Wisconsin Breast Cancer

INTRODUCTION

Wisconsin Breast Cancer Database is part of the *mlbench* package in R. It has the data of the Biopsy procedure done on 699 women having breast cancer using FNAC to check the extremity of the cancer. The data was collected in between Jan 1989 to Nov 1991, whereas it was published in July 1992. The objective of the test was to determine whether the cancer was benign or malignant. There were 9 easily-assessed cytological characteristics collected, such as uniformity of cell size and shape, were measured for each tissue sample on a 1-10.

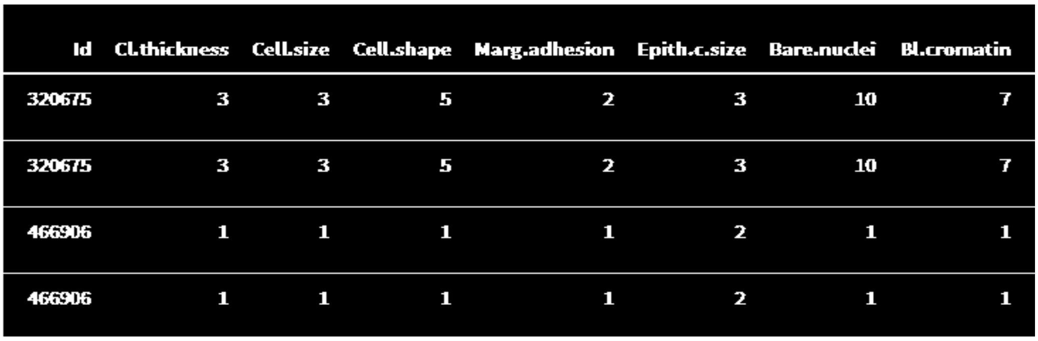
OBJECTIVE

The objective of the project is to build a classifier for **Class** - *benign or malignant* of a tissue sample based on some or all cytological characteristics.

EXPLANOTARY DATA ANALYSIS

Here, n = 699 and p = 11 which means that **n>p**, we can say that *Wisconsin Breast Cancer Database* is a *tall data*.

Based on variable *Id*, 645 observations out of 699 patients are unique identifiers. There are observations which are redundant and occur more than once like "897471". Since, these are the observations collected in 8 different periods and we can assume that some patients were evaluated several times during the campaign.



Id	Cl.thickness	Cell.size	Cell.shape	Marg.adhesion	Epith.c.size	Bare.nuclei	Bl.cromatin
320675	3	3	5	2	3	10	7
320675	3	3	5	2	3	10	7
466906	1	1	1	1	2	1	1
466906	1	1	1	1	2	1	1

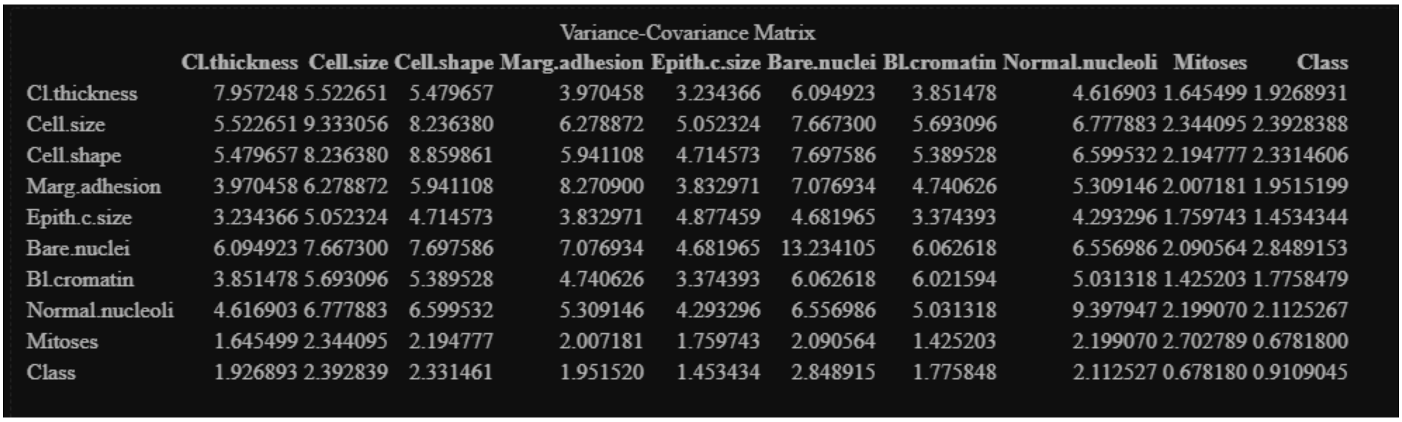
There are 9 cytological characteristics collected which are *Cl.thickness, Cell.size, Cell.shape, Marg.adhesion, Epith.c.size, Bare.nuclei, Bl.cromatin, Normal.nucleoli, Mitoses* in the data set which are on the measured on the range of 0-10. They will be treated as predictor variables. Since they are in the form of factor, we will convert them into quantitative variables. Finally, *Class* variable; a response variable is a factor with 2 levels - benign or malignant based on the predictor variables.

