Classifiers for Breast Cancer

Introduction

The objective of this project is to develop a classifier based on the nine cytological criteria to determine whether a tissue sample is benign or malignant. Since the dataset is based on the real-life situation, several classifiers are compared and analysed. For example, it can be distributed as 3 big parts which are subset selection, regularisation and discriminant analysis. Specifically, in subset selection, the best subset selection is conducted with three different methods – Adjusted R^2, Mallow's C_p Statistics and Bayes Information Criterion (BIC). Also, ridge regression and the LASSO are compared for regularisation and linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) for discriminant analysis.

Data Preparation

Formatting Data According to the *sapply()* function, except *Id* attributes all the columns are factor type. However, since it is hard to plot a graph with factor type, it is required to convert them into numeric variables. Using *as.numeric()* changes all the factor columns without *Class* attribute since the values are either 'benign' or 'malignant' which is not numeric.

Cleaning Data In addition, to improve data quality, it is essential to remove missing values from dataset. Thus, by implementing *drop_na()* all the rows with *NA* are deleted from the dataset.

Data Exploration

Numerical Summaries

As you can see from Appendix A numerical summaries, from *Cl.thickness* to *Mistoses* is numerical values and *Class* attribute is categorical. Since the purpose of the project is classifying the breast cancer type, it can be interpreted that *Class* attribute is the response variable and others – except *Id* – are predictor variables.

Graphical Summaries

Scatter Matrix

Since the dataset is multivariate data, scatterplot matrix is plotted between all predictor variables as a pair. Even though it is difficult to recognise an outstanding feature, the second variable *Cell.size* and the third variable *Cell.shape* describes a strong positive linear relationship compared to other pairs (Figure 1).

