SINGAPORE MANAGEMENT UNIVERSITY

DSA211 Statistical Learning with R Project Report

Project Title:

Fighting the War on Diabetes

Group 1

Group Members

Name	Student ID
Aniket Bhat	01334947
Justin Wong Yi Jie	01333192
Bryan Lum Jin Yoong	01340099
Shawn Lim Hong Hao	01336698
Lau Jun Xiang	01335910

Executive Summary

Data analytics in healthcare can allow healthcare professionals and operators to gain insights into healthcare management, diagnoses and costs. The team utilized statistical modelling to perform diagnostic and predictive analytics.

Type 2 diabetes mellitus (T2DM) is characterized by a relative insulin deficiency or insulin resistance. It is also associated with a cluster of metabolic abnormalities, including hyper-tension and dyslipidemia. (Ogedengbe, 2016) The team examined variables that described Type 2 diabetes in women and created a predictive model that can reliably predict whether or not an individual has diabetes.

Concepts applied include Multiple Regression Model, Logistic Regression Model, Best Subset Selection, K-fold Cross Validation and Lasso. Additionally, the team supported their decisions by using confusion matrices and the Receiver Operating Characteristic curve (ROC), which connects all thresholds according to their respective False negative and False positive rate. The team explored a total of 4 models. Models 1, 2 and 3 are obtained using the Best Subset Selection Method based on the AIC, BIC and CV score criteria respectively. Model 4 uses the lasso method.

1. Motivation for Project

1.1 War on Diabetes in Singapore

1.1.1 Prevalence of diabetes in Singapore

In 2016, the World Health Organisation (WHO) stated that the number of adults estimated to be living with diabetes has nearly increased by 4 times over 35 years. A report in 2015 by the International Diabetes Federation or IDF revealed that Singapore has the second-highest proportion of diabetics among developed nations. (Lim, 2016)

Singapore has a rapidly ageing population, putting it at higher risk to chronic diseases. As such, it is essential for a country like Singapore, with limited resources and manpower, to study how susceptible our population are to these chronic illnesses, so that the supply for healthcare services is optimised and prepared for the rise in demand for healthcare.

Diabetes has been identified as one of the most prevalent chronic diseases in Singapore and the number of cases are rapidly increasing. In 2014, there were a total of 440,000 Singaporeans who have diabetes, and this number is expected to rise to 1 million by 2050. If current trends of diabetes continue, the prevalence of Type 2 diabetes is expected to be 1 in 2 adults (Phan et al., 2014). The graph below, adapted from the Ministry of Health Singapore, shows the significance of diabetes in Singapore.

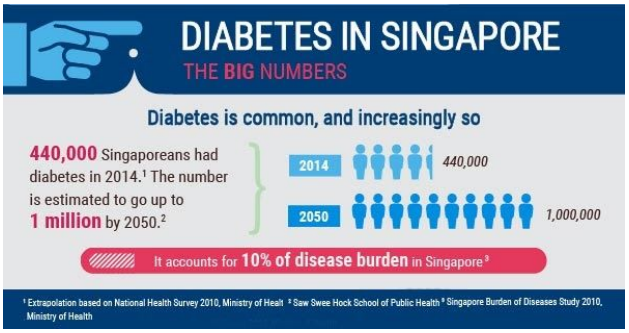


Figure 1: Diabetes in Singapore (Ministry of Health, 2010)

1.1.2 Prevalence of diabetes amongst women

With such a high expected prevalence of diabetes in Singapore, it is essential to identify the significant causes resulting in more cases of diabetes locally. Around 19,000 people are diagnosed with Type 1 or Type 2 diabetes annually in Singapore and many Type 2 diabetics are diagnosed to have contracted diabetes due to poor lifestyle choices. (Goh.Y.H, 2020)

The graph below shows that the prevalence of diabetes among Singapore residents aged 18 to 69 years has increased amongst both females and males from 2004 to 2010. For females, the growth rate between 2004 and 2010 is 36.8%. (HealthHub, 2019)

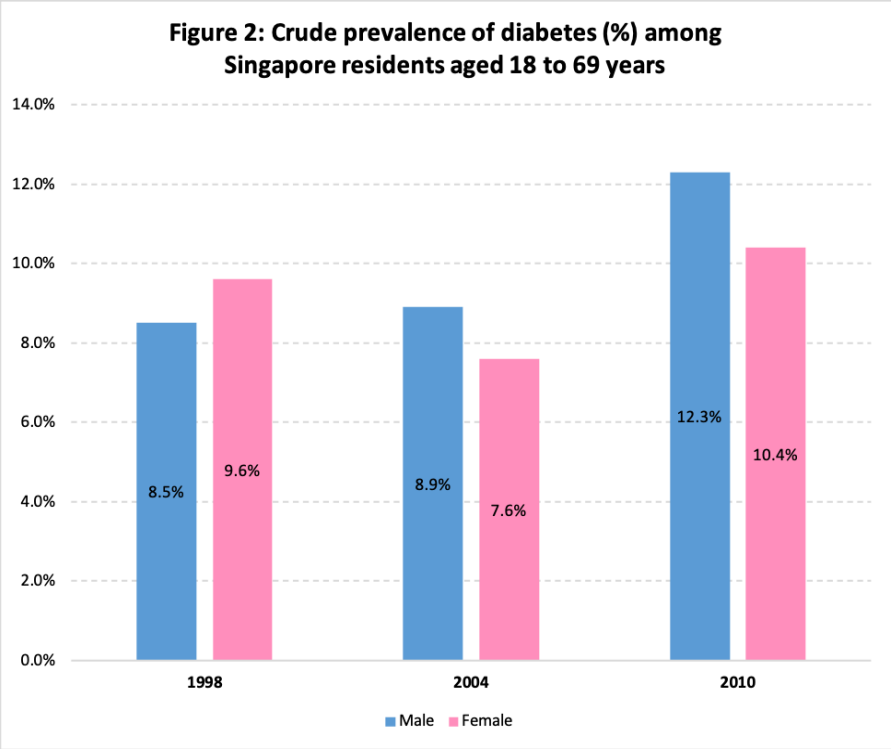


Figure 2: Crude prevalence of diabetes (%) among Singapore residents aged 18 to 69 years

With the increasing numbers amongst women, it is crucial to understand the different types of diabetes affecting women as the treatment and cause will be deferred from the general causes of Type 1 and Type 2 diabetes. Our group has identified pregnant women as our target of this study.

1.1.3 Impact of diabetes on pregnant women

Gestational diabetes is Type 2 diabetes that happens for the first time when a woman is pregnant. In most cases, gestational diabetes goes away after a woman delivers her baby. In some cases, women who have been diagnosed with gestational diabetes have an increased risk for developing Type 2 diabetes post-delivery. In addition, the child is also at risk for obesity and Type 2 diabetes. (medlineplus,2020)

Singapore has one of the highest rates of gestational diabetes in the world. (SingHealth, 2019) It is estimated that more than 6,000 pregnant women suffer from the condition every year, of which more than 4,000 will develop diabetes in their lifetime. Gestational diabetes is very common and affects about one in five pregnant women in Singapore. (NUHS,2017)

This study aims to identify the significance of pregnancy as a variable affecting the probability of being diagnosed with diabetes. Once a pregnant woman is diagnosed with gestational diabetes, it is crucial for them to control their blood sugar. Failure to do so can lead to health complications for the pregnant woman and the baby, which in some cases, can be severe and/or long-term. The impact on the child and the mother is summarised in Figure 3 below:

Implications	**Implication(s) on Mother**	**Implication(s) on Child**
Overweight Infant	● Results in discomfort to the mother towards the end of the pregnancy ● C-section	● Over-fed ● Type 2 diabetes from birth ● Could be born with nerve damage due to pressure on shoulder during delivery
C-Section (Cesarean Section)	● Longer to recover from childbirth	
High Blood Pressure (Preeclampsia)	● Could lead to seizures or stroke during labour and delivery	● Premature birth
Low Blood Sugar (Hypoglycemia)	● Could lead to fatality	● Could lead to fatality

Figure 3: Summary of Implications of Gestational Diabetes on Mother and Child (CGC.gov, 2020)

1.2 Objective of Project

Firstly, the project aims to identify significant variables that explain Type 2 diabetes amongst women. The healthcare ministry, policy makers and Singapore's healthcare sector can further expand the use of the model to include more relevant variables, such as smoking, drinking, and race, to develop a more accurate model that caters for different genders and demographic groups.

Secondly, the project aims to develop an accurate prediction model to provide female patient risk stratification for Type 2 diabetes in Singapore. With an accurate prediction model for Type 2 diabetes in women, healthcare professionals can utilise the prediction model and improve the accuracy of diagnosing a female with Type 2 diabetes.

1.2.1 Flow of the project

After developing the objective of the project, the project is planned in 4 main phases: **research, data analysis, development of prediction model, application of model**. The flowchart in Figure 4 illustrates the flow of the entire project, providing a visual aid behind the process of this study.

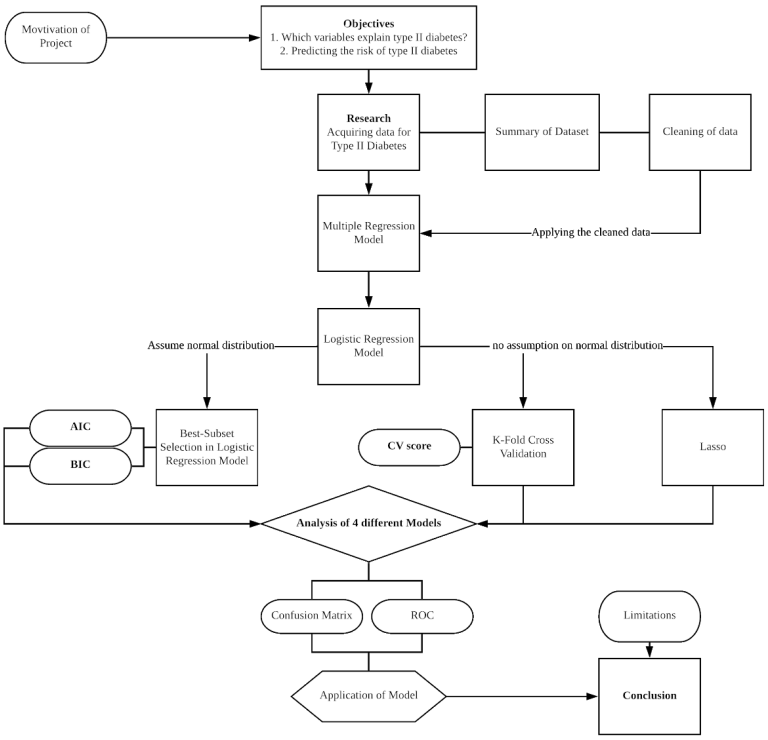


Figure 4: Project Flowchart

2. Methodology

2.1 Origin of Data

We obtained our dataset from a medical study done on the Pima Indian population near Phoenix, Arizona. The population has been identified and under continuous study since 1965 by the National Institute of Diabetes and Digestive and Kidney Diseases because of its high incidence rate of diabetes.

In the research paper, the study was conducted on females above 21 years of age over a period of 5 years. Amongst a pool of examinations, there was a criteria selection, whereby patients found to have diabetes within a year were removed as they were potentially easier to forecast.

The patients were given 75gm of a carbohydrate solution and their plasma glucose concentration was taken 2 hours after, and if it was > 200mg/dl, they were medically diagnosed to have diabetes. In addition, they recorded 8 variables which were identified as significant risk factors for diabetes in the medical field. These 8 observed variables are recorded in Figure 5 in Section 3.2.

