Prediction of Coronary Heart Disease using Logistic Regression

Inder Kaur
Gourab Karmakar
Akashdip Sen
Department of Statistics, University of Kalyani

Under the guidance of
Soumyadeep Das,
Assistant Professor, Department of Statistics
Bidhannagar College

July 2020

Abstract

In the present 21st century, the way an individual leads his life is a major factor leading to numerous medical problem. One of the major health issues is Coronary Heart Disease (CHD) that might be nascent since the early age. Today machine learning is extensively applied in the medical industry. The objective of this project is to use technology to predict the possibility of CHD based on the lifestyle of a person. Consequently that would help the patients to take appropriate preventive measures and the doctors to determine their course of action. This will decrease the medical expense incurred by the patient and would help him in getting the right treatment at an early stage which would decrease the rate of mortality due to CHDs. Logistic regression has been applied to serve the purpose of this study to find the features that are statistically significantly associated with Coronary Heart Disease .

Keywords- *Logistic Regression, Machine learning, Coronary Heart Disease*

Contents

1 Introduction

In today's fast growing world, Coronary Heart Diseases have become the leading cause of mortality. 17.9 million people die each year from Coronary vascular diseases, an estimated 31% of all death worldwide. Despite of wide variation in the factors leading to coronary heart diseases, tobacco use, high cholesterol level, rapid fluctuations in blood pressure, education, diabetes, low fruit and vegetable intake in the lower socioeconomic backgrounds have emerged out as a major causes of this disease. To counter this epidemic, the development of strategies like formulation and effective implementation of evidence based policy, early detection, reinforcement of health systems are required. The main motivation for choosing this case study is to predict using the past history of the patient, the chances of him/her developing coronary heart disease in the long run. This will help the patient to take the required precautions from before-hand. Hence this shall help in decreasing the rate of mortality due to coronary diseases in the future.

The entire study is divided into 3 folds with the 1st fold describes the data, 2nd fold dealing with the development of the theory and the 3rd fold deals with the analysis.

2 Data Description

The dataset is a part of the ongoing cardiovascular study conducted on the residents of the town of Framingham, Massachussetts. The Main aim is to classify two groups of people, i.e., those who are at risk of the coronary Heart Disease in ten years and those who are not, taking into account several factors that aid the disease, using the Logistic Regression Classifier.

The Dataset provides the information on 4238 patients with 16 attributes. The dataset has been downloaded from Kaggle.

2.1 Variable Description

Each of the attributes presented in the Dataset is a potential Risk Factor.

- **Sex**: (Male or Female) Whether the patient is a male or a female.(binary variable with 1 for male and 0 for female)
- **Age**: Age of the patients rounded to the nearest integer.(continuous)
- **currentSmoker**: Whether the patient under study is a current smoker or not.(binary variable with 1 for being a current smoker and 0 for being a non-smoker)
- **cigsPerDay**: Number of cigarettes consumed per day on an average by the patient.(continuous)
- **BPMeds**: Whether the patient was under blood pressure medication or not.
- **prevalentStroke**: Whether the patient had stroke in the past or not.(binary variable with 1 indicating that the patient had stroke and 0 indicating that the patient didn't had stroke)
- **prevalentHyp**: Whether the patient was hypertensive or not.(binary variable with 1 indicating that the patient was hypertensive and 0 indicating that the patient wasn't)
- **Diabetes**: Whether the patient had diabetes or not.(binary variable with 1 indicating presence of diabetes and 0 indicating absence)

- **totChol**: Total cholesterol level of the patient.(continuous variable)
- **sysBP**: Systolic Blood pressure.(continuous variable)
- **diaBP**: Diastolic Blood pressure.(continuous variable)
- **BMI**: Body Mass Index of the patient.(continuous variable)
- **heartRate**: Heart rate of the patient.(continuous variable)
- **glucose**: Glucose level of the patient.(continuous variable)
- **TenYearCHD(Target variable)**: Ten year risk of Coronary Heart Disease of the patient.(binary variable with 1 indicating that the patient is at risk of CHD and 0 indicating that the patient is not at risk)

2.2 Missing values

In real world, it is quite usual to encounter the problem of missing values in datasets. The current dataset on which the model is to be built also has some missing values corresponding to the different attributes described earlier which is provided in the following table:-

Variables	Missing value Counts
male	0
age	0
education	105
currentSmoker	0
cigsPerDay	29
BPmeds	53
prevalentStroke	0
prevalentHyp	0
diabetes	0
totChol	50
sysBP	0
diaBP	0
BMI	19
heartRate	1
glucose	388
TenYearCHD	0

In total there are about 645 missing observations across all the attributes in the study and it is impossible to fill those missing values by means of forward filling, backward filling or filling the missing observation with the mean value of the corresponding attribute. So the model is built by dropping the patients with missing values to avoid any kind of fallacy that may affect the model development in further steps.