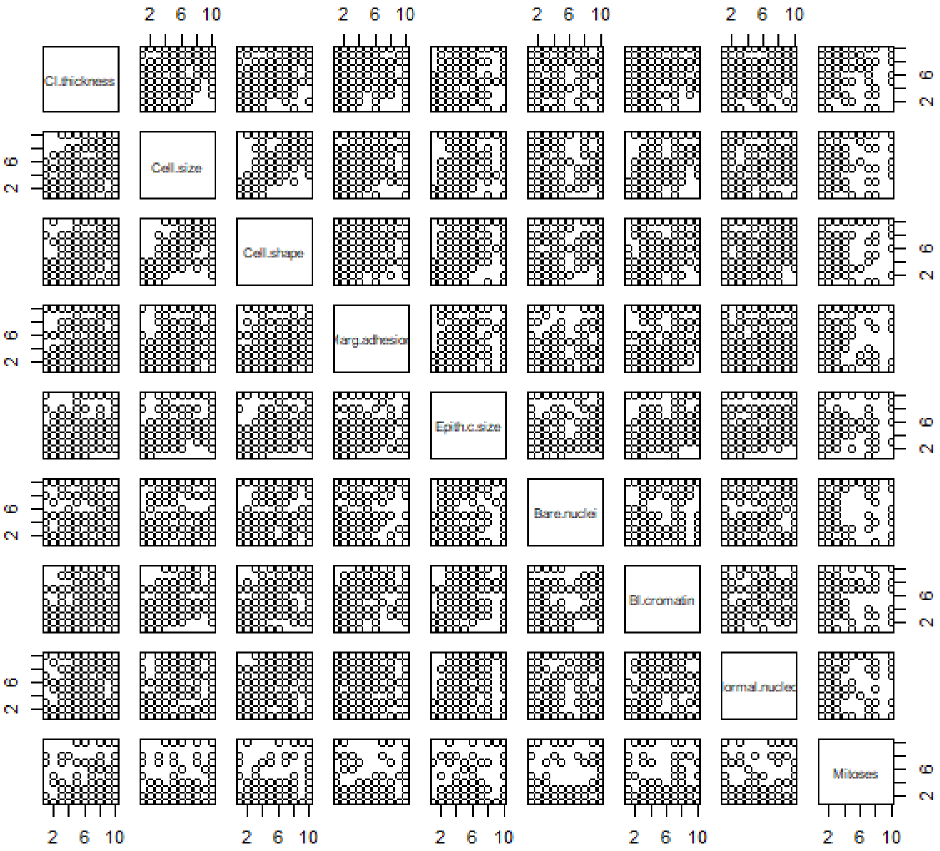
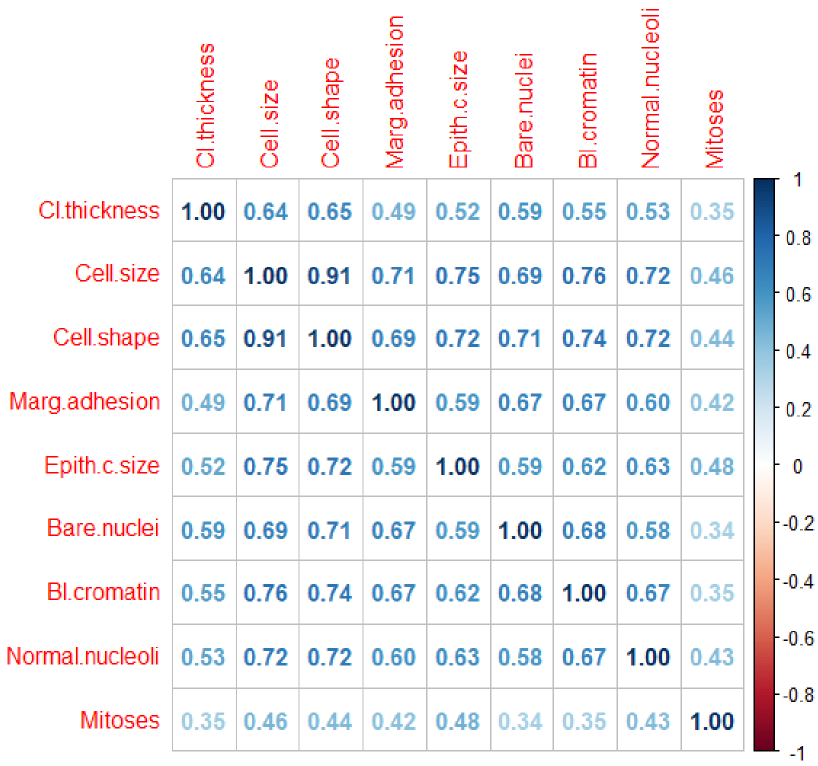


Figure 1: Scatter Matrix of predictor variables

Correlation Matrix

For the more precise investigation, correlation matrix is drawn with the values. As Figure 2 shows, *Cell.size* and *Cell.shape* has the highest correlation of 0.91 among others so it is confirmed that the findings from previous paragraph was correct. Furthermore, based on the numeric values, it shows other strong correlations that are not identified in scatter matrix. For instance, *Cell.size* and *Bl.cromatin* has the second highest value of 0.76.



	Cl.thickness	Cell.size	Cell.shape	Marg.adhesion	Epith.c.size	Bare.nuclei	Bl.cromatin	Normal.nucleoli	Mitoses
Cl.thickness	1.00	0.64	0.65	0.49	0.52	0.59	0.55	0.53	0.35
Cell.size	0.64	1.00	0.91	0.71	0.75	0.69	0.76	0.72	0.46
Cell.shape	0.65	0.91	1.00	0.69	0.72	0.71	0.74	0.72	0.44
Marg.adhesion	0.49	0.71	0.69	1.00	0.59	0.67	0.67	0.60	0.42
Epith.c.size	0.52	0.75	0.72	0.59	1.00	0.59	0.62	0.63	0.48
Bare.nuclei	0.59	0.69	0.71	0.67	0.59	1.00	0.68	0.58	0.34
Bl.cromatin	0.55	0.76	0.74	0.67	0.62	0.68	1.00	0.67	0.35
Normal.nucleoli	0.53	0.72	0.72	0.60	0.63	0.58	0.67	1.00	0.43
Mitoses	0.35	0.46	0.44	0.42	0.48	0.34	0.35	0.43	1.00

Figure 2: Correlation Matrix of predictor variables

Subset Selection

Among Best Subset Selection and Automated (stepwise) Selection, this project decides to apply best subset selection. Generally, dataset with large number of variables – over 40 variables – is suggested to execute automated (stepwise) selection. However, the dataset given here has only 9 predictor variables so that it is able to apply best subset selection. Additionally, automated (stepwise) selection does not consider all the possible models so using best subset selection would be beneficial in this case.

