

Analysis of the Risk Factors in Heart Disease

Li Sun

INTRODUCTION

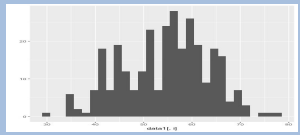
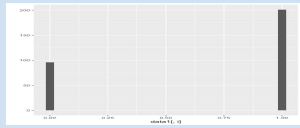
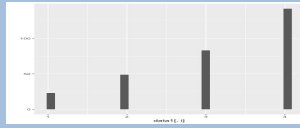
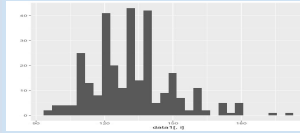
In this project, a heart disease dataset downloaded from <https://www.kaggle.com/ronitf/heart-disease-uci> is analyzed to identify the risk factors which have important effects on the presence of heart disease.

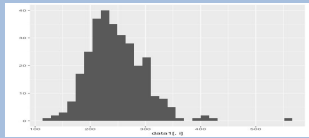
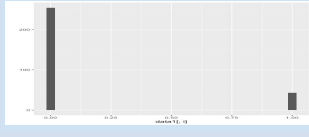
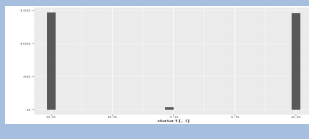
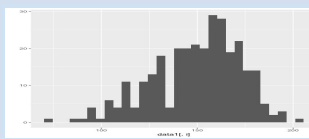
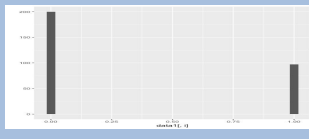
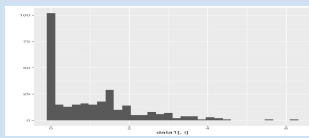
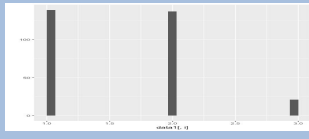
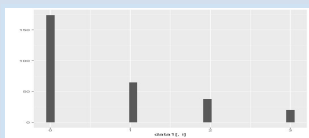
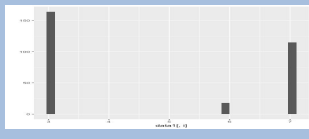
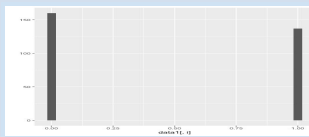
DATA SET DESCRIPTION

The dataset used in the research contains 297 observations with a total number of 14 attributes which are closely linked to heart disease. The data were observed from the patients with heart disease and normal patients. They are distributed from various ages and contain both genders. These 14 effects includes age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise induced angina, ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, number of major vessels colored, and Thallium stress test result. The values of variable target, 0 and 1, representing the absence or presence of heart disease respectively.

The 14 attributes are described in the table 1 below. The last attribute target is the one to be predicted. The listed information includes attribute, description, type and histogram plot.

