Analysis of Presence of Heart Disease using Statistical Methods

Introduction:

Since cardiovascular diseases are one of the world's leading causes of mortality, it is vital to detect and analyze heart disease early. The Heart Disease dataset offers important insights into a number of the variables associated to the occurrence of heart disease and is frequently employed in medical research. Two main research questions are addressed in this study:

1. Which clinical variables are significantly associated with heart disease?
2. How effectively can a logistic regression model predict the presence of heart disease?

The answers to these questions and significant issues could help in the improvement of early intervention plans and diagnostic techniques for heart disease.

Methods:

To understand the main features of the dataset, first applied preliminary analysis, which includes graphical representations, descriptive statistics, and frequency table. The aim of this project is want to estimate and predict heart disease so we applied logistic regression model in the main analysis.

Dataset:

The UCI Heart Disease dataset, which included 14 variables (e.g., cholesterol, resting blood pressure), diagnostic indicators (e.g., maximum heart rate achieved, exercise-induced angina, and type of chest pain), and demographic characteristics (e.g., age, sex) and 303 observations was used in the analysis. These variables include clinical variables. The target variable, "heart disease," is binary: 0 denotes the absence of the disease, and 1 denotes it is present.

Graphical Representations:

One of the most familiar graphical representation to check the normality assumption is histogram, which we are applied in this study.

Descriptive Statistics:

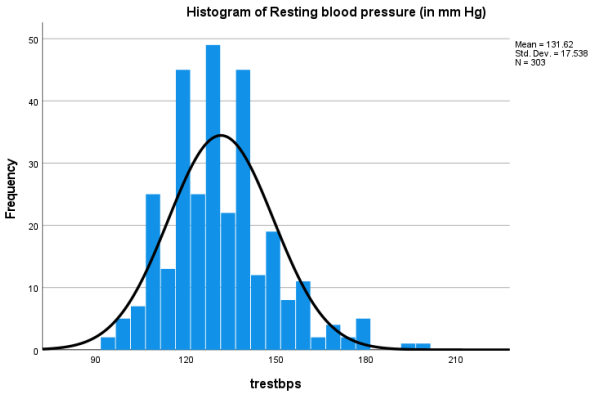
If the data is qualitative then we need to use frequency table instead of descriptive statistics to check the main feature of the data.

Logistic Regression Model:

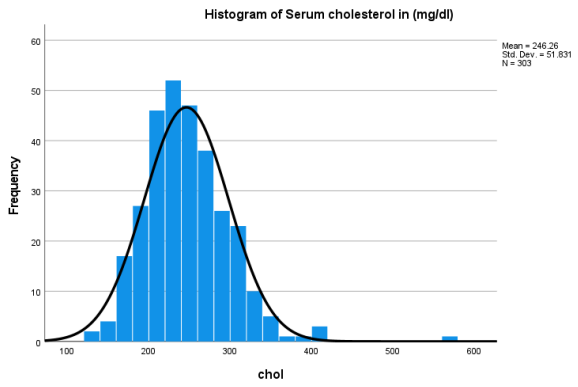
Our dependent variable “presence of heart disease” is a qualitative variable, then we used logistic regression model.

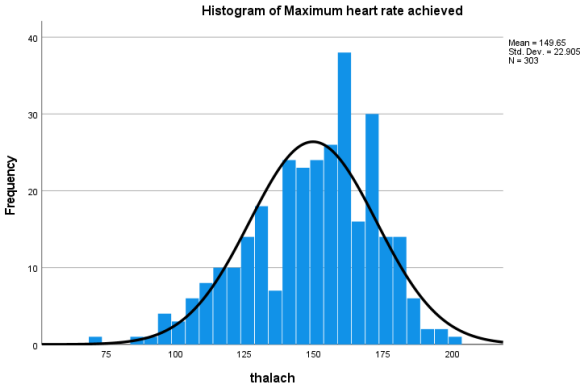
Results:

To check the normality of the quantitative data using the following histogram.