Best Subset Selection

Training/Validation set

Before implementing best subset selection, the whole dataset is split into training set and validation set. In order to achieve precise estimates of the test error, it is crucial to apply best subset selection to only training set. Especially, at the later stage, the cross-validation will not be accurate if best subset selection is conducted on the entire dataset.

Methods

This paper will compare the performance of three different methods of best subset selection – such as Adjusted R^2, Mallow's C_p Statistics and Bayes Information Criterion (BIC) – to select the best subset. From Figure 3, the highlighted dot represents the best performance with specific number of predictor variables. As Figure 3 describes, Adjusted R^2 suggests the model with 7 variables, Mallow's C_p with 7 variables and BIC with 5 variables. Nevertheless, it is hard to tell which number of variables is better since the result is different depending on the model.

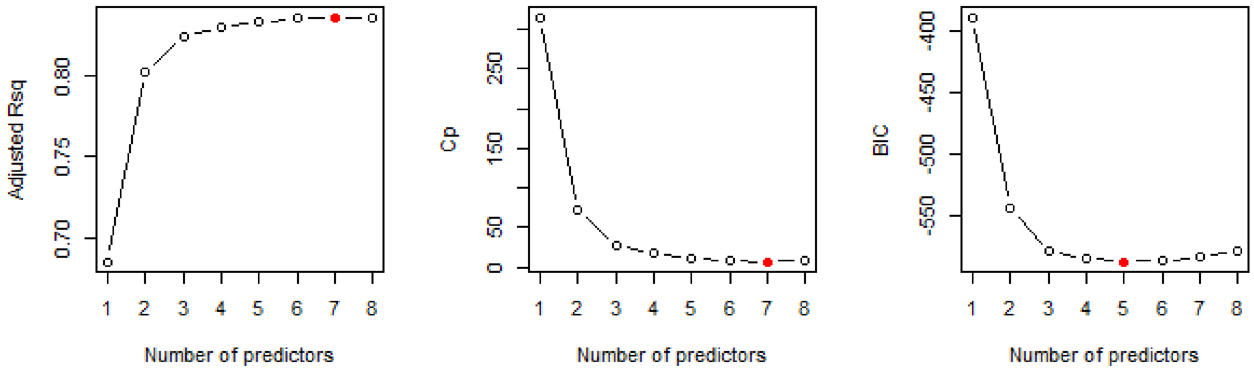


Figure 3: Performance depending on number of predictors (Left: Adjusted R^2, Middle: Mallow's C_p statistics, Right: BIC)

Cross-Validation

In order to derive the best subset, cross-validation is used for model selection by comparing the different size of models. Supposed that k is each of the number of training set, $k = 10$ folds are applied for the experiment. As a result, 5-predictor model reveals to be the best model according to cross-validation (Figure 4). To specify these 5 variables, best subset selection is generated with full data set and it turns out they were *Cl.thickness, Cell.size, Bare.nuclei, Bl.cromatin* and *Normal.nucleoli*.

