International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue XI, Nov 2018- Available at www.ijraset.com

Measurement of Risk Factor for Heart Disease using Correlation and Chi Squared Test

Madhu. G[1], K. P. Nelavigi[2]
[1, 2]*Department of Statistics, Mes Degree College, Bangalore*

Abstract: ***In modern world, Heart disease is a biggest hurdle for the human society. The World Health Organization (WHO) analyzed that 17.3 million deaths occur worldwide due to heart disease. Latest statistics suggest that in India, there are 30 million heart patients. One person dies every 33 seconds owing to a heart attack in India and 2 lakh surgeries are being performed every year. In this paper an attempt is made to determine risk factors for heart disease by using statistical tools such as correlation and chi-squared tests from the data of "Cleveland Clinic Foundation (USA)".***
Keywords: Age, Gender, Correlation, Chi-squared test, Exercise induced angina, Chest pain and maximum heart rate.

I. INTRODