of freedom
Residual deviance: 102.90  on 673  degrees of freedom
AIC: 122.9

Number of Fisher Scoring iterations: 8
```

Fig 11 : Summarizing the fit model

Inspecting the p-value column in the table. We see that Bare.nuclei, Cl.thickness, Marg.adhesion, Bl.cromation has a coefficient which is significantly different from zero when testing at the 0.1% and 1% level. In other words, if we label Cl.thickness,……, as X1,X2,….,X8 and mitoses has X9 and then if we perform 9 hypothesis tests, we would only reject the null hypothesis at the 0.1% level for j=2, when testing the effect of Cl.thickness and Bare.nuclei. This suggest that for the other variables, given a model that already contains all the other predictors in question adds very little in terms of forecasting whether the Breast cancer is benign or malignant.

Test error for logistic regression on full dataset

```
[1] 0.03513909
```

To improve the model, we can apply the best subset selection method, Applying the best subset selection AIC, BIC and CV and extracting the models minimizing the AIC, BIC and CV.

```
Morgan-Tatar search since family is non-gaussian.
Morgan-Tatar search since family is non-gaussian.
[1] 7
[1] 5
[1] 4
```

When we apply AIC, we are getting model with 7 variables, when BIC is applied we are getting model with 5 variables and when CV is applied we are getting model with 4 variables. In order to choose a single best model, we can plot to show the criteria vary with the numbers of predictors:

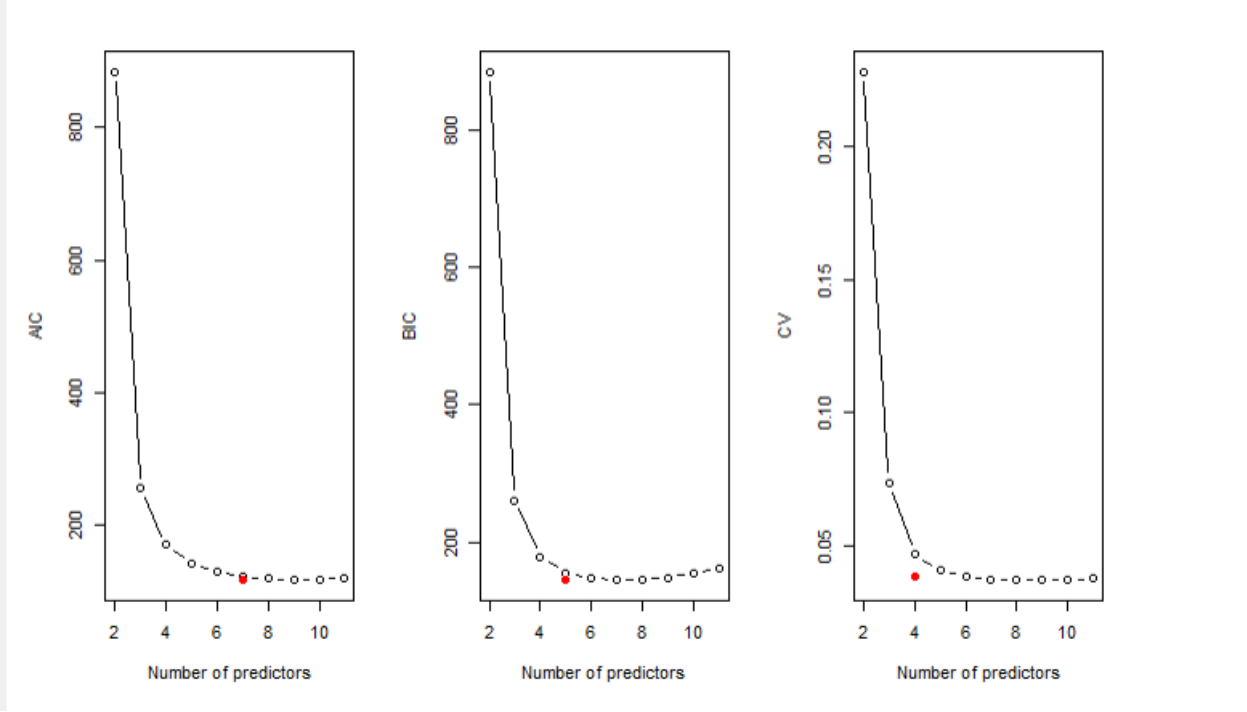

Fig 12: Best subset selection for the Breast Cancer data

From the plot it seems like the model with 5 predictors is best to predict the response variables (Class) for the Breast cancer data. We can also clearly see the elbow at k=5. The variables which are dropped out are Cell.size, Cell.shape, Epith.size, Mitoses.

Fitting the model to the 5 variables obtained by BIC (Cl.thickness, Marg.adhesion, Bare.nuclei, Bl.cromatin, Normal.nucleoli) and doing cross validation on the dataset and obtaining the test error.