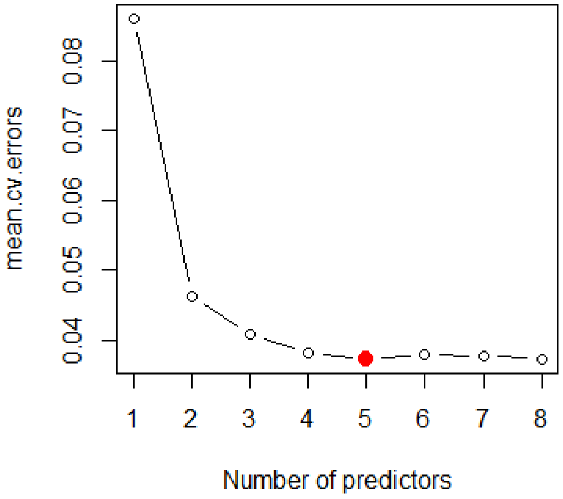


Figure 4: Performance with Cross-Validation

Regularisation

Regularisation involves applying a penalty to the model in order to limit the model's freedom and prevent overfitting. Especially, with linear models, the penalty is applied to the coefficients which multiply each predictor variables.

As mentioned previously, it is important to split the dataset into training and validation set for accurate estimation. Thus, regularisation also uses split dataset for both Ridge Regression and the LASSO.

Ridge Regression

Tuning Parameter

Ridge Regression is fitted with various range of the tuning parameter λ. The range of parameter is set from $\lambda = 10^5$ to $\lambda = 10^{-3}$ and the large value of lambda shrinks a lot and the small value does very little (Figure 5). In Figure 5, each line corresponds to the coefficient for a distinct variable and generally the lines shrink gradually.

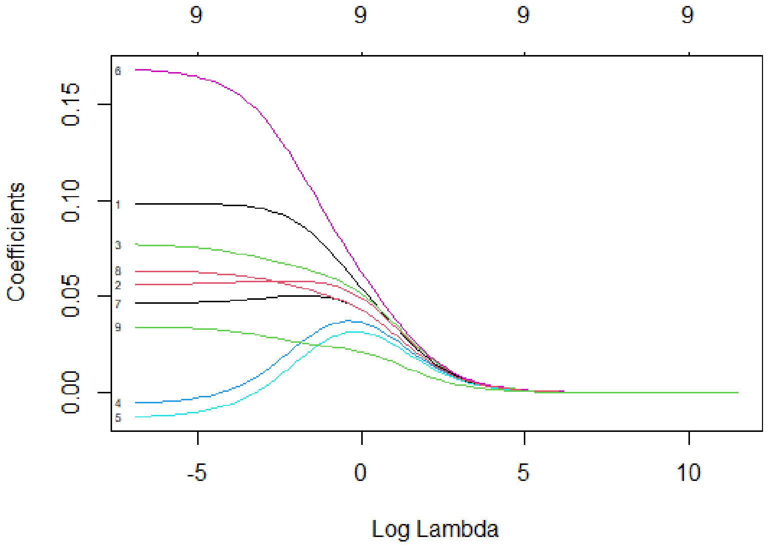


Figure 5: [Ridge Regression] Coefficients depending on value of tuning parameter

Cross-Validation

To find the best value for tuning parameter lambda, ten-fold cross validation is executed. Figure 6 displays the mean MSE across for each value of and error bars around which represent the plus and minus standard error. As Figure 6 shows, the standard error range is quite small as the error bar area is narrow. According to the result of cross-validation, the best lambda value is $\lambda = 0.046$ and the corresponding test mean squared error is 3.7%.

In consequence, the coefficients result by ridge regression fitted by the best lambda derived is shown in Table 1. The difference from best subset selection is that ridge regression does not show the final variables, instead it shows the coefficient value for each of them. As a result, only *Mitoses* has the smallest value which makes a difference from other values. Thus, *Mitoses* can be dropped from predictor variables.