When working with categorical data, we to adopt the more standard 0 and 1 numerical labels. Since on the data origin server, it is prescribed to use 2 and 4 numerical labels, we will adapt the customised(data related approach).

- **2 - Benign**
- **4 - Malignant**

There are 16 observations on *Bare.nuclei* variable, encoded as **NA**. For the purpose of the project we remove those responses, after which data counts 683.

However, there is a concerning issue, 8 observations in data set are redundant which could lead to bias our analysis. This can be sorted by removing those observations, post which we will be left with *675* observations.



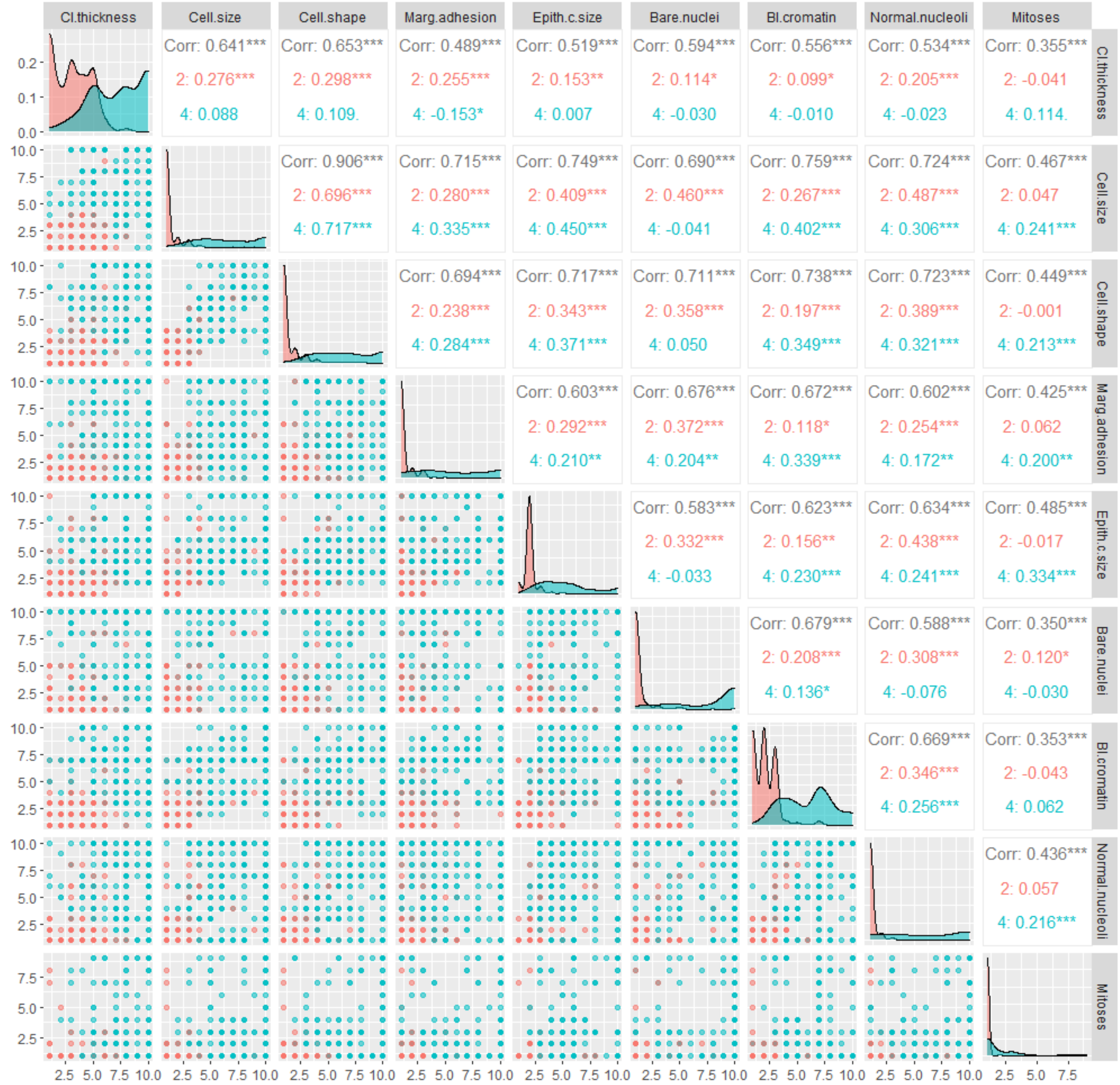
Variance-Covariance Matrix

	Cl.thickness	Cell.size	Cell.shape	Marg.adhesion	Epith.c.size	Bare.nuclei	Bl.cromatin	Normal.nucleoli	Mitoses	Class
Cl.thickness	7.957248	5.522651	5.479657	3.970458	3.234366	6.094923	3.851478	4.616903	1.645499	1.9268931
Cell.size	5.522651	9.333056	8.236380	6.278872	5.052324	7.667300	5.693096	6.777883	2.344095	2.3928388
Cell.shape	5.479657	8.236380	8.859861	5.941108	4.714573	7.697586	5.389528	6.599532	2.194777	2.3314606
Marg.adhesion	3.970458	6.278872	5.941108	8.270900	3.832971	7.076934	4.740626	5.309146	2.007181	1.9515199
Epith.c.size	3.234366	5.052324	4.714573	3.832971	4.877459	4.681965	3.374393	4.293296	1.759743	1.4534344
Bare.nuclei	6.094923	7.667300	7.697586	7.076934	4.681965	13.234105	6.062618	6.556986	2.090564	2.8489153
Bl.cromatin	3.851478	5.693096	5.389528	4.740626	3.374393	6.062618	6.021594	5.031318	1.425203	1.7758479
Normal.nucleoli	4.616903	6.777883	6.599532	5.309146	4.293296	6.556986	5.031318	9.397947	2.199070	2.1125267
Mitoses	1.645499	2.344095	2.194777	2.007181	1.759743	2.090564	1.425203	2.199070	2.702789	0.6781800
Class	1.926893	2.392839	2.331461	1.951520	1.453434	2.848915	1.775848	2.112527	0.678180	0.9109045

All the covariances are positive in sign which suggests that variables vary in same direction which implies positive increase of a variable means positive increase on the other variable. Eg: The increase in Cell size means increase in the marginal adhesion.

The correlation matrix and plot suggest that:

- Benign class has comparatively lower measure of most variables such as cell size and cell thickness, etc whereas Malignant tend to have higher measure.