2.3 A Glimpse into the data

Here is a glimpse of the first 10 rows of the data on which the analysis is to be carried-

male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
1	39	4	0	0	0	0	0	0	195	106	70	26.97	80	77	0
0	46	2	0	0	0	0	0	0	250	121	81	28.73	95	76	0
1	48	1	1	20	0	0	0	0	245	127.5	80	25.34	75	70	0
0	61	3	1	30	0	0	1	0	225	150	95	28.58	65	103	1
0	46	3	1	23	0	0	0	0	285	130	84	23.1	85	85	0
0	43	2	0	0	0	0	1	0	228	180	110	30.3	77	99	0
0	63	1	0	0	0	0	0	0	205	138	71	33.11	60	85	1
0	45	2	1	20	0	0	0	0	313	100	71	21.68	79	78	0
1	52	1	0	0	0	0	1	0	260	141.5	89	26.36	76	79	0
1	43	1	1	30	0	0	1	0	225	162	107	23.61	93	88	0

3 Logistic Regression

Classification problems mainly deal with the problem of identifying to which category or sub category a new observation belongs to, on the basis of the training set of instances belonging to the same population whose category is already known. Basically it is used to predict the class or category of a new observation based on the training set.

Linear regression is the most extensively used statistical technique for predictive analysis. It is used to explain the relationship between dependent and independent variables using a straight line.

Logistic regression is a statistical technique used in research designs that call for analyzing the relationship of an outcome or dependent variable (categorical in nature) to one or more predictors or independent variables while the dependent variable is either

- **Dichotomous (binomial)** : having only two categories. for example, whether one uses drugs(illegal) (no or yes).
- **Unordered polychotomous** : which is a nominal scale variable with three or more categories, for example, political party identification (Democrat, Republican, other, or none).
- **Ordered Polychotomous** : which is an ordinal scale variable with three or more categories, for example, level of education completed (e.g., less than elementary school, elementary school, high school, an undergraduate degree, or a graduate degree).

3.1 Preference of Logistic Regression over Linear Regression

Logistic Regression is an instance of classification technique that can be used to predict a qualitative response. Predicting a qualitative response for an observation can be referred to as classifying that observation, since it involves classifying the observation to a category. On the other hand, the methods that are often used for classification first predict the probability of each of the categories of a qualitative variable, as the basis for making the classification. A simple linear regression model, with Y as the response variable and X as the covariate, is not capable of predicting the probability. If linear regression is used to model a binary response variable, the resulting model may restrict

the predicted Y values within 0 and 1. Here's where logistic regression comes into play where the probability score reflects the probability of the occurrence of the event.

3.2 Logistic Model

Consider the study where the outcome variable Y is dichotomous, i.e, it is coded as 1 indicating the presence of a characteristic and 0 indicating the absence of a characteristic and the quantities $X_1, X_2, \ldots, X_p$ denotes the p-independent variables. In any regression problem, the key quantity is the mean value of the outcome variable given the value of the independent variable which is denoted as $E(Y|\mathbf{X}=\mathbf{x})$. This quantity is read as the "expected value of Y for a given value of $X_1, X_2, \ldots, X_p$". In linear regression this conditional mean may be expressed as

$$E(Y|\mathbf{X} = \mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

, where $\mathbf{X}'=(X_1, X_2, \ldots, X_p)$

This expression indicates that it is possible for $E(Y|x)$ to take any value between $-\infty$ and $+\infty$. But with dichotomous data, the conditional mean must be greater than or equal to 0 and less than or equal to 1. When $E(Y|\mathbf{x})$ is plotted against the different values of $\mathbf{x}'=(x_1, x_2, \ldots, x_p)$ it is seen that the conditional mean approaches 0 and 1 gradually and the plot resembles the shape of an *S-shaped curve* which resembles the cumulative distribution function of a Logistic distribution . Many distributions have been proposed for use in the analysis of a dichotomous outcome variable but the logistic distribution is preferred as it is quite flexible and lends itself to a clinically meaningful interpretation. In order to simplify the notation, the quantity $\pi(\mathbf{x})$ is used to represent the conditional mean of Y given $\mathbf{x}$ when the logistic distribution is used. The specific form of the logistic distribution used for a dichotomous outcome variable and p predictor variables is

$$\pi(\mathbf{x}) = \frac{e^{\beta_0 + \beta' \mathbf{x}}}{1 + e^{\beta_0 + \beta' \mathbf{x}}}$$

A transformation of $\pi(\mathbf{x})$ that is central to the study of logistic regression is the logit transformation. This transformation is defined in terms of $\pi(\mathbf{x})$ as

$$g(\mathbf{x}) = ln\left[\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right] = \beta_0 + \beta'\mathbf{X}$$

where $\beta' = (\beta_1, \beta_2, \ldots, \beta_p)$.

In case of p covariates $X_1, X_2, \ldots\ldots, X$, the logit of the multiple regression model is given by

$$g(\mathbf{x}) = \beta_0 + \beta'\mathbf{X}$$

Where $g(\mathbf{x})$ is defined as above and we have

$$\pi(\mathbf{x}) = P[Y = 1|\mathbf{X} = \mathbf{x}]$$

Vectors $\mathbf{X}$ and β are defined as usual.

3.3 Estimation of parameters:

The parameters of the logistic regression model can be obtained by the method of Maximum Likelihood Estimation.