The total sample size of the dataset originally consisted of 768 observations. Preliminary analysis shows us that there is incomplete data where certain entries = 0 or NA. We assume that these data were not recorded by the principal researchers and have chosen to remove them from our dataset to reduce systematic error. Examples of this includes Triceps Skin fold, Diastolic Blood Pressure, which should not take on values of 0. As a result, our cleaned dataset consists of a sample size of 532. (refer to R code in Appendix for cleaning of data)

2.2 Dataset

Variables	No. of Categories	Categories
Pregnancies: Number of times pregnant	3	(0,1,2) (3,4,5,6) (>7)
Glucose: Plasma Glucose Concentration at 2 hours in an oral Glucose Tolerance Test (GTT)	6	(0~89.1)(89.1~107.1) (blank, 107.2~123.1), (123.2~143.1) (143.2~165.1)(>165.2)
Blood Pressure: Diastolic Blood Pressure(mm Hg)	4	(blank) (1~76.1) (76.2~98.1) (>98.2)
Skin Thickness: Triceps Skin Fold	4	(blank) (1~25) (26~32) (>33)
Insulin: 2HR Serum Insulin(µU/ml)	5	(blank) (1~110) (111~150)(151~240) (>241)
BMI: Body Mass Index(Weight in kg /(Height in m)2	5	(1~22.814)(22.815~26.84) (blank~ 26.841~33.55) (33.551~35.563)(>35.564)
Diabetes Degree Pedigree Function[1]	5	(0~0.244)(0.245~0.525)(0.526~0.805)(0.806~ 1.11)(>1.11)
Age(years)	5	(21~24) (25~30) (31~40) (41~55) (>55)
Outcome: Takes the value of 1 if the individual has diabetes, 0 otherwise	2	(0)(1)

Figure 5: Input Variables

Source: Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus

2.3 Summary and Nature of Data

	Number of times pregnant	Plasma Glucose Concentration at 2 hours in an oral Glucose Tolerance Test (GTT)	Diastolic Blood Pressure (mm Hg)	Triceps Skin Fold	2HR Serum Insulin(µU/ml)	Body Mass Index(Weight in kg /(Height in m)2	Diabetes Degree Pedigree Function	Age(years)	Independent
Min.	0.00	56	24	7	0	18.20	0.0850	21	0
1st Qu.	1.00	98.75	64	22	0	27.88	0.2587	23	0
Median	2.000	115	72	29.00	91.5	32.8	0.416	28.00	0
Mean	3.52	121.0301	71.50564	29.18233	114.9887	32.89023	0.5029662	31.61466	0.3327
3rd Qu.	5.00	141.25	80.00	36	165.2	36.90	0.6585	38.00	1
Max.	17.000	199	110	99.00	846.0	67.10	2.42	81.00	1.0000

Figure 6: Summary Statistics of Dataset

[1] *The degree pedigree function(DPF) provides a measure of the expected genetic influence of affected and unaffected relatives on the subjects eventual diabetes risk.*

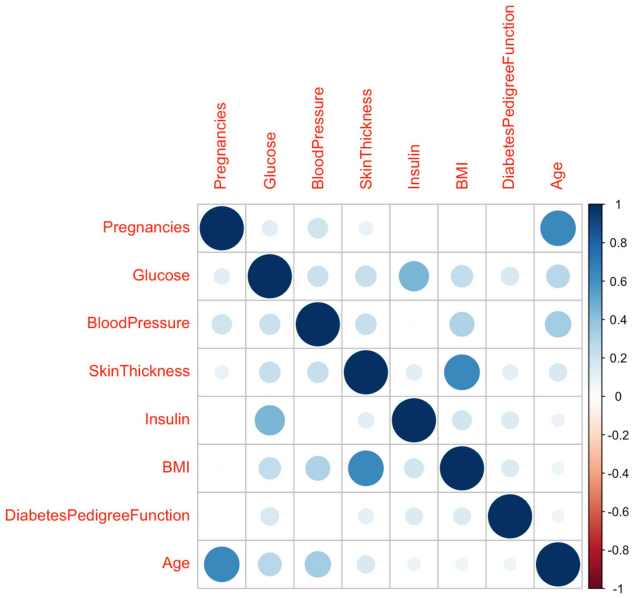


Figure 7: Correlation plot between each variable

In the correlation plot above, we observe significant positive correlations between (1) *Glucose* and *Insulin*, (2) *BMI* and *Skin Thickness*, (3) *Age* and *Pregnancy*.

(1) A rise in sugar triggers one's pancreas to release insulin into the bloodstream.

(2) Skin thickness is strongly associated with BMI. Skin layers become progressively thicker with increasing BMI (José et al., 2014)

(3) Age and Pregnancy is highly correlated due to biological factors affecting the fertility of a woman. Female fertility tends to decrease while a woman is in her early 30s, and the fall picks up after age 35. (Department of Health & Human Services, 2014)

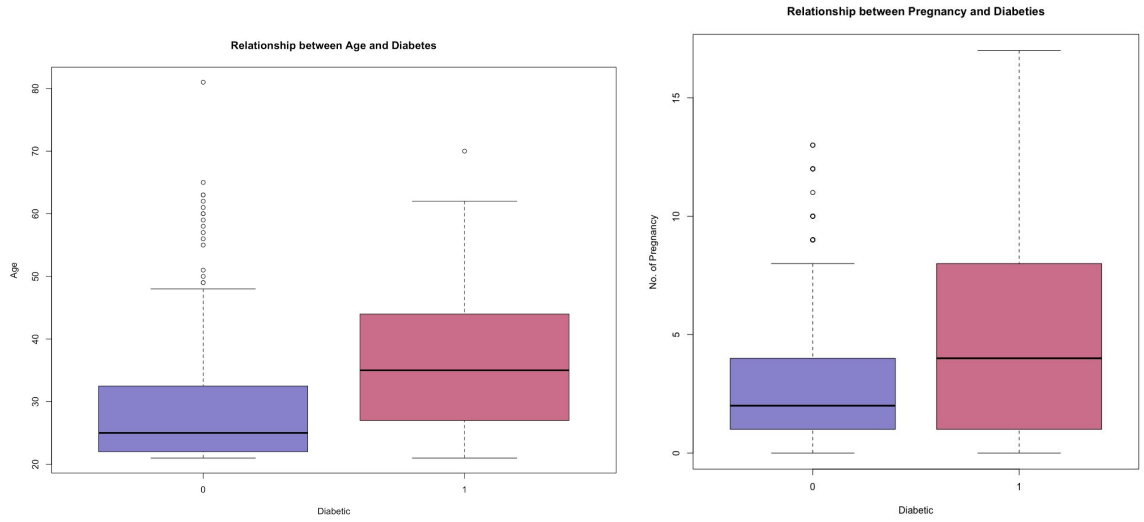


Figure 8: Boxplots of Age, Pregnancy against Diabetes

As seen from the boxplots above, we can see that older women and women with a higher number of pregnancies are more likely to have diabetes.

2.4 Limitations of the dataset

The dataset is specific to women who were Native American. However, while the sample may be limited to one specific race, the model we derive focuses on other significant variables that have been identified as possible significant variables contributing to Type 2 diabetes. A possible extension of our model could be to include race as a variable with Singapore's healthcare data, which the general public has restricted access to. In the BMJ Journal, the prevalence of Type 2 diabetes in Singapore was categorised according to Age-specific, gender-specific, and ethnicity-specific prevalence estimates and forecasts of (diagnosed and undiagnosed) Type 2 diabetes. (Phan, 2014) Medical experts in Singapore have noted a significant difference in the prevalence of diabetes between Chinese, Malays, and Indians. (Chiang, 2011)

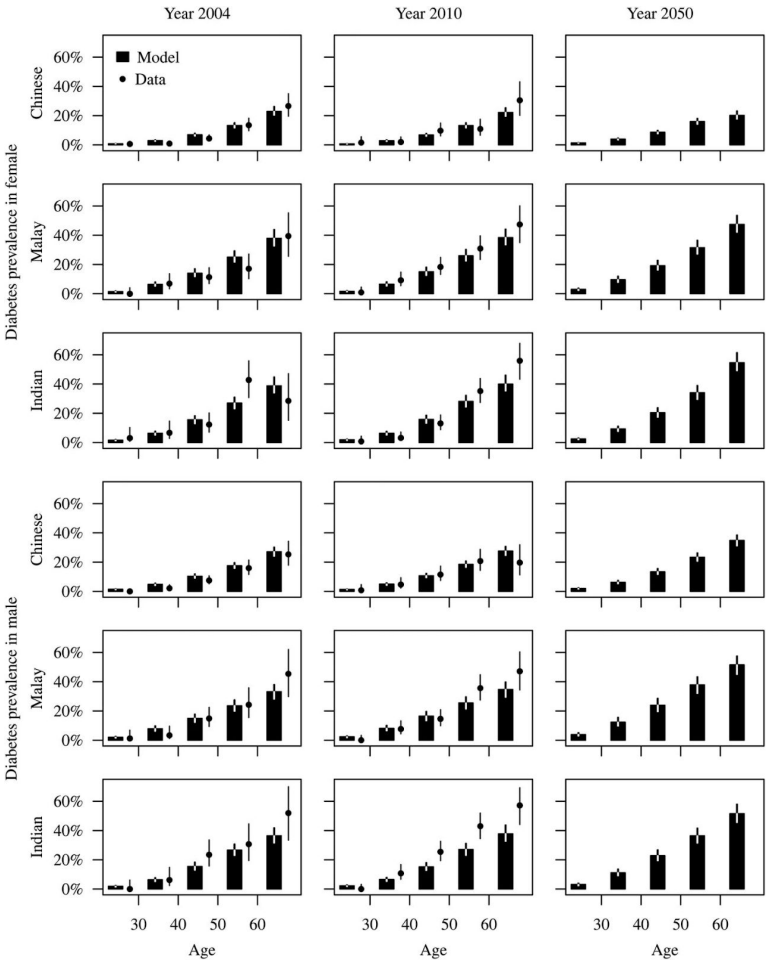


Figure 9: Prevalence of type 2 mellitus diabetes in Singapore by BMJ Journal

3. Modelling

3.1 Multiple OLS Linear Regression Model

We first regressed Outcome against the 8 other independent variables in a linear model. However as seen in the appendix, only Glucose, Pregnancies, BMI and Diabetes Pedigree function were shown to be significant.

Y : *Outcome*

$X_1 : Pregnancies$; $X_2 : Glucose$; $X_3 : BloodPressure$; $X_4 : SkinThickness$; $X_5 : Insulin$; $X_6 : BMI$

$X_7 : DiabetesPedigreeFunction$; $X_8 : Age$

Lm1:

$$Outcome = -1.0519076 + 0.0198196\,X_1 + 0.0063119\,X_2 - 0.0010991\,X_3 + 0.0005243\,X_4 - 0.0001592\,X_5 + 0.0122568\,X_6 + 0.1837957\,X_7 + 0.0043336\,X_8$$

Based on the residual analysis, and observation of the Q-Q and P-P plots, we have no reason to doubt the linear regression's assumptions. However, as the linear model was difficult to interpret we then did the logistic model (GLM).

3.2 Logistic Regression Model

$X_1 : Pregnancies$; $X_2 : Glucose$; $X_3 : BloodPressure$; $X_4 : SkinThickness$; $X_5 : Insulin$; $X_6 : BMI$
$X_7 : DiabetesPedigreeFunction$; $X_8 : Age$

$$P(Diabetes) = \frac{exp(-9.68 + 0.121\,X_1 + 0.0374\,X_2 - 0.00932\,X_3 + 0.00634\,X_4 - 0.00105\,X_5 + 0.086\,X_6 + 1.34\,X_7 + 0.0264\,X_8)}{1 + exp(-9.68 + 0.121\,X_1 + 0.0374\,X_2 - 0.00932\,X_3 + 0.00634\,X_4 - 0.00105\,X_5 + 0.086\,X_6 + 1.34\,X_7 + 0.0264\,X_8)}$$

$$log(\frac{P(Diabetes)}{1-P(Diabetes)}) = -9.68 + 0.121\,X_1 + 0.0374\,X_2 - 0.00932\,X_3 + 0.00634\,X_4 - 0.00105\,X_5 + 0.086\,X_6 + 1.34\,X_7 + 0.0264\,X_8$$

Interpretation of variables:

- A unit increase in *Pregnancies* will lead to a 12.9% increase in odds of getting diabetes, holding the other variables constant.
- A unit increase in *Glucose* will lead to a 3.8% increase in odds of getting diabetes, holding the other variables constant.
- A unit increase in *BloodPressure* will lead to a 0.93% decrease in odds of getting diabetes, holding the other variables constant.
- A unit increase in *SkinThickness* will lead to a 0.64% increase in odds of getting diabetes, holding the other variables constant.
- A unit increase in *Insulin* will lead to a 0.11% decrease in odds of getting diabetes, holding the other variables constant.
- A unit increase in *BMI* will lead to a 9.0% increase in odds of getting diabetes, holding the other variables constant.
- A unit increase in *DiabetesPedigreeFunction* will lead to a 280% increase in odds of getting diabetes, holding the other variables constant.
- A unit increase in *Age* will lead to a 2.68% increase in odds of getting diabetes, holding the other variables constant.

3.2.1 Logistic Regression Model assumptions

A logistic regression requires there to be little or no multicollinearity among the independent variables. This means that the independent variables should not be too highly correlated with each other. In order to test for multicollinearity, we would need to calculate the variance inflation factors (VIF) of the regressors. Multicollinearity is considered severe when the VIF value is greater than 10.

Running the *vif* function, we are able to retrieve the VIF values of the regressors.
Pregnancies: 1.892387 ; Glucose: 1.378937 ; BloodPressure: 1.191287 ; SkinThickness: 1.638865
Insulin: 1.384038 ; BMI: 1.832416 ; DiabetesPedigreeFunction: 1.031715 ; Age: 1.974053

As seen from the results, there are no signs to indicate that multicollinearity is present or severe.

3.3 Model Selection

For model selection, we employ the Best Subset Selection approach as it is feasibly exhaustive for the number of independent variables in the dataset. To carry out the Best Subset Selection process, we utilised the bestglm function as it is tailored for this purpose and possesses high functionality. It offers a variety of custom choice inputs and methods to cater for the various criteria available for choosing the best model.

3.3.1 Model 1: Best Subset Selection: AIC criteria

The first metric criterion we examine is the AIC, which reflects the amount of information lost by a model and estimates exogenous prediction error. A model carrying a low AIC value would thus be ideal as it would reflect less predictive information lost and consequently less predictive error. Observing AIC will help to reduce underfitting and high bias in the model. The bestglm function runs all possible models at the different combinations of total numbers of independent variables before gradually eliminating those with higher AIC values for each total number of independent variables in the model. The function narrows down to the lowest AIC model for each total number of independent variables before eventually selecting the model with the lowest AIC value as the best model. Figure 10 below illustrates the lowest AIC value models at each total number of independent variables included:

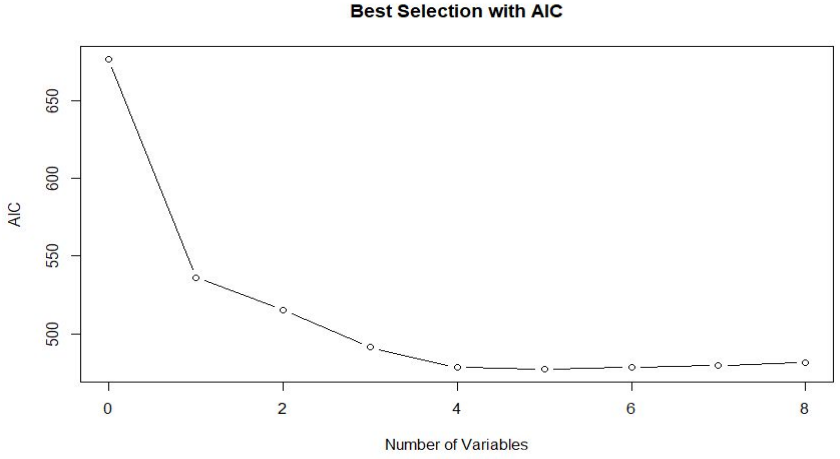


Figure 10: Best Selection with AIC

The best model, with the lowest relative AIC value of 477.0785 and standalone AIC value of 479.08, generated by the function contains the following 5 independent variables: Pregnanices, Glucose, BMI, DiabetesPedigreeFunction and Age. The output of the best AIC model is as follows:

$$log(\frac{P(Diabetes)}{1-P(Diabetes)}) = -9.879992 + 0.123887Pregnancies^{**} + 0.035026Glucose^{***} + 0.085123BMI^{***} + 1.321554DiabetesPedigreeFunction^{***} + 0.023844Age^{>}$$

Significance levels: '***' : 0 or effectively 0.0001, '**' : 0.001, '*' : 0.01, > : 0.05

3.3.2 Model 2: Best Subset Selection BIC criteria

The second metric criterion we examine is the BIC, which penalises additional and excessive parameters in a model and more so than AIC. A model carrying a low BIC value would thus be ideal as it would reflect less overfitting and variance in the model. The bestglm function runs all possible models at the different combinations of total numbers of independent variables before gradually

eliminating those with higher BIC values for each possible total number of independent variables in the model. The function narrows down to the lowest BIC model for each number of total independent variables before eventually selecting the model with the lowest BIC value as the best model. Figure 11 below illustrates the lowest BIC value models at each total number of independent variables included:

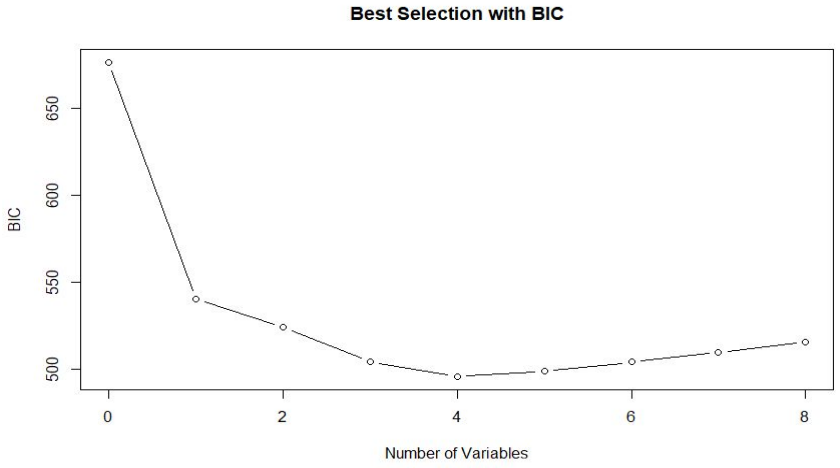


Figure 11: Best Selection with BIC

The best model, with the lowest BIC value of 495.4028, generated by the function contains the following 4 independent variables: Pregnanices, Glucose, BMI and DiabetesPedigreeFunction. The output of the best BIC model is as follows:

$$log(\frac{P(Diabetes)}{1-P(Diabetes)}) = -9.449410 + 0.172876\,Pregnancies^{***} + 0.036482 Glucose^{***} + 0.084232 BMI^{***} + 1.361746 DiabetesPedigreeFunction^{***}$$

Significance levels: '***' : 0 or effectively 0.0001, '**' : 0.001, '*' : 0.01, > : 0.05

3.3.3 Model 3: Best Subset Selection CV Score Criteria with 10 Fold Cross Validation

Lastly, we observe CV Score as a criterion from performing 10 Fold Cross Validation on the data at each combination of total number of independent variables to include in the model. The bestglm function allows us to apply the Best Subset Selection approach while simultaneously incorporating CV Score as a criterion metric to determine the best model. In this way we may combine the perks of minimal test error as well as exhaustive simulation from both Cross Validation and Best Subset Selection respectively. It is important to note that the bestglm function does not select the best model based on the conventional smallest CV error. Instead, bestglm computes a CV Score as well as an interval $CV \mp s/\sqrt{K}$, where CV is the mean of the 10 fold CV errors, s is the sample standard deviation of the 10 fold CV errors and K is the number of folds, in this case 10. Bestglm then finds the best CV Score at each total number of independent variables and the model with the best CV Score that is still within this interval, will be selected as the best model. This improves the stability of the K-fold Cross Validation method over picking the best model over the smallest mean 10 fold CV error. From this process, the function outputs the following best model:

$$log(\frac{P(Diabetes)}{1-P(Diabetes)}) = -8.831722 + 0.167525\,Pregnancies^{***} + 0.036666\,Glucose^{***} + 0.086421\,BMI^{***}$$

Significance levels: '***' : 0 or effectively 0.0001, '**' : 0.001, '*' : 0.01, > : 0.05

3.3.4 Model 4: Shrinkage approach Lasso

Lastly, we use the lasso shrinkage approach as a method of variable selection. The lasso approach shrinks the coefficient estimates towards zero but has the effect of forcing some of the coefficients to be exactly zero when the tuning parameter, λ is sufficiently large. To find the appropriate lambda, we split the data into a training and test set. We then perform the k-fold cross validation of the lasso with default k = 10 on the training set. We then select the λ which gives us the lowest binomial deviance, and obtain λ = 0.01. Next, we apply it to our test set to find the accuracy in the test set. Once satisfied, we write a function with the non-zero coefficient estimates to get our model:

$$log(\frac{P(Diabetes)}{1-P(Diabetes)} = -8.798973375 + 0.102940318Pregnancies + 0.032295005Glucose + 0.003706411SkinThickness + 0.068979711BMI + 1.063010860DiabetesPedigreeFunction + 0.021402651Age$$

We use a function to obtain an estimated probability of diabetes which we will save for use later in determining the accuracy of the model.

4. Analysing the results

4.1 Confusion Matrix

We derive the confusion matrix of all 4 models by obtaining the fitted values and using a low threshold of 0.25. We use a relatively lower threshold as in the field of healthcare, we err on the side of caution, if a patient shows even a low probability of diabetes, we would classify them as having diabetes and send them for further testing or treatment.

At the threshold of 0.25, the confusion matrices are as such:

Model 1: Accuracy: 74.81% (OER:25.19%)
False Negative Rate: 16.38%
False Positive Rate: 29.58%

pred1	0	1
Diabetes	105	148
No Diabetes	250	29

Model 2: Accuracy: 74.43% (OER: 25.57%
False Negative Rate: 16.95%
False Positive Rate: 29.86%

pred2	0	1
Diabetes	106	147
No Diabetes	249	30

Model 3: Accuracy: 72.93% (OER:27.07%)
False Negative Rate: 18.64%
False Positive Rate: 31.27%

pred3	0	1
Diabetes	111	144
No Diabetes	244	33

Model 4: Accuracy: 74.24% (OER:25.76%)
False Negative Rate: 15.25%
False Positive Rate: 30.99%

pred4	0	1
Diabetes	110	150
No Diabetes	245	27

From the confusion matrix table at a threshold of 0.25, it would seem that model 1, the AIC approach is the most accurate as it has the lowest Overall Error Rate of 25.19% in predicting whether the women had Diabetes or No Diabetes. However, in the sensitive field of medicine, having a low Overall Error Rate itself is not sufficient. More importantly, the False Negative Rate must be kept minimal as a high False Negative Rate would mean patients who actually have Diabetes are being misdiagnosed or misclassified as having No Diabetes and would thus be left untreated.

While having a high False Positive Rate could also be dangerous if measures like insulins shots are being given to people who in fact are healthy and do not have Diabetes, due to being asymptomatic and further tests, it is easier to identify and rectify the misclassifications of the False Positive. Hence, Model 4, the Lasso Method, would arguably be the best model as it has the lowest False Negative Rate amongst all the models while compromising just 0.57% of its accuracy (as compared to the lowest False positive ratemodel: Model 1) for 1.13% less in False Negative Rate.

4.2 Receiver Operating Characteristic: Area Under Curve (AUC)

To determine the predictive power of the 4 models across all thresholds, we generate the Receiver Operating Characteristic curve, which connects all thresholds according to their respective False negative and False positive rate. The area under the curve then quantifies the discriminative power of the 4 regression models. 100% AUC means that the model has full and perfect predictive power, while the 45 degree line from origin is the 50% AUC line, which suggests its predictive power is similar to a coin toss. The model which yields the highest AUC is determined to be the model with the highest predictive power. According to the AUC, Model 1 is the Best AIC model with 86.1%, following closely behind at 86.0% is the lasso model. The ROC curve is shown below:

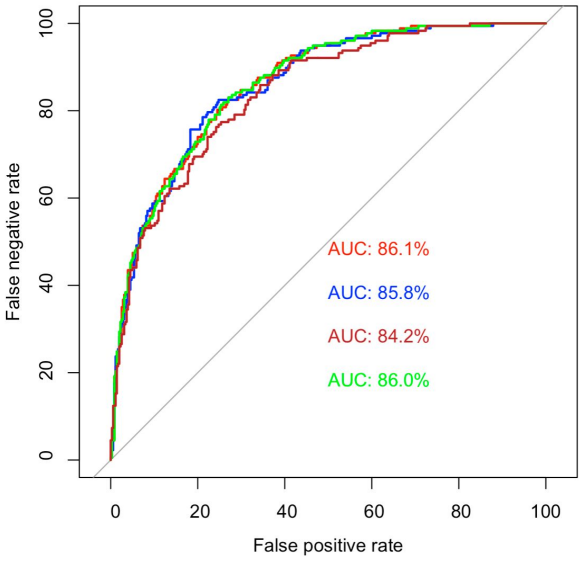


Figure 12: ROC Curve of the 4 Models
(Model 1 = Red, Model 2 = Blue, Model 3 = Brown, Model 4 = Green)

5. Recommendation

5.1 Selecting the Best model

Earlier in the report, our group mentioned the 2 objectives of this project:

1. Identify significant variables that explain Type 2 diabetes amongst women, and
2. To develop an accurate prediction model to provide female patient risk stratification for Type 2 diabetes in Singapore.

The results of objective 1: Model 1 (Best AIC) remains to be the most explanatory model as it retains more information by having more explanatory variables than Model 2 and 3, while also maintaining its fit of the data by retaining more error minimizing coefficient estimates. The latter's strength of the retainment of its efficient coefficient estimates is unlike what we see from the Lasso model

which comprises its explanatory power for lower prediction error. This is because the effects of the independent variables in the model are forcibly shrunk to reduce the variance of the model, as seen in the confusion matrix and ROC comparisons. Therefore, we determine that Model 1 would be the best model in determining significant variables that explain Type 2 diabetes amongst women.
Model 1:

$$log(\frac{P(Diabetes)}{1-P(Diabetes)}) = -9.879992 + 0.123887Pregnancies + 0.035026Glucose + 0.085123BMI + 1.321554DiabetesPedigreeFunction + 0.023844Age$$

Healthcare professionals can take extra note of an individual's number of pregnancies, glucose levels, BMI, Diabetes pedigree function and age are the variables that are more likely to cause diabetes. Notably, this model is in *ln(odds ratio)* and requires conversion to obtain the probability of diabetes.

