زانکۆی سەڵاحەدین-هەولێر

Salaheddin university-Erbil

Applications of Logistic Regression Analysis

By Using: R programming

Research Project

Submitted to the department of (mathematics)in partial fulfillment of the requirements for the degree of BSc in (mathematic)

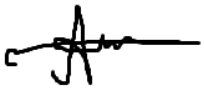
Prepared By: Zhyan Haidar Perdoowd

Supervised by: Dr. Awaz Kakamam Muhammad

April -2023

Certification of the supervisors:

I certify that this work was prepared under my supervision at the department of mathematics /college of education/Salaheddin university –Erbil in partial fulfilment of the requirements for the degree of bachelor of philosophy of science in mathematics.

Signature: 

Supervisor: Dr. Awaz Kakamam Muhammad

Scientific grade: Lecturer

In view of the available recommendations, forward this work for debate by the examining committee.

Signature:

Name: Dr. Rashad R. Haji

Scientific grade: Assistant Professor

Chairman of the Mathematics Department

Acknowledgement

Primarily, I would like to thanks my god for helping me to complete this research with success.

Then I would like to express special of my supervisor Dr. Awaz Kakamam whose valuable to guidance has been the once helped me to completing my research. Words can only inadequately express my gratitude to my supervisor for patiently helping me to think clearly and consistently by discussing every point of this project with me.

I would also like to extend my gratitude to the head of the Mathematic department Assist. Prof. Dr. Rashad Rashid.

I would like to thank my family, friend and library staff whose support has helped me to conceive this research

Abstract

The heart disease is one of the most common diseases in the world that many people suffer from, so far there are several articles that have been published related to heart disease in various disciplines of science and social context [1]. In this work, we applied logistic regression to product model on a dataset with is contains around 303 instances, each having 13 features which are used to infer the presence (values 1, 2, 3, 4) or absence (value 0) of heart disease. we take 4 variables to prediction such as: age, sex, cholesterol (chol), chest pain (cp) with heart disease.

The goal of this project is to explore logistic regression. logistic regression models have been used to predictions. we found that the relation between (chol) and (heart disease) is 0.434. We can say that there is very strong positive correlation between (chol) and (heart disease). The correlation coefficient between Age and heart disease is -0.2254 which is suggests a negative correlation. The correlation coefficient between (sex) and (chol) is a -0.1979 which is week negative correlation between them.

Contents

Introduction:

Heart disease describes a range of conditions that affect the heart. Diseases under the umbrella term heart disease include:(Cardiovascular disease, Heart arrhythmia, Congenital heart disease, Cardiomyopathy, Heart disease caused by heart infections, heart valve disease), first described in 1768 by William Heberden, it was believed by many to have something to do with blood circulating in the coronary arteries, though others though it was a harmless condition according , Heart disease (Ischaemic Heart disease ;CHD) it's estimated around 200 million people are living 110 million mean and 80 million women have coronary heart disease , heart disease is significantly high in the world especially in countries India, China , Indonesia Russia, Iran and turkey [1]

Symptoms of heart disease, beating swelling and slowing of heart beats, pain in the chest and feeling uncomfortable, severe breathing, dizziness and burrowing, yellow ingestion and disintegration of the skin, stains on the legs and throat and around the eyes. That they have a lot of danger to human life

Now we are talk about what is logistic regression what are the type of logistic what are the step of logistic for analysis data how do logistic regression use to analysis data of heart disease and what are the regression between each (chest paint (cp), Age, cholesterol (chol) and sex with heart disease).[2]

logistic regression is a logistic model, a model that assumes a logistic relationship between the input variables (x) and the single output variable (y). logistic regression analysis is used to predict the value of a variable based on the value of another variable.

In this work, I applied logistic regression to predict model on a dataset that which I received this data from a trusted website [3]. This data includes 303 examples (cases) with 4 variables (Age, sex, cp, Chol, and heart disease). logistic regression model has been used to predictions. We use R programing [4] and Excel to this analysis.

This project contained two chapters. In chapter one, I describe a logistic regression model. logistic regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable (often called the 'outcome' or 'response' variable). The variable we are using to predict the other variable's value is called the independent variable (often called 'predictors', 'covariates', 'explanatory variables', 'attributes 'or 'features').

Chapter two includes some applications on the dataset by using simple logistic regression.

