Western Michigan University - Department of Statistics

Predicting Heart Disease Risk Factors Using Logistic Regression Model

Ruvini Jayamaha

June 29, 2022

1 Introduction

Heart disease is the leading cause of death in the USA, causing 25% of deaths each year. The most common type of heart disease is coronary artery disease, in which blood flow to the heart is restricted. Certain medical conditions and lifestyle factors can put someone at a higher risk of heart disease, such as diabetes, obesity, physical inactivity, diet, smoking, and alcohol consumption. This study aims to build a best model that accurately predicts if a patient has heart disease based on the demographic, behavioral and medical risk factors.

2 Data and Methodology

The data are collected from 4 different medical clinical places. The source data set contains 303 patients records with 14 measurements with missing values and available in uci website[1]. For this analysis we use 207 rows and 14 columns which are the most effected to the response variable. The variables used for this analysis are:

Variable	Description	Variable Type	Factor levels
age	Age in years	Numerical	
sex	Sex	Factor	M - Male F- Female
cp	chest pain	Factor	1 - typical angina 2 - atypical angina 3 - non angina pain 4 - asymptomatic
trestbps	resting blood pressure in mm Hg	Numerical	
chol	serum cholesterol in mg/dl	Numerical	
fbs	fasting blood sugar in mg/dl	Factor	1 - if measurement is > 120 mg/dl 0 - if measurement is < 120 mg/dl
restecg	resting electrocardiographic results	Factor	1 - normal 2 - having ST-T wave abnormality 3 - probable or definite left ventricular hypertrophy
thalach	maximum heart rate	Numerical	
exang	exercise induced angina	Factor	1 - yes 0 - no
oldpeak	ST depression induced by exercise relative to rest	Numerical	
slope	slope of the peak exercise ST segment	Factor	1 - upsloping 2 - flat 3 - down sloping
ca	number of major vessels (0-3) colored by fluoroscopy	Factor	0 - absence 1,2,3 - presence
thal	short of thallium heart scan	Factor	3 - normal 6 - fixed defect 7 - reversible defect
hd	diagnosis of heart disease	Factor	0 - less than or equal to 50% diameter narrowing 1 - greater than 50% diameter narrowing

Table 2.1: Variables Description

Analysis is based on the regression approach. In this stage, we consider data pre-processing before applying regression methods. Mainly target to handle the missing values and create a new subset data set from original data set taking into account with interesting variables. Use pairwise comparison plots to visualize the data set and perform descriptive analysis.Further, perform statistical analysis based on chi squared tests, odds ratio, and confidence intervals. Secondly, check key assumptions and fit the logistic regression models and select the best fitted model. R software is used for this analysis.

$$log\left(\frac{p(x)}{1-p(x)}\right)=\beta_0+X.\beta \tag{2.1}$$

3 Simulation and Results

In data preparation stage, we cleared the missing values and prepared a subset from the source dataset including the factors we are interested in this analysis. The response variable is "hd" diagnosis of heart disease which is a binary. 207 rows and 14 columns are used for this analysis.

3.1 Exploratory Analysis

In this phase, we try to identify the relationship between the response variable presence of heart disease and some selected explanatory variables numerically and graphically.

3.1.1 Descriptive Statistics

Variable	Mean	Standard Deviation	Min	Max
"age" - Age	54.54	9.049736	29.0	77.0
"trestbps" - resting blood pressure in mmHg	131.7	17.76281	94.0	200.0
"chol" - serum cholesterol in mg/dl	247.4	51.99758	126.0	564.0
"thalach" - maximum heart rate	149.6	22.94156	71.0	202.0

Table 3.1: Summary Statistics

3.1.2 Prevalence of Heart disease across age

The distribution of the presence of heart disease is left skewed while the distribution of the absence of heart disease appears more normally distributed. These graphics suggests that age has a relationship with heart disease and there are more older people with heart disease than younger people with heart disease.