Let $P[Y_i = 1] = \pi_i$ and $P[Y_i = 0] = 1 - \pi_i$, $i = 1, 2, \ldots, n$

The likelihood function is

$$L(\beta|Y_1, Y_2, \ldots, Y_n) = \prod_1^n \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

where $\beta' = (\beta_1, \beta_2, \ldots, \beta_p)$ and $\pi_i = \frac{\exp(\beta_0 + \sum_1^p x_{ji}\beta_j)}{1+\exp(\beta_0 + \sum_1^p x_{ji}\beta_j)}$.

Taking Log on both sides and then differentiating the likelihood with respect to β_j, we have-

$$\frac{\partial log L(\beta|Y_1, Y_2, \ldots, Y_n)}{\partial \beta_j} = \mathbf{x_j'}(\mathbf{y} - \pi), j = 0, \ldots, p$$

where $\mathbf{x_j'} = (x_{j1}, x_{j2}, \ldots, x_{jp})$ & $\pi = (\pi_1, \pi_2, \ldots, \pi_n)$.

Upon further Calculation we obtain,

$$\frac{\partial}{\partial\beta} logL(\beta|Y_1, Y_2, \ldots, Y_n) = \mathbf{X}'(\mathbf{Y} - \pi)$$

where vectors $\mathbf{X}, \pi$ are defined as usual and $\mathbf{Y}' = (Y_1, Y_2, \ldots, Y_p)$.

The above system of equations of the order cannot be solved analytically. But it can be solved by Newton-Raphson method. The convergence is reached when $\beta_i \simeq \beta_{i-1}$

4 Model Evaluation

4.1 Akaike's Information Criteria

Suppose that there is a statistical model of some data. Let p be the number of estimated parameters in the model. For a p-variate linear regression model with $\varepsilon \sim N(0, \sigma^2 I)$, the likelihood function is

$$L(y, \beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[\frac{-(y - X\beta)'(y - X\beta)}{2\sigma^2}\right]$$

whose log-likelihood is

$$lnL(y, \beta, \sigma^2) = -\frac{n}{2} ln2\pi - \frac{n}{2} ln\sigma^2 - \frac{1}{2}\frac{(y - X\beta)'(y - X\beta)}{\sigma^2}$$

So, Akaike's Information Criteria is defined as

$$AIC = 2p - 2lnL(y, \widehat{\beta}, \widetilde{\sigma^2})$$

where $\widehat{\beta}$ is the MLE of β & $\widetilde{\sigma^2} = \frac{n-p}{n}\widehat{\sigma^2}$ is the unbiased estimator of $\widehat{\sigma^2}$, $\widehat{\sigma^2}$ being the MLE of σ^2.

Given a set of candidate models for the data, the preferred model is the one with the minimum AIC value. Thus, AIC rewards goodness of fit (as assessed by the likelihood function), but it also includes a penalty that is an increasing function of the number of estimated parameters. The penalty

discourages overfitting, because increasing the number of parameters in the model almost always improves the goodness of the fit.

4.2 Receiver Operating Characteristic(ROC)

A Receiver Operating Character curve or the R.O.C curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The ROC curve is created by plotting the True Positive Rate (TPR) against the False Positive Rate at various threshold settings. The True Positive Rate is also known as Sensitivity, Recall or Probability of detection in machine learning. The false positive rate is known as the probability of false alarm and can be calculated as (1 – specificity). It can also be thought of as the plot of the power as a function of the Type 1 error of the decision rule.

4.3 Area under the ROC curve (AUC)

To compare different classifier, it can be useful to summarize the performance of each classifier into a single measure. One common approach is to calculate the area under the ROC curve which is abbreviated as AUC.
A classifier with higher AUC can occasionally score worse in a specific region than another classifier with lower AUC. But in practice, the AUC performs well as a general measure of predictive accuracy.