From this work we found that the relation between (cp) and heart disease is a 0.4338, we can say that there is very strong positive correlation between them. In this project, the result show that correlation coefficient between chol. and heart disease is -0.0852. It means they have week negative relation between them. Moreover, we found that the correlation coefficient between age and heart disease is -0.2254 the correlation coefficient of -0.2254 suggests a negative correlation between age and heart disease . The correlation coefficient between sex and chol. is a-0.1979 which is week negative correlation between them. for this data analysis we use R programing [4] and Excel programing.

Chapter one:

Logistic Regression

In this chapter we will show some basic definition and concepts about logistic regression.

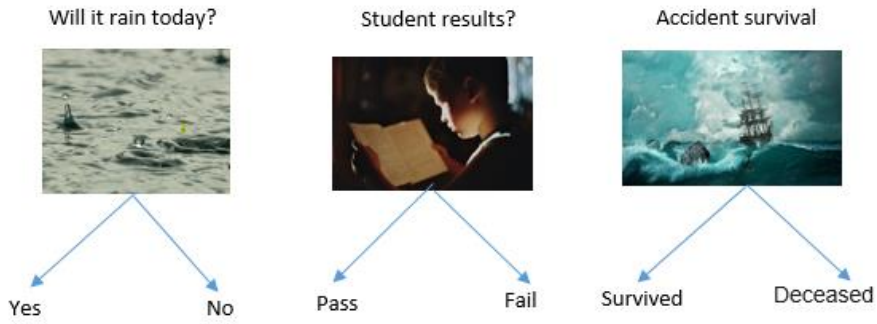
What is regression?

regression is a statistical procedure which attempts to predict the values of a given variable, (termed the dependent, outcome, or response variable) based on the values of one or more other variables (called independent variables, predictors, or covariates). The result of a regression is usually an equation which summarizes the relationship between the dependent and independent variable(s). Typically, the model is accompanied by summary statistics describing how well the model fits the data, the amount of variation in the outcome accounted for by the model, and a basis for comparing the existing model to other similar models. By comparing these statistics across multiple models, the user is able to determine a combination and order of independent variables that most satisfactorily predict the values of the outcome. Numerous forms of regression have been developed to predict the values of a wide variety of outcome measures. Since the focus of regression modeling is on the response variable, the type of regression you use will be dictated by the type of response variable you are analyzing and by your eventual analytic goal.

Definition of Logistic Regression:

The logistic regression statistic modeling technique is used when we have a binary outcome variable. For example: given the parameters, will the student pass or fail? Will whether rain or not?

So, though we may have continuous or categorical independent variables, we can use the logistic regression modeling technique to predict the outcome when the outcome variable is binary.



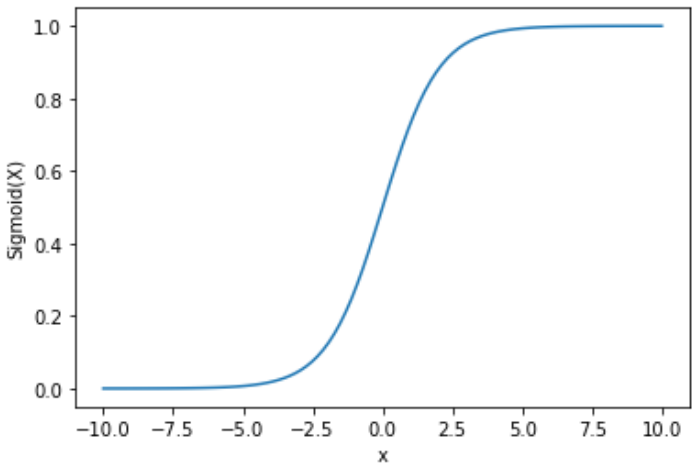
Let's see how the algorithm differs from linear regression. Linear regression statistical model is used to predict continuous outcome variables, whereas logistic regression predicts categorical outcome variables. Linear regression model regression line is highly susceptible to outliers. So, it will not be appropriate for logistic regression.

Below is the function for logistic regression:

$$f(x) = \frac{1}{1 + e^{-x}}$$

e : is log base

x : is the numerical value that needs to be transformed.



If we feed an output value to the sigmoid function, it will return the probability of the outcome between 0 and 1. If the value is below 0.5, then the output is return as No/Fail/Deceased. If the value is above 0.5, then the output is returned as Yes/Pass/Deceased.

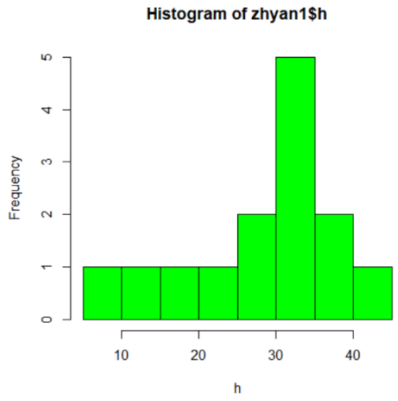
Examples to logistic regression:

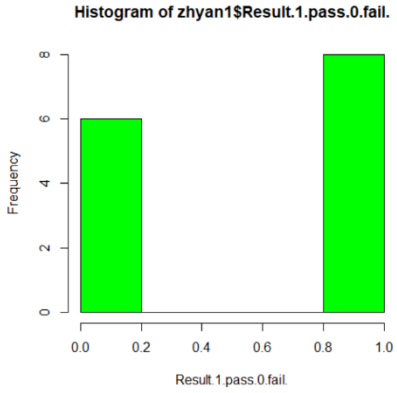
In this example we applied logistic regreisstion to predict model. Data includes 14 examples with dependent variables and independent variables in which they study hours within a week and the result is pass or fail , see table 1.1 Hours studies: - the number of these hours that the students have studied Result(1=pass,0=fail): -[1=pass]it means they are out of the test and [0=fail]it means they didn't pass the test.

Table1.1: hours studies, result(1=pass,0=fail).

Hours studies (h)	Result(1=pass,0=fail)
29	0
15	0
33	1
28	1
39	1
44	1
31	1
19	0
9	1
24	0
32	0
31	0
37	1
35	1

We will reprsenete a graph for the data set and describe the distrbution frequency table of it.





Fiuger1.1 the students' hours of study are in the data

Fuger1.2 student results compared to school hours

Assumptions of Logistic regression:

1. Independent variables show a linear relationship with the log of output variables.
2. Non-Collinearity between independent variables. That is, independent variables are independent of each other.
3. Output variable is binary.

Type of logistic regression:

There are three main types of logistic regression binary, multinomial and ordinal. They differ in execution and theory. Binary regression deals with two possible values, essentially: yes or no, multinomial logistic regression deals with three or more values. And ordinal logistic regression deals with three or more classes in a predetermined.

1) Binary logistic regression

Binary logistic regression is just two possible outcome answers. This concept is typically represented as 0 or a 1 in coding.

Binary logistic regression is a type of regression analysis where the dependent variable is a dummy variable (coded 0, 1)

Why not just use ordinary least squares?

Y = a + bx. You would typically get the correct answers in terms of the sign and significance of coefficients. However, there are three problems:

The error terms are heteroskedastic (variance of the dependent variable is different with different values of the independent variables. The error terms are not normally distributed, and most importantly, for purpose of interpretation, the predicted probabilities can be greater than 1 or less than 0, which can be a problem for subsequent analysis.

The coefficients of the multiple regression model are estimated using sample data with k independent variable.

$$\hat{Y}_i = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_k X_k$$

2) Ordinal Logistic Regression

Ordinal logistic regression is also a model where there are multiple classes that an item can be classified as however in this case an ordering of classes is required classes do not need to be proportion the distance between each class.

3) Multinomial logistic regression

Multinomial logistic regression is a model where there are multiple classes that an item can be classified as. There is a set of three or more predefined classes set up prior to running the model.

Chapter two

Some Application of logistic regression

In this chapter we applied logistic regression to predict model. We have the data from [1] this data includes 303 examples with dependent and independent variables. The independent variables are: age, sex, cp, and Chol where

Age: - {age in year}

Sex: -{1=male;0=female}

Cp: - {chest pain type}

Chol: - {serum cholesterol in mg/dl}

The dependent variable is heart disease, (output).