- Relationship among the variables is positive overall, whereas there are a few variables if classified based on class which tend to have negative relation.
- Cell.size and Cell.shape has 0.906 correlation value overall which implies that they have good relationship. It appears to have decent value for benign and malignant class.
- Some variables like Bare.nuclei, Cl.thickness, Bl.cromatin, etc. seem to create a clear separation between benign and malignant tumors.

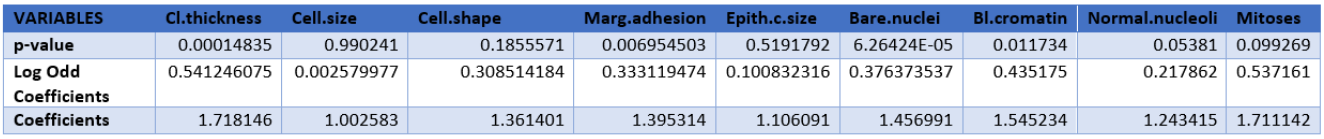


Correlation Matrix between Response Variables

LOGISTIC REGRESSION

Since all predictor variables are on the same 1-10 scale (1 meaning low and 10 high), we do not need to scale the data the variables.

Inspecting the coefficient table:



VARIABLES	Cl.thickness	Cell.size	Cell.shape	Marg.adhesion	Epith.c.size	Bare.nuclei	Bl.cromatin	Normal.nucleoli	Mitoses
p-value	0.00014835	0.990241	0.1855571	0.006954503	0.5191792	6.26424E-05	0.011734	0.05381	0.099269
Log Odd Coefficients	0.541246075	0.002579977	0.308514184	0.333119474	0.100832316	0.376373537	0.435175	0.217862	0.537161
Coefficients	1.718146	1.002583	1.361401	1.395314	1.106091	1.456991	1.545234	1.243415	1.711142

The column of p-values (i.e. Pr(>|t|) column) for the t-tests :

$$H_0 : \beta_i = 0 \quad \text{versus} \quad H_1 : \beta_i \neq 0$$

If we label Cl.thickness..Mitoses as X1..X9 then perform 9 hypothesis test. For exploratory variables Cell.size, Cell.shape and Epith.c.size have p-value greater than 0.5. This suggests that if we consider the predictor variables one-at-a-time, each of the predictors contribute very little to the model.

Therefore, all the subsets are not needed to be included in the regression. We can use subset selection method to choose the predictors.

IMPROVING THE MODEL

In Logistic Regression section, we saw that the full model contains more predictors than it needs. This suggests we might be able to produce a model that generates predictions which,

would show less variability (under repeated sampling), by employing one of the dimension-reduction/subset selection or regularisation techniques.

1. SUBSET SELECTION

Subset selection is used to find all the explanatory variables that are "best" at predicting the response variables. This technique reduces the variance in our parameter estimates and can therefore improve predictive performance.