Table 1: Attribute Information

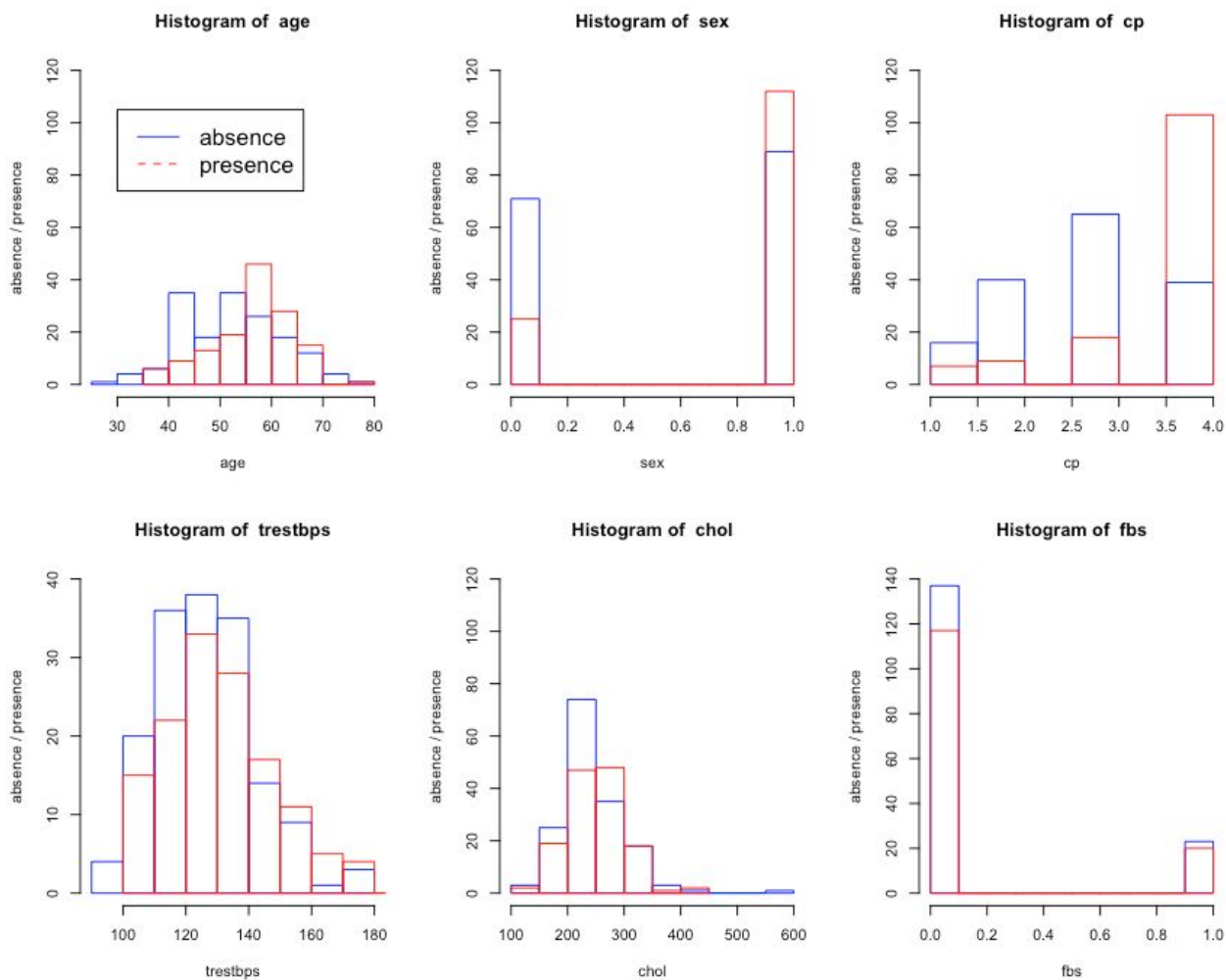
ATTRIBUTE	DESCRIPTION	TYPE	HISTOGRAM PLOT
age	age in years	INTEGER	
sex	sex	CATEGORICAL 0 = female 1 = male	
cp	chest pain type	CATEGORICAL Value 1: typical angina Value 2: atypical angina Value 3: non-anginal pain Value 4: asymptomatic	
restbps	resting blood pressure	CONTINUOUS	

chol	serum cholesterol in mg/d	CONTINUOUS	
fbss	fasting blood sugar > 120 mg/dl	CATEGORICAL 1 = true 0 = false	
restecg	resting electrocardiographic results	CATEGORICAL Value 0: normal Value 1: having ST-T wave abnormality Value 2: probable or definite left ventricular hypertrophy	
thalach	maximum heart rate achieved	CONTINUOUS	
exang	exercise induced angina	CATEGORICAL 1 = yes 0 = no	
oldpeak	ST depression induced by exercise relative to rest	CONTINUOUS	
slope	the slope of the peak exercise ST segment	CATEGORICAL Value 1: upsloping Value 2: flat Value 3: downsloping	
ca	number of major vessels (0-3) colored by fluoroscopy	INTEGER	
Thal	Thallium stress test result	CATEGORICAL 3 = normal 6 = fixed defect 7 = reversible defect	
target	diagnosis of heart disease 0 absence, 1 presence	INTEGER	