Graphical presentation:

We will Represented a graph for the data set and describe the distribution frequency table of it

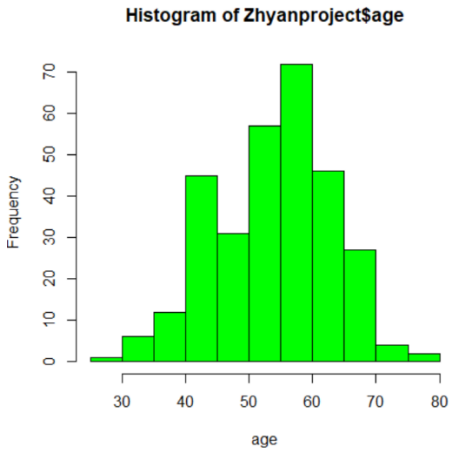
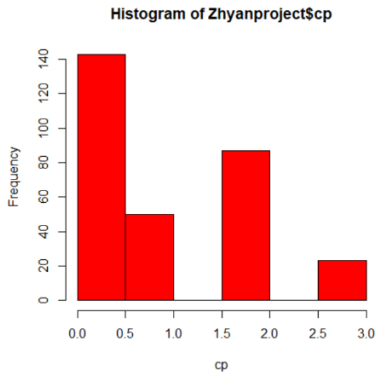
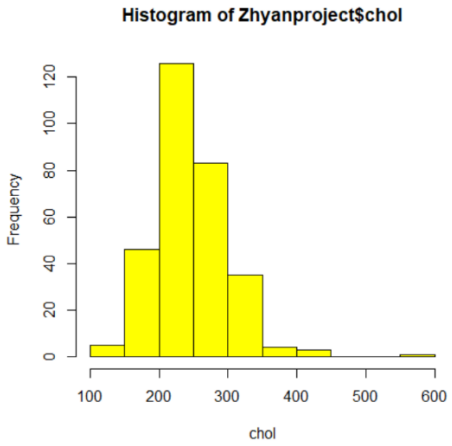


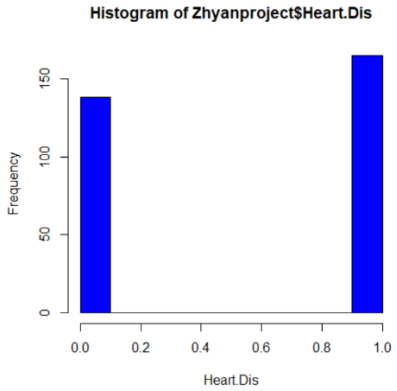
Figure 2.1: histogram for age



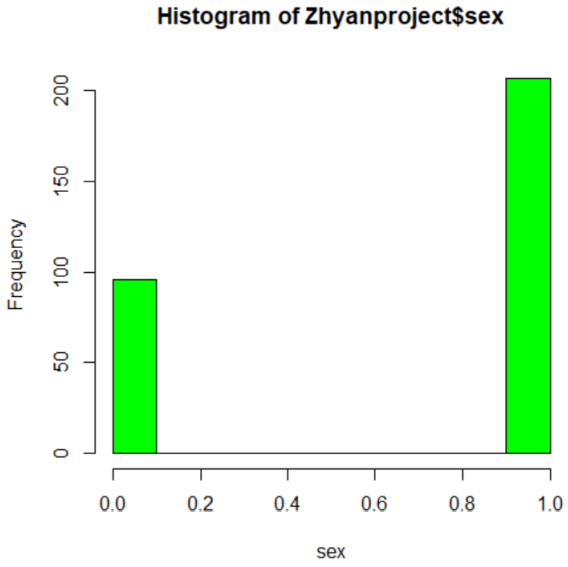
Figuer2.2: histogram for chest pian (cp)



Fiuger2.3: histogram of dataset for Chol (cholesterol)



Figuer2.4 histogram of dataset for heart disease



Figuer2.5: histogram of dataset for sex

Correlation coefficient:-

Definition:(Correlation coefficient): the correlation coefficient is an indicator of measuring the dependent between attributes. The Pearson's correlation coefficient (PCC) can be defined as the covariance of two random variables divided by the product of the individual standard deviation. Let as consider two variable X and Y written xi and yi , where i=1,2,3 ..,n. the Pearson's r is defined as:

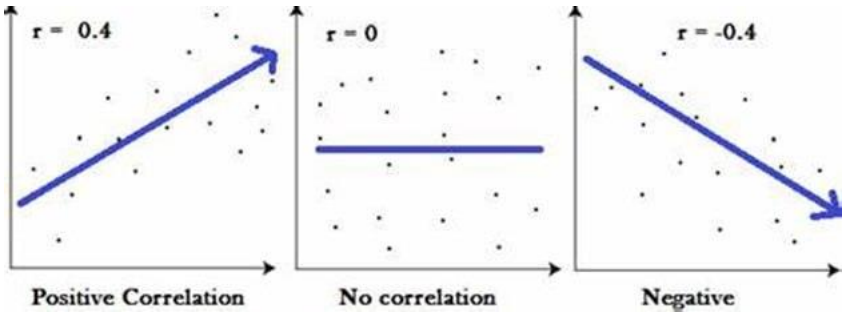
$$r=\frac{\sum_{i=1}^{n}(xi-\bar{x})(yi-\bar{y})}{\sqrt{\sum_{i=1}^{n}(xi-\bar{x})^2\sum_{i=1}^{n}(yi-\bar{y})^2}} \text{ where } -1 < r < 1.$$

***Notation*:**

If Y tends to increase as X increases, the correlation is called positive, or direct, correlation.

If Y tends to decrease as X increases, the correlation is called negative, or inverse, correlation.

If there is no relationship indicated between the variables, we say that there is no correlation between them (i.e., they are uncorrelated).



In this project we put r=0.4 very strong correlation between the features in this data set and r= 0.3 strong, and r= 0.1 is week correlation.

Difference between correlation and regression:

In this table below we will show some basic differences between Correlation and regression.

Table 2.1: Differences between Correlation and regression.

Basis for comparison	Correlation	regression
meaning	Correlation is a statical measure that determines the association or co-relation between two variables	Regression describes how to numerically relate an independent variable to the dependent variables
usage	To represent a linear relationship between two variables	To fit the best line and to estimate one variable based on another
Dependent and independent variables	No difference	Both variables are different
indicates	Correlation coefficient indicates the extent to which two variables move together	Regression indicates the impact of a change of unit on the estimated variables(y)in the know variables(x)
objective	To find a numerical value expressing between variables	To estimate value of random variables on the basis of the value of fixed variables

The main problem with Figure 2.6 is that the variability in heart disease at all ages is large. This makes it difficult to see any functional relationship between Age and heart disease. One common method of removing some variation, while still maintaining the structure of the relationship between the dependent and the independent variable, is to create intervals for the independent variable and compute the mean of the outcome variable within each group. We use this strategy by grouping age into the categories (Age Group) defined in Table 2.2. Table 2.1 contains, for each age group, the frequency of occurrence of each outcome, as well as the percent with heart disease.

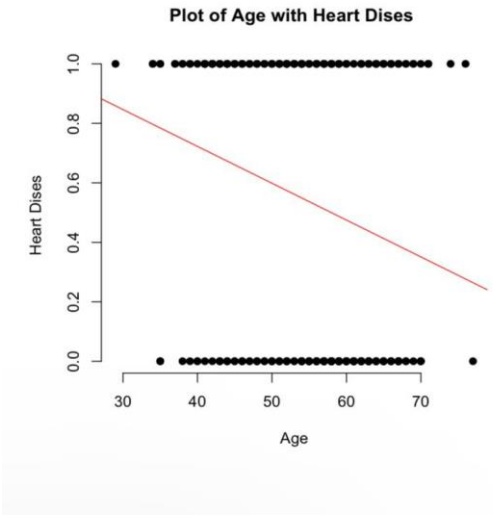


Figure2.6: scatterplot of "yes" or "no" of heart disease by age.

Table 2.1 illustrate frequency table of Age Group by heart disease. I divide Age feature for 10 subset and putted in 10 class intervals. Each class have size 5.

Table 2.2: Frequency Table of Age Group by Heart disease.

N	Age group	n	Heart disease(yes)	Heart disease(no)	Mean
1	29-33	1	1	0	1
2	34-38	11	8	3	0.73
3	39-43	33	25	8	0.76
4	44-48	38	25	13	0.66
5	49-53	46	32	14	0.70
6	54-58	71	32	39	0.45
7	59-63	53	15	38	0.28
8	64-68	39	20	19	0.52
9	69-73	10	6	4	0.6
10	74-78	3	2	1	0.66