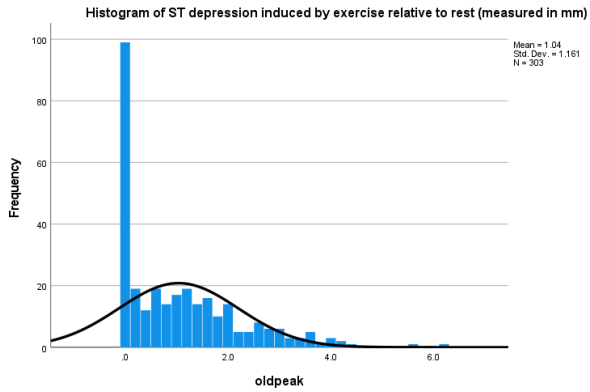


A normal distribution is indicated by the symmetrical age data. Leptokurtic resting blood pressure indicates that it does not have a normal distribution.





Maximum heart rate is negatively skewed and serum cholesterol is positively skewed; both are not in line with the normal distribution.



"ST depression" deviates from the normal distribution and is positively skewed.

Descriptive Statistics:

To check the main features of quantitative variables using the following descriptive statistics.

Descriptive Statistics				
Variable	**Minimum**	**Maximum**	**Mean**	**Std. Deviation**
Age of the individual (in years)	29	77	54.37	9.082
Resting blood pressure (in mm Hg)	94	200	131.62	17.538
Serum cholesterol in mg/dl	126	564	246.26	51.831
Maximum heart rate achieved	71	202	149.65	22.905
ST depression induced by exercise relative to rest (measured in mm)	.0	6.2	1.040	1.1611

According to the descriptive statistics, the average age is 54.37 years, the average blood pressure (trestbps) is 131.62 mm Hg, and the average cholesterol was 246.26 mg/dl. The dataset's variation in cardiovascular health indicators is seen in the ST depression (oldpeak) mean of 1.04 mm and the maximum heart rate (thalach) average of 149.65 bpm.

Frequency Table:

The qualitative variables are represented by using the following frequency tables.

Frequency Table of Gender of the individual			
Categories	**Frequency**	**Percent**	**Cumulative Percent**
0	96	31.7	31.7
1	207	68.3	100.0

According to frequency table Compared to females (0), male are more frequent (1).

Frequency Table of Chest pain type			
Categories	**Frequency**	**Percent**	**Cumulative Percent**
0	143	47.2	47.2
1	50	16.5	63.7
2	87	28.7	92.4
3	23	7.6	100

The most common type of angina is typical (0), which is followed by atypical angina (1), non-anginal pain (2), and asymptomatic cases (3).

Frequency Table of Fasting Blood Sugar			
Categories	**Frequency**	**Percent**	**Cumulative Percent**
0	258	85.1	85.1
1	45	14.9	100

Forty-five people had increased fasting blood sugar, while the majority (0) have normal measurements (≤120 mg/dl).

Frequency Table of Resting Electrocardiographic Results			
Categories	**Frequency**	**Percent**	**Cumulative Percent**
0	147	48.5	48.5
1	152	50.2	98.7
2	4	1.3	100

While left ventricular hypertrophy (2) is relatively rare, normal findings (0) and ST-T abnormalities (1) are nearly equally distributed (4).

Frequency Table of Exercise-induced Angina			
Categories	**Frequency**	**Percent**	**Cumulative Percent**
0	204	67.3	67.3
1	99	32.7	100

99 people have angina during activity (1), compared to the majority (0) who do not.

Frequency Table of Slope of the Peak Exercise ST Segment			
Categories	**Frequency**	**Percent**	**Cumulative Percent**
0	21	6.9	6.9
1	140	46.2	53.1
2	142	46.9	100

At 142 and 140, downsloping (2) and flat slopes (1) are almost equal, although upsloping (0) occurs less frequently (21).