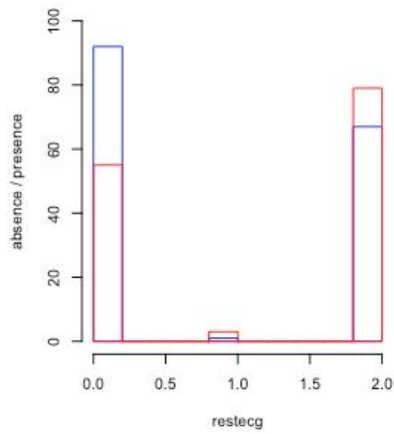
DATA ANALYSIS

In order to obtain inspection about distribution of the attributes over the corresponding subsamples (160 patients without disease, 139 patients present disease), histogram and bar charts of 14 attributes in the heart disease data are plotted in Figure 1 below. In the figure, corresponding to each attribute, the number of patients with absence of heart disease are shown in blue, and the number of patients with presence of heart disease are shown in red.

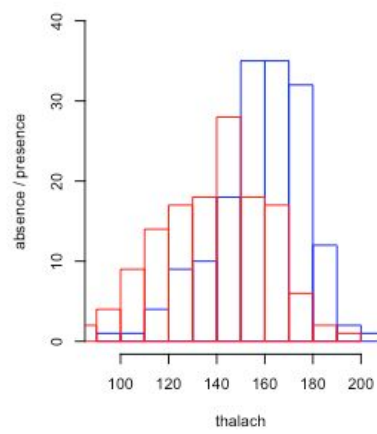
Figure 1: Plots of Heart Disease Data



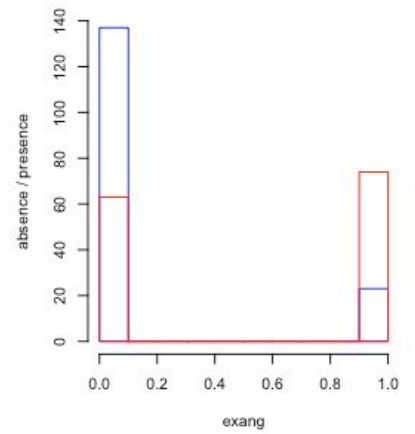
Histogram of restecg



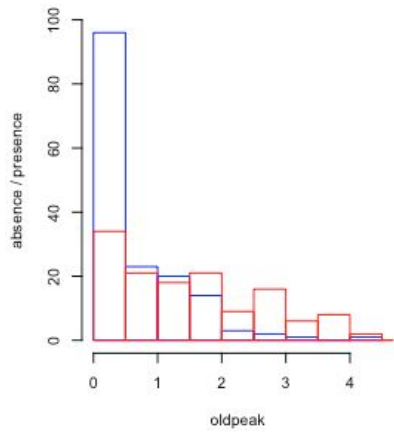
Histogram of thalach



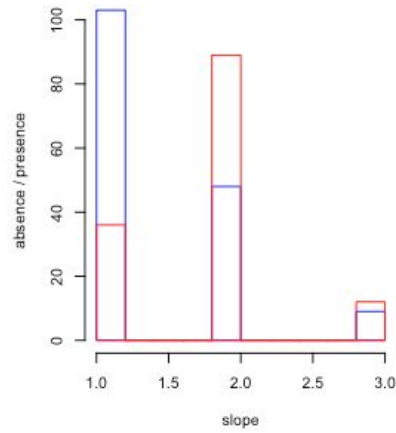
Histogram of exang



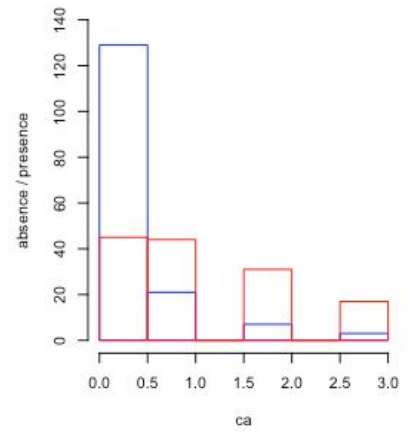
Histogram of oldpeak



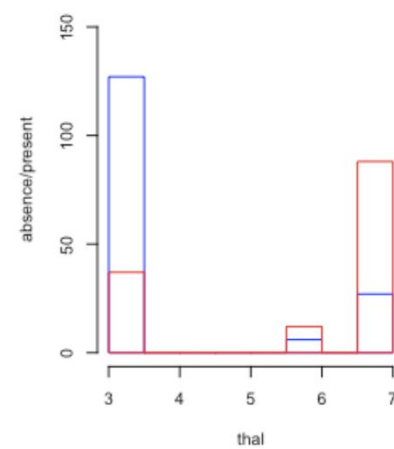
Histogram of slope



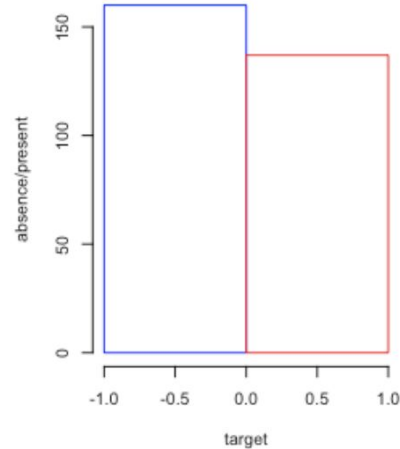
Histogram of ca



Histogram of thal



Histogram of target



These plots show that comparing with the people without the disease, those who are older (*age*) and male (*sex* = 1), with asymptomatic chest pain (*cp* = 4), higher blood pressure (*trestbps*), higher cholesterol level (*chol*) and lower maximum heart rate (*thalach*), with higher chances of exercises induced angina (*exang* = 1), higher ST depression induced by exercise (*oldpeak*), the flat or down slope of the peak exercise ST segment (*slope* = 2,3), with colored major vessels (*ca*) and thallium stress test with fixed or reversible defect (*thal* = 6, 7), are likely to suffer from heart disease.