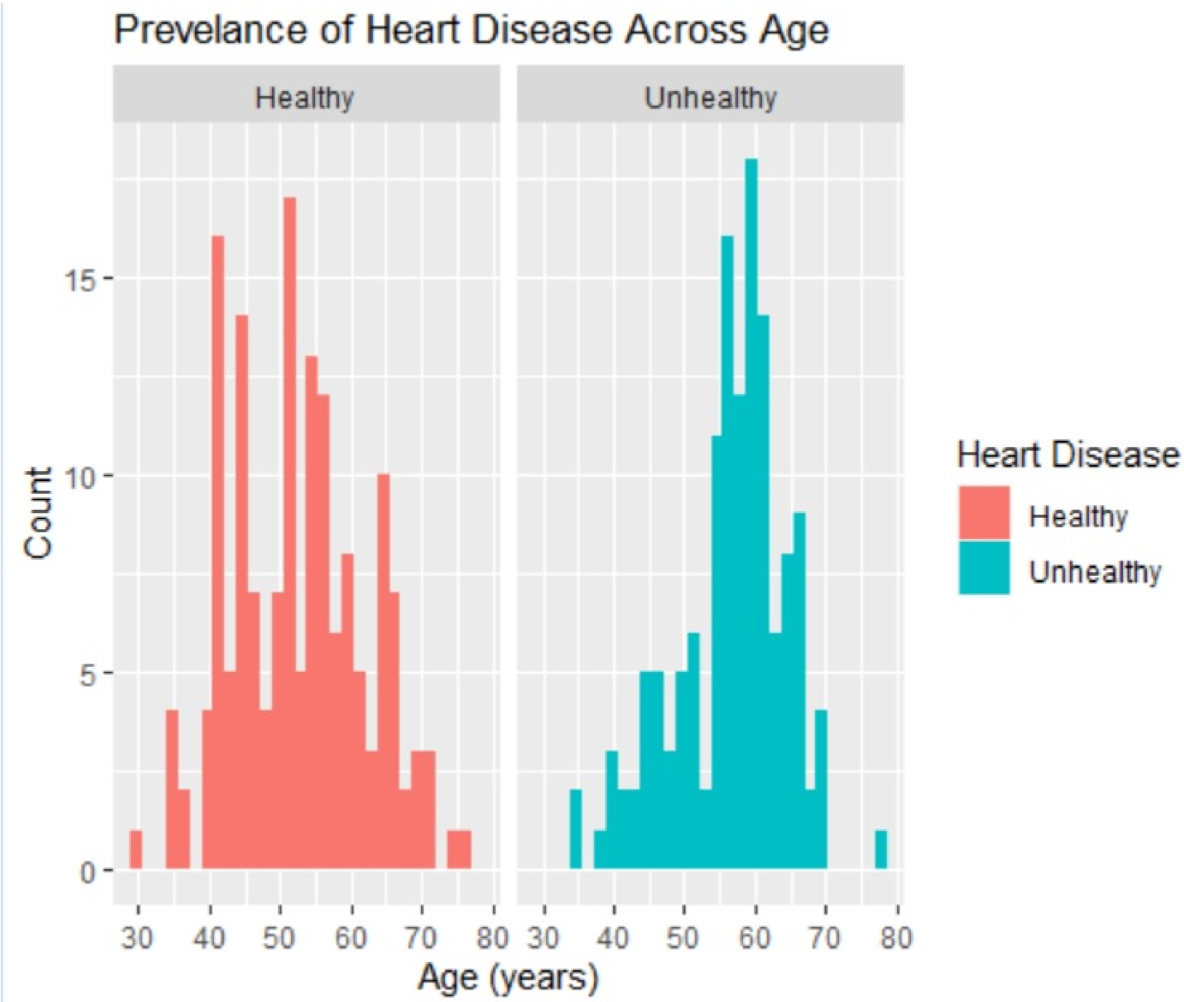

Figure 3.1: Heart disease vs age

3.1.3 Prevalence of Heart disease across different chest pain types

Relationship between presence of heart disease and chest pain, number 4 category that is Asymptotic chest pain type has the highest chance of presence of heart disease.