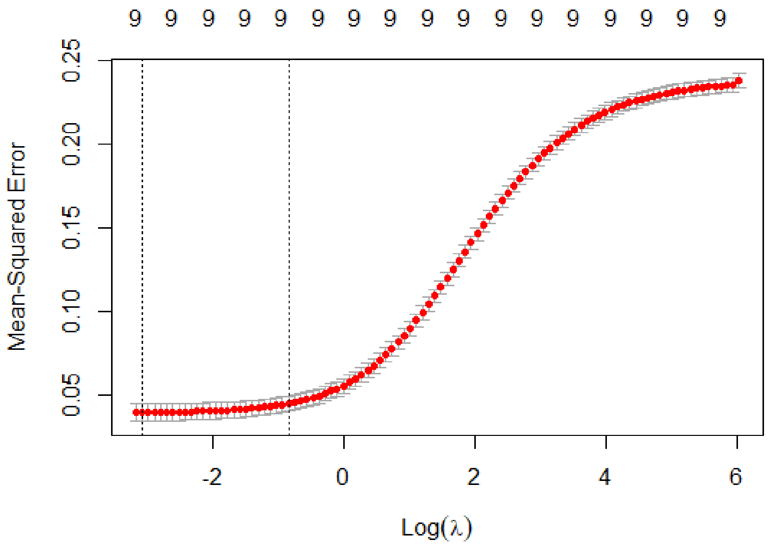


Figure 6: Cross-Validation score of ridge regression

Table 1: Ridge Regression – Coefficients with best lambda parameters

Cl.thickness	**Cell.size**	**Cell.shape**	**Marg.adhesion**	**Epith.c.size**	**Bare.nuclei**	**Bl.cromatin**	**Normal.nucleoli**	**Mitoses**
0.085373169	0.059655045	0.055477182	0.031992364	0.027792731	0.142086767	0.051430964	0.053873749	0.003704901

The LASSO

Tuning Parameter

In the same context as ridge regression, tuning parameter is performed by fitting the LASSO regression with each value of the set parameter. In this case, the parameter lambda is set in the range from $\lambda = 10^5$ to $\lambda = 10^{-3}$. As Figure 7 displays, except two cases, all the lines drop suddenly compared to ridge regression. Also, unlike ridge regression, the LASSO can make some of coefficients shrink to zero.

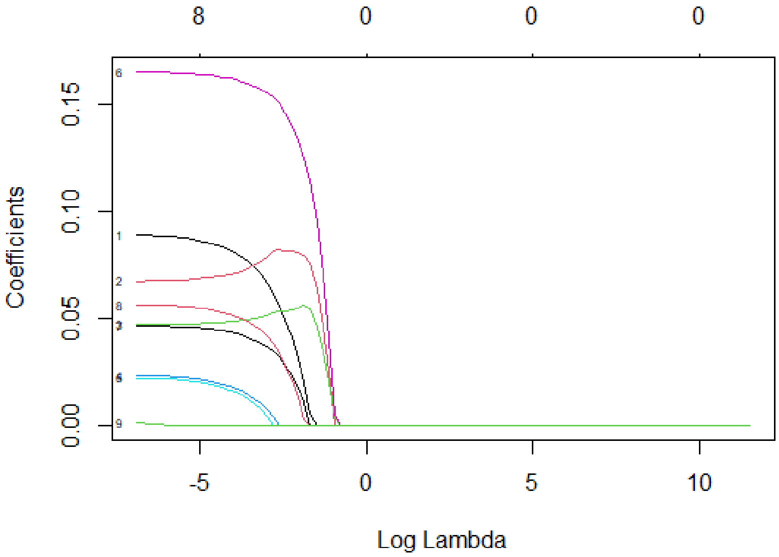


Figure 7: [The LASSO] Coefficients depending on value of tuning parameter

Cross-Validation

Ten-fold cross validation is implemented with the LASSO to find the smallest lambda value. As Figure 8 tells, the error bar range is not very large. As a result, the best lambda value which minimises cross-validation is $\lambda = 0.005$ and the test mean squared error with this lambda is 3.7% which is very small.

To conclude, the derived value $\lambda = 0.005$ from cross-validation is applied to the original dataset to examine the coefficient estimates. Table 2 explains the result of coefficients and *Mistoses* turn out to be zero, in other words, only 8 variables are chosen out of 9 by cross-validation.