As for Objective 2, AIC still performs well in predicting the outcome of diabetes as seen by high accuracy in the confusion matrix and high AUC for ROC. However, when we want to predict what will happen in the future, given a new patient, it is more appropriate to use the cross validation of the lasso. Furthermore, the difference in AUC between Model 1 and 4 is only 0.1 %. In regularisation methods such as the lasso approach, the models will give us the smallest least square errors, or in the case of logistic regression, the lowest binomial deviance. In addition, it doesn't require the assumption of a normal distribution. Lastly, as mentioned in the previous section, Model 4 has the lowest False Negative Rate amongst the 4 models. In the real medical field, professionals would rather compromise the 0.57% in Overall Error Rate for 1.13% saved in False Negative Rate. Hence, Model 4 as a predictive model is recommended in the healthcare sector over Model 1. The lasso model is as such:

$$log(\frac{P(Diabetes)}{1-P(Diabetes)} = -8.798973375 + 0.102940318Pregnancies + 0.032295005Glucose + 0.003706411SkinThickness + 0.068979711BMI + 1.063010860DiabetesPedigreeFunction + 0.021402651Age$$

6. Conclusion

In conclusion, using the AIC model we identified that the more significant factors that result in diabetes in women were Pregnancy, Glucose, BMI, DiabetesPedigreeFunction and Age. As seen in 5.1, their respective coefficients were $0.123887Pregnancies$, $0.035026Glucose$, $0.085123BMI$ $1.321554DiabetesPedigreeFunction\,(DPF)$, $0.023844Age$. As such, it can be identified that the order of significance would be DPF, Pregnancies, BMI, Glucose and Age. This is useful information as it allows healthcare individuals to know that it would be more important to allocate resources to women with higher DPF or that are pregnant, as compared to women with high BMI or high glucose level, or elderly women.

Additionally, we believe that our lasso model (Model 4), will help healthcare professionals better identify patients who are at risk of developing diabetes. Firstly, by reducing the number of independent variables to a small number of 6, we believe that its predictive power is relatively high, as seen by the AUC = 86.0% and OER = 25.76%. Additionally, the low binomial deviance of λ = 0.01, indicates that the variance of our model is relatively low and should be relatively accurate.

7. References

Chiang, P. P. C., Lamoureux, E. L., Cheung, C. Y., Sabanayagam, C., Wong, W., Tai, E. S., ... Wong, T. Y. (2011, September 1). Racial Differences in the Prevalence of Diabetes but Not Diabetic Retinopathy in a Multi-ethnic Asian Population. Retrieved from https://iovs.arvojournals.org/article.aspx?articleid=2165746

Department of Health & Human Services. (2014, February 28). Age and fertility. Retrieved from https://www.betterhealth.vic.gov.au/health/conditionsandtreatments/age-and-fertility

Derraik, J. G. B., Rademaker, M., Cutfield, W. S., Pinto, T. E., Tregurtha, S., Faherty, A., ... Hofman, P. L. (2014, January 21). Effects of age, gender, BMI, and anatomical site on skin thickness in children and adults with diabetes. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3897752/

Diabetes in Singapore. (n.d.). Retrieved from https://www.healthhub.sg/a-z/diseases-and-conditions/626/diabetes

Gestational Diabetes and Pregnancy. (2020, February 27). Retrieved from https://www.cdc.gov/pregnancy/diabetes-gestational.html

Gestational Diabetes . (2020, February 13). Retrieved from https://medlineplus.gov/diabetesandpregnancy.html

Han, G. Y. (2020, February 26). Parliament: 19,000 diagnosed with diabetes yearly, more expected to be diagnosed in short term, says MOH. Retrieved from https://www.straitstimes.com/politics/parliament-19000-diagnosed-with-diabetes-yearly-more-expected-to-be-diagnosed-in-short-term

José G. B. Derraik, Marius Rademaker, Wayne S. Cutfield, Teresa E. Pinto, Sheryl Tregurtha, Ann Faherty, Jane M. Peart, Paul L. Drury, Paul L. Hofman. (2014, January 21). Effects of Age, Gender, BMI, and Anatomical Site on Skin Thickness in Children and Adults with Diabetes. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3897752/

Ogedengbe, S., Ezeani, I. U., & Aihanuwa, E. (2016, January). Comparison of clinical and biochemical variables in type 2 diabetes mellitus patients and their first-degree relatives with metabolic

syndrome in Benin City, Nigeria: A cross sectional case controlled study. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/27560635

Phan, T. P., Alkema, L., Tai, S., Tan, K. H. X., Qian Yang, W.-Y. L., Yik Ying Teo, C.-Y. C., ... Cook, A. R. (2014, June 1). Forecasting the burden of type 2 diabetes in Singapore using a demographic epidemiological model of Singapore. Retrieved from https://drc.bmj.com/content/2/1/e000012

Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S. (1988, November 9). Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2245318/

War on Diabetes. (2019, February). Retrieved from https://www.singhealth.com.sg/rhs/keep-well/war-on-diabetes

Xu, C., & McLeod, A. I. (2010). bestglm: Best Subset GLM. Western Ontario, Canada. Retrieved from http://www2.uaem.mx/r-mirror/web/packages/bestglm/vignettes/bestglm.pdf

7.2 R Codes

```
library(dplyr)
library(fitdistrplus)
library(ggplot2)
library(pROC)
library(car)
library(PresenceAbsence)
library(leaps)
library(glmnet)
library(bestglm)
library(corrplot)
library(dismo)

Original_Diabetes <- read.csv("~/Desktop/Original_Diabetes.csv")
summary(Original_Diabetes)

Diabetes <- Original_Diabetes %>%
  filter(SkinThickness!=0 & BloodPressure != 0 & BMI!= 0 & Glucose!=0)

summary(Diabetes)

boxplot(Diabetes$Pregnancies~Diabetes$Outcome, col=c(rgb(0.1,0.1,0.7,0.5),
rgb(0.8,0.1,0.3,0.6)),
      xlab="Diabetic", ylab="No. of Pregnancy", main="Relationship between Pregnancy and
Diabeties")
boxplot(Diabetes$Age~Diabetes$Outcome, col=c(rgb(0.1,0.1,0.7,0.5), rgb(0.8,0.1,0.3,0.6)),
      xlab="Diabetic", ylab="Age", main="Relationship between Age and Diabetes")

names(Diabetes)
pairs(Diabetes[-9])
correlations <- cor(Diabetes[,1:8])
corrplot(correlations, method="circle")

X1 <- Diabetes$Pregnancies
X2 <- Diabetes$Glucose
X3 <- Diabetes$BloodPressure
X4 <- Diabetes$SkinThickness
X5 <- Diabetes$Insulin
X6 <- Diabetes$BMI
X7 <- Diabetes$DiabetesPedigreeFunction
X8 <- Diabetes$Age
Y<- Diabetes$Outcome
#Multiple Linear Regression Model

mlm1 <- lm(Outcome~.,data=Diabetes)
summary(mlm1)
```

```
resid <- residuals(mlm1)

plot(X1, residuals(mlm1), main="Relationship between
    Pregnancies and residuals",
    xlab="Pregnancies", ylab="Residuals")
plot(X2, residuals(mlm1), main="Relationship between
    Glucose and residuals",
    xlab="Glucose", ylab="Residuals")
plot(X3, residuals(mlm1), main="Relationship between
    BP and residuals",
    xlab="BloodPressure", ylab="Residuals")
plot(X4, residuals(mlm1), main="Relationship between
    SkinThickness and residuals",
    xlab="SkinThickness", ylab="Residuals")
plot(X5, residuals(mlm1), main="Relationship between
    Insulin and residuals",
    xlab="Insulin", ylab="Residuals")
plot(X6, residuals(mlm1), main="Relationship between
    BMI and residuals",
    xlab="BMI", ylab="Residuals")
plot(X7, residuals(mlm1), main="Relationship between
    DiabetesPedigreeFunction and residuals",
    xlab="DiabetesPedigreeFunction", ylab="Residuals")
plot(X8, residuals(mlm1), main="Relationship between
    Age and residuals",
    xlab="Age", ylab="Residuals")

mlm1 <- glm(Outcome~., data=Diabetes)
summary(mlm1)
fnorm <- fitdist(resid,"norm")
summary(fnorm) #reject null hypothesis means regression is linear
plot(fnorm)

#multiple logistic regression
mlog.diabetes <- glm(Outcome~., data=Diabetes, family=binomial)
summary(mlog.diabetes)

exp(coef(mlog.diabetes))
exp(cbind(OR=coef(mlog.diabetes),confint(mlog.diabetes)))

pairs(Diabetes, col=Diabetes$Outcome)
vif(mlog.diabetes)

#bestglm - Subset Selection based on AIC, BIC, CV

bmodelAIC <- bestglm(Diabetes,IC="AIC", family = binomial)
summary(bmodelAIC$BestModel)
```

```
bmodelBIC <- bestglm(Diabetes,IC="BIC",family = binomial)
summary(bmodelBIC$BestModel)

RNGkind(sample.kind = "Rounding")
set.seed(100)

bmodelCV <- bestglm(Diabetes,IC="CV",CVArgs = list(Method="HTF",K=10,REP=1),family =
binomial)
summary(bmodelCV$BestModel)

bmodelAIC$Subsets
bmodelBIC$Subsets

plot1 = plot(x=c(0:8),bmodelAIC$Subsets$AIC, xlab = "Number of Variables",ylab = "AIC",
        main = "Best Selection with AIC", type = "b")
plot2 = plot(x=c(0:8),bmodelBIC$Subsets$BIC,xlab = "Number of Variables",ylab = "BIC",
        main = "Best Selection with BIC", type = "b")

#lasso prediction

RNGkind(sample.kind = "Rounding")
set.seed(100)
grid <- 10^seq(10,-2,length = 100)

train <- sample(1:nrow(Diabetes),nrow(Diabetes)/2)
test <- (-train)
Diabetes.train <- Diabetes[train,]
Diabetes.test <- Diabetes[test,]

X <- model.matrix(Diabetes$Outcome~.,Diabetes)[,-1]
Y.test <- Y[test]
Y.train <- Y[train]

lasso.mod <- glmnet(X[train,], Y[train], alpha = 1, lambda = grid, family = "binomial")
plot(lasso.mod)

RNGkind(sample.kind = "Rounding")
set.seed(100)

cv.out <- cv.glmnet(X[train,] , Y[train], alpha = 1 ,lambda = grid, family = "binomial")
cv.out
plot(cv.glmnet(X[train,] , Y[train], alpha = 1,family="binomial"))

bestlam <- cv.out$lambda.min
bestlam
```

```
#test
bmodellasso <- glmnet(X,Y, alpha = 1, family = "binomial",lambda = bestlam)

X.test <- model.matrix(Outcome ~., Diabetes.test)[,-1]

probabilities <- bmodellasso%>% predict(newx = X.test)
predicted.classes <- ifelse(probabilities > 0.25, 1, 0)

table(Diabetes.test$Outcome,predicted.classes)
mean(predicted.classes==Diabetes.test$Outcome)

lasso.coef <- predict(bmodellasso, type = "coefficients", s = bestlam)
lasso.coef
lasso.coef[lasso.coef!=0]

prob.lasso <- NULL
for (i in 1:nrow(Diabetes)){
  w <- exp(-8.798973375 + 0.102940318 * X1[i] + 0.032295005 * X2[i] + 0.003706411 * X4[i]
+ 0.068979711* X6[i] + 1.063010860 * X7[i] + 0.021402651 * X8[i])
  prob.lasso[i]<- w/(1+w)
}


#accuracy

prob1 <- predict(bmodelAIC$BestModel, type="response")
pred1 <- rep("No Diabetes", 532)
pred1[prob1 > 0.25] <- "Diabetes"
table1 <- table(pred1, Diabetes$Outcome)
table1
accuracy1 <- (table1[1,2]+table1[2,1])/length(Diabetes$Outcome)
accuracy1

prob2 <- predict(bmodelBIC$BestModel, type="response")
pred2 <- rep("No Diabetes", 532)
pred2[prob2 > 0.25] <- "Diabetes"
table2 <- table(pred2, Diabetes$Outcome)
table2
accuracy2 <- (table2[1,2]+table2[2,1])/length(Diabetes$Outcome)
accuracy2

prob3 <- predict(bmodelCV$BestModel, type="response")
pred3 <- rep("No Diabetes", 532)
pred3[prob3 > 0.25] <- "Diabetes"
table3 <- table(pred3, Diabetes$Outcome)
table3
accuracy3 <- (table3[1,2]+table3[2,1])/length(Diabetes$Outcome)
```

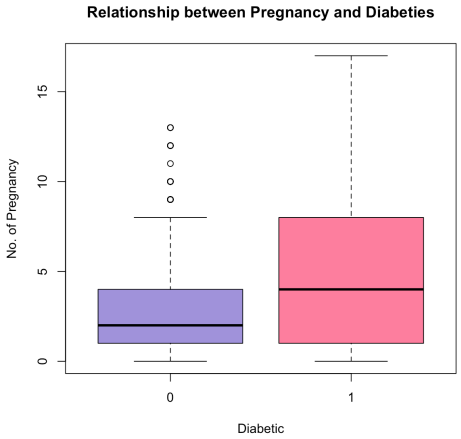
```
accuracy3

pred4 <- rep("No Diabetes", 532)
pred4[prob.lasso > 0.25] <- "Diabetes"
table4 <- table(pred4, Diabetes$Outcome)
table4
accuracy4 <- (table4[1,2]+table4[2,1])/length(Diabetes$Outcome)
accuracy4

accuracylist <- c(accuracy1, accuracy2, accuracy3, accuracy4)
accuracylist

bmodelAIC$BestModel
bmodelBIC$BestModel
bmodelCV$BestModel



#ROC
roc(Diabetes$Outcome, bmodelAIC$BestModel$fitted.values, plot = TRUE, percent = TRUE,
xlab = "False positive rate", ylab = "False negative rate", legacy.axes = TRUE, col = "Red",
print.auc = TRUE)
plot.roc(Diabetes$Outcome, bmodelBIC$BestModel$fitted.values, add = TRUE , col =
"Blue",print.auc = TRUE, print.auc.y = 40, percent = TRUE)
plot.roc(Diabetes$Outcome, bmodelCV$BestModel$fitted.values, add = TRUE , col =
"Brown",print.auc = TRUE, print.auc.y = 30, percent = TRUE)
plot.roc(Diabetes$Outcome, prob.lasso, add = TRUE , col = "Green",print.auc = TRUE,
print.auc.y = 20, percent = TRUE)
```

7.3 Outputs

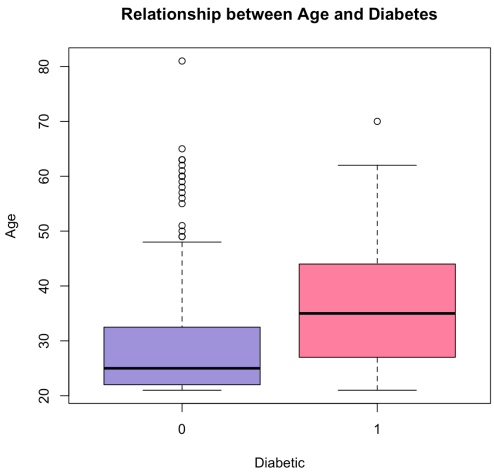
```
> library(dplyr)
> library(fitdistrplus)
> library(ggplot2)
> library(pROC)
> library(car)
> library(PresenceAbsence)
> library(leaps)
> library(glmnet)
> library(bestglm)
> library(corrplot)
> library(dismo)

> Original_Diabetes <- read.csv("~/Desktop/Original_Diabetes.csv")
> summary(Original_Diabetes)

 Pregnancies      Glucose      BloodPressure    SkinThickness      Insulin          BMI
 Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00   Min.   :  0.0   Min.   : 0.00
 1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00   1st Qu.:  0.0   1st Qu.:27.30
 Median : 3.000   Median :117.0   Median : 72.00   Median :23.00   Median : 30.5   Median
:32.00
 Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.54   Mean   : 79.8   Mean   :31.99
 3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:32.00   3rd Qu.:127.2   3rd
Qu.:36.60
 Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00   Max.   :846.0   Max.   :67.10
 DiabetesPedigreeFunction      Age          Outcome
 Min.   :0.0780           Min.   :21.00   Min.   :0.000
 1st Qu.:0.2437           1st Qu.:24.00   1st Qu.:0.000
 Median :0.3725           Median :29.00   Median :0.000
 Mean   :0.4719           Mean   :33.24   Mean   :0.349
 3rd Qu.:0.6262           3rd Qu.:41.00   3rd Qu.:1.000
 Max.   :2.4200           Max.   :81.00   Max.   :1.000
>

> Diabetes <- Original_Diabetes %>%
+   filter(SkinThickness!=0 & BloodPressure != 0 & BMI!= 0 & Glucose!=0)

> summary(Diabetes)
 Pregnancies      Glucose      BloodPressure    SkinThickness      Insulin          BMI
 Min.   : 0.000   Min.   : 56.00   Min.   : 24.00   Min.   : 7.00   Min.   :  0.0   Min.   :18.20
 1st Qu.: 1.000   1st Qu.: 98.75   1st Qu.: 64.00   1st Qu.:22.00   1st Qu.:  0.0   1st Qu.:27.88
 Median : 2.000   Median :115.00   Median : 72.00   Median :29.00   Median : 91.5   Median
:32.80
 Mean   : 3.517   Mean   :121.03   Mean   : 71.51   Mean   :29.18   Mean   :115.0   Mean
:32.89
 3rd Qu.: 5.000   3rd Qu.:141.25   3rd Qu.: 80.00   3rd Qu.:36.00   3rd Qu.:165.2   3rd
Qu.:36.90
```

```
Max.   :17.000   Max.   :199.00   Max.   :110.00   Max.   :99.00   Max.   :846.0   Max.   :67.10
DiabetesPedigreeFunction      Age             Outcome
Min.   :0.0850          Min.   :21.00   Min.   :0.0000
1st Qu.:0.2587          1st Qu.:23.00   1st Qu.:0.0000
Median :0.4160          Median :28.00   Median :0.0000
Mean   :0.5030          Mean   :31.61   Mean   :0.3327
3rd Qu.:0.6585          3rd Qu.:38.00   3rd Qu.:1.0000
Max.   :2.4200          Max.   :81.00   Max.   : 1.0000
```

```
> boxplot(Diabetes$Pregnancies~Diabetes$Outcome, col=c(rgb(0.1,0.1,0.7,0.5),
rgb(0.8,0.1,0.3,0.6)),
+       xlab="Diabetic", ylab="No. of Pregnancy", main="Relationship between Pregnancy
and Diabeties")
```

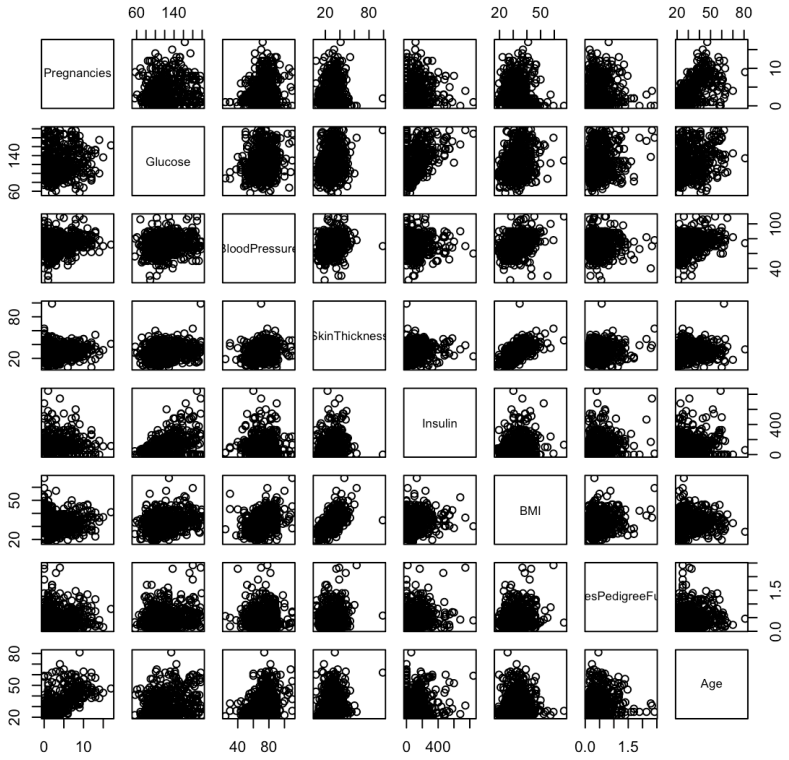



```
> boxplot(Diabetes$Age~Diabetes$Outcome, col=c(rgb(0.1,0.1,0.7,0.5), rgb(0.8,0.1,0.3,0.6)),
+       xlab="Diabetic", ylab="Age", main="Relationship between Age and Diabetes")
```

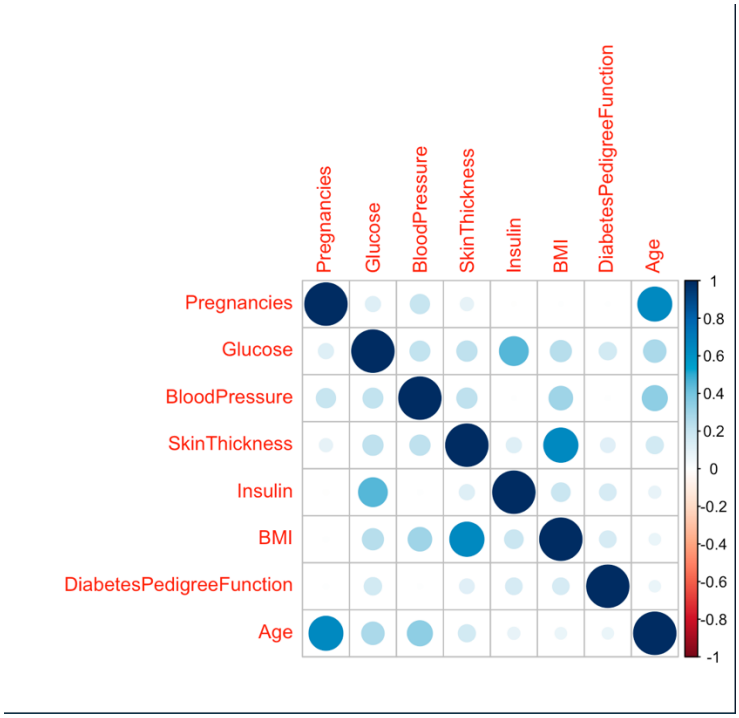



```
> names(Diabetes)
[1] "Pregnancies"          "Glucose"              "BloodPressure"
[4] "SkinThickness"        "Insulin"              "BMI"
```

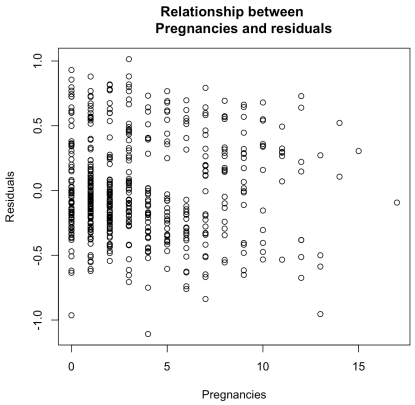
[7] "DiabetesPedigreeFunction" "Age" "Outcome"
> pairs(Diabetes[-9])



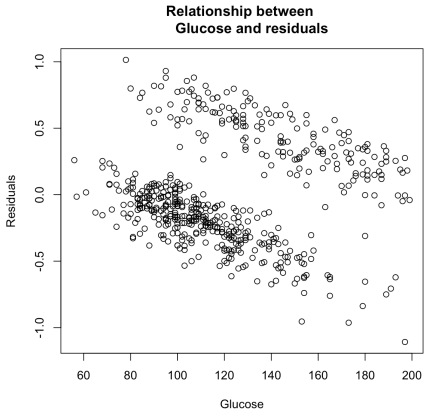
> correlations <- cor(Diabetes[,1:8])
> corrplot(correlations, method="circle")



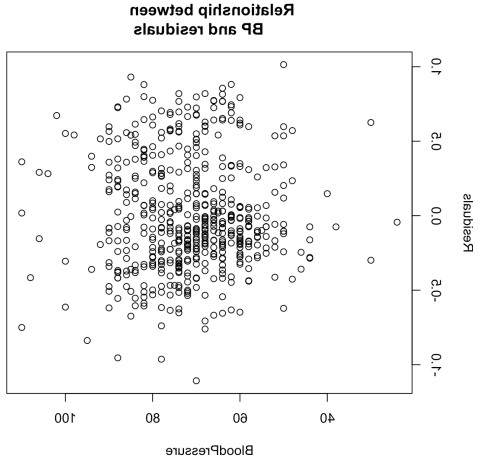
```
> X1 <- Diabetes$Pregnancies
> X2 <- Diabetes$Glucose
> X3 <- Diabetes$BloodPressure
> X4 <- Diabetes$SkinThickness
> X5 <- Diabetes$Insulin
> X6 <- Diabetes$BMI
> X7 <- Diabetes$DiabetesPedigreeFunction
> X8 <- Diabetes$Age
> Y<- Diabetes$Outcome

> mlm1 <- lm(Outcome~.,data=Diabetes)
> summary(mlm1)

Call:
lm(formula = Outcome ~ ., data = Diabetes)

Residuals:
    Min      1Q  Median      3Q     Max
-1.1081 -0.2604 -0.0748  0.2660  1.0141

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -1.0519076  0.1197594  -8.784  < 2e-16 ***
Pregnancies              0.0198196  0.0065775   3.013 0.002710 **
Glucose                  0.0063119  0.0006463   9.766  < 2e-16 ***
BloodPressure           -0.0010991  0.0015310  -0.718 0.473158
SkinThickness            0.0005243  0.0021035   0.249 0.803272
Insulin                 -0.0001592  0.0001550  -1.027 0.305103
BMI                      0.0122568  0.0033400   3.670 0.000268 ***
DiabetesPedigreeFunction 0.1837957  0.0495993   3.706 0.000233 ***
Age                      0.0043336  0.0021673   1.999 0.046072 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3834 on 523 degrees of freedom
Multiple R-squared:  0.3493,  Adjusted R-squared:  0.3393
F-statistic: 35.09 on 8 and 523 DF,  p-value: < 2.2e-16

> resid <- residuals(mlm1)
>

> plot(X1, residuals(mlm1), main="Relationship between
+     Pregnancies and residuals",
+     xlab="Pregnancies", ylab="Residuals")
```



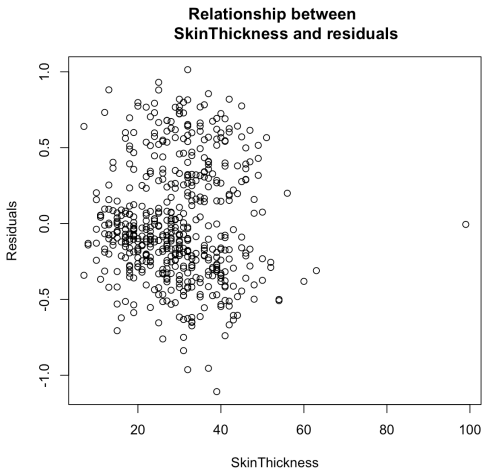

```
> plot(X2, residuals(mlm1), main="Relationship between
+      Glucose and residuals",
+      xlab="Glucose", ylab="Residuals")
```



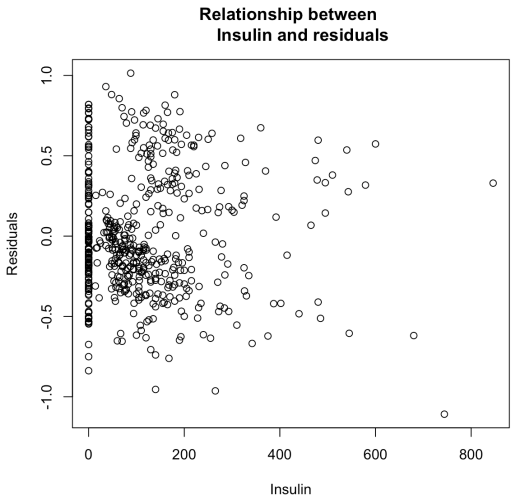

```
plot(X3, residuals(mlm1), main="Relationship between
+      BP and residuals",
+      xlab="BloodPressure", ylab="Residuals")
```



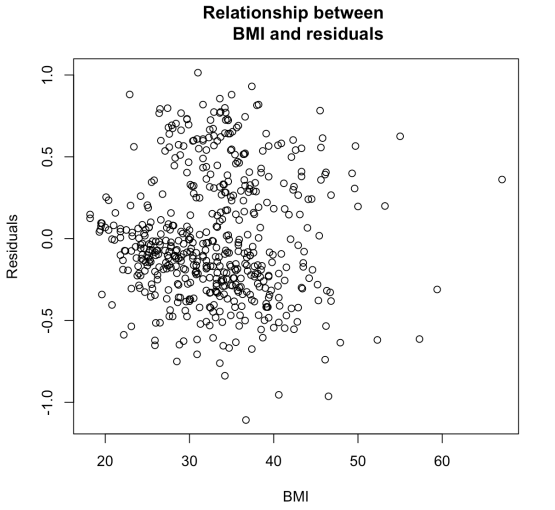
```
> plot(X4, residuals(mlm1), main="Relationship between
+      SkinThickness and residuals",
+      xlab="SkinThickness", ylab="Residuals")
```



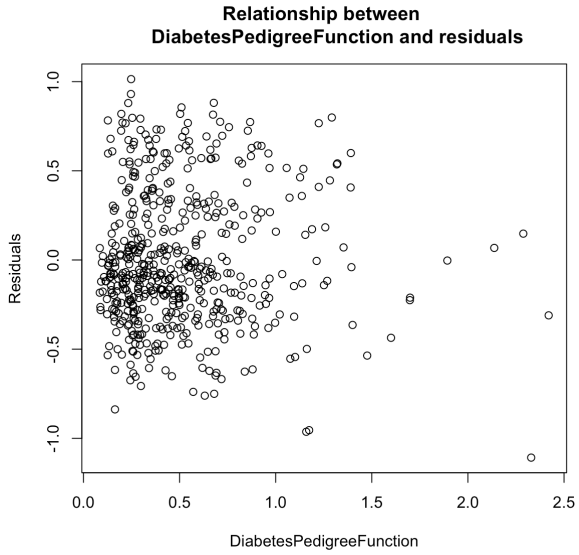

```
> plot(X5, residuals(mlm1), main="Relationship between
+     Insulin and residuals",
+     xlab="Insulin", ylab="Residuals")
```



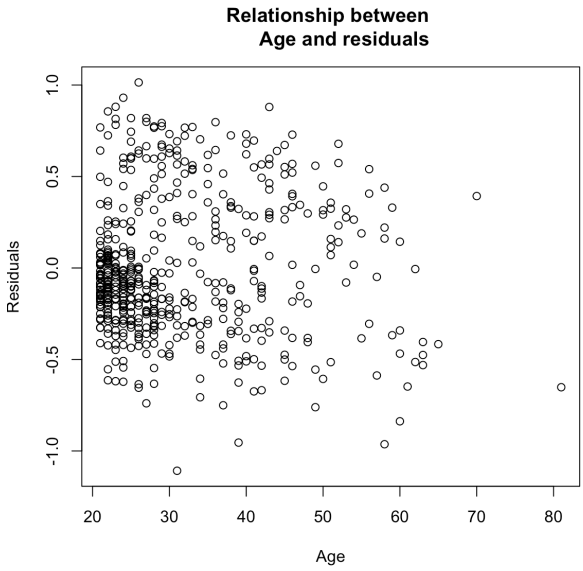

```
> plot(X6, residuals(mlm1), main="Relationship between
+     BMI and residuals",
+     xlab="BMI", ylab="Residuals")
```




```
> plot(X7, residuals(mlm1), main="Relationship between
+      DiabetesPedigreeFunction and residuals",
+      xlab="DiabetesPedigreeFunction", ylab="Residuals")
```

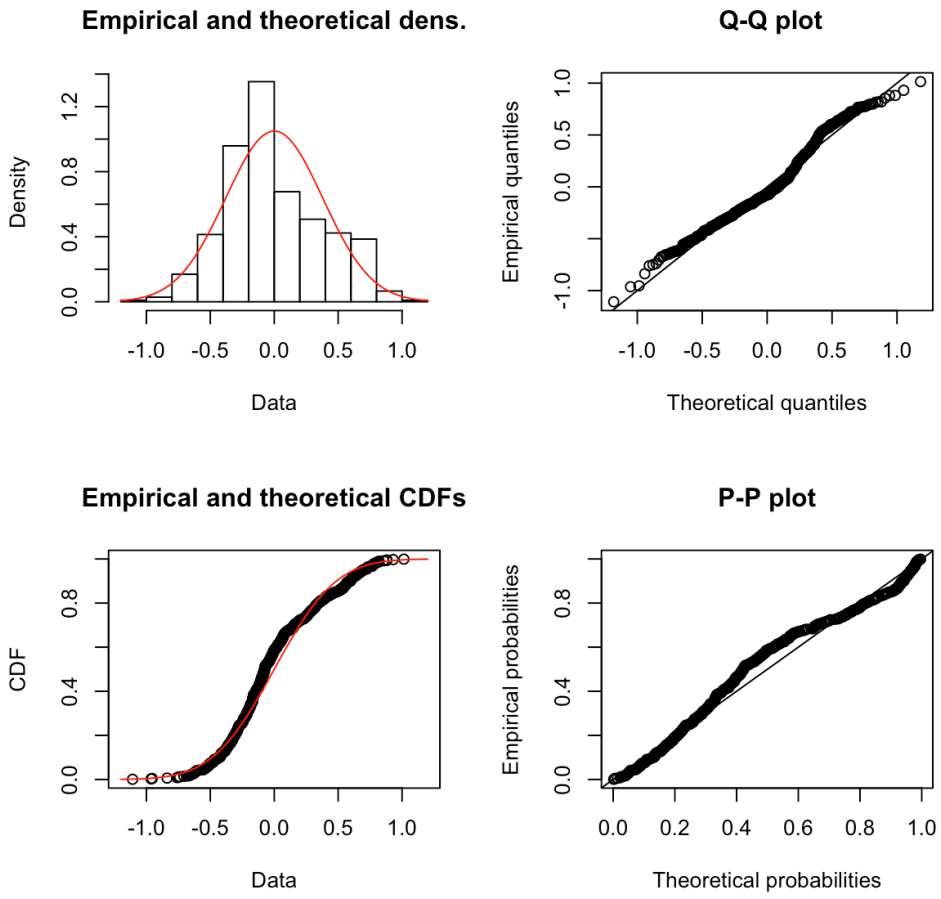



```
> plot(X8, residuals(mlm1), main="Relationship between
+      Age and residuals",
+      xlab="Age", ylab="Residuals")
```



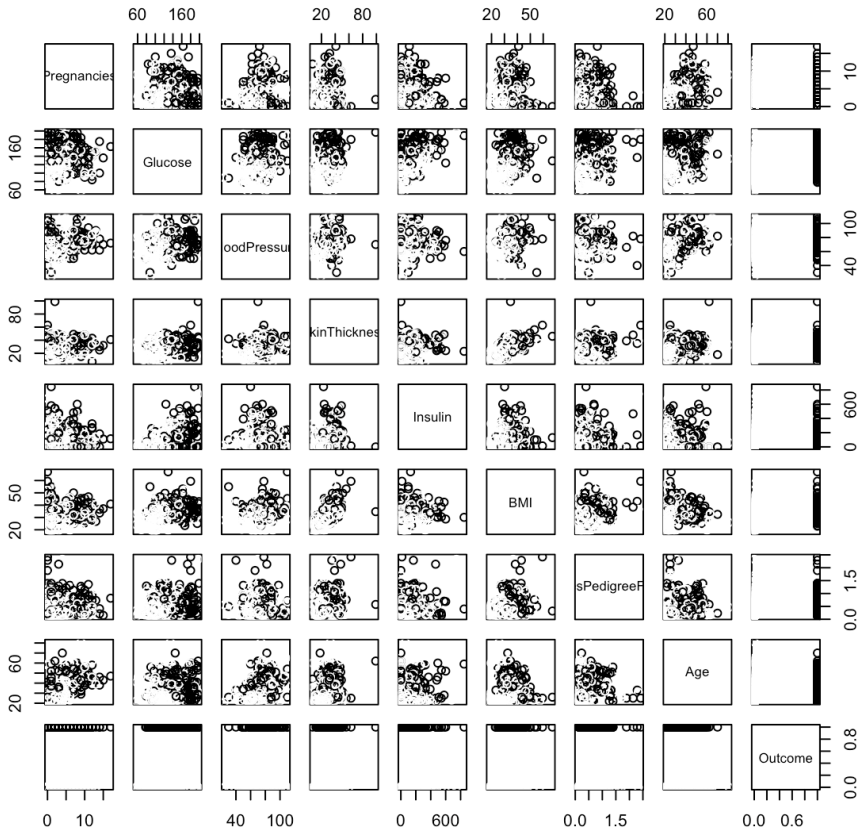

```
> mlm1 <- glm(Outcome~., data=Diabetes)
> summary(mlm1)

Call:
glm(formula = Outcome ~ ., data = Diabetes)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1081  -0.2604  -0.0748   0.2660   1.0141

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -1.0519076  0.1197594  -8.784  < 2e-16 ***
Pregnancies              0.0198196  0.0065775   3.013 0.002710 **
Glucose                  0.0063119  0.0006463   9.766  < 2e-16 ***
BloodPressure           -0.0010991  0.0015310  -0.718 0.473158
SkinThickness            0.0005243  0.0021035   0.249 0.803272
Insulin                 -0.0001592  0.0001550  -1.027 0.305103
BMI                      0.0122568  0.0033400   3.670 0.000268 ***
DiabetesPedigreeFunction 0.1837957  0.0495993   3.706 0.000233 ***
Age                      0.0043336  0.0021673   1.999 0.046072 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.1469581)

    Null deviance: 118.111  on 531  degrees of freedom
Residual deviance:  76.859  on 523  degrees of freedom
AIC: 500.51

Number of Fisher Scoring iterations: 2
```

```
> fnorm <- fitdist(resid,"norm")
> summary(fnorm) #reject null hypothesis means regression is linear
Fitting of the distribution ' norm ' by maximum likelihood
Parameters :
       estimate Std. Error
mean 1.150100e-17 0.01647919
sd   3.800946e-01 0.01165219
Loglikelihood:  -240.2531   AIC:  484.5061   BIC:  493.0594
Correlation matrix:
     mean sd
mean    1  0
sd      0  1

> plot(fnorm)
>
```




```
> mlog.diabetes <- glm(Outcome~., data=Diabetes, family=binomial)
> summary(mlog.diabetes)

Call:
glm(formula = Outcome ~ ., family = binomial, data = Diabetes)
```

```
Deviance Residuals:
   Min      1Q  Median      3Q     Max
-2.8627  -0.6639  -0.3672   0.6347   2.4942

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -9.677562   1.005400  -9.626  < 2e-16 ***
Pregnancies              0.121235   0.043926   2.760 0.005780 **
Glucose                  0.037439   0.004765   7.857 3.92e-15 ***
BloodPressure           -0.009316   0.010446  -0.892 0.372494
SkinThickness            0.006341   0.014853   0.427 0.669426
Insulin                 -0.001053   0.001007  -1.046 0.295651
BMI                      0.085992   0.023661   3.634 0.000279 ***
DiabetesPedigreeFunction 1.335764   0.365771   3.652 0.000260 ***
Age                      0.026430   0.013962   1.893 0.058371 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 676.79  on 531  degrees of freedom
Residual deviance: 465.23  on 523  degrees of freedom
AIC: 483.23

Number of Fisher Scoring iterations: 5

>
> exp(coef(mlog.diabetes))
            (Intercept)             Pregnancies                 Glucose           BloodPressure
           6.267413e-05            1.128890e+00            1.038148e+00            9.907271e-01
          SkinThickness                 Insulin                     BMI DiabetesPedigreeFunction
           1.006361e+00            9.989476e-01            1.089798e+00            3.802899e+00
                    Age
           1.026782e+00
> exp(cbind(OR=coef(mlog.diabetes),confint(mlog.diabetes)))
                             OR       2.5 %       97.5 %
(Intercept)            6.267413e-05 8.025160e-06 0.0004168335
Pregnancies            1.128890e+00 1.036704e+00 1.2320360221
Glucose                1.038148e+00 1.028816e+00 1.0482603112
BloodPressure          9.907271e-01 9.706059e-01 1.0113139150
SkinThickness          1.006361e+00 9.777702e-01 1.0363314278
Insulin                9.989476e-01 9.969664e-01 1.0009272582
BMI                    1.089798e+00 1.041124e+00 1.1426497905
DiabetesPedigreeFunction 3.802899e+00 1.875747e+00 7.8783186343
Age                    1.026782e+00 9.992068e-01 1.0556961454
>pairs(Diabetes, col=Diabetes$Outcome)
```