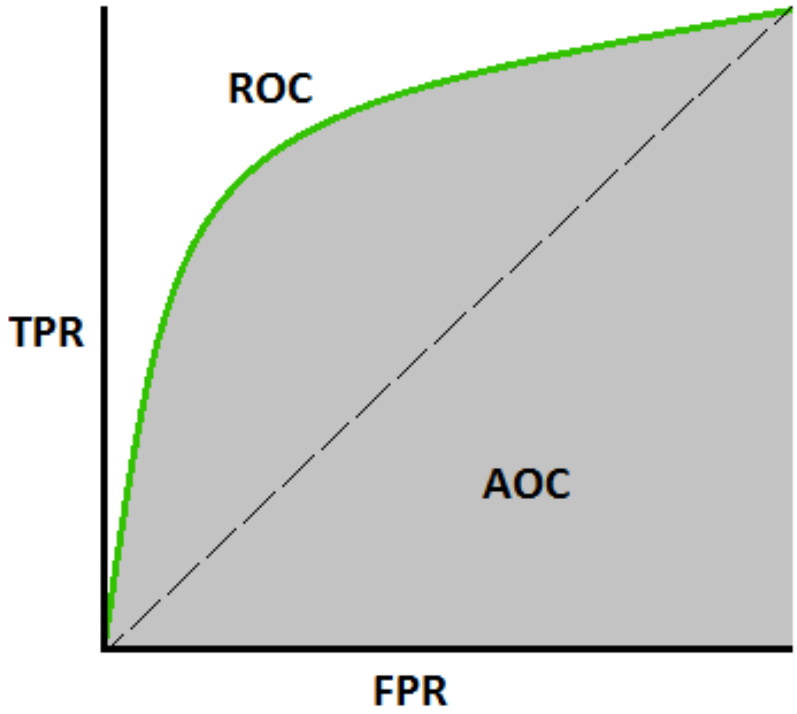


Figure 1: Example of an ROC AUC curve

5 Dealing With Multicollinearity-

5.1 Variance Inflation Factor

Interpretation of the multiple regression equation depends implicitly on the assumption that the covariates are not strongly inter-related. I is usual to interpret a regression coefficient as measuring the change in the response variable when the corresponding covariate is increased by one unit and all the other covariates are held constant. This interpretation may not be valid if there are strong linear relationships among the covariates. This problem is addressed as multi-collinearity. A thorough investigation of multi-collinearity will involve examining the value of R^2 that results from regressing each of the covariates against all the others. The relationship between the predictor variables can be evaluated by examining a quantity called the variance inflation factor (VIF). Let R_j^2 be the square of the multiple correlation coefficient that results when the covariate X_j is regressed against all the other covariates. Then the variance inflation for X_j is

$$VIF_j = \frac{1}{1 - R_j^2}, j = 1, 2, \ldots, p$$

The Variance Inflation (VIF) is the quotient of the variance in a model with multiple terms by the variance of a model with one term alone. It quantifies the severity of multi-collinearity, i.e., the correlation among the covariates, in an ordinary least square regression analysis . It provides an index that measures how much the variance of an estimated regression coefficient is increased because of collinearity.
If the VIF of the ith estimated parameter of the multiple linear regression equation is greater than 5 , then there is some moderate collinearity but if the VIF of the ith estimated parameter is greater than 10, then there is some severe multi-collinearity in the model.

6 Confusion Matrix

In the field of machine learning and specifically the problem of statistical classification, a confusion matrix or the error matrix is specific table layout that allows visualizations of the performance of the algorithm. Each row of the matrix represents the instances in an actual class (or vice versa). The name stems from the fact that it makes easy to see if the system is confusing two classes. It is special kind of contingency table with two dimensions ("actual and "predicted"), and identical sets of classes in both dimensions. It is extremely useful for measuring the Recall, precision, specificity, accuracy and most importantly AUC-ROC curve.
The form of the confusion matrix is given:

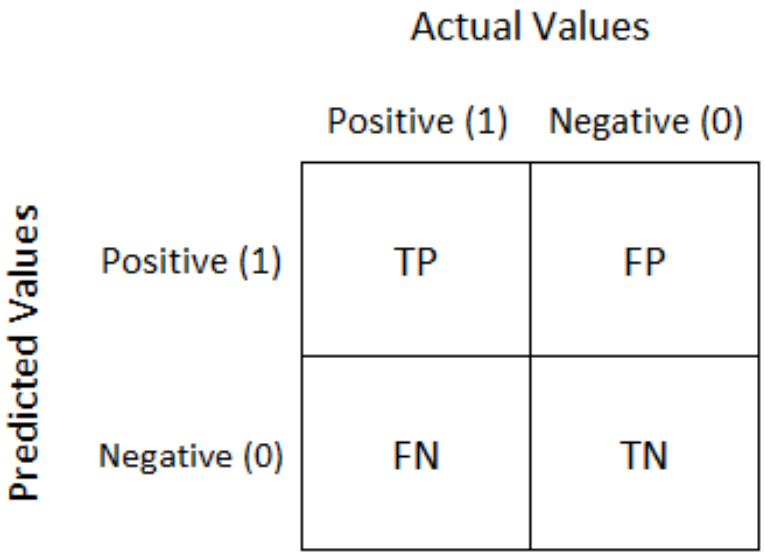


Figure 2: Example of a Confusion Matrix

The quantities are described as follows:

1. **True Positive (TP)** -Observation is positive and is predicted to be positive.

2. **False Negative (FN)** - Observation is positive but is predicted to be negative.

3. **True Negative (TN)** - Observation is negative and is predicted to be negative.

4. **False Positive (FP)** - Observation is negative but is predicted to be positive.