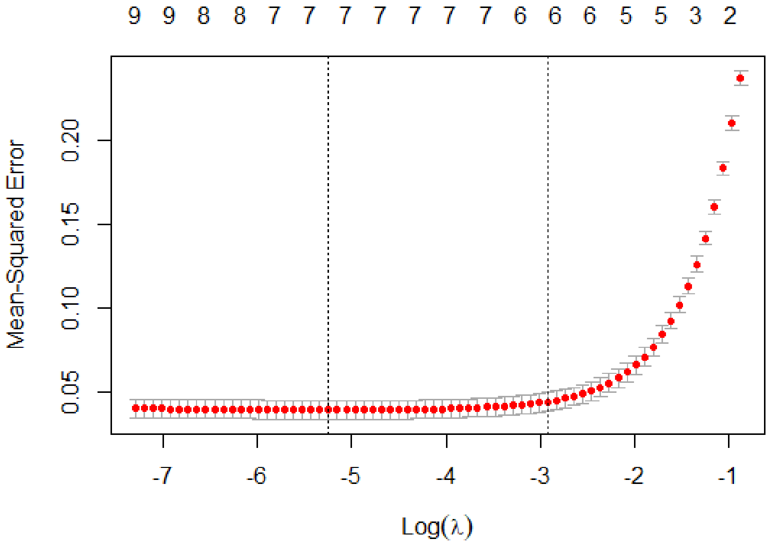


Figure 8: Cross-Validation score of the LASSO

Table 2: The LASSO - Coefficients with best lambda parameters

Cl.thickness	Cell.size	Cell.shape	Marg.adhesion	Epith.c.size	Bare.nuclei	Bl.cromatin	Normal.nucleoli	Mitoses
0.08704066	0.06847259	0.04743192	0.02209497	0.02073954	0.16424830	0.04570367	0.05518531	0.00000000

Discriminant Analysis

This paper investigates two different discriminant analysis, linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA). Generally, LDA and QDA both are derived from normal distribution. However, LDA assumes that there exists a common covariance matrix between all variables, on the contrary, QDA assumes that, instead of common covariance matrix, each variable has their own covariance matrix.

LDA and QDA calculate the group means and the likelihood that a given variable will belong to each group. The group with the greatest probability score is the one that is thereafter affected by the variable. Table 3 and 4 represents the group means of each variable and each class. On the left side, the class name 1 represents the benign and class 2 is malignant. As you can see, the group means for both LDA and QDA are the same.

Table 3: Group means of LDA

	CL.THICKNESS	CELL.SIZE	CELL.SHAPE	MARG.ADHESION	EPITH.C.SIZE	BARE.NUCLEI	BL.CROMATIN	NORMAL. NUCLEOLI	MITOSES
1	-0.5498028	-0.6009862	-0.6015684	-0.5159557	-0.5297767	-0.6076586	-0.5561858	-0.5255979	-0.3255607
2	0.9715256	1.1820348	1.1878516	1.0246090	0.9702928	1.1367901	0.9861525	0.9827057	0.5853208

Table 4: Group means of QDA

	CL.THICKNESS	CELL.SIZE	CELL.SHAPE	MARG.ADHESION	EPITH.C.SIZE	BARE.NUCLEI	BL.CROMATIN	NORMAL.NUCLEOLI	MITOSES
1	-0.5498028	-0.6009862	-0.6015684	-0.5159557	-0.5297767	-0.6076586	-0.5561858	-0.5255979	-0.3255607
2	0.9715256	1.1820348	1.1878516	1.0246090	0.9702928	1.1367901	0.9861525	0.9827057	0.5853208

In addition, the test error is computed for both of methods. Before calculation, it is expected to see lower error rate in QDA, however, LDA got 4.7% and QDA got 5.0%. Although the QDA results in higher rate, it is a slight difference of only 0.3%.

Furthermore, to obtain the better performance for both LDA and QDA, cross-validation approach is conducted. As Figure 9 describes, both methods show the precise prediction for each class. LDA predicted 226 right out of 230 for class 1 and 105 out of 115 accurate for class 2. In the same vein, QDA got the right 215 predictions of 230 for class 1 and 111 right out of 115 (Figure 10).