vif(mlog.diabetes)

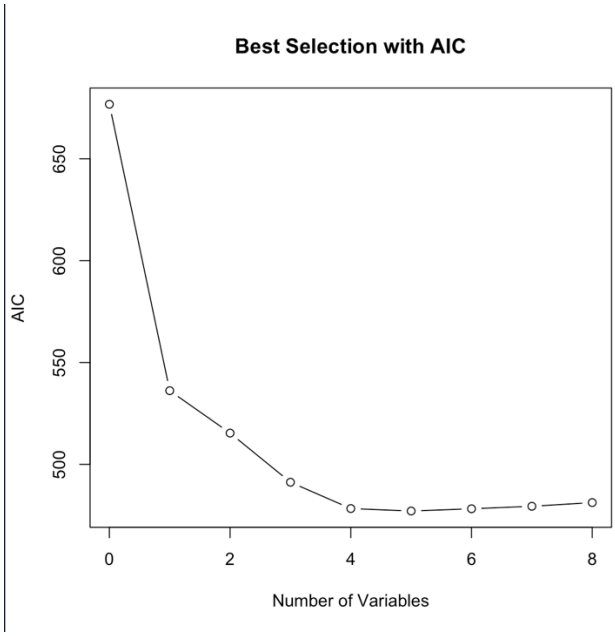
```
> vif(mlog.diabetes)
       Pregnancies            Glucose      BloodPressure      SkinThickness
          1.687511           1.314813           1.261718           1.604702
           Insulin                BMI DiabetesPedigreeFunction                Age
          1.294280           1.771704           1.020968           1.806439
```

```
> bmodelAIC <- bestglm(Diabetes,IC="AIC", family = binomial)
Morgan-Tatar search since family is non-gaussian.
> summary(bmodelAIC$BestModel)

Call:
glm(formula = y ~ ., family = family, data = Xi, weights = weights)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-2.9894  -0.6530  -0.3700   0.6442   2.5417

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)        -9.879992   0.918614 -10.755  < 2e-16 ***
Pregnancies         0.123887   0.043514   2.847 0.004412 **
```

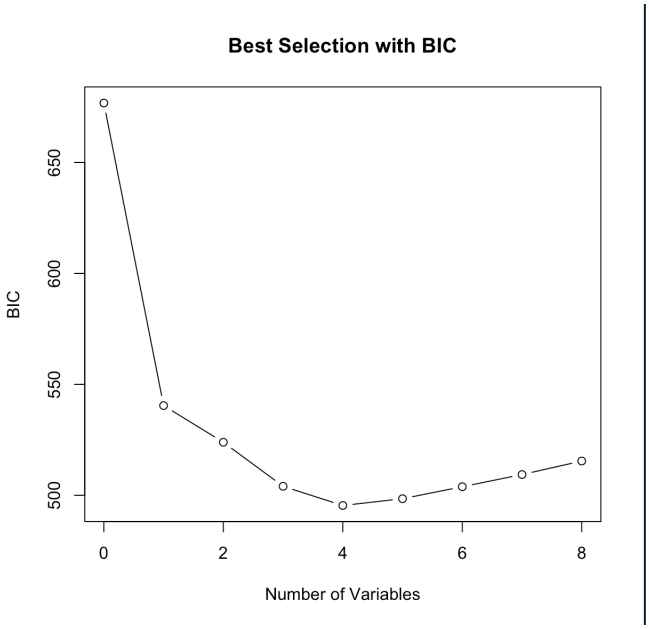
```
Glucose                  0.035026   0.004201   8.338  < 2e-16 ***
BMI                      0.085123   0.018110   4.700  2.6e-06 ***
DiabetesPedigreeFunction  1.321554   0.362538   3.645 0.000267 ***
Age                      0.023844   0.013311   1.791 0.073244 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 676.79  on 531  degrees of freedom
Residual deviance: 467.08  on 526  degrees of freedom
AIC: 479.08

Number of Fisher Scoring iterations: 5

>

> bmodelBIC <- bestglm(Diabetes,IC="BIC",family = binomial)
Morgan-Tatar search since family is non-gaussian.
> summary(bmodelBIC$BestModel)

Call:
glm(formula = y ~ ., family = family, data = Xi, weights = weights)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0663  -0.6606  -0.3794   0.6282   2.5214

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)             -9.449410   0.874874 -10.801  < 2e-16 ***
Pregnancies              0.172876   0.034474   5.015 5.31e-07 ***
Glucose                  0.036482   0.004136   8.822  < 2e-16 ***
BMI                      0.084232   0.018089   4.657 3.21e-06 ***
DiabetesPedigreeFunction  1.361746   0.361589   3.766 0.000166 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 676.79  on 531  degrees of freedom
Residual deviance: 470.30  on 527  degrees of freedom
AIC: 480.3

Number of Fisher Scoring iterations: 5

>
```

```
> RNGkind(sample.kind = "Rounding")
Warning message:
In RNGkind(sample.kind = "Rounding") : non-uniform 'Rounding' sampler used
> set.seed(100)
>
> bmodelCV <- bestglm(Diabetes,IC="CV",CVArgs = list(Method="HTF",K=10,REP=1),family =
binomial)
Morgan-Tatar search since family is non-gaussian.
> summary(bmodelCV$BestModel)


Call:
glm(formula = y ~ ., family = family, data = data.frame(Xy[,
    c(bestset[-1], FALSE), drop = FALSE], y = y))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2567  -0.6637  -0.3976   0.6680   2.3874

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.831722   0.819250 -10.780  < 2e-16 ***
Pregnancies  0.167525   0.033772   4.960 7.03e-07 ***
Glucose      0.036666   0.004039   9.079  < 2e-16 ***
BMI          0.086421   0.017780   4.860 1.17e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 676.79  on 531  degrees of freedom
Residual deviance: 485.22  on 528  degrees of freedom
AIC: 493.22

Number of Fisher Scoring iterations: 5

>

> bmodelAIC$Subsets
   Intercept Pregnancies Glucose BloodPressure SkinThickness Insulin   BMI
DiabetesPedigreeFunction
0       TRUE      FALSE   FALSE       FALSE       FALSE   FALSE FALSE
FALSE
1       TRUE      FALSE    TRUE       FALSE       FALSE   FALSE FALSE
FALSE
2       TRUE       TRUE    TRUE       FALSE       FALSE   FALSE FALSE
FALSE
3       TRUE       TRUE    TRUE       FALSE       FALSE   FALSE  TRUE
FALSE
4       TRUE       TRUE    TRUE       FALSE       FALSE   FALSE  TRUE
 TRUE
```

```
5*     TRUE      TRUE    TRUE       FALSE       FALSE   FALSE  TRUE              TRUE
6      TRUE      TRUE    TRUE       FALSE       FALSE    TRUE  TRUE              TRUE
7      TRUE      TRUE    TRUE        TRUE       FALSE    TRUE  TRUE              TRUE
8      TRUE      TRUE    TRUE        TRUE        TRUE    TRUE  TRUE              TRUE
    Age logLikelihood     AIC
0  FALSE    -338.3940 676.7880
1  FALSE    -267.0794 536.1587
2  FALSE    -255.6742 515.3485
3  FALSE    -242.6113 491.2226
4  FALSE    -235.1481 478.2963
5*  TRUE    -233.5392 477.0785
6   TRUE    -233.0979 478.1959
7   TRUE    -232.7065 479.4130
8   TRUE    -232.6150 481.2300
> bmodelBIC$Subsets
  Intercept Pregnancies Glucose BloodPressure SkinThickness Insulin   BMI
DiabetesPedigreeFunction
0      TRUE      FALSE   FALSE       FALSE       FALSE   FALSE FALSE             FALSE
1      TRUE      FALSE    TRUE       FALSE       FALSE   FALSE FALSE             FALSE
2      TRUE       TRUE    TRUE       FALSE       FALSE   FALSE FALSE             FALSE
3      TRUE       TRUE    TRUE       FALSE       FALSE   FALSE  TRUE             FALSE
4*     TRUE       TRUE    TRUE       FALSE       FALSE   FALSE  TRUE              TRUE
5      TRUE       TRUE    TRUE       FALSE       FALSE   FALSE  TRUE              TRUE
6      TRUE       TRUE    TRUE       FALSE       FALSE    TRUE  TRUE              TRUE
7      TRUE       TRUE    TRUE        TRUE       FALSE    TRUE  TRUE              TRUE
8      TRUE       TRUE    TRUE        TRUE        TRUE    TRUE  TRUE              TRUE
    Age logLikelihood     BIC
0  FALSE    -338.3940 676.7880
1  FALSE    -267.0794 540.4354
2  FALSE    -255.6742 523.9018
3  FALSE    -242.6113 504.0525
4* FALSE    -235.1481 495.4028
5   TRUE    -233.5392 498.4617
6   TRUE    -233.0979 503.8558
7   TRUE    -232.7065 509.3495
8   TRUE    -232.6150 515.4431

> plot1 = plot(x=c(0:8),bmodelAIC$Subsets$AIC, xlab = "Number of Variables",ylab = "AIC",
+           main = "Best Selection with AIC", type = "b")
>
```

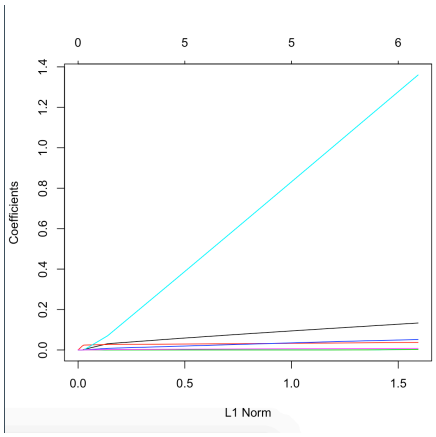



```
> plot2 = plot(x=c(0:8),bmodelBIC$Subsets$BIC,xlab = "Number of Variables",ylab = "BIC",
+           main = "Best Selection with BIC", type = "b")
```