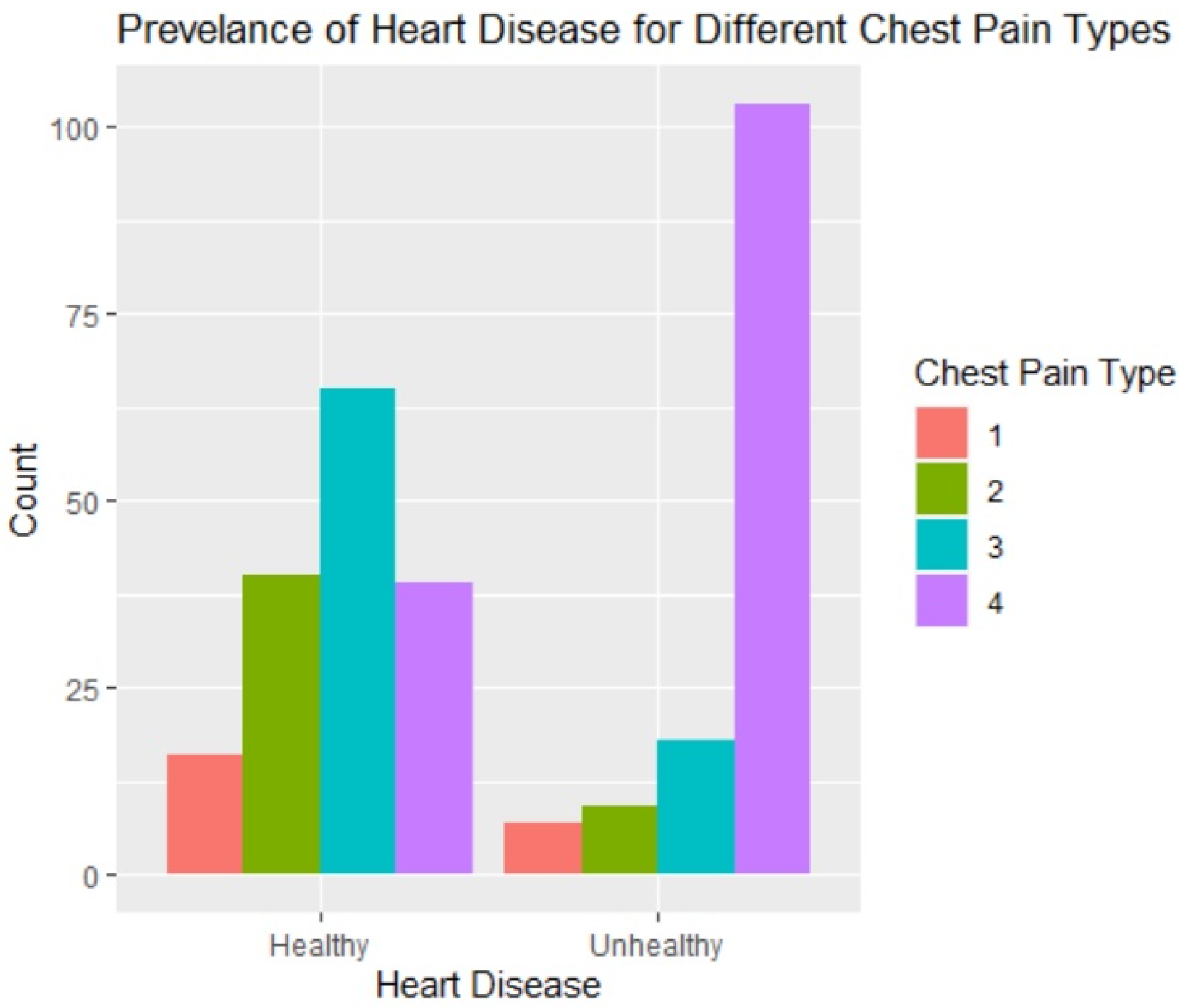

Figure 3.2: Heart disease vs chest pain

3.1.4 Prevalence of Heart disease across sex, fasting blood sugar and exercises induced angina

Figure 3.3 shows a higher proportion of males with heart disease than females with heart disease, suggesting that there is a relationship between sex and heart disease. Figure 3.4 fasting blood sugar levels do not appear to have a correlation with heart disease, as there appears to be a similar proportion of presence and absence of heart disease for people with high and low fasting blood sugar levels. Figure 3.5 implies that there is an even larger distinction for heart disease as it relates to exercise induced angina, for a much higher proportion of people with exercise induced angina have heart disease compared to people without exercise induced angina.

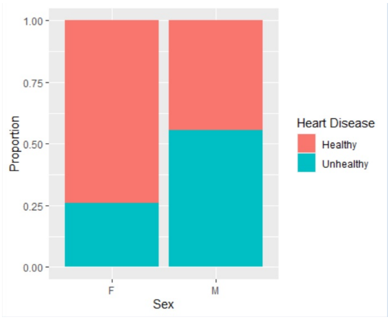

Figure 3.3: hd vs sex

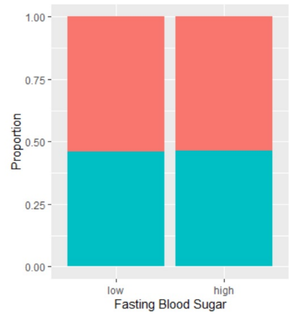

Figure 3.4: hd vs fbs

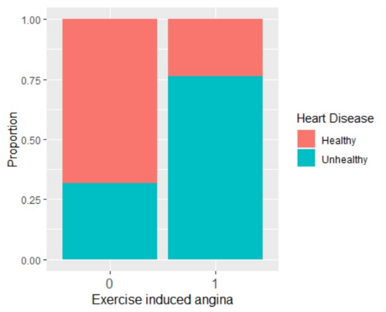

Figure 3.5: hd vs exang

3.1.5 Analysis of highest correlation variables

From the following pairwise comparison plot and R output, implies there is highest correlation between the variables,

- Age vs Serum Cholesterol level
- Age vs Resting Blood pressure
- Maximum heart rate vs Serum Cholesterol level

Further, without any specific fitting data visualizations shows that the relationship between above three cases have some curve fitting.

	data.age	data.trestbps	data.chol	data.thalach
data.age	1.0000000	0.29047626	2.026435e-01	-3.945629e-01
data.trestbps	0.2904763	1.00000000	1.315357e-01	-4.910766e-02
data.chol	0.2026435	0.13153571	1.000000e+00	-7.456799e-05
data.thalach	-0.3945629	-0.04910766	-7.456799e-05	1.000000e+00

Figure 3.6: Correlation Values

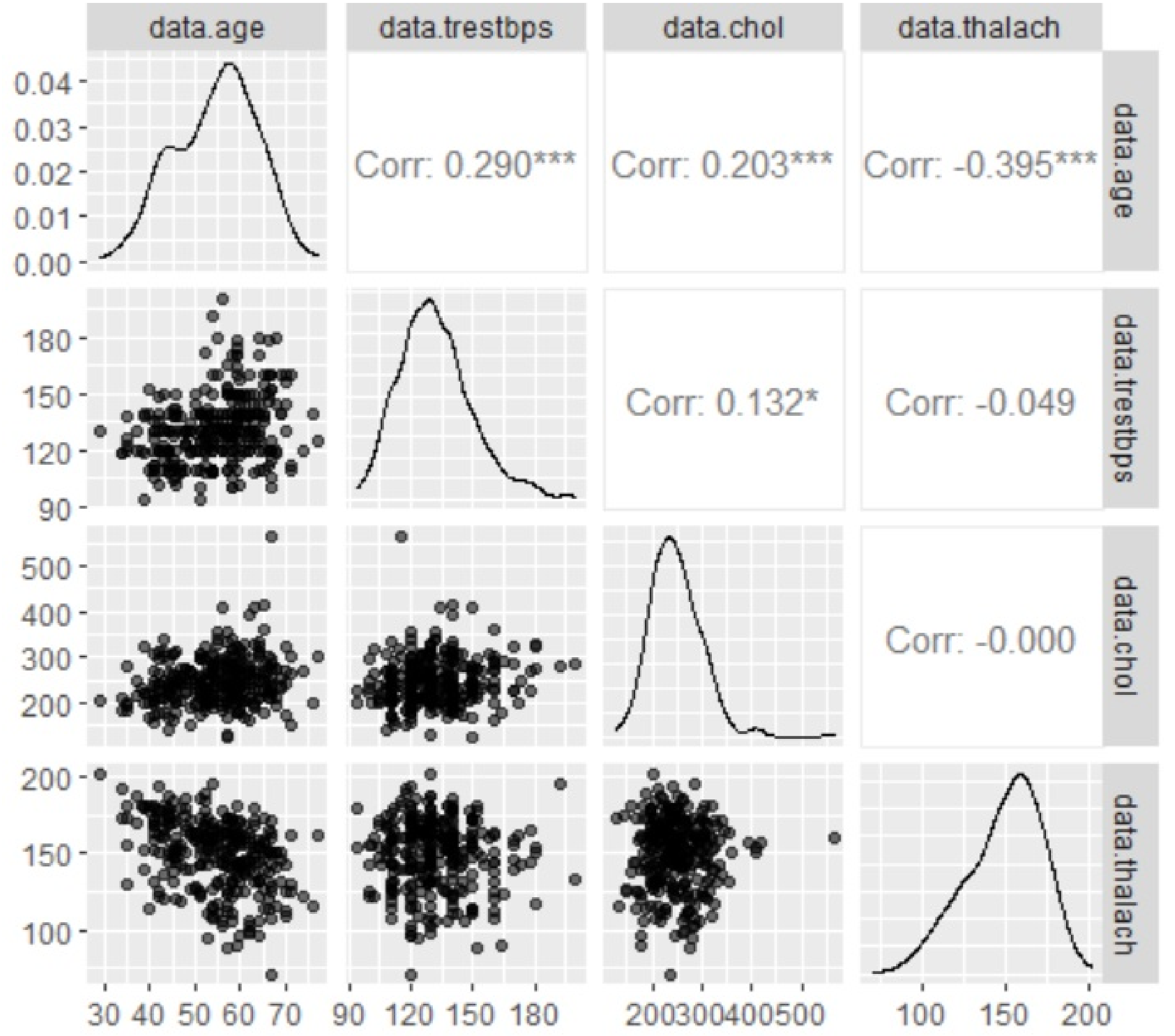

Figure 3.7: Pairwise Comparison Plot

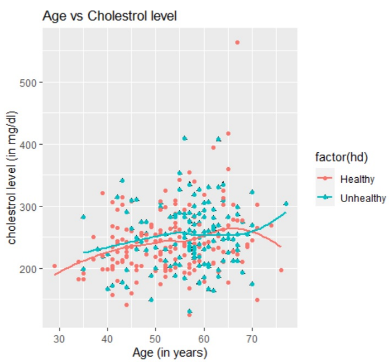

Figure 3.8: Age vs Cholesterol level

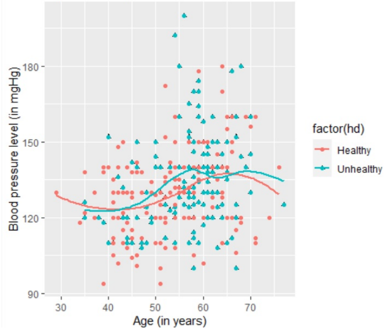

Figure 3.9: Age vs blood pressure

Figure 3.10: Heart rate vs Cholesterol level

3.2 Chi square Test Analysis

In this section, we compare the following variables with presence of heart disease and perform the independence test. Consider the following hypothesis for each cases,

$H0: X_1$ and X_2 are independent Vs $H1: X_1$ and X_2 are dependent

Heart disease presence compared with	Chi-squared value, df	P value	Conclusion ($\alpha = 0.05$)
Sex	21.852, df=1	2.946e-06	Dependent
Chest pain	77.276, df=3	2.2e-16	Dependent
Fasting Blood sugar	1.9997e-31, df=1	1	Independent
Resting electrocardiograph results	9.5755, df=2	0.008331	Dependent
Exercise induced angina	50.943, df=1	9.511e-13	Dependent
Slope of the peak exercise ST segment	43.473, df=2	3.63e-10	Dependent
Number of major vessels colored by fluoroscopy	72.301, df=3	1.373e-15	Dependent
Short of thallium heart scan	82.46, df=2	2.2e-16	Dependent

Table 3.2: Chi-Squared Test

From above analysis, presence of heart disease with the variables; Sex, Chest pain, resting electrocardiograph result, exercise induced angina, slope of the peak exercise ST segment, number of major vessels colored by fluoroscopy and short of thallium heart scan are associated at 5% significance level.

3.3 Odd Ratio and 95% Odd Ratio Confidence Interval

Heart disease presence compared with	Odd Ratio	95% OR CI
Sex	3.573933	(2.073452 , 5.990964)
Chest pain	0.5142857	(0.1691412 , 1.5744659)
Fasting Blood sugar	1.018209	(0.5374749 , 1.9388468)
Resting electrocardiographic results	2.017112	(1.264710 , 3.185909)
Exercise induced angina	6.996549	(3.957047 , 11.908679)
Slope of the peak exercise ST segment	5.069688	(3.045050 , 8.228052)
Number of major vessels colored by fluoroscopy	8.507527	(4.934884 , 14.154749)
Short of thallium heart scan	10.40131	(5.978033 , 17.403718)

Table 3.3: Odd ratio comparison

From odd ratio also suppose that being male is likely to increase the odds in presence of heart disease. By comparing 95% Confidence intervals, variables sex, Resting electrocardiographic results, exercise induced angina, slope of the peak exercise ST segment, number of major vessels colored by fluoroscopy, short of thallium heart scan are associated with presence of heart disease.

3.4 Fitting Logistic Regression model

Logistic Regression is a modeling technique commonly used in the biological sciences for categorical outcomes. Logistic Regression computes the probability of a discrete event and classifies each observation based on this probability.

3.4.1 Model 1: Using all variables

```
Call:
glm(formula = hd ~ ., family = "binomial", data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0490  -0.4847  -0.1213   0.3039   2.9086

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.253978   2.960399  -2.113 0.034640 *
age         -0.023508   0.025122  -0.936 0.349402
sexM         1.670152   0.552486   3.023 0.002503 **
cp2          1.448396   0.809136   1.790 0.073446 .
cp3          0.393353   0.700338   0.562 0.574347
cp4          2.373287   0.709094   3.347 0.000817 ***
trestbps     0.027720   0.011748   2.359 0.018300 *
chol         0.004445   0.004091   1.087 0.277253
fbs1        -0.574079   0.592539  -0.969 0.332622
restecg1     1.000887   2.638393   0.379 0.704424
restecg2     0.486408   0.396327   1.227 0.219713
thalach     -0.019695   0.011717  -1.681 0.092781 .
exang1       0.653306   0.447445   1.460 0.144267
oldpeak      0.390679   0.239173   1.633 0.102373
slope2       1.302289   0.486197   2.679 0.007395 **
slope3       0.606760   0.939324   0.646 0.518309
ca1          2.237444   0.514770   4.346 1.38e-05 ***
ca2          3.271852   0.785123   4.167 3.08e-05 ***
ca3          2.188715   0.928644   2.357 0.018428 *
thal6       -0.168439   0.810310  -0.208 0.835331
thal7        1.433319   0.440567   3.253 0.001141 **
---
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 409.95  on 296  degrees of freedom
Residual deviance: 183.10  on 276  degrees of freedom
AIC: 225.1

Number of Fisher Scoring iterations: 6
```

By considering p-values of the output, we can conclude that the explanatory variables; sex, chest pain, resting blood pressure, number major vessels colored by fluoroscopy, slope of the peak exercise and short of thallium heart scan are significant at 5% significance level.

3.4.2 Model 2: Reducing variables from model 1

```
Call:
glm(formula = hd ~ sex + cp + trestbps + ca + slope + thal, family = "binomial",
    data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9626  -0.4746  -0.1170   0.3944   2.9125

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.907350   1.873754  -4.754 2.00e-06 ***
sexM         1.515713   0.494127   3.067 0.002159 **
cp2          1.109543   0.780235   1.422 0.155008
cp3          0.267203   0.698587   0.382 0.702097
cp4          2.647832   0.681891   3.883 0.000103 ***
trestbps     0.026831   0.010469   2.563 0.010382 *
ca1          2.207236   0.474859   4.648 3.35e-06 ***
ca2          3.066687   0.686580   4.467 7.95e-06 ***
ca3          2.288734   0.865330   2.645 0.008171 **
slope2       1.914643   0.430960   4.443 8.88e-06 ***
slope3       1.492420   0.714306   2.089 0.036678 *
thal6        0.002464   0.730261   0.003 0.997307
thal7        1.563587   0.414943   3.768 0.000164 ***
---
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 409.95  on 296  degrees of freedom
Residual deviance: 197.85  on 284  degrees of freedom
AIC: 223.85

Number of Fisher Scoring iterations: 6
```

There is no highly effect from interaction terms. Therefore, interaction terms are not taking into account to build the model 2. Thus, we can conclude that model 2 is the best fitted model for this data set. The summary of the model as follows,

$$\begin{aligned}\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = & -8.907350 + (1.515713)\text{Sex Male} \\ & + (2.647832)\text{ Chest pain asymptotic} + (0.026831)\text{resting blood pressure} \\ & + (2.207236)\text{ one major vessels colored by fluoroscopy} \\ & + (3.066687)\text{two major vessels colored by fluoroscopy} \\ & + (2.288734)\text{three major vessels colored by fluoroscopy} \\ & + (1.914643)\text{flat slope of peak exercise ST segment} \\ & + (1.492420)\text{down slope of peak exercise ST segment} \\ & + (1.563587)\text{reversible defect thallium heart scan}\end{aligned} \tag{3.1}$$

3.5 Logistic regression model assumption

In this section we check the key assumptions of logistic regression model that perform by model 2. The response variable is presence of heart disease which is binary. It satisfies the key assumption.

3.5.1 Observation's Independence

There is no systematic pattern in the following residual plot. It implies that observations are independent.

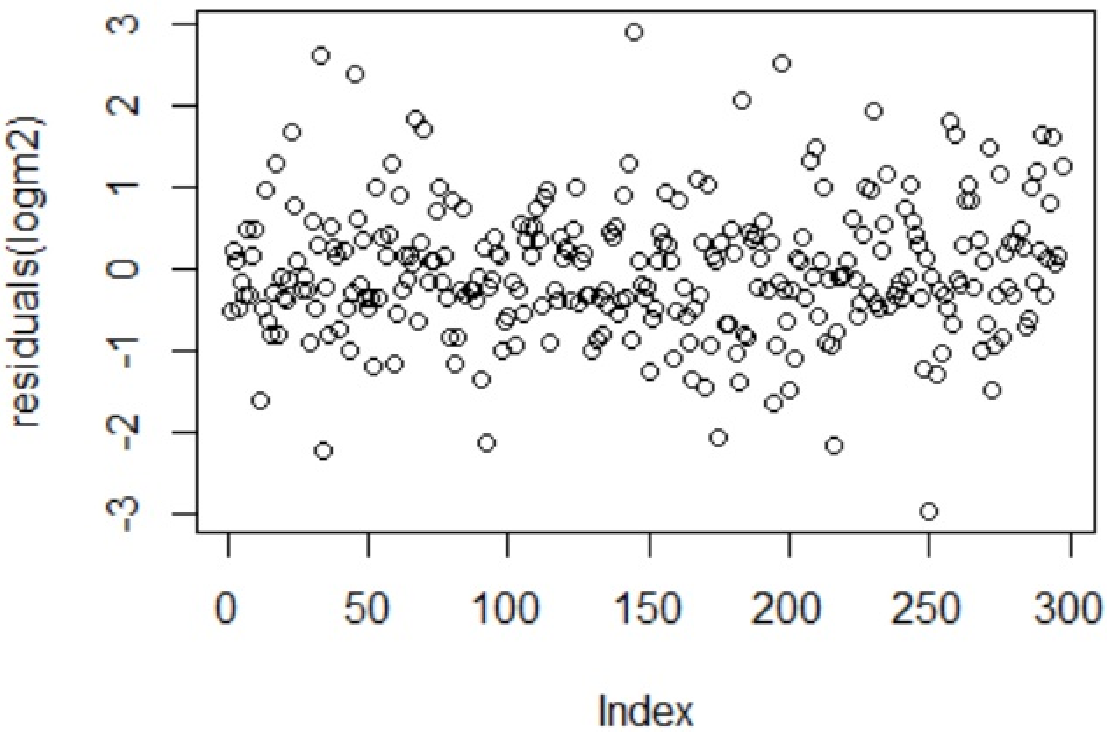

Figure 3.11: Residual Plot

3.5.2 Multicollinearity

To check the multicollinearity among the explanatory variables, use the VIF (Variation Inflation Factor) Test. As a rule of thumb, a VIF value that exceeds 4 or 10 indicates a problematic amount of collinearity. In our result, all the variables have the VIF value below 4 which is implies no multicollinearity.

```
> vif(logm2)
             GVIF Df GVIF^(1/(2*Df))
sex      1.479301  1        1.216265
cp       1.479235  3        1.067430
trestbps 1.128058  1        1.062101
ca       1.359681  3        1.052542
slope    1.366078  2        1.081107
thal     1.339684  2        1.075847
```

Figure 3.12: VIF Test

3.5.3 Extreme Outliers

Logistic regression assume that there are no extreme outliers or influential observation in the data set. The most common way to test this by calculating cook's distance which is,

$$D_i = \left(\frac{r_i^2}{p.MSE}\right)\left(\frac{h_{ii}}{(1-h_{ii})^2}\right) \tag{3.2}$$

r_i - i^{th} residual
p - number of coefficients in the regression model
MSE - mean squared error
h_{ii} - i^{th} leverage value

In this analysis, there are 3 extreme outliers, this is a 2% compared with the model data set. It can be ignored.