		Predicted Group	
		1	2
Actual Group	1	226	4
	2	10	105

Figure 9: Confusion matrix of LDA

		Predicted Group	
		1	2
Actual Group	1	215	15
	2	4	111

Figure 10: Confusion matrix of QDA

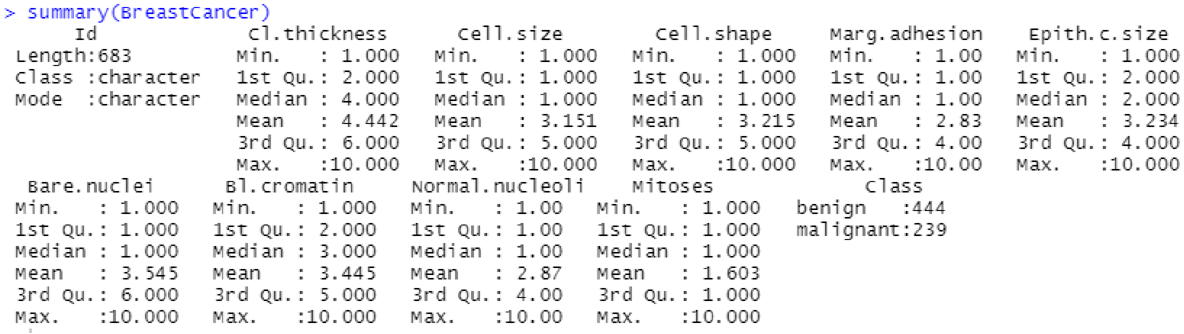
Conclusion

Among the results of test error, ridge regression and the LASSO regression have 3.7% rate and LDA and QDA have 4.7% and 5.0% respectively. Based on this mean squared error, ridge regression and the LASSO seems to be the best classifier. However, ridge regression cannot shrink to zero so it does not drop any predictor variables so the LASSO could be the best classifier among these two.

Moreover, the LASSO only drops one variable *Mitoses* which made the final dataset with 8 predictor variables. Also, as the best subset selection shows, three of methods – Adjusted R^2 and Mallow's C_p Statistics and BIC – selected 7 or 5 variables instead of whole 9 variables.

Appendix

Appendix A: Numerical Summaries of Dataset



```
> summary(BreastCancer)
      Id             Cl.thickness      Cell.size        Cell.shape      Marg.adhesion    Epith.c.size   
 Length:683         Min.   : 1.000   Min.   : 1.000   Min.   : 1.000   Min.   : 1.00   Min.   : 1.000  
 Class :character   1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.: 1.00   1st Qu.: 2.000  
 Mode  :character   Median : 4.000   Median : 1.000   Median : 1.000   Median : 1.00   Median : 2.000  
                    Mean   : 4.442   Mean   : 3.151   Mean   : 3.215   Mean   : 2.83   Mean   : 3.234  
                    3rd Qu.: 6.000   3rd Qu.: 5.000   3rd Qu.: 5.000   3rd Qu.: 4.00   3rd Qu.: 4.000  
                    Max.   :10.000   Max.   :10.000   Max.   :10.000   Max.   :10.00   Max.   :10.000  
  Bare.nuclei      Bl.cromatin     Normal.nucleoli    Mitoses             Class    
 Min.   : 1.000   Min.   : 1.000   Min.   : 1.00   Min.   : 1.000   benign   :444  
 1st Qu.: 1.000   1st Qu.: 2.000   1st Qu.: 1.00   1st Qu.: 1.000   malignant:239  
 Median : 1.000   Median : 3.000   Median : 1.00   Median : 1.000                  
 Mean   : 3.545   Mean   : 3.445   Mean   : 2.87   Mean   : 1.603                  
 3rd Qu.: 6.000   3rd Qu.: 5.000   3rd Qu.: 4.00   3rd Qu.: 1.000                  
 Max.   :10.000   Max.   :10.000   Max.   :10.00   Max.   :10.000                  
```