Heart Disease Risk Factor Exploration Using Logistic Regression

J Kou

1. Objective

Heart disease is the leading cause of death around the world. It is recommended to prevent heart disease early in life. There are many factors may cause heart disease. The sooner we manage the risk factors the higher the chance of leading a heart-healthy life. Thus, identifying factors that increase a person's risk for heart disease is crucial. The main purpose of this project intends to figure out the relevant risk factors of heart disease using logistic regression.

2. Data

2.1 Data source

The dataset was downloaded from Kaggle (https://www.kaggle.com/ronitf/heart-disease-uci). This dataset provides 297 patients' information. The names and social security numbers of the patients were removed from the database, replaced with dummy values.

2.2 Variables

There are 14 variables in the dataset, the "target" refers to the presence of heart disease in the patients where "1" indicates presence and "0" indicates absent of heart disease. Other variables are the predictors to be used in this project, six of them are continues variables which are age, trestbps (resting blood pressure (in mm Hg on admission to the hospital)), chol (serum cholestoral in mg/dl), thalach (maximum heart rate achieved), oldpeak (ST depression induced by exercise relative to rest) and ca (number of major vessels colored by fluoroscopy), seven are categorical variables which are sex, cp (chest pain type), fbs (fasting blood sugar), restecg (resting electrocardiographic results), exang (exercise induced angina), slope (the slope of the peak exercise ST segment), and thal (Thalassemia). More detailed information of categorial variables is shown in table 1.

Table 1. Description of Categorial Variables

name	description	levels

target	Heart disease	1=presence; 0=absence
sex	gender	1 = male; 0 = female
cp	chest pain type	1=typical angina; 2=atypical angina; 3=non-anginal pain; 4=asymptomatic
fbs	fasting blood sugar > 120 mg/dl	1 = true; 0 = false
restecg	resting electrocardiographic	0=normal; 1=having ST-T wave abnormality; 2=showing probable or definite left ventricular hypertrophy
exang	exercise induced angina	1 = yes; 0 = no
slope	the slope of the peak exercise ST segment	1=upsloping; 2=flat; 3=down sloping
thal	Thalassemia	3 = normal; 6 = fixed defect; 7 = reversible defect

3. Methods

3.1 Logistic regression model selection

Logistic regression (LR) method was used in this project to predict the relationship between the presence of heart disease and predictionary factors. Firstly, the preliminary model was selected using stepwise method. Then Anova test was performed on the selected model to test the significance of each predictionary variable. Any variables with p-value >0.05 were removed in order to get the final logistic model.

3.2 Model evaluation

The fitted logistic regression model was evaluated by plotting ROC curve (the line of true positive rate v.s. false positive rate) using ten-fold cross validation method. And the area under ROC curve (AUC) was also calculated to measure classification accuracy.

3.3 Model interpretation (log odds ratio and odds ratio)

Odds Ratio is a measure of association between exposure and an outcome. It represents the odds that an outcome will occur given a particular condition, compared to the odds of the outcome occurring in the absence of that condition. The odds ratio of different situations are calculated as following methods.

3.3.1 log odds ratio of single variable

For variable selected in the final model works independently with other variables. The log odds ratio will be the coefficient of that variable.

3.3.2 log odds ratio of interaction between two binary variables

For a binary categorial variable Xi interact with another binary categorical variable Yi:

$$logit\,(P) = \alpha + \beta 1 Xi + \beta 2 Y i + \gamma Xi Y i$$

the log odds ratio is calculated as table 2.

Table 2. interpretation of interaction between two binary variables

	X=0	X=1	Log odds ratio
Y=0	α	$\alpha+\beta_1$	β_1
Y=1	$\alpha+\beta_2$	$\alpha+\beta_1+\beta_2+\gamma$	$\beta_1+\gamma$
Log odds ratio	β_2	$\beta_2+\gamma$	

3.3.3 log odds ratio of interaction between continuous and categorical variables

For a continuous variable X interact with a binary variable Yi:

$$logit\,(P) = \alpha + \beta 1X + \beta 2Yi + \gamma XYi$$

The log odds ratio is calculated by the following formula:

$$log(odds\ ratio) = log\,(odds(P = 1|x = x0 + 1,\ = false,\ y = k))/(odds(P = 1|\,x = x0,\ y = k)) = \beta 1 + \gamma$$

3.3.4 odds ratio

Odds ratio is calculated using the formula:

$$odds\ ratio = e^{log(odds\ rratio)}$$

3.4 Software

All the methods were performed using R-studio.

4. Results

4.1 Data cleaning of categorial variables of original dataset

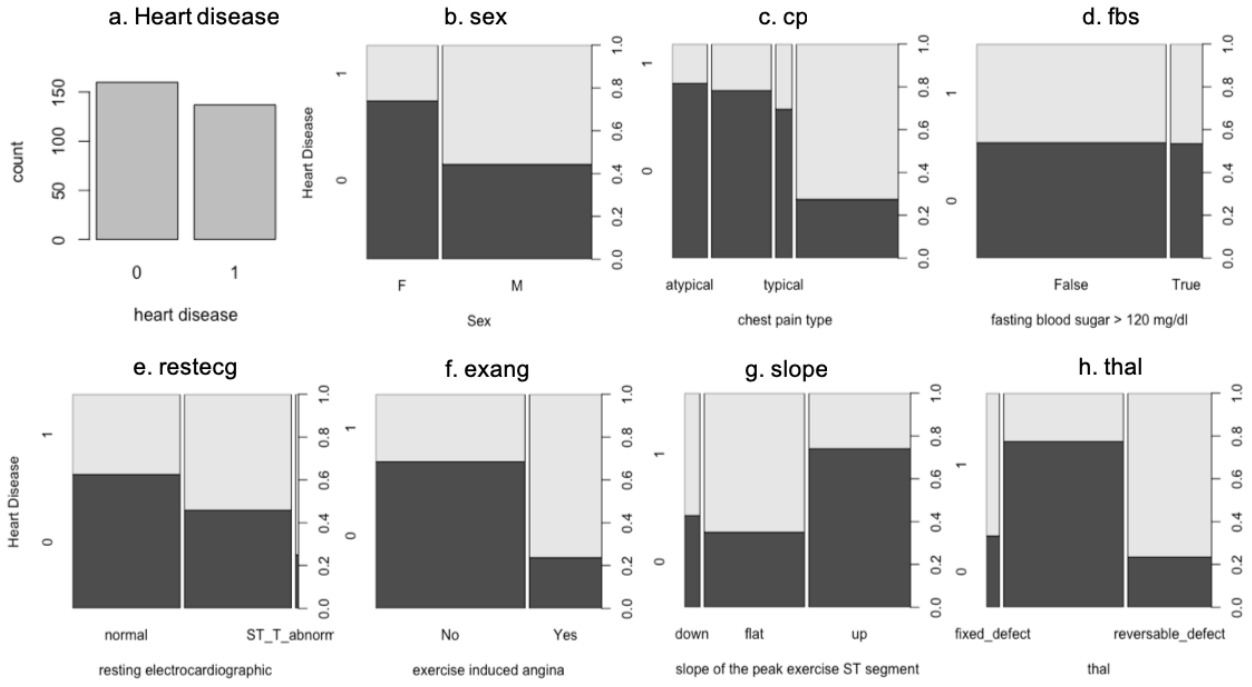

Figure 1. Plotting of categorical variables of original dataset.

It is not proper to fit logistic regression with some variables which have limited observation in some level, thus the first step for logistic regression is to eliminate this problem. According to figure 1, variable cp has little observations in “typical angina” level, thus “typical angina” and “atypical angina” were combined together as “angina” for future analysis. The variable restecg has few observations at level “ST-T abnormality”, thus “ST-T abnormality” and “probable or definite left hypertrophy” were combined as “abnormal” for the following logistic regression model selection. Similarly, the “down” of slope was combined with “flat” as “not up”, and “fixed defect” and “reversible defect” of thal were combined as “defect” for future analysis. The updated dataset was plotted as shown in Figure 2.

4.2 Data exploration after data cleaning

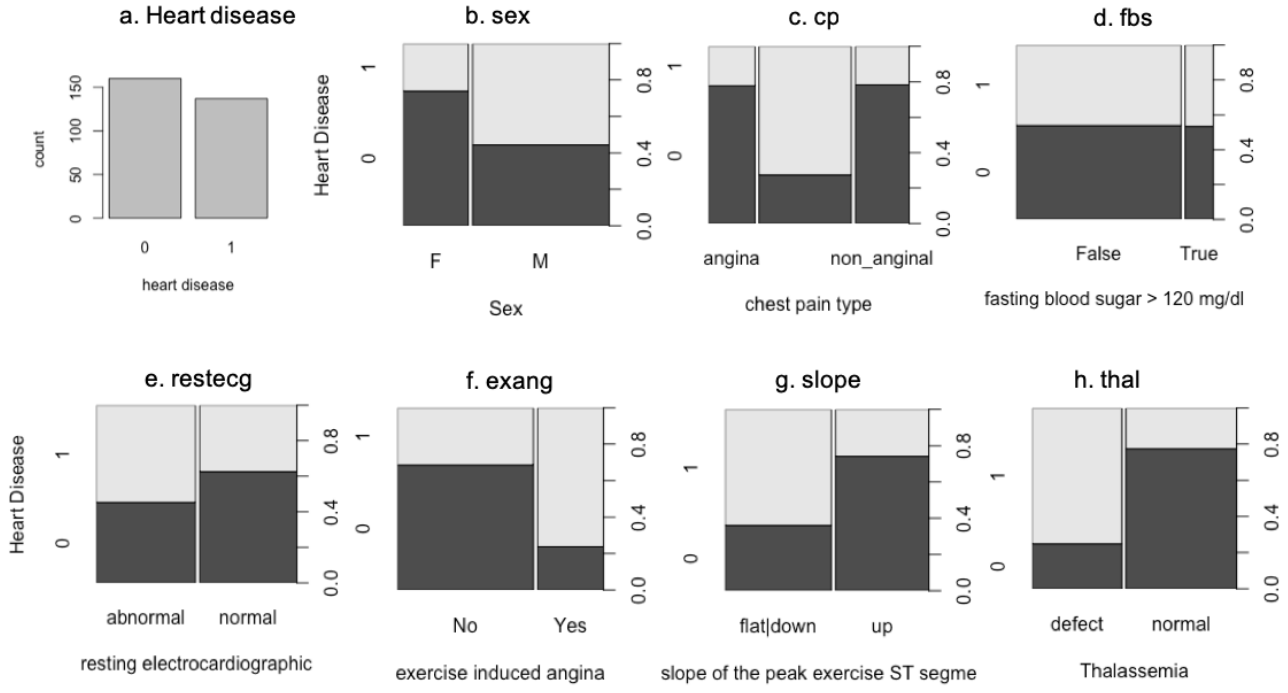

Figure 2a. Categorical variable v.s heart disease

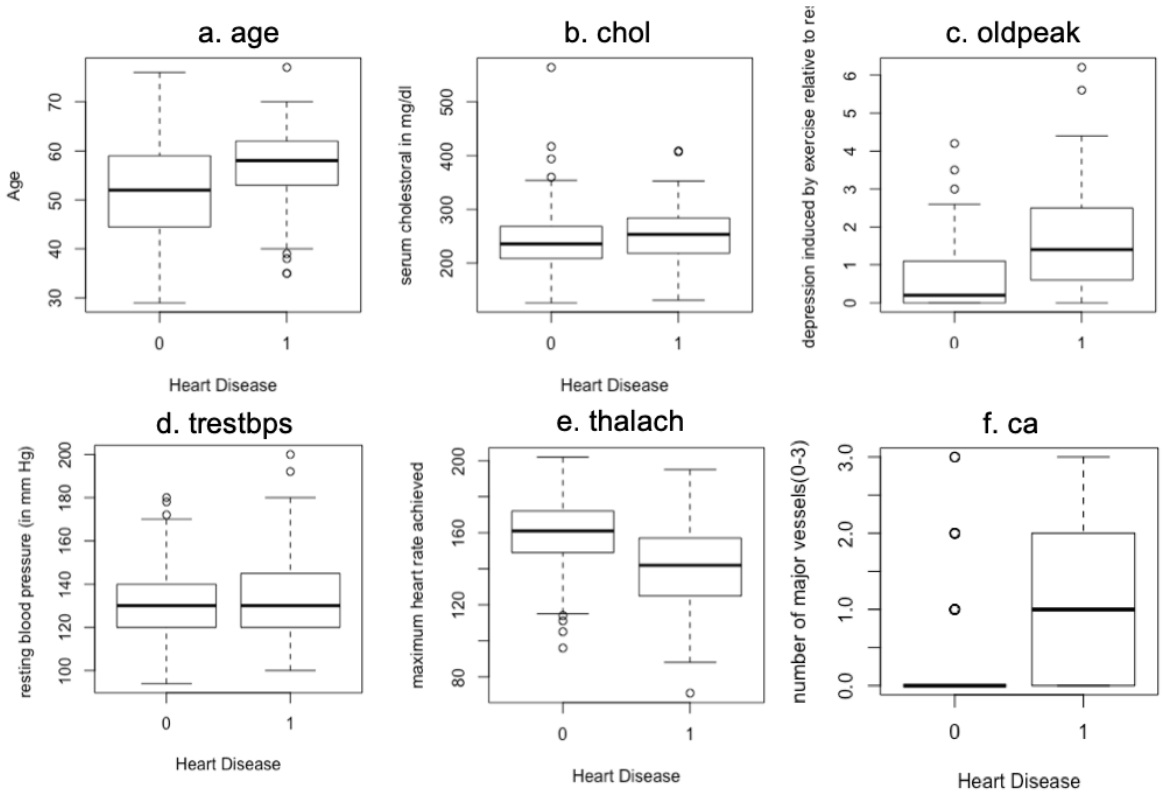

Figure 2b. Continuous variables v.s. heart disease

As shown in figure 2a, this dataset of patients with and without heart disease is almost the same. And after data cleaning most of the variables have balanced observations in each level of effect. Figure 2b shows the relationship of each continuous variable with heart disease. Most of the continuous variables have balanced observations in the present or absent of heart disease except variable ca. However, it seems patients without heart disease have fewer number of ca compare to patients with heart disease, which may indicate ca works significantly in predicting of heart disease. Thus, ca was not removed at the beginning of the model selection.

4.3 Logistic regression model selection

Logistic regression is the method for predicting a binary outcome with one or more predictionary variables. In this project, it measures the probability of the presence or absence of heart disease using a logistic function. Due to large numbers of candidate risk factors in the dataset, it is time consuming to consider the interaction between variables at the beginning of the model selection. Thus, it is a good strategy to figure out the risk factors without considering interactions first, then take account interactive effects among these selected risk factors.

4.3.1 Model without interaction

The forward method was performed to do major risk factor selection. The starting model is target~1 (all these factors have no influence on the presence of heart disease), and the full model is target~. (contains all the factors). The selection process was terminated when the current model reached the minimum AIC (Akaike information criterion, the smaller the AIC, the better the model). Model.1 (AIC = 224.6) selected without interaction factors is:

Model 1:

$$Target \sim sex + cp + trestbps + chol + fbs + thalach + exang + slope + ca + thal$$

4.3.2 Model with interaction

The forward method was performed to select the model with interaction items. The starting model is model.1, the full model is the one which contains all the interactions between each pair of variables: target~(model.1)^2. Model.2 (AIC: 214.6) selected containing interactions is:

Model 2:

$$target \sim sex + cp + trestbps + chol + fbs + thalach + exang + slope + ca + thal + exang : slope + sex : fbs$$

4.3.3 Final model

Anova test was performed on model.2 to test the significance of each predictor variable. According to Anova table (see appendix 1), among all the major risk factors chol and thalach with p-value bigger than 0.05. Thus, chol, thalach and related interaction items were removed in the final model (AIC=215.36):

Final model:

$$target \sim sex + cp + trestbps + fbs + exang + slope + ca + thal + exang : slope + sex : fbs + fbs : slope + ex$$

Final model with coefficients:

$$logitP\ (heart\ disease = present) = -1.96 - 0.01 trestbps + 1.25 ca + 0.14 \chi_{male}^{sex} - 0.14 \chi_{non-anginal\ pain}^{cp} + 1.43 \chi_{asymptomatic}^{cp} + 6$$

4.4 Model evaluation

ROC (Receiver Operator Characteristic) is plotted between True Positive Rate (Y axis) and False Positive Rate (X Axis). A strong model should strive to attain the highest possible true positive rate while keeping the false positive rate relatively low, a model dose so will

have a high arch and thus a high area under ROC curve (AUC). Higher the AUC, better the model, which means greater classification accuracy on the training dataset. A model with random prediction results in an area of 50%, as shown in figure 3, this model has an area 91% which means this model performs good when predicting heart disease.

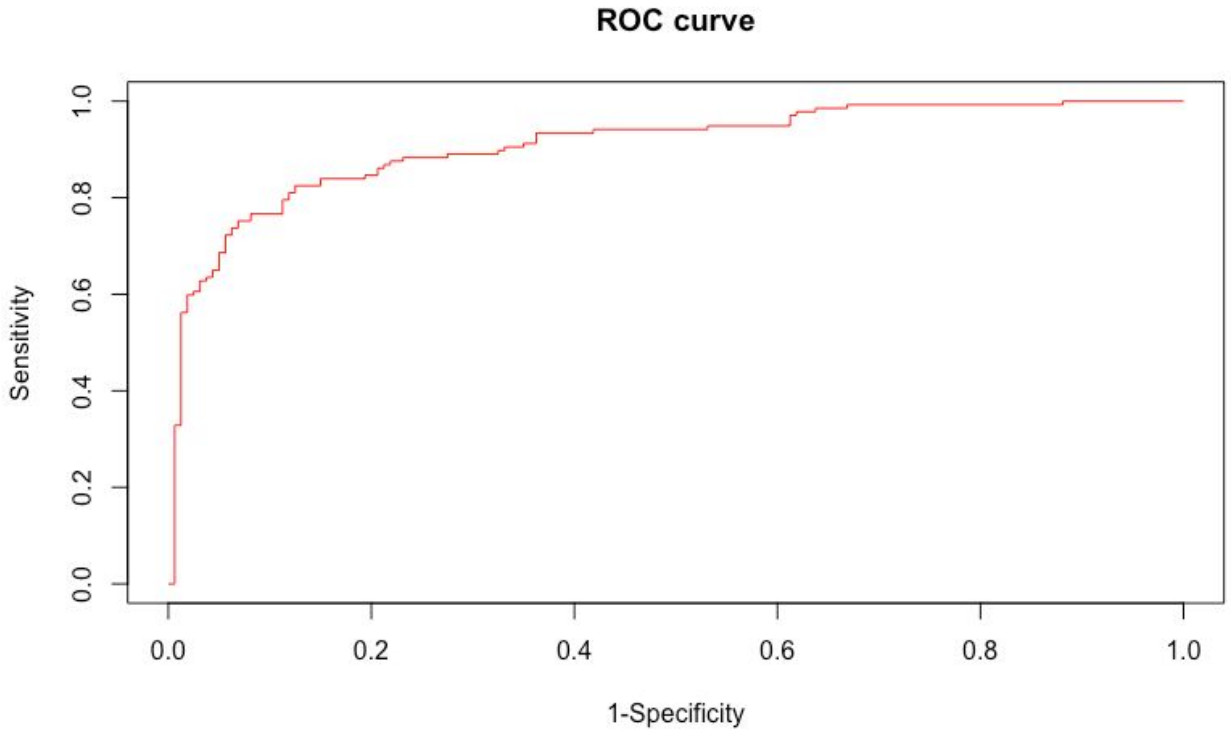

Figure 3. ROC curve of logistic regression model

4.5 Model interpretation

According to the fitted final model, the odds ratios of risk factors under different conditions are summarized in table 3.

Table 3. Summary of logistic regression

Factor	Log odds ratio (current condition/reference level)	Log odds ratio	Odds ratio [(b)]
Sex (Ref. level: $Sex = F$)	$Sex = M,\ Fbs = false,\ Thal = defect$ [(a)] $Sex = M,\ Fbs = true,\ Thal = defect$ $Sex = M,\ Fbs = false,\ Thal = normal$ $Sex = M,\ Fbs = true,\ Thal = normal$	0.14 -7.65 3.16 -5.63	1.15 0.00 23.57 0.00
Cp (Ref. level: $Cp = angina$)	$Cp = non\_anginal$ $Cp = asymptomatic$	-0.14 1.43	0.87 4.18
Trestbps (Unit: mmHg)	$Trestbps\ (increase\ 1\ unit),\ Exang = no$ $Trestbps\ (increase\ 1\ unit), Exang = yes$	0.01 0.07	1.01 1.07

Fbs (Ref. level: $Fbs = false$)	$Fbs = true, Sex = F,\ Slope = not\_up$ $Fbs = true, Sex = M,\ Slope = not\_up$ $Fbs = true, Sex = F,\ Slope = up$ $Fbs = true, Sex = M,\ Slope = up$	6.10 -1.69 -1.19 -8.98	447.51 0.18 0.30 0.00
Exang (Ref. level: $Exang = no$)	$Exang = yes,\ Slope = not\_up$ $Exang = yes,\ Slope = up$	-6.48 -8.82	0.00 0.00
Slope (Ref. level: $Slope = not\_up$)	$Slope = up,\ Exang = no$ $Slope = up,\ Exang = yes$	-0.85 -3.19	0.43 0.04
Ca (Unit: 1)	$Ca\ (increase\ 1\ unit),\ Exang = no,\ Fbs = false$ $Ca\ (increase\ 1\ unit),\ Exang = yes,\ Fbs = false$ $Ca\ (increase\ 1\ unit),\ Exang = no,\ Fbs = true$ $Ca\ (increase\ 1\ unit),\ Exang = yes,\ Fbs = true$	1.25 3.68 3.05 5.48	3.49 39.65 21.12 239.85
Thal (Ref. level: $Thal = defect$)	$Thal = normal,\ Sex = F$ $Thal = normal,\ Sex = M$	-3.22 -1.20	0.04 0.30

$(a)\ Sex = M,\ Fbs = false,\ Thal = defec$ t means the log odds ratio was calculated in the following way: $log(odds\ ratio) = log\frac{odds(heart\ disease=presence|Sex=M,\ Fbs=false,\ Thal=defect}{odds(heart\ disease=presence|Sex=F,\ Fbs=false,\ Thal=defect}$. All the following indexes indicate the similar information, that is $log(odds\ ratio) = log\frac{odds(heart\ diease=presence|condition=1)}{odds(heart\ disease=presence|condition=reference\ level)}$

(b) odd ratio= $e^{log(odds\ ratio)}$

4.5.1 Male and Female perform differently on heart disease

According to table 3, variable Sex has interaction with **fbs** (fasting blood sugar) and **thal** (Thalassemia). When patients' fbs is less than 120 mg/dl and have defect thal, the odds ratio is 1.15, which means the odds of male to have heart disease is 1.15 times the odds of female. In terms of percentage, the odds of male are 15% higher than female, however, its 95% confidence interval covers 0 (-2.31 to 2.38, see appendix 2) and the center of 95% closes to 0, which means this phenomenon is not significant. When patients don't have thal and their fbs is not higher than 120 mg/dl, the odds of male is 2257% higher than female. However, when patients' fbs is higher than 120 mg/dl, regardless the patients' status of thal, the odds of male are always 100% less than female.

4.5.2 Patients with chest pain have higher chance to gain heart disease

Similarly, for **cp** (chest pain type), patients without anginal pain is 13% less likely for patients with anginal pain to have heart disease. Moreover, patients with asymptomatic angina are 318% higher to have heart disease compare to patients with anginal pain.

4.5.3 Exang promotes the effect of high trestbps which contributes to heart disease

Trestbps (resting blood pressure) is associate with **exang** (exercise induced angina), when patients have no symbol of exang and when trestbps increase 1 mm Hg, the odds of patients getting heart disease will increase by 1%. When patients have exang, the percentage will increase to 7%, which means the presence of exang will enhance the effect of the trestbps.

4.5.4 Fbs works differently on heart disease under different conditions

The odds ratio of fbs says that, when patients gender is female and **slope** (the slope of the peak exercise ST segment) is not increasing, the odds of patients with high blood sugar level will have bigger chance to get heart disease compare with patients with fbs less than 120 mg/ml (odds ratio=447.51). However, when patients are male or the slope is up, the chance for patience with high blood sugar level to get heart disease will be less than patients with blood sugar less than 120 mg/ml.

4.5.5 The present of exang, the up of slope, and the absent of thal lower the chance of getting heart disease

For the exang, it interacts with the factor slope, however, whenever the slope is up or not, the odds of patients with exang to get heart disease is always 100% less than those without exang. Similarly, whenever patients have exang (odds ratio=0.04) or not (odds ratio=0.43), the odds of those patients with slope up to get heart disease is always lower than those with slope down or flat. In addition, although thal interacts with sex, the absent of thal (normal) has lower chance to get heart disease no matter the gender is female (odds ratio=0.04) or male (odds ratio=0.30) compare to patients have thal.

4.5.6 the increasing number of major vessels (**ca**) coordinates with exang and fbs enhance the probability of patients to get heart disease.

Ca is the number of major vessels colored by fluoroscopy, according to table 3, this variable interacts with exang and fbs. To be more specific, when patients have no exang and fbs less than 120 mg/dl, the odds of patients to get heart disease will increase by 249% (odds ratio=3.49) when the number of major vessels increased by 1. Moreover, factors exang and fbs will facilitate the influence of the increasing number of ca, specifically, the exist of exang will increase the odds ratio to 39.65, and the high fbs will increase the odds ratio to 21.12. When exang and high fbs are both present, the patient's odds ratio will

change to 239.85, which means the odds of those patients to have heart disease will increase 23885% when the number of major vessels increase by 1.

Summary

This project figured out several major risk factors for heart disease: In most of the cases, male patients have higher chance to have heart disease than females except when patients have Thalassemia; In addition, patients with Thalassemia have higher probability to have heart disease; Patients with chest pain have high risk of having heart disease; If the peak of exercise ST segment is not upsloping, this may show a presence of heart disease; The increasing of resting blood pressure may increase the risk of getting heart disease; High number of major vessels is a strong indicator of heart disease, moreover, the existence of high fasting blood sugar and exercise induced angina will contribute the effect of increasing number of major vessels. There is an interesting fact from this model that aging is not a contributor of heart disease, which means heart disease can occur at any age. Thus, it is never too young to pay attention to these heart disease risk factors. Heart disease prevention is critical, it should begin in early life.