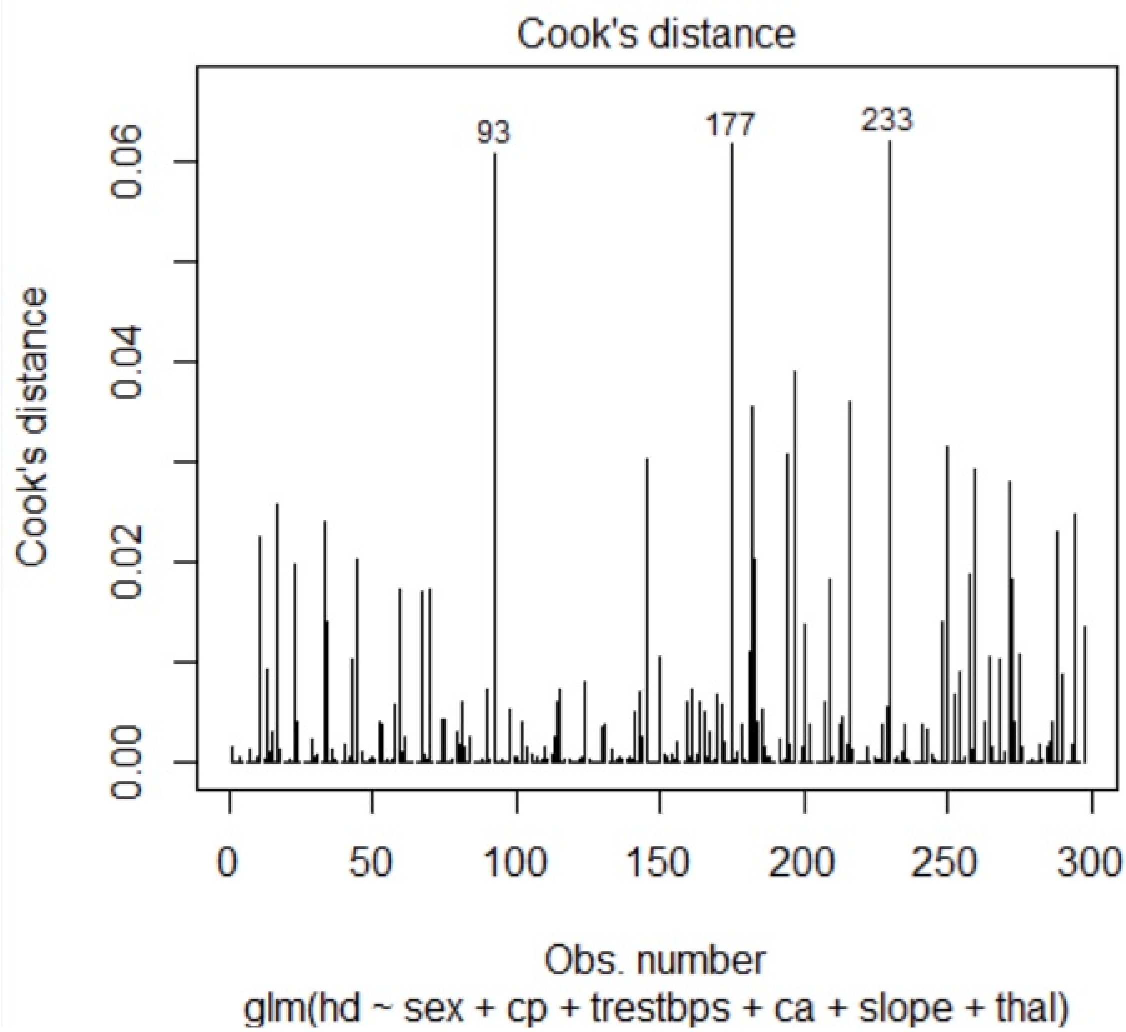

Figure 3.13: cook's distance

4 Conclusion

In conclusion, Some of the predictors used in this model are not controllable, but others may be used to dictate a person's lifestyle choices. The logistic regression model proved that, Sex, chest pain, resting blood pressure, number of major vessels (0-3) colored by fluoroscopy, slope of the peak exercise ST segment and short of thallium heart scan are significantly associated with presence of heart disease. Further, Male has higher proportion of having heart disease than female and Asymptotic chest pain type has the highest chance of presence of heart disease. Age, cholesterol, and fasting blood sugar were found to be insignificant variables in this data model. We can assume that these factors highly dependent in person's lifestyle and the genetic variation of the people selected for the analysis. This model showed a substantial influence of sex on the prediction of heart disease, so it may be interesting to investigate an interaction between symptoms and sex for sufferers of heart disease and can improve by taking into other factors such has smoking, obesity, sleep apnea, heart defects present at birth etc.

5 References

- Statistics for Epidemiology Nicholas P JewellBoca Raton Chapman Hall/CRC, 2004.
- Prediction of Coronary Heart Disease Using Risk Factor Categories, Peter W.F. Wilson, Ralph B. D'Agostino, Daniel Levy, Albert M. Belanger, Halit Silbershatz

- https://www.statology.org/assumptions-of-logistic-regression/
- http://archive.ics.uci.edu/ml/datasets/Heart+Diseases