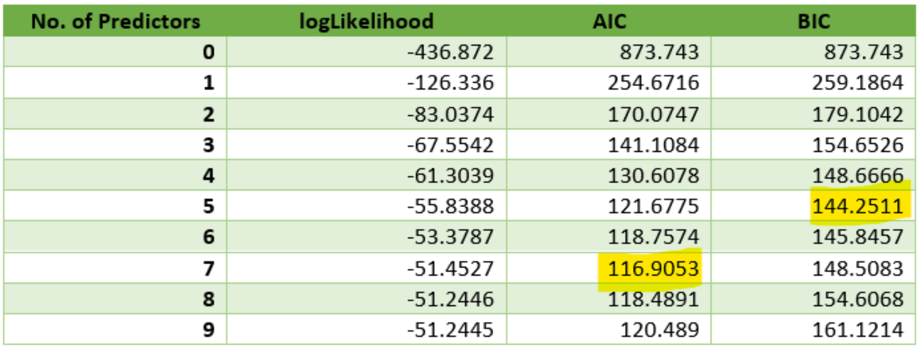
Best Subset Selection :

We will apply best subset selection using :

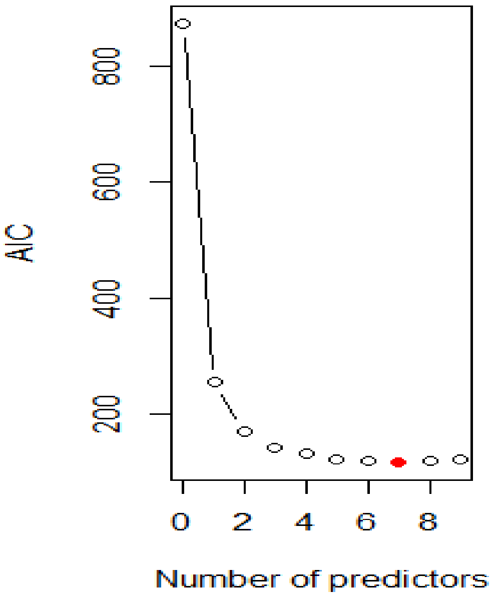
1. AIC (Akaike’s Information Criteria) :   A measure of the goodness of fit of any estimated statistical model.

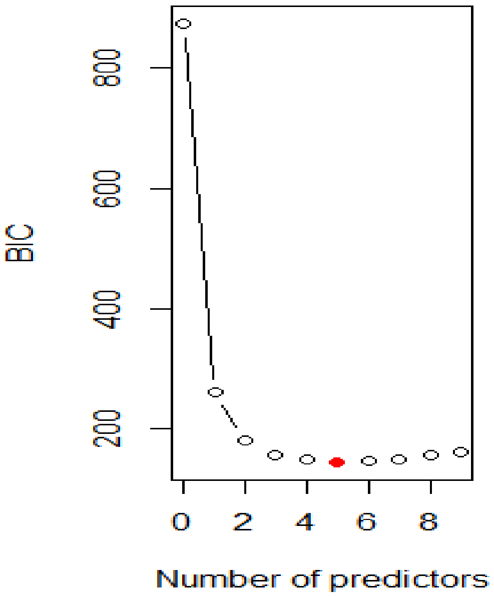
2. BIC (Bayes Information Criteria) : A type of model selection among a class of parametric models with different   numbers of parameters.

When comparing the BIC and the AIC, penalty for additional parameters is more in BIC than AIC.



No. of Predictors	logLikelihood	AIC	BIC
0	-436.872	873.743	873.743
1	-126.336	254.6716	259.1864
2	-83.0374	170.0747	179.1042
3	-67.5542	141.1084	154.6526
4	-61.3039	130.6078	148.6666
5	-55.8388	121.6775	144.2511
6	-53.3787	118.7574	145.8457
7	-51.4527	116.9053	148.5083
8	-51.2446	118.4891	154.6068
9	-51.2445	120.489	161.1214





From the plot above, we can depict that optimal value of variables in AIC and BIC is 7 and 5, respectively. It also seems like the model with six predictor is a good compromise. Hence, We extract the variables from the best-fitting 6-predictor model.

Logistic Regression on 6-Predictor Model :

We will construct a reduced data set containing only selected parameters to obtain the coefficients, p-value, and other data for the model.

Inspecting the coefficient table:

	Cl.thickness	Cell.shape	Marg.adhesion	Bare.nuclei	Bl.cromatin	Normal.nucleoli
Pr(>|z|)	5.34333E-06	0.037467	0.003222	7.1E-05	0.005619	0.022585
Log Odd Coefficients	0.6279049	0.343156	0.341267	0.372701	0.46128	0.24917
Coefficients	1.873681	1.409388	1.406729	1.45165	1.586102	1.28296

In the above, we can see that the selected model uses Cl.thickness..Normal.nucleoli variables and coefficients of the predictors in the 6-predictor model is significantly different from zero. All the predictors coefficients are also positive which suggests that higher value of the cytological characters are associated with a higher probability of a malignant class. All the p-vale seems to be lower than 0.5 which rejects the null hypothesis of dependency.

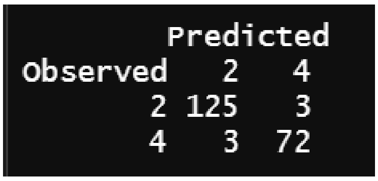
Logistic Regression uses logit link function to bend our line of best fit and convert the classification problem to regression. In order to interpret logistic regression coefficient, we exponentiate log odd coefficients. The column for coefficient (i.e Estimate) for Cl.thickness and Bl.cromatin seem to be significant variables, with higher values leading to a higher probability of malignant tumor.

Out-of-sample validation approach:

We will be checking the model's performance by Out-of-sample validation approach.

1. Validation Set Approach:

We estimate Test Error associated with predictor models based on random split of 70:30 into training and test data.



```
        Predicted
Observed   2   4
       2 125   3
       4   3  72
```

Confusion Matrix

In test data, we see that for cancer class - malignant, we correctly predict this on 72 out of 75 samples. However, cancer class - benign, 125 out of 128 were correctly predict. In both False Positive and False Negative is in 3 samples.

Test error is the proportion of misclassified observations, which is *0.02955665* . Therefore, **2.96%** observations are misclassified for 6-predictor model using Validation Set Approach.

Accuracy/success rate of the model is the proportion of correctly classified observations, which is *0.9704433*. Therefore, 97.04% samples are classified correctly.

2. k-Fold Cross Validation:

Test error can be estimated accurately by using k-fold cross validation. In this, we will be using 10 folds to do the validate the Best Subset Logistic Regression as we have limited number of samples. In it data will be split into 10 sub-sets and trained on remaining subset, calculating test error each time. We take average of the test error as the final test error.

Test Error : 0.03250658

It depicts that **3.25%** of the samples were mis-classified. 10-Fold Cross Validation seems be having higher test error than validation set approach.

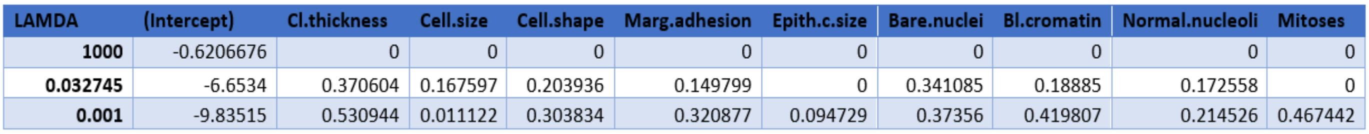
2. REGULARISATION

Regularisation is a technique used to reduce the error by fitting a function appropriately on the given training set and avoid overfitting.

LASSO

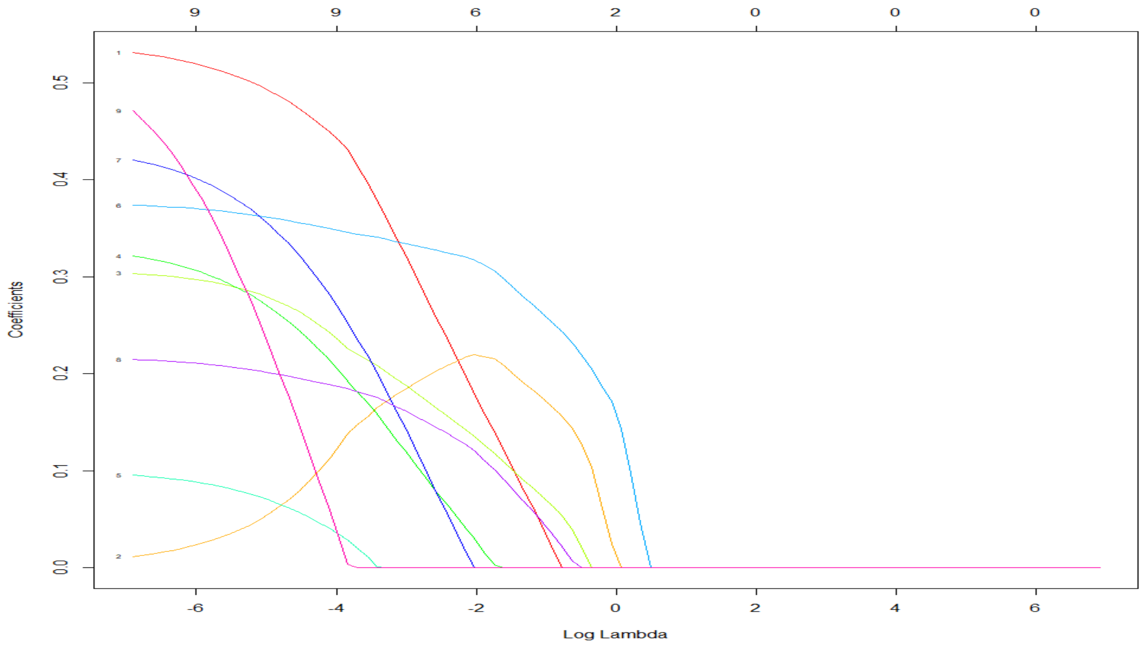
LASSO produces a regression model that is penalised with LASSO Regression. The consequence of this is that it shrinks coefficients and sets some coefficients to zero. It adds a penalty equivalent to the absolute magnitude of regression coefficients and tries to minimise them. We choose to incline towards LASSO over ridge regression is because the fitted model will always include all p explanatory variables in ridge regression. Also, the coefficients which are responsible for large variance are converted to zero in LASSO regularisation.