```
[1] 0.03074671
```

Fitting the model with LASSO penalty:

In LASSO we add a penalty to the loss function which, in the case of logistic regression is the negative log-likelihood value.

We begin by specifying a grid of values for the tuning parameter, and then fitting the model with LASSO penalty for each value in the grid. Plotting the function to examine how the coefficients of each variable changes as the tuning parameter is increased.

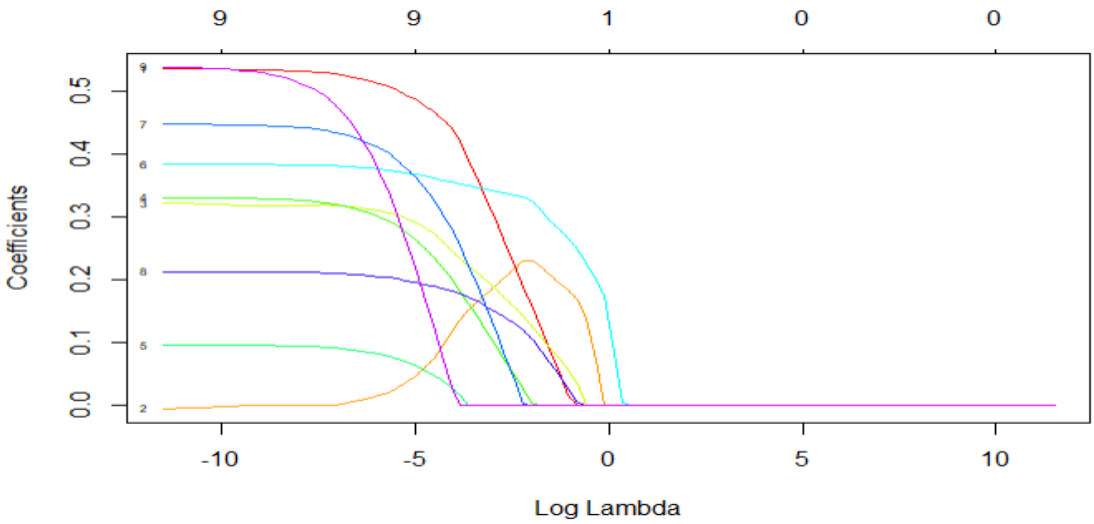

Fig 13: The effect of varying the tuning parameter in the logistic regression model with LASSO penalty for the Breast Cancer data.

From the plot we see that first one to drop out is Mitoses, followed by Epith.size, Bl.cromation and so on, last one to drop out is Bare.nuclei.

In order to choose the single value for the tuning parameter we will use cross validation, In order to visualize graphically how the test error varies with tuning parameter, we will pass the object returned by the cv.glmnet function to the plot function.

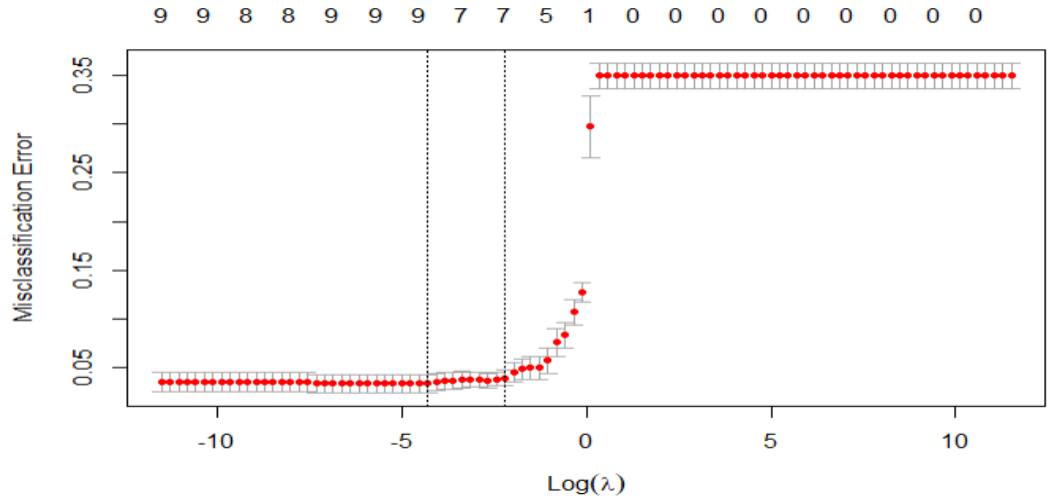

Fig 14: Cross-validation scores for the Breast cancer data using logistic regression with LASSO penalty.

We can now identify the optimal value for the tuning parameter and applying the cross the validation at the optimal value to find out the test error.

Optimal value:

```
[1] 0.01072267
[1] 70
```

Regression coefficients obtained by performing the lasso with chosen values of lambda minimum value are:

```
10 x 1 sparse Matrix of class "dgCMatrix"
                      s1
(Intercept)     -8.29080949
Cl.thickness     0.46662313
Cell.size        0.07406160
Cell.shape       0.27426482
Marg.adhesion    0.23672310
Epith.c.size     0.04874647
Bare.nuclei      0.36146492
Bl.cromatin      0.32806258
Normal.nucleoli  0.18923253
Mitoses          0.12827339
```

Fig 15: Regression coefficients obtained by performing the lasso with chosen values of lambda minimum value

From the above we see none of the variables are getting dropped, LASSO method is picking all the variables to build a model.

Test error:

```
[1] 0.03367496
```

Bayes classifier for linear discriminant analysis (LDA):

```
Call:
lda(y ~ ., data = X1)

Prior probabilities of groups:
        0         1
0.6500732 0.3499268

Group means:
  Cl.thickness Cell.size Cell.shape Marg.adhesion Epith.c.size Bare.nuclei Bl.cromatin Normal.nucleoli  Mitoses
0     2.963964  1.306306   1.414414      1.346847     2.108108    1.346847    2.083333        1.261261 1.065315
1     7.188285  6.577406   6.560669      5.585774     5.326360    7.627615    5.974895        5.857741 2.543933

Coefficients of linear discriminants:
                      LD1
Cl.thickness    0.182556105
Cell.size       0.125687035
Cell.shape      0.090054130
Marg.adhesion   0.047213478
Epith.c.size    0.057570551
Bare.nuclei     0.261447573
Bl.cromatin     0.110626289
Normal.nucleoli 0.106511431
Mitoses         0.008591172
```

Fig 16: Groups means for LDA

The LDA output indicates that our prior probabilities are 0.65 and 0.35, in other words 65% of the data is for benign and 35% data is for malignant. The LDA also provides group means which is the average of each predictor within the class, Which implies that female suffering with malignant on average has a Cl.thickness of 7.18, Cell.size of 6.5, Cell.shape of 6.5, Marg.adhesion of 5.5 etc. The female suffering with benign on average has a Cl.thickness of 2.96, Cell.size of 1.3, Cell.shape of 1.414 etc.

Bayes classifier for quadratic discriminant analysis (QDA):

```
Call:
qda(y ~ ., data = X1)

Prior probabilities of groups:
        0         1
0.6500732 0.3499268

Group means:
  Cl.thickness Cell.size Cell.shape Marg.adhesion Epith.c.size Bare.nuclei Bl.cromatin Normal.nucleoli  Mitoses
0     2.963964  1.306306   1.414414      1.346847     2.108108    1.346847    2.083333        1.261261 1.065315
1     7.188285  6.577406   6.560669      5.585774     5.326360    7.627615    5.974895        5.857741 2.543933
```

Fig 18: Groups means for QDA

The QDA output indicates that our prior probabilities are 0.65 and 0.35, in other words 65% of the data is for benign and 35% data is for malignant. The LDA also provides group means which is the average of each predictor within the class, Which implies that female suffering with malignant on average has a Cl.thickness of 7.18, Cell.size of 6.5, Cell.shape 6.5, Marg.adhesion 5.5 etc. The female suffering with benign on average has a Cl.thickness of 2.96, Cell.size of 1.3, Cell.shape of 1.414 etc.

From above we see that LDA and QDA are producing the same results

Test error for LDA using cross validation:

```
[1] 0.03806735
```

Test error for QDA using cross validation:

```
[1] 0.04685212
```

Table to represent different test error for different methods:

Description: df [5 x 2]

methods <chr>	Test_errors <dbl>
logistic regression on full dataset	0.03513909
best subset selection	0.03221083
LASSO	0.03367496
LDA	0.03806735
QDA	0.04685212

5 rows

Fig 19: Data representing different test errors

Test error on the least square using best subset selection is best method as it produces least test error value among other 5 methods. Using only 5 variables we are able to build a model which helps in predicting the response variable with low test error among 4 methods.