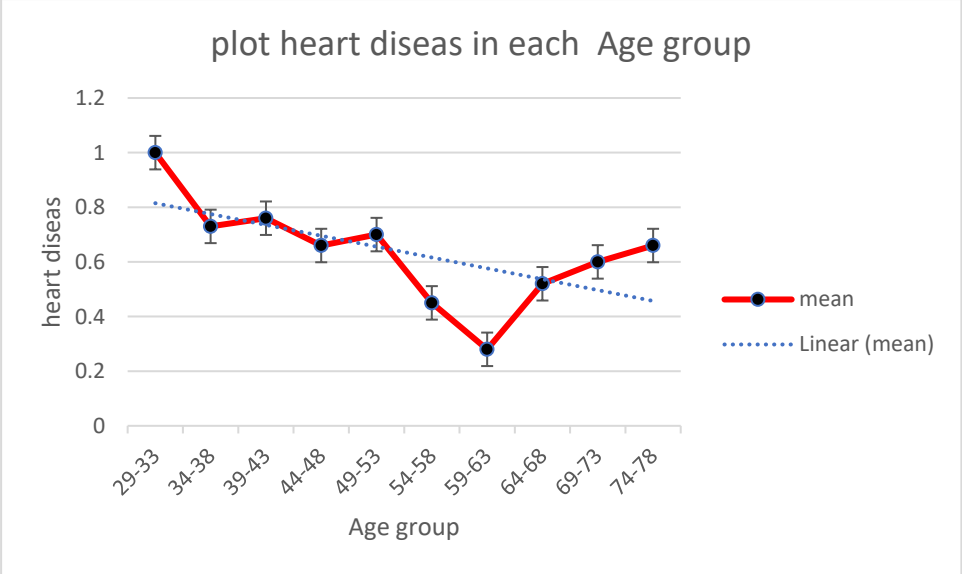


Figure 2.7: plot of the percentage of subjects with hard disease in each age group.

Moreover, from figures 2.6 and 2.7, I can have same study for another variable (age group and hard disease)