Classification rate or accuracy is defined by –

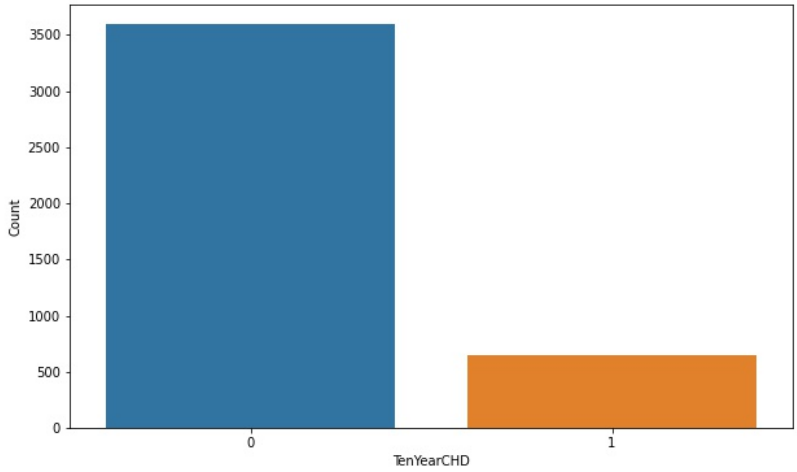
$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

7 Analysis

7.1 Exploratory Data Analysis

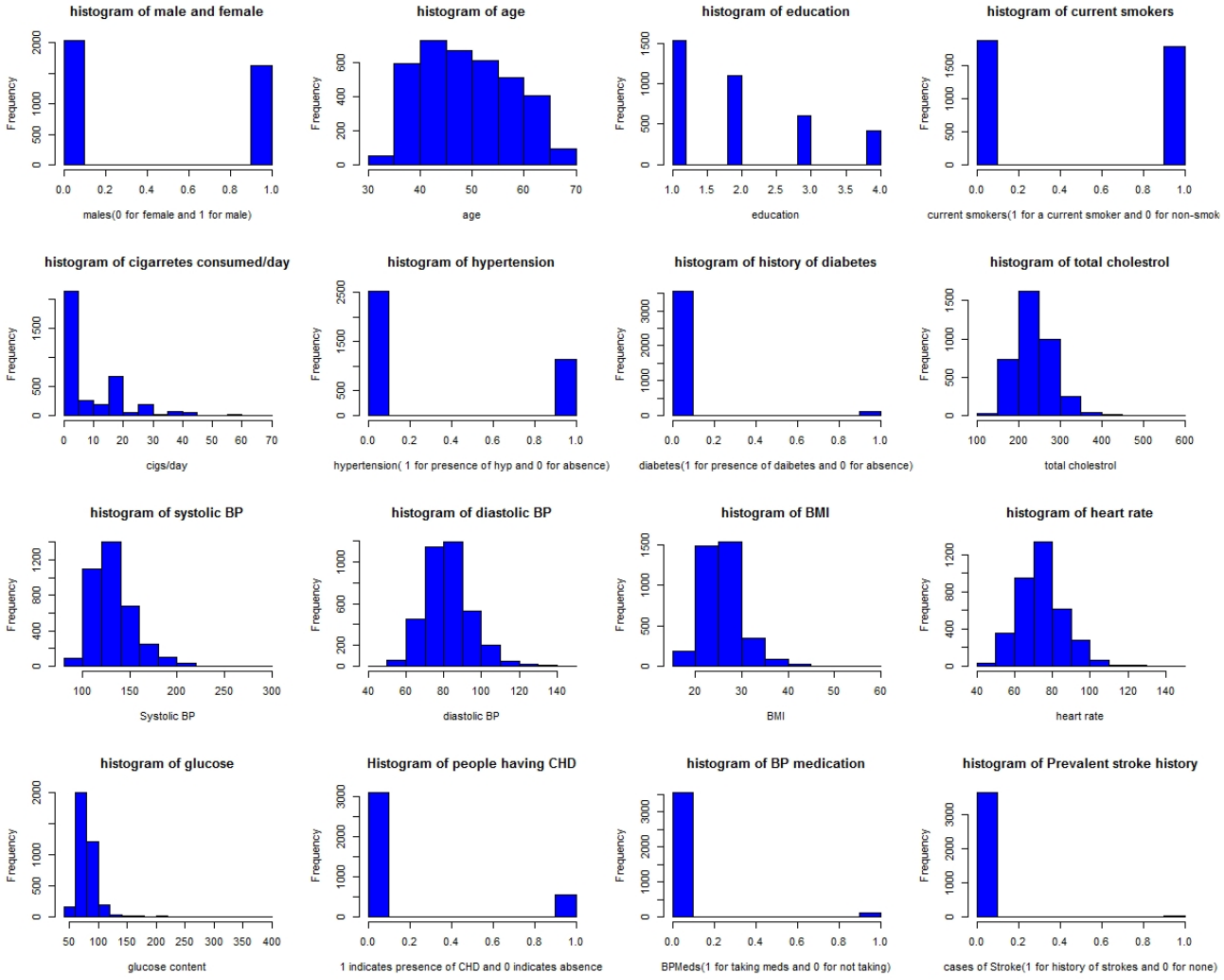
7.1.1 Bar Plot for the number of people having CHD

First we examine the number of people who have Coronary Heart Disease with the help of a basic bar plot-



It is clear that majority of the people in the dataset don't have any Heart Disease.

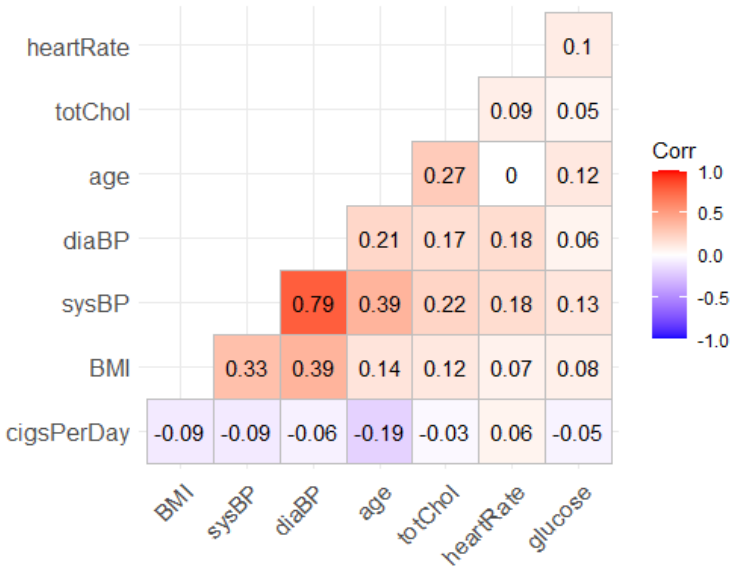
7.1.2 Univariate analysis



Some of the points to be noted from the histogram-

- The female subjects are more than the number of male subjects in the study.
- Most of the subjects dont consume cigarettes at all.
- Most of the subjects dont have any history hypertension, diabetes, strokes and BP medication.

7.1.3 Bivariate Analysis

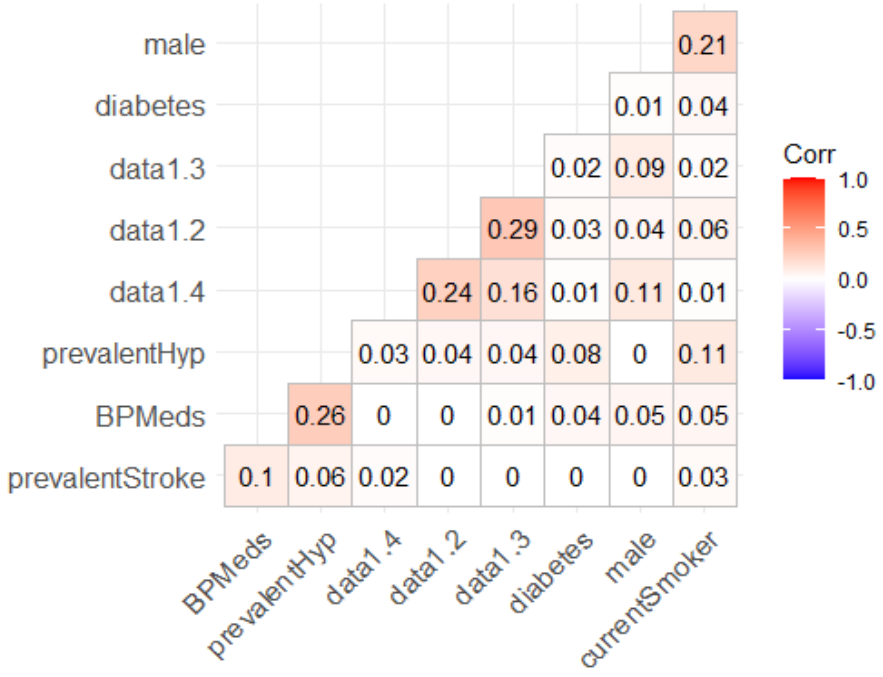


From the correlation matrix of the numerical variables pairs with correlation more than 0.6 are checked . It is clear that the variables-

- diaBP and sysBP are correlated

So the variables sysBP is dropped.

The association matrix of the categorical variables by using the cramer's V is also represented-



From the association matrix of the categorical variables, it is evident that there is no significant association among the categorical variables. So, all the categorical variables are retained.

7.2 Model Fitting

The dataset is then randomly divided into the training set and the test set with the split ratio being 75:25 because the model can be trained with the training set so that the model sees and learns from the data. The test set is used to provide an unbiased evaluation of a final model fit on the training dataset.

After careful analysis of the correlation matrix, the variables currentSmoker, diabetes, sysBP and prevalentHyp are dropped from the features.

7.2.1 Fitting the Logistic Regression model-

At first, the model is fitted on the features that were left after removal of the variables that were causing the problem of multi-collinearity.

From the output provided in Figure 3, it is already clear that most of the explanatory variables are not significant and may not be beneficial for the purpose of the study.

The significant variables are retained using the *Backward Selection* technique and the output of the final iteration of the procedure giving the reduced model is provided in Figure 4.

Clearly, the variables retained are all significant at 10% level of significance to the purpose of the study.

```
Call:
glm(formula = TenYearCHD ~ ., family = binomial, data = training_set)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8994  -0.5909  -0.4257  -0.2837   2.8616

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     -8.841069   0.832892 -10.615  < 2e-16 ***
male             0.415397   0.126880   3.274  0.00106 **
age              0.060482   0.007829   7.726 1.11e-14 ***
currentSmoker    0.098592   0.181112   0.544  0.58619
cigsPerDay       0.018389   0.007084   2.596  0.00944 **
BPMeds           0.189663   0.268429   0.707  0.47984
prevalentStroke  0.859823   0.551487   1.559  0.11897
prevalentHyp     0.297466   0.159922   1.860  0.06288 .
diabetes        -0.067457   0.362069  -0.186  0.85220
totChol          0.002701   0.001287   2.098  0.03589 *
sysBP            0.011415   0.004361   2.617  0.00886 **
diaBP            0.001803   0.007474   0.241  0.80934
BMI              0.024966   0.014658   1.703  0.08851 .
heartRate       -0.002343   0.004824  -0.486  0.62728
glucose          0.008188   0.002717   3.013  0.00259 **
data1.2         -0.265149   0.144320  -1.837  0.06618 .
data1.3         -0.156227   0.176130  -0.887  0.37508
data1.4          0.086904   0.184163   0.472  0.63701
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2341.3  on 2741  degrees of freedom
Residual deviance: 2062.8  on 2724  degrees of freedom
AIC: 2098.8

Number of Fisher Scoring iterations: 5
```

Figure 3: output after fitting the retained variables

```
Call:
glm(formula = TenYearCHD ~ male + age + cigsPerDay + prevalentStroke +
    prevalentHyp + totChol + sysBP + BMI + glucose + data1.2,
    family = binomial, data = training_set)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9804  -0.5953  -0.4289  -0.2839   2.8913

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -8.940873   0.688222 -12.991  < 2e-16 ***
male             0.437881   0.124067   3.529 0.000417 ***
age              0.060433   0.007560   7.994  1.3e-15 ***
cigsPerDay       0.020836   0.004797   4.343  1.4e-05 ***
prevalentStroke  0.911572   0.545083   1.672 0.094454 .
prevalentHyp     0.304982   0.157122   1.941 0.052252 .
totChol          0.002693   0.001281   2.103 0.035462 *
sysBP            0.012216   0.003321   3.678 0.000235 ***
BMI              0.025419   0.014107   1.802 0.071565 .
glucose          0.007653   0.002025   3.780 0.000157 ***
data1.2         -0.243801   0.134532  -1.812 0.069953 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2341.3  on 2741  degrees of freedom
Residual deviance: 2065.3  on 2731  degrees of freedom
AIC: 2087.3

Number of Fisher Scoring iterations: 5
```

Figure 4: output for the reduced model

7.2.2 Checking for Multicollinearity-

It is important to check whether the variables retained have correlations among them via the Variance Inflation Factors-

```
     male       age  cigsPerDay prevalentStroke  prevalentHyp   totChol
 1.209023  1.224891    1.254077        1.009764      1.933669  1.061635
    sysBP       BMI     glucose         data1.2
 2.079038  1.130361    1.020580        1.033994
```

It is clear that there is no multicollinearity in the reduced model.

7.2.3 Checking the Accuracy of the model-

After the model is fit, it is important to check whether there is an instance of overfitting or underfitting of the data.

The confusion matrices for both the training set and the test set are provided-

Accuracy of the model on the training set-

```
 y_pred_train
     0    1
0 2307   17
1  383   35

               Accuracy : 0.8541
                 95% CI : (0.8403, 0.8671)
```

It can be seen that there is 85.41% of accuracy rate of the model in the training set along with 400 misclassifications.

Accuracy of the model on the test set-

```
 y_pred__test
    0   1
0 768   7
1 129  10

               Accuracy : 0.8512
                 95% CI : (0.8265, 0.8737)
```

It can be seen that there is almost same accuracy on the test set and it can be said that there is no instance of over-fitting of the data, with the model making about 136 mis-classifications in the test set.

Model Adequacy check-

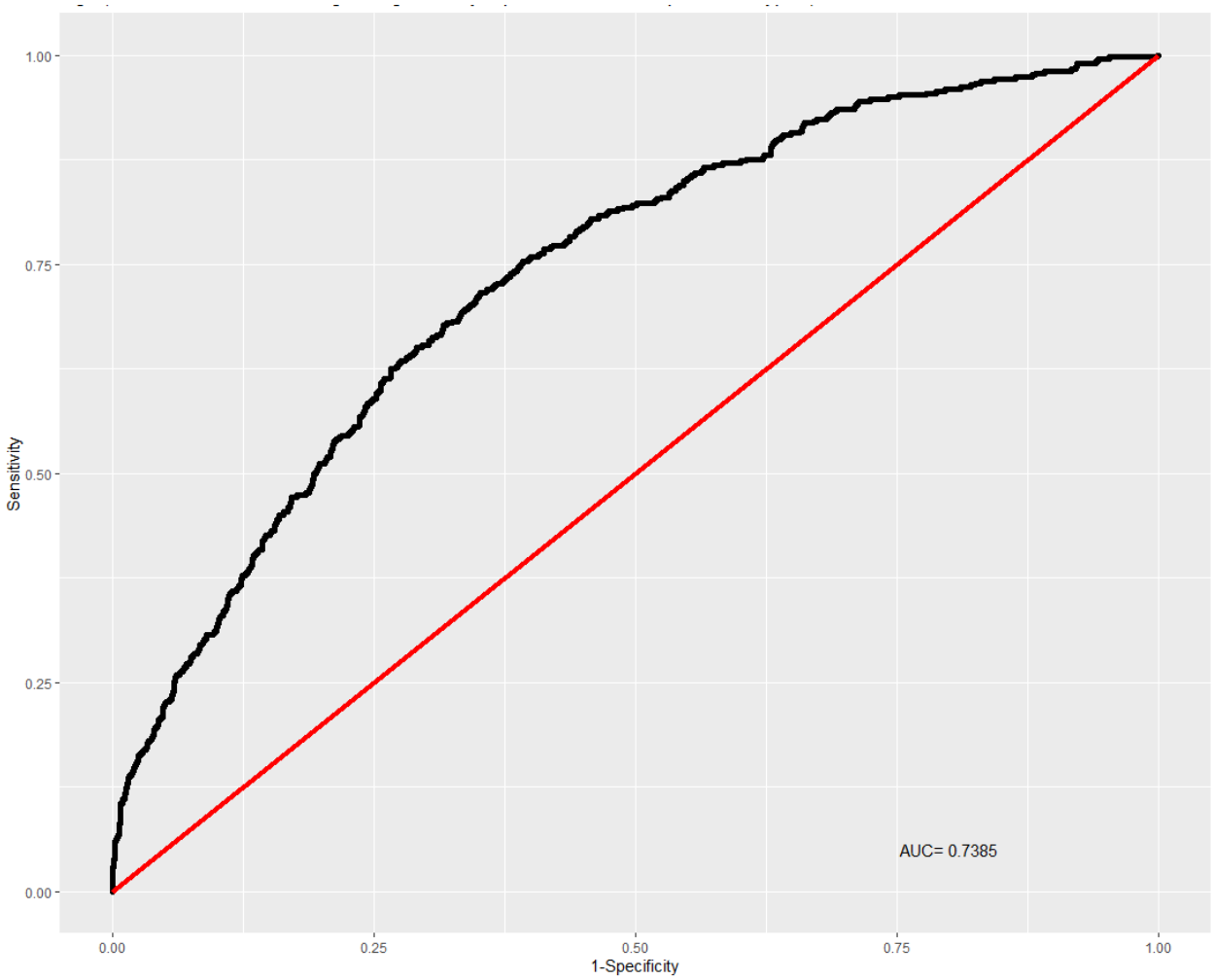
It is important to check the performance of the model based on factors like sensitivity, specificity. The output is produced as follows-

```
            Sensitivity : 0.58824
            Specificity : 0.85619
         Pos Pred Value : 0.07194
         Neg Pred Value : 0.99097
             Prevalence : 0.01860
         Detection Rate : 0.01094
   Detection Prevalence : 0.15208
      Balanced Accuracy : 0.72221

       'Positive' Class : 1
```

It can be observed from the above figures that logistic regression model is more specific rather than sensitive.

7.2.4 Plotting the ROC AUC curve-



The ROC-AUC curve quantifies the model accuracy, higher the area greater is the model classification accuracy. An AUC value of 73.85% means acceptable discrimination which means the model can perform well with more sample size.

8 Conclusion

It is very important to find the factors that are significant for Coronary Heart Disease. The variables retained in the study are all significant and can be counted upon as the factors which are associated to the risk of Coronary Heart Disease. Also it is found that men are more susceptible to Coronary Heart Disease(as evident from the log of odds in the summary table). The factors like smoking, strokes and increasing age all have a significant role in causing Coronary Heart Disease. Except age, the other factors, if controlled, can reduce the possibility of Coronary Heart Disease in future. The accuracy of the model is 85.12% which can be increased with more data and by including new factors related to the hygiene as well as the lifestyle of a person.

Acknowledgement

Before we get into the thick of things, we would like to thank our project guide Soumyadeep Das, Assistant Lecturer, Bidhannagar College who guided us at every step with his wisdom, knowledge as well as his constant support and energy. Without his insight and proper guidance the project wouldn't have been completed.

We would also like to thank Dr. Sisir Kumar Samanta, *Head*, *Department of Statistics*, *University of Kalyani* for allowing us to work on this topic "**Prediction of Coronary Heart Disease using Logistic Regression**" due to which we got to know a lot about the factors that are responsible for Heart diseases as well as build our concept on Logistic Regression and Machine learning. It is to these people that we owe our deepest gratitude.

References

[1] *Agresti, A. An Introduction to Categorical Data analysis, 2007, John Wiley & Sons, Inc.*

[2] *Dataset is obtained from-*
`https://www.kaggle.com/amanajmera1/framingham-heart-study-dataset/data`

[3] *Fox, J. Applied Regression analysis & Generalized Linear models, 2016, SAGE publications*

[4] *Hosmer, D.W, Lemeshow, S. Applied logistic Regression, 2000, John Wiley & Sons, Inc.*

[5] *James, G, Witten, D, Hastie, T, Tibshirani, R. An Introduction to-Statistical Learning with applications in R, 2017, Springer*