Frequency Table of Thalassemia (a blood disorder)			
Categories	**Frequency**	**Percent**	**Cumulative Percent**
0	2	0.7	0.7
1	18	5.9	6.6
2	166	54.8	61.4
3	117	38.6	100

Normal cases (1) are uncommon, while fixed problems (2) are most prevalent (166), followed by reversible defects (3) at 117.

Frequency Table of Gender of the individual			
Categories	**Frequency**	**Percent**	**Cumulative Percent**
0	175	57.8	57.8
1	65	21.5	79.2
2	38	12.5	91.7
3	20	6.6	98.3
4	2	1.7	100

Four vessels (4) are uncommon (5), while zero vessels (0) are the most common (175) and decrease as the number of vessels rises.

Frequency Table of the Presence of Heart Disease			
Categories	**Frequency**	**Percent**	**Cumulative Percent**
0	138	45.5	45.5

1	165	54.5	100

There are 138 people without heart diseases and 165 people with heart disease.

Initial Logistic Regression Model:

The initial logistic regression model in which all variables are included as independent variables is given below.

Model Summary	
Cox & Snell R Square	**Nagelkerke R-Square**
0.494	0.660

Variables in Equation		
Heart Disease	**Odds Ratio**	**Sig.**
Age of the Individual (in years)	.995	.832
Gender of the Individual	.172	.000
Chest Pain Type	2.363	.000
Resting Blood Pressure (in mm Hg)	.981	.060
Serum Cholesterol (in mg/dl)	.995	.221
Fasting Blood Sugar	1.036	.947
Resting Electrocardiographic Results	1.594	.181
Maximum Heart Rate Achieved	1.023	.026
Exercise-induced Angina	.375	.017
ST depression induced by Exercise relative to Rest (measured in mm)	.583	.012
The slope of the Peak Exercise ST Segment	1.785	.098
Number of Major Vessels	.461	.000
Thalassemia (a blood disorder)	.406	.002
	31.515	.180

The variable age, chol, fbs, and slope have insignificant impact on heart disease, so we removed these variables from the model, and the estimate the following logistic regression model.

Final Logistic Regression Model:

The final logistic regression model is given below.

Model Summary	
Cox & Snell R Square	**Nagelkerke R-Square**
0.487	0.650

Variables in Equation		
Heart Disease	**Odds Ratio**	**Sig.**
Gender of the Individual	.000	.226
Chest Pain Type	.000	2.329
Resting Blood Pressure (in mm Hg)	.046	.980
Resting Electrocardiographic Results	.077	1.815
Maximum Heart Rate Achieved	.005	1.026
Exercise-induced Angina	.012	.363
ST depression induced by Exercise relative to Rest (measured in	.000	.487
mm)	.000	.489
Number of Major Vessels	.001	.394
Thalassemia (a blood disorder)	.198	11.817

Male sex (OR = 0.226), higher ST depression (OR = 0.487), exercise-induced angina (OR = 0.363), severe thalassemia (OR = 0.394), and more main vessels (OR = 0.489) decrease the risks of heart disease, although higher chest pain types (OR = 2.329) and abnormal ECG (OR = 1.815) raise them. There are also significant impacts from resting blood pressure (OR = 0.98) and maximum heart rate (OR = 1.026).

Discussion:

This study's main goal was to find important clinical predictors of heart disease and evaluate how well a logistic regression model could predict its occurrence.

The findings showed that a number of variables, such as the type of pain in the chest, maximum heart rate, and ST depression, are important indicators of heart disease. People with abnormal ECG readings and unusual chest pain types were more likely to have the situation. On the other hand, a lower risk of heart disease was associated with to higher maximal heart rates and higher ST depression levels. By emphasizing important variables and excluding less significant ones, the logistic regression model proved helpful for making predictions.

To expand on this study, future studies might take into account:

1. Extending the dataset to include a wider range of populations and take genetic and geographic variances into consideration.

2. Including other indicators, such genetic characteristics, habits of eating, or levels of physical activity.

3. The dataset does not include information on comorbid conditions, lifestyle factors (e.g., smoking, diet, physical activity), or genetic predispositions, which are important contributors to heart disease.