We choose grid values in the range of -3 to 3 for the tuning parameter.



LAMDA	(Intercept)	Cl.thickness	Cell.size	Cell.shape	Marg.adhesion	Epith.c.size	Bare.nuclei	Bl.cromatin	Normal.nucleoli	Mitoses
1000	-0.6206676	0	0	0	0	0	0	0	0	0
0.032745	-6.6534	0.370604	0.167597	0.203936	0.149799	0	0.341085	0.18885	0.172558	0
0.001	-9.83515	0.530944	0.011122	0.303834	0.320877	0.094729	0.37356	0.419807	0.214526	0.467442

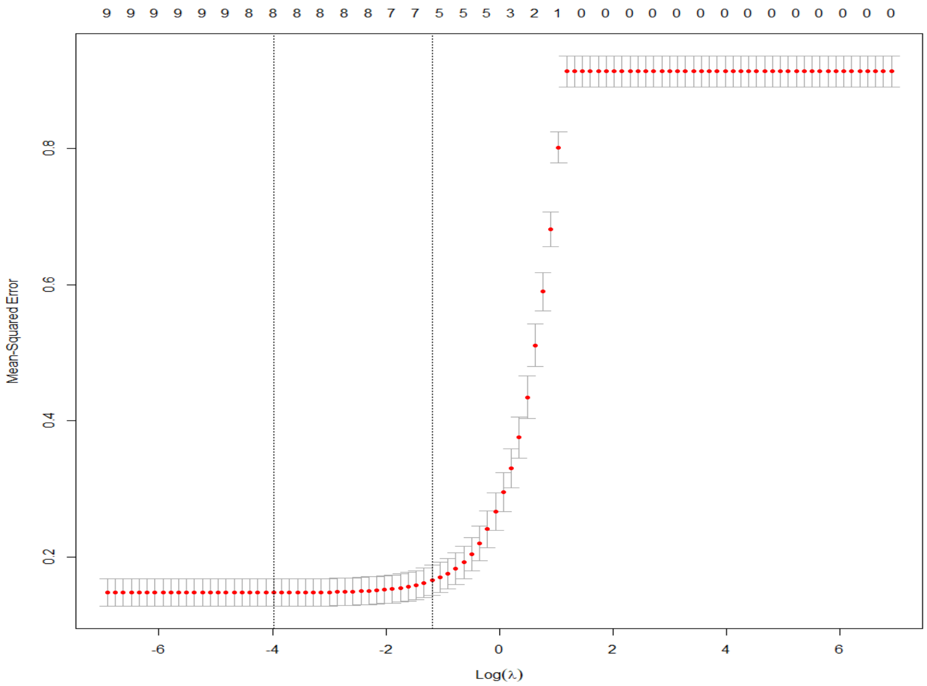
1. When λ = 1000, the regression coefficients for each explanatory variable are equal to zero. This is the null model which only includes an intercept term.
2. When λ = 0.032745, some of the coefficients are exactly equal to zero and the non-zero coefficients.
3. When λ = 0.001, we obtain a solution very close to the estimates for Logistic Regression.



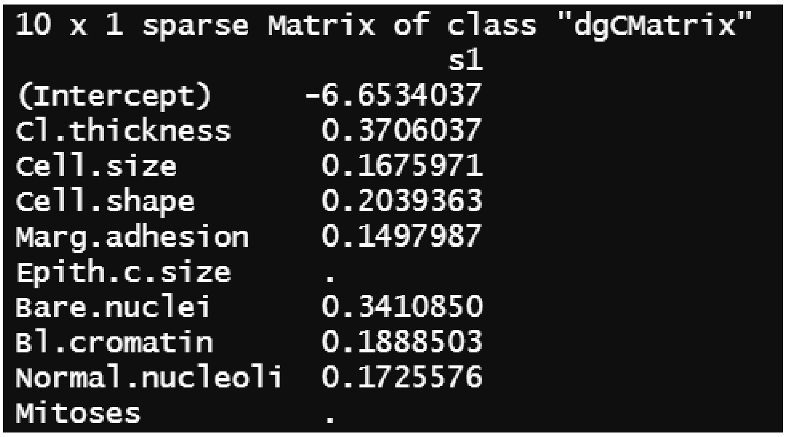
Each line represents the regression coefficient for a different variable. As the LASSO performs variable selection, in addition to shrinkage, we see variables drop from the model as the tuning parameter increases. The graph illustrates, as the value of tuning parameter, λ, increases we see that the ninth variable, Mitoses (magenta) is the first one to drop out of the model, followed by the fifth variable, Epith.c.size(teal) and the last one to drop is sixth variable, Bare.nuclei (light blue).

In order to select a single value for tuning parameter, we use Cross-validation.

We apply 10-fold cross validation with a random 10-fold partition.



The optimal value for tuning parameter is **0. 03274549**, below is the table for Parameter Estimate associated it. The graph above depicts the coefficients value through lamba value.



```
10 x 1 sparse Matrix of class "dgCMatrix"
                        s1
(Intercept)     -6.6534037
Cl.thickness     0.3706037
Cell.size        0.1675971
Cell.shape       0.2039363
Marg.adhesion    0.1497987
Epith.c.size     .
Bare.nuclei      0.3410850
Bl.cromatin      0.1888503
Normal.nucleoli  0.1725576
Mitoses          .
```

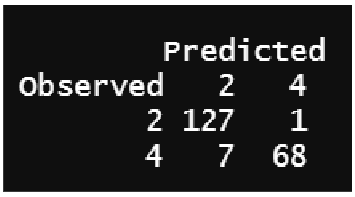
Like Best Subset Selection, predictors in LASSO like Epith.c.size and Mitoses are shrunk to zero. When we compare corresponding 6-predictor model (with no penalty) with 7-predictor model (with penalty), coefficients that have retained like Cell.size, Bl.cromatin has been shrunk more towards zero.

Out-of-sample validation approach:

We will be checking the model's performance by Out-of-sample validation approach.

1. Validation Set Approach:

We estimate Test Error associated with predictor models based on random split of 70:30 into training and test data.



```
        Predicted
Observed   2   4
       2 127   1
       4   7  68
```

Confusion Matrix

In test data, we see that for cancer class - malignant, we correctly predict this on 68 out of 75 samples. However, cancer class - benign, 127 out of 128 were correctly predict. In both False Positive and False Negative is in 1 and 7 samples respectively.

Test error is the proportion of misclassified observations, which is *0.03940887* . Therefore, **3.94%** observations are misclassified for 7-predictor model using Validation Set Approach.

2. k-Fold Cross Validation:

Test error can be estimated accurately by using k-fold cross validation. In this, we will be using 10 folds to do the validate the LASSO Logistic Regression as we have limited number of samples.

Test Error : 0.04450698

It depicts that **4.45%** of the samples were mis-classified. 10-Fold Cross Validation seems be having higher test error than validation set approach.

DISCRIMINANT ANALYSIS

Discriminant analysis models the distribution of the explanatory variables, X separately in each of the response variable, and then uses Bayes theorem to flip these around into estimates for the probability of the response category given the value of X.

Linear Discriminant Analysis

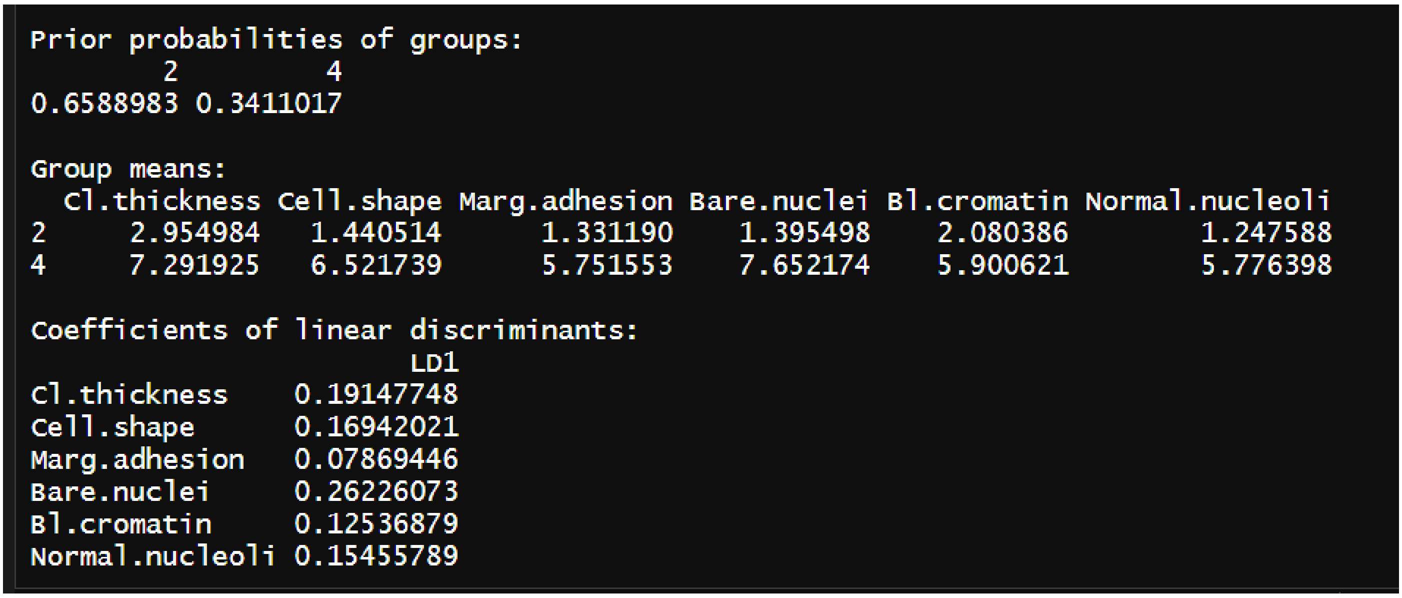
LDA is used to determine group means and also for each individual, it tries to compute the probability of the individual belonging to a different groups. Hence, that particular individual acquires the group with the highest probability score.

We will continue using 6-predictor model variables as predictor which we got from Best Subset Selection. We will use training data from Validation Set Approach for the same to build Bayes Classifier for LDA.

Discriminant Functions:

Q1 = -2.5353 +0.7379* x1 - 0.0322 * x2 + 0.1185 * x3 + 0.2123* x4 + 0.6525* x5 + 0.1585 * x6

Q2 = -23.4248 + 1.6683 * x1 + 0.8554 * x2 + 0.5009 * x3 + 1.4866 * x4 + 1.2617 * x5 + 0.9095 * x6



```
Prior probabilities of groups:
        2         4 
0.6588983 0.3411017 

Group means:
  Cl.thickness Cell.shape Marg.adhesion Bare.nuclei Bl.cromatin Normal.nucleoli
2     2.954984   1.440514      1.331190    1.395498    2.080386        1.247588
4     7.291925   6.521739      5.751553    7.652174    5.900621        5.776398

Coefficients of linear discriminants:
                       LD1
Cl.thickness    0.19147748
Cell.shape      0.16942021
Marg.adhesion   0.07869446
Bare.nuclei     0.26226073
Bl.cromatin     0.12536879
Normal.nucleoli 0.15455789
```

Prior probabilities of group: These represent the probability of each tumor class in the training set. It means that 65.89% of all observations in the training set were of class - benign and 34.11% were of malignant.

Group means: The mean values for each predictor variable for each class. For example, we see a clear difference between the proportion of Bare.nuclei (1.39 vs 7.65) for their tumor class. Therefore, More the difference between the mean, easier it is to classify observation.

Coefficients of linear discriminants: The linear combination of predictor variables that are used to form the decision rule of the LDA model. This defines the coefficient of the linear equation that is used to classify the response classes. Since, There are 2 response class in the model, only one set of LDA coefficients (LDA) is there

Proportion of trace: These display the percentage separation achieved by each linear discriminant function. Since, There is only 1 dimension for Linear Discriminant function, it is 1 or 100%.

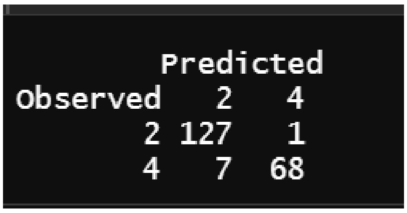
Out-of-sample validation approach:

We will be checking the model’s performance by Out-of-sample validation approach.

1. Validation Set Approach:

We estimate Test Error associated with predictor models based on random split of 70:30 into training and test data.

Confusion Matrix



```
        Predicted
Observed   2   4
       2 127   1
       4   7  68
```

In test data, we see that for cancer class - malignant, we correctly predict this on 68 out of 75 samples. However, cancer class - benign, 127 out of 128 were correctly predict. In both False Positive and False Negative is in 1 and 7 samples respectively.

Test error is the proportion of misclassified observations, which is *0.03940887*. Therefore, **3.94%** observations are misclassified for 6-predictor model using Validation Set Approach.

2. k-Fold Cross Validation:

Test error can be estimated accurately by using k-fold cross validation. In this, we will be using 10 folds to do the validate the using Linear Discriminant function as we have limited number of samples.

Test Error : 0.04438104

It depicts that **4.44%** of the samples were mis-classified. 10-Fold Cross Validation seems be having higher test error than validation set approach.

CONCLUSION :

In both the approach, Validation Set Approach and K-Fold Cross Validation – Best Subset Logistic Regression seems to have perform better with lower test error. It includes 6-predictor variables which are dependent on the better classification/prediction of the predictor variable.