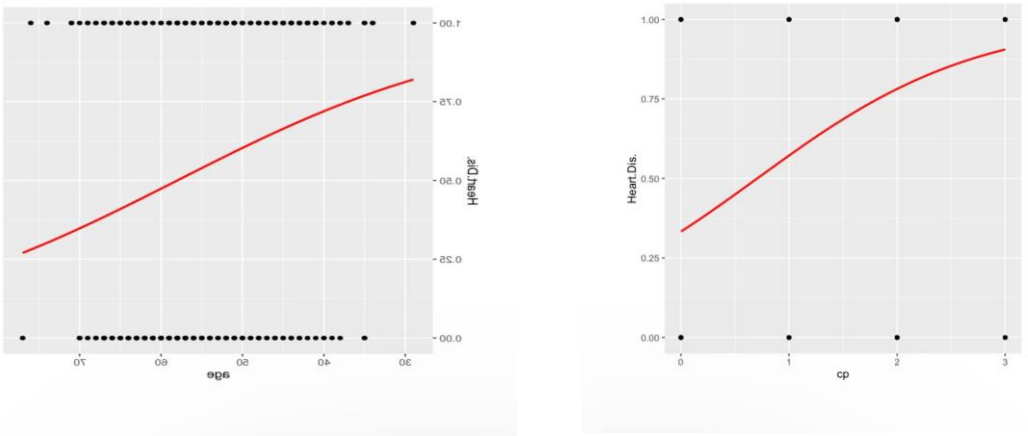


Figure 2.8: a) plot of age with hard disease **b) plot of cp with hard disease**

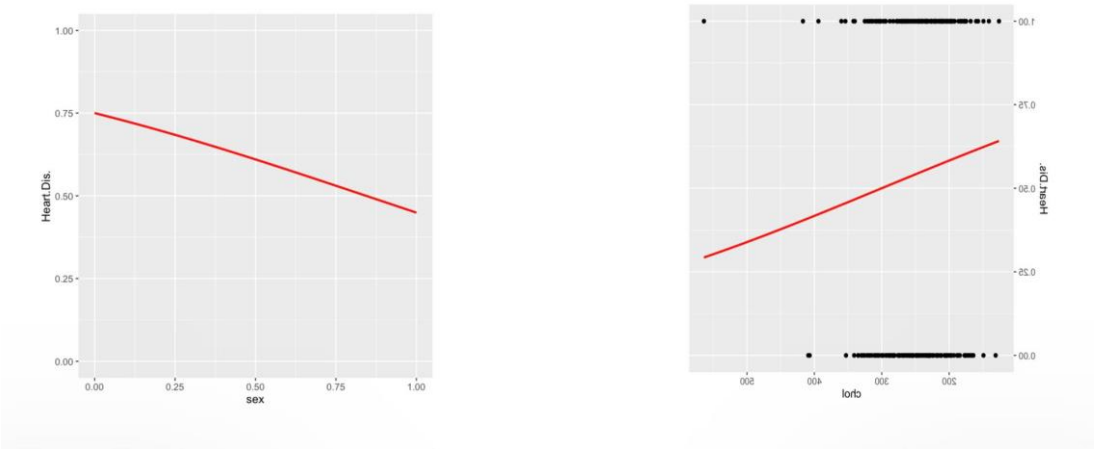


Figure 2.9: a) plot of sex with hard disease b) plot of Chol with hard disease.

Logistic regression model:

In this project, we applied regression method on the dataset. first, I want to find regression model to predict independent variable age with dependent variable heart disease.

Between age and heart disease

glm(formula = Heart.Dis. ~ age, data = Zhyanproject)
Deviance Residuals:

Min	1Q	Median	3Q	Max
-0.7843	-0.4996	0.2899	0.4385	0.7233

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.217731	0.170000	7.163	6.09e-12 ***
age	-0.012382	0.003084	-4.015	7.52e-05 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘’ 1
(Dispersion parameter for gaussian family taken to be 0.2369743)
Null deviance: 75.149 on 302 degrees of freedom
Residual deviance: 71.329 on 301 degrees of freedom
AIC: 427.61

Between (chest pain) cp and heart disease

glm(formula = Heart.Dis. ~ cp, data = Zhyanproject)				
Deviance Residuals:				
Min	1Q	Median	3Q	Max
-0.97082	-0.34180	0.02918	0.44853	0.65820
Coefficients:				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.34180	0.03547	9.636	2e-16 ***
cp	0.20967	0.02510	8.353	2.47e-15 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for gaussian family taken to be 0.2026811)
Null deviance: 75.149 on 302 degrees of freedom
Residual deviance: 61.007 on 301 degrees of freedom
AIC: 380.25
Number of Fisher Scoring iterations: 2

Between sex and heart disease

glm(formula = Heart.Dis. ~ sex, data = Zhyanproject)				
Deviance Residuals:				
Min	1Q	Median	3Q	Max
-0.7500	-0.4493	0.2500	0.5507	0.5507
Coefficients:				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.75000	0.04894	15.324	< 2e-16 ***
sex	-0.30072	0.05921	-5.079	6.68e-07 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for gaussian family taken to be 0.2299581)
Null deviance: 75.149 on 302 degrees of freedom
Residual deviance: 69.217 on 301 degrees of freedom
AIC: 418.5
Number of Fisher Scoring iterations: 2

Between cholestrol and heart disease

glm(formula = Heart.Dis. ~ chol, data = Zhyanproject)				
Deviance Residuals:				
Min	1Q	Median	3Q	Max
-0.6391	-0.5370	0.4027	0.4499	0.7161
Coefficients:				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.7465813	0.1390865	5.368	1.6e-07 ***
chol	-0.0008204	0.0005527	-1.484	0.139

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for gaussian family taken to be 0.2478489)
Null deviance: 75.149 on 302 degrees of freedom
Residual deviance: 74.603 on 301 degrees of freedom
AIC: 441.2
Number of Fisher Scoring iterations: 2

Conclusion: -

?????

References:

Barrett-Connor, E. L. I. Z. A. B. E. T. H., & Khaw, K. T. (1984). Family history of heart attack as an independent predictor of death due to cardiovascular disease. *Circulation*, *69*(6), 1065-1069.

Erkan, A., & Yildiz, Z. (2014). Parallel lines assumption in ordinal logistic regression and analysis approaches. *International Interdisciplinary Journal of Scientific Research*, *1*(3), 8-23.

UCI Machine Learning Repository: Heart Disease Data Set. https://archive.ics.uci.edu/ml/datasets/heart+disease

Guido, J. J., Winters, P. C., & Rains, A. B. (2006). Logistic regression basics. MSc University of Rochester Medical Center, Rochester, NY.

Tranmere, M., & Elliot, M. (2008). Binary logistic regression. Cathie Marsh for census and survey research, paper, 20.

Janosi, A., Steinbrunn, W., Pfisterer, M., & Detrano, R. (1988). UCI machine learning repository-heart disease data set. *School Inf. Comput. Sci., Univ. California, Irvine, CA, USA.*