In this project, a logistic regression method is used to determine the final model. The final model is a logistic regression model with interaction between continuous variables and categorical variables. It is represented as follows:

Logit P(target = presence) =

$$\begin{aligned}
 & -3.392 - 1.427 * ca - 0.735 * oldpeak + 0.027 * trestbps - 0.034 * thalach \\
 & + 1.655 * \chi_{male}^{sex} - 6.149 * \chi_{fixed\ defect}^{thal} + 0.833 * \chi_{reversible\ defect}^{thal} \\
 & + 1.664 * \chi_{atypical\ angina}^{cp} + 0.444 * \chi_{non-angina\ pain}^{cp} + 3.070 * \chi_{asymptomatic}^{cp} \\
 & - 3.455 * \chi_{flat}^{slope} + 14.791 * \chi_{downsloping}^{slope} \\
 & + 0.0186 * ca * thalach \\
 & + 20.183 * \chi_{fixed\ defect}^{thal} * ca - 0.756 * \chi_{reversible\ defect}^{thal} * ca \\
 & + 0.029 * \chi_{flat}^{slope} * trestbps - 0.124 * \chi_{downsloping}^{slope} * trestbps \\
 & + 2.508 * \chi_{fixed\ defect}^{thal} * oldpeak + 1.266 * \chi_{reversible\ defect}^{thal} * oldpeak \\
 & + 1.362 * \chi_{flat}^{slope} * oldpeak + 1.574 * \chi_{downsloping}^{slope} * oldpeak
 \end{aligned}$$

DATA INTERPRETATION

Based on the final model, we find that 8 out of 13 variables have important effects on the risk of heart disease. They are *ca*, *oldpeak*, *thresbps*, *thalach*, *thal*, *sex*, *cp* and *slope*.

First, I will interpret the final model without considering interaction.