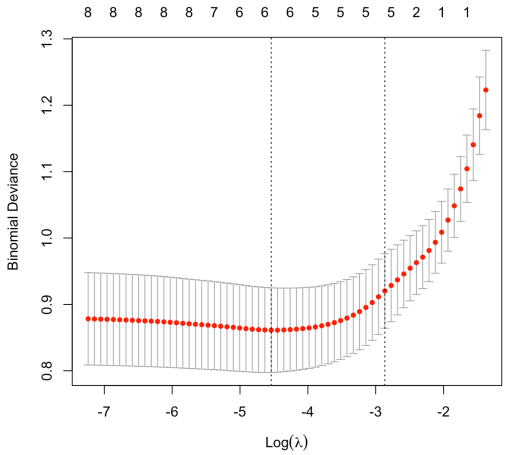



```
> RNGkind(sample.kind = "Rounding")
Warning message:
In RNGkind(sample.kind = "Rounding") : non-uniform 'Rounding' sampler used
> set.seed(100)
> grid <- 10^seq(10,-2,length = 100)
> train <- sample(1:nrow(Diabetes),nrow(Diabetes)/2)
> test <- (-train)
> Diabetes.train <- Diabetes[train,]
```

```
> Diabetes.test <- Diabetes[test,]
>
> X <- model.matrix(Diabetes$Outcome~.,Diabetes)[,-1]
> Y.test <- Y[test]
> Y.train <- Y[train]
> X <- model.matrix(Diabetes$Outcome~.,Diabetes)[,-1]
> Y.test <- Y[test]
> Y.train <- Y[train]
>
> lasso.mod <- glmnet(X[train,], Y[train], alpha = 1, lambda = grid, family = "binomial")
> plot(lasso.mod)
Warning message:
In regularize.values(x, y, ties, missing(ties)) :
  collapsing to unique 'x' values
```




```
> RNGkind(sample.kind = "Rounding")
Warning message:
In RNGkind(sample.kind = "Rounding") : non-uniform 'Rounding' sampler used
> set.seed(100)
>
> cv.out <- cv.glmnet(X[train,] , Y[train], alpha = 1 ,lambda = grid, family = "binomial")
> cv.out

Call:  cv.glmnet(x = X[train, ], y = Y[train], lambda = grid, alpha = 1,      family = "binomial")

Measure: Binomial Deviance

    Lambda Measure      SE Nonzero
min 0.01000  0.8450 0.08224       6
1se 0.05337  0.9025 0.06089       5
> plot(cv.glmnet(X[train,] , Y[train], alpha = 1,family="binomial"))
```

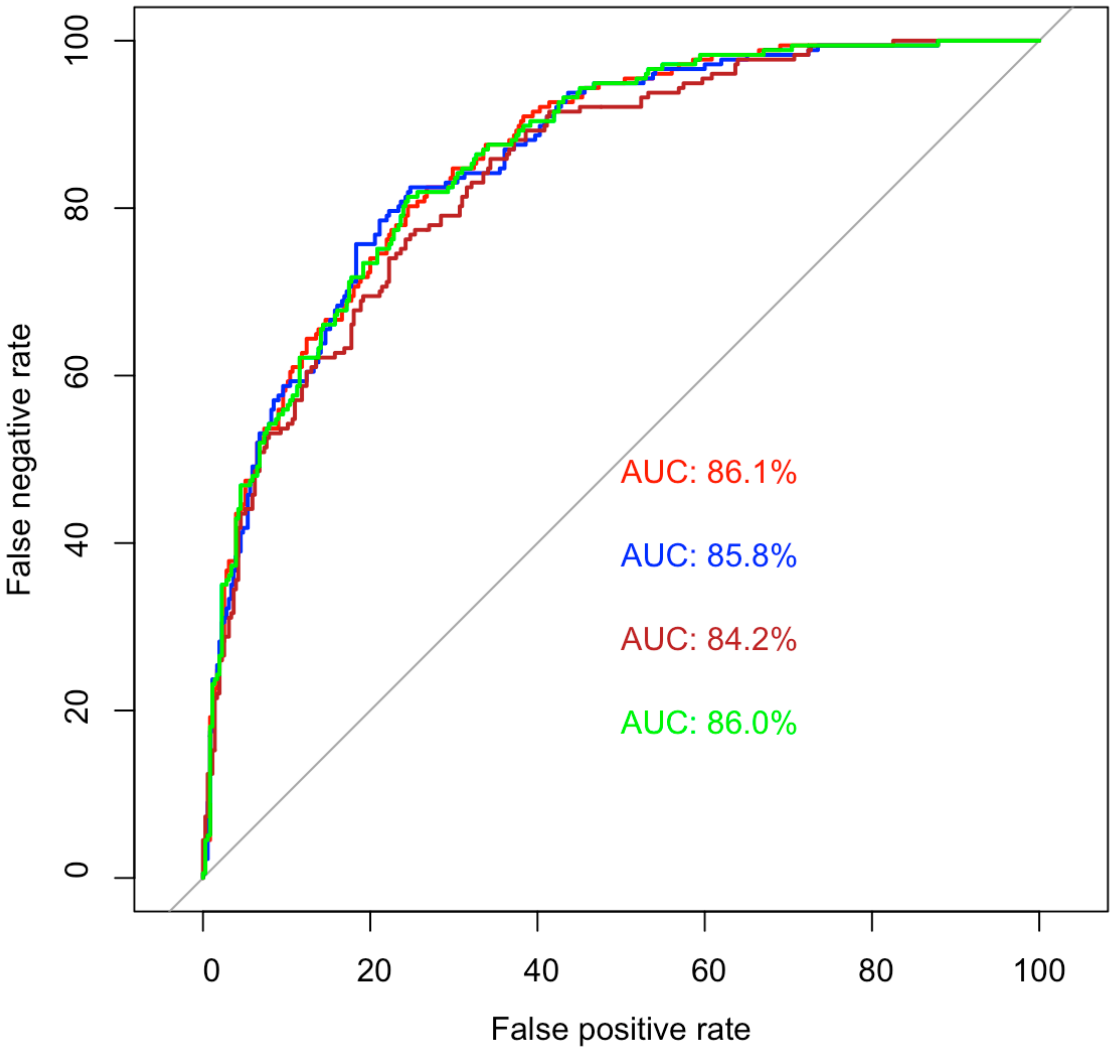



```
> bestlam <- cv.out$lambda.min
> bestlam
[1] 0.01

> probabilities <- bmodellasso%>% predict(newx = X.test)
> bmodellasso <- glmnet(X,Y, alpha = 1, family = "binomial",lambda = bestlam)
>
> X.test <- model.matrix(Outcome ~., Diabetes.test)[,-1]
>
> probabilities <- bmodellasso%>% predict(newx = X.test)
> predicted.classes <- ifelse(probabilities > 0.25, 1, 0)
>
> table(Diabetes.test$Outcome,predicted.classes)
  predicted.classes
    0   1
 0 161   7
 1  53  45
> mean(predicted.classes==Diabetes.test$Outcome)
[1] 0.7744361
>

> lasso.coef <- predict(bmodellasso, type = "coefficients", s = bestlam)
> lasso.coef
9 x 1 sparse Matrix of class "dgCMatrix"
                           1
(Intercept)          -8.798973375
Pregnancies           0.102940318
Glucose              0.032295005
BloodPressure         .
SkinThickness         0.003706411
Insulin              .
BMI                  0.068979711
DiabetesPedigreeFunction  1.063010860
```

```
Age                  0.021402651
> lasso.coef[lasso.coef!=0]
<sparse>[ <logic> ] : .M.sub.i.logical() maybe inefficient
[1] -8.798973375  0.102940318  0.032295005  0.003706411  0.068979711  1.063010860
0.021402651
>
> prob.lasso <- NULL
> for (i in 1:nrow(Diabetes)){
+   w <- exp(-8.798973375 + 0.102940318 * X1[i] + 0.032295005 * X2[i] + 0.003706411 *
X4[i] + 0.068979711* X6[i] + 1.063010860 * X7[i] + 0.021402651 * X8[i])
+   prob.lasso[i]<- w/(1+w)

> prob1 <- predict(bmodelAIC$BestModel, type="response")
> pred1 <- rep("No Diabetes", 532)
> pred1[prob1 > 0.25] <- "Diabetes"
> table1 <- table(pred1, Diabetes$Outcome)
> table1

pred1         0   1
 Diabetes    105 148
 No Diabetes 250  29
> accuracy1 <- (table1[1,2]+table1[2,1])/length(Diabetes$Outcome)
> accuracy1
[1] 0.7481203
>
> prob2 <- predict(bmodelBIC$BestModel, type="response")
> pred2 <- rep("No Diabetes", 532)
> pred2[prob2 > 0.25] <- "Diabetes"
> table2 <- table(pred2, Diabetes$Outcome)
> table2

pred2         0   1
 Diabetes    106 147
 No Diabetes 249  30
> accuracy2 <- (table2[1,2]+table2[2,1])/length(Diabetes$Outcome)
> accuracy2
[1] 0.7443609
>
> prob3 <- predict(bmodelCV$BestModel, type="response")
> pred3 <- rep("No Diabetes", 532)
> pred3[prob3 > 0.25] <- "Diabetes"
> table3 <- table(pred3, Diabetes$Outcome)
> table3

pred3         0   1
 Diabetes    111 144
 No Diabetes 244  33
```

```
> accuracy3 <- (table3[1,2]+table3[2,1])/length(Diabetes$Outcome)
> accuracy3
[1] 0.7293233
>
> pred4 <- rep("No Diabetes", 532)
> pred4[prob.lasso > 0.25] <- "Diabetes"
> table4 <- table(pred4, Diabetes$Outcome)
> table4

pred4         0   1
  Diabetes    110 150
  No Diabetes 245  27
> accuracy4 <- (table4[1,2]+table4[2,1])/length(Diabetes$Outcome)
> accuracy4
[1] 0.7424812
>
> accuracylist <- c(accuracy1, accuracy2, accuracy3, accuracy4)
> accuracylist
[1] 0.7481203 0.7443609 0.7293233 0.7424812
>
> bmodelAIC$BestModel

Call:  glm(formula = y ~ ., family = family, data = Xi, weights = weights)

Coefficients:
          (Intercept)           Pregnancies               Glucose                   BMI
             -9.87999               0.12389               0.03503               0.08512
DiabetesPedigreeFunction                   Age
              1.32155               0.02384

Degrees of Freedom: 531 Total (i.e. Null);  526 Residual
Null Deviance:     676.8
Residual Deviance: 467.1        AIC: 479.1
> bmodelBIC$BestModel

Call:  glm(formula = y ~ ., family = family, data = Xi, weights = weights)

Coefficients:
          (Intercept)           Pregnancies               Glucose                   BMI
             -9.44941               0.17288               0.03648               0.08423
DiabetesPedigreeFunction
              1.36175

Degrees of Freedom: 531 Total (i.e. Null);  527 Residual
Null Deviance:     676.8
Residual Deviance: 470.3        AIC: 480.3
> bmodelCV$BestModel
```

```
Call:  glm(formula = y ~ ., family = family, data = data.frame(Xy[,
    c(bestset[-1], FALSE), drop = FALSE], y = y))

Coefficients:
(Intercept)  Pregnancies      Glucose          BMI
   -8.83172      0.16752      0.03667      0.08642

Degrees of Freedom: 531 Total (i.e. Null);  528 Residual
Null Deviance:     676.8
Residual Deviance: 485.2        AIC: 493.2
>

> roc(Diabetes$Outcome, bmodelAIC$BestModel$fitted.values, plot = TRUE, percent =
TRUE, xlab = "False positive rate", ylab = "False negative rate", legacy.axes = TRUE, col =
"Red", print.auc = TRUE)
Setting levels: control = 0, case = 1
Setting direction: controls < cases

Call:
roc.default(response = Diabetes$Outcome, predictor =
bmodelAIC$BestModel$fitted.values,     percent = TRUE, plot = TRUE, xlab = "False positive
rate",     ylab = "False negative rate", legacy.axes = TRUE, col = "Red",     print.auc = TRUE)

Data: bmodelAIC$BestModel$fitted.values in 355 controls (Diabetes$Outcome 0) < 177
cases (Diabetes$Outcome 1).
Area under the curve: 86.06%
> plot.roc(Diabetes$Outcome, bmodelBIC$BestModel$fitted.values, add = TRUE , col =
"Blue",print.auc = TRUE, print.auc.y = 40, percent = TRUE)
Setting levels: control = 0, case = 1
Setting direction: controls < cases
> plot.roc(Diabetes$Outcome, bmodelCV$BestModel$fitted.values, add = TRUE , col =
"Brown",print.auc = TRUE, print.auc.y = 30, percent = TRUE)
Setting levels: control = 0, case = 1
Setting direction: controls < cases
> plot.roc(Diabetes$Outcome, prob.lasso, add = TRUE , col = "Green",print.auc = TRUE,
print.auc.y = 20, percent = TRUE)
Setting levels: control = 0, case = 1
Setting direction: controls < cases
>
```


False negative rate
0
20
40
60
80
100
False positive rate
0
20
40
60
80
100
AUC: 86.1%
AUC: 85.8%
AUC: 84.2%
AUC: 86.0%