Among these main factors, the risk of heart disease decreases while the number of major vessels colored by fluoroscopy (*ca*), ST depression induced by exercise (*oldpeak*), or maximum heart rate (*thalach*) increases. If the number (≤ 4) of the major vessels colored by fluoroscopy increases by 1, the odds of risk of heart disease will decrease by a factor $\exp(1.427) = 4.17$, that is a 317% decrease. For 1 unit increase in ST depression induced by exercise, the odds of risk of heart disease will decrease by a factor $\exp(0.735) = 2.09$, that is a 109% decrease. For 1 beat rate increase in maximum heart beat rate, the odds of risk of heart disease will decrease by a factor $\exp(0.034) = 1.034$, that is a 3.4% decrease. On the contrary, the risk of heart disease increases while resting blood pressure (*trestbp*) increases. If the resting blood pressure increases 1 mmHg, the odds of risk of heart disease will increase by a factor $\exp(0.027) = 1.027$, that is a 2.7% increase.

The data in our final model tells that males (*sex*) are more likely to get heart disease than females. For thallium stress test result (*thal*), the person with reversible defect has the higher risk of heart disease compared with the person with the result of fixed defect or normal. If a person has chest pain that is type asymptomatic (*cp*), he or she has the highest chance of suffering from heart disease in the chest pain group. The result that the slope of the peak exercise ST segment (*slope*) is downsloping is a sign of having a higher risk of heart disease than the result of up-slope, and flat-sloping has a lower risk than upsloping.

When we take the interaction between variables into account, the intercorrelation among multiple attributes are analyzed here as well.

The interaction model represents that the variables *ca* (*number of major vessels colored by fluoroscopy*) and *thalach* (*maximum heart rate*) are correlated. So the change of odds ratio of *ca* depends on the value of *thalach*, and vice versa. First let us look at the variable *ca*. With one unit increase in the number of major vessels colored by fluoroscopy, the odds ratio of *ca* will transit from decrease to increase when maximum heart rate is greater than 76. Since the range of *thalach* is from 71 to 202, the maximum heart rate for the majority of the people is greater than 76. In this phenomenon, for people with maximum heart rate greater than 76, the more the number of major vessels colored by fluoroscopy, the higher risk of heart disease. Secondly we focus on the variable *thalach*. The value of *ca* will affect the odds ratio of *thalach* as well. With one unit increase of maximum heart rate, the odds of *thalach* will decrease if the number of major vessels colored by fluoroscopy is 0 or 1, and the odds of *thalach* will change to increase if the number is 2 or 3.

Variable *ca* (*number of major vessels colored by fluoroscopy*) correlates with *thal* (*thallium stress test result*). From the discussion previously, if we only consider variable *ca* alone, we know the odds of risk of heart disease will decrease by 317% with 1 major vessels increase. But in practice, variable *thal* has an influence on variable *ca*, and the risk of heart disease will change. Moreover, two thallium stress tests, fixed defect and reversible defect, affect an opposite way. With 1 major vessel increase in vessels colored by fluoroscopy, and with a fixed defect test result, the odds of risk of heart disease will increase, instead of decrease, dramatically by a factor $1.4e^8$. In contrast, if the test result is reversible defect, the odds of risk of heart disease will decrease more than doubling, that is 787%.

Variables between *slope* (*slope of the peak exercise ST segment*) and *trestbps* (*resting blood pressure*) are correlated. If a person has a flat slope of peak exercise ST segment, the risk of heart disease will increase more while resting blood pressure increases, compared with the situation considering *trestbps* alone. In this situation, if this person's resting blood pressure increases by 1 mmHg, the odds of risk of heart disease will increase 5.8%, instead of 2.8%. In contrast, If a person has a down slope of the peak exercise ST segment, the odds of risk of heart disease will decrease while his/her resting blood pressure increases, which is the opposite of the situation considering *trestbps* alone. If the person's resting blood pressure increases by 1 mmHg, the odds of risk of heart disease will decrease by 9.2%, instead of increasing by 2.8%.

Variable *oldpeak* (*ST depression induced by exercise*) correlates with *thal* (*thallium stress test result*) and *slope* (*the slope of the peak exercise ST segment*). If we only consider *oldpeak* alone, the risk of heart disease decreases while ST depression induced by exercise increases. However, with *thal* and *slope* involved, the odds ratio of *oldpeak* depends on the value of *thal* or *slope*, rather than the value of itself. As a result, the risk of heart disease will increase while ST depression induced by exercise increases for a person with a fixed defect or reversible defect of thallium stress test result, or with flat slope or downsloping slope of the peak exercise ST segment. With one unit increase of ST depression induced by exercise, their odds of risk of heart disease will increase 2461%, 70%, 87.2%

or 131.4%, respectively.

CONCLUSION

In the paper, a heart disease data set is used to identify the risk factors of the presence of heart disease. After model selection, a logistic regression model is applied. From the analysis, It is concluded that the major risk factors are ca, oldpeak, thresbps, thalach, thal, sex, cp and slope. Further, how these factors affect the risk of heart is discussed as well.

APPENDIX

```
> ### Data1
> data1 = read.csv('~/.GSU/Categorical Analysis/Project1/project.1.data.1.csv', header = T)
> str(data1)
'data.frame':   297 obs. of  14 variables:
 $ age      : int  63 67 67 37 41 56 62 57 63 53 ...
 $ sex      : int  1 1 1 1 0 1 0 0 1 1 ...
 $ cp       : int  1 4 4 3 2 2 4 4 4 4 ...
 $ trestbps : int  145 160 120 130 130 120 140 120 130 140 ...
 $ chol     : int  233 286 229 250 204 236 268 354 254 203 ...
 $ fbs      : int  1 0 0 0 0 0 0 0 0 1 ...
 $ restecg  : int  2 2 2 0 2 0 2 0 2 2 ...
 $ thalach  : int  150 108 129 187 172 178 160 163 147 155 ...
 $ exang    : int  0 1 1 0 0 0 0 1 0 1 ...
 $ oldpeak  : num  2.3 1.5 2.6 3.5 1.4 0.8 3.6 0.6 1.4 3.1 ...
 $ slope    : int  3 2 2 3 1 1 3 1 2 3 ...
 $ ca       : int  0 3 2 0 0 0 2 0 1 0 ...
 $ thal     : int  6 3 7 3 3 3 3 3 7 7 ...
 $ target   : int  0 1 1 0 0 0 1 0 1 1 ...
>
> sum(is.na(data1)) # check if there are missing values
[1] 0
> ## print the histogram of each variables separately
> library(ggplot2)
> for (i in (1:14))
+ {print(qplot(data1[,i], geom = 'histogram', xlab = names(data1[i])))}
>
> ## plot histogram of subsamples
> op = par(mfrow = c(2,3))
> for(i in 1:6)
+ {if(i == 6) {yy = 140} else{if (i== 4) {yy = 40} else {yy =120}}
+   hist(data1[,i][data1$target==0], border = 'blue', main =
+     paste('Histogram of ', names(data1[i])), xlab = names(data1[i]),
+     ylab = 'absence / presence',ylim = c(0,yy))
+   hist(data1[,i][data1$target>0], border = 'red',
+     ylab = 'presence', add=T)
+   if (i == 1)
```

```

+   {legend(30,105, legend=c("absence", "presence"),
+         col=c("blue", "red"), lty=1:1.5, cex=0.9)
+   }
+ }
>
> op = par(mfrow = c(2,3))
> for(i in 7:12)
+ {if(i==9 | i == 12) {yy = 140} else{if (i == 8) {yy = 40} else {yy =100}}
+   hist(data1[,i][data1$target==0], border = 'blue', main =
+     paste('Histogram of ', names(data1[i])), xlab = names(data1[i]),
+     ylab = 'absence / presence',ylim = c(0,yy))
+   hist(data1[,i][data1$target>0], border = 'red',
+     ylab = 'presence', add=T)
+ }
>
> op = par(mfrow = c(2,3))
> for(i in 13:14)
+ { if (i == 14) {xx = c(-1,1)} else {xx = c(3,7)}
+   hist(data1[,i][data1$target==0], border = 'blue', main =
+     paste('Histogram of ', names(data1[i])), xlab = names(data1[i]),
+     ylab = 'absence/present', xlim = xx, ylim = c(0,160))
+   hist(data1[,i][data1$target>0], border = 'red',
+     ylab = 'presence', add=T)
+ }
>
>
> ## Data processing
> data1$sex[data1$sex == 0] = 'Female'
> data1$sex[data1$sex == 1] = 'Male'
> data1$sex = as.factor(data1$sex)
>
> data1$cp[data1$cp == 1] = 'Typical Angina'
> data1$cp[data1$cp == 2] = 'Atypical Angina'
> data1$cp[data1$cp == 3] = 'Non-Angina pain'
> data1$cp[data1$cp == 4] = 'Asymptomatic'
> data1$cp = as.factor(data1$cp)
>
> data1$fbs[data1$fbs == 0] = 'False'
> data1$fbs[data1$fbs == 1] = 'True'
> data1$fbs = as.factor(data1$fbs)
>
> data1$restecg[data1$restecg == 0] = 'Normal'
> data1$restecg[data1$restecg == 1] = 'ST-T Abnormal'
> data1$restecg[data1$restecg == 2] = 'LV Hypertrophy'
> data1$restecg = as.factor(data1$restecg)
>
> data1$exang[data1$exang == 0] = 'No'
> data1$exang[data1$exang == 1] = 'Yes'
> data1$exang = as.factor(data1$exang)
>
> data1$slope[data1$slope == 1] = 'Up'
> data1$slope[data1$slope == 2] = 'Flat'
> data1$slope[data1$slope == 3] = 'Down'
> data1$slope = as.factor(data1$slope)
>
> data1$thal[data1$thal == 3] = 'Normal'

```



```

> data1$thal[data1$thal == 6] = 'Fixed defect'
> data1$thal[data1$thal == 7] = 'Reversible defect'
> data1$thal = as.factor(data1$thal)
>
> # Relevel reference level for each categorical variable
> data1$sex = relevel(data1$sex, 'Female')
> data1$cp = relevel(data1$cp, 'Typical Angina')
> data1$fbs = relevel(data1$fbs, 'False')
> data1$restecg = relevel(data1$restecg, 'Normal')
> data1$exang = relevel(data1$exang, 'No')
> data1$slope = relevel(data1$slope, 'Up')
> data1$thal = relevel(data1$thal, 'Normal')
>
> ## model selection using forward method
> fit.full = glm(target~., data = data1, family = binomial)
> fit.null = glm(target~1, data = data1, family = binomial)
> select1.1 = step(fit.null, scope = list(lower = fit.null, upper = fit.full),
+               direction = 'forward')
> .....
>
> # Step: AIC=223.98
> # target ~ thal + ca + cp + oldpeak + slope + sex + trestbps +
> #   exang + thalach
> #
> # Df Deviance    AIC
> # <none>      195.98 223.98
> # + chol      1  194.47 224.47
> # + fbs       1  195.13 225.13
> # + age       1  195.85 225.85
> # + restecg   2  193.96 225.96
>
>
> ## generate formula with interaction
> data1.1 = (select1.1$model)
>
> pred = colnames(data1.1)[-1]
> m = length(pred)
> pred.names = pred
>
> for (i in 1:(m-1)){
+   for (j in (i+1):m) {
+     pred.names = c(pred.names, paste(pred[i], ':', pred[j]))
+   }
+ }
>
> Formula = formula(paste('target ~ ', paste(pred.names, collapse = '+')))
>
> ## interaction model selection using forward method
> fit.full.1 = glm(Formula, data = data1.1, family = binomial)
> fit.null.1 = glm(target ~ 1, data = data1.1, family = binomial )
> select1.2 = step(fit.null.1, scope = list(lower=fit.null.1, upper = fit.full.1), direction = 'forward')

Start: AIC=411.95
target ~ 1
.....
Step: AIC=213.73

```

```
target ~ thal + ca + cp + oldpeak + slope + sex + trestbps +
  thalach + thal:ca + slope:trestbps + thal:oldpeak + oldpeak:slope +
  ca:thalach
```

	Df	Deviance	AIC
<none>		169.73	213.73
+ ca:oldpeak	1	167.88	213.88
+ sex:trestbps	1	167.90	213.90
+ slope:sex	2	166.96	214.96
+ exang	1	169.06	215.06
+ thal:sex	2	167.16	215.16
+ cp:slope	6	159.22	215.22
+ sex:thalach	1	169.30	215.30
+ ca:trestbps	1	169.40	215.40
+ thal:thalach	2	167.40	215.40
+ ca:slope	2	167.40	215.40
+ ca:sex	1	169.42	215.42
+ oldpeak:trestbps	1	169.53	215.53
+ oldpeak:thalach	1	169.67	215.67
+ trestbps:thalach	1	169.68	215.68
+ oldpeak:sex	1	169.72	215.72
+ cp:oldpeak	3	165.85	215.85
+ cp:thalach	3	166.09	216.09
+ ca:cp	3	166.12	216.12
+ cp:trestbps	3	166.87	216.87
+ slope:thalach	2	169.12	217.12
+ thal:trestbps	2	169.66	217.66
+ cp:sex	3	169.17	219.17
+ thal:slope	4	167.72	219.72
+ thal:cp	6	165.05	221.05

There were 50 or more warnings (use warnings() to see the first 50)

```
>
> ## final logistic regression model
> fit.final = glm(target ~ thal + ca + cp + oldpeak + slope + sex + trestbps +
+   thalach + thal:ca + slope:trestbps + thal:oldpeak + oldpeak:slope +
+   ca:thalach, data = data1, family = binomial)
>
> summary(fit.final)
```

Call:

```
glm(formula = target ~ thal + ca + cp + oldpeak + slope + sex +
  trestbps + thalach + thal:ca + slope:trestbps + thal:oldpeak +
  oldpeak:slope + ca:thalach, family = binomial, data = data1)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.58848	-0.41275	-0.08489	0.27507	2.69529

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-3.39241	3.04882	-1.113	0.265839
thalFixed defect	-6.14884	5.81670	-1.057	0.290465
thalReversible defect	0.83283	0.62916	1.324	0.185597
ca	-1.42748	1.89726	-0.752	0.451815
cpAsymptomatic	3.06995	0.83403	3.681	0.000232 ***
cpAtypical Angina	1.66399	0.91469	1.819	0.068882 .

cpNon-Angina pain	0.44442	0.81160	0.548	0.583977
oldpeak	-0.73481	0.55965	-1.313	0.189191
slopeDown	14.79102	6.31005	2.344	0.019076 *
slopeFlat	-3.45546	3.13347	-1.103	0.270133
sexMale	1.65494	0.54560	3.033	0.002419 **
trestbps	0.02694	0.01514	1.780	0.075144 .
thalach	-0.03436	0.01419	-2.421	0.015460 *
thalFixed defect:ca	20.18278	949.50011	0.021	0.983041
thalReversible defect:ca	-0.75612	0.53565	-1.412	0.158072
slopeDown:trestbps	-0.12392	0.05614	-2.207	0.027295 *
slopeFlat:trestbps	0.02866	0.02310	1.241	0.214774
thalFixed defect:oldpeak	2.50750	2.92706	0.857	0.391632
thalReversible defect:oldpeak	1.26651	0.53649	2.361	0.018239 *
oldpeak:slopeDown	1.57370	1.15494	1.363	0.173013
oldpeak:slopeFlat	1.36201	0.61818	2.203	0.027577 *
ca:thalach	0.01863	0.01238	1.504	0.132483

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: 169.73 on 275 degrees of freedom
 AIC: 213.73

Number of Fisher Scoring iterations: 17

```
>
>
> coef(fit.final)
```

(Intercept)	thalFixed defect
-3.39240904	-6.14884037
thalReversible defect	ca
0.83283464	-1.42748231
cpAsymptomatic	cpAtypical Angina
3.06994689	1.66398718
cpNon-Angina pain	oldpeak
0.44442205	-0.73480834
slopeDown	slopeFlat
14.79102276	-3.45545726
sexMale	trestbps
1.65494196	0.02694373
thalach	thalFixed defect:ca
-0.03435759	20.18277870
thalReversible defect:ca	slopeDown:trestbps
-0.75611840	-0.12392157
slopeFlat:trestbps	thalFixed defect:oldpeak
0.02866177	2.50750261
thalReversible defect:oldpeak	oldpeak:slopeDown
1.26651496	1.57370447
oldpeak:slopeFlat	ca:thalach
1.36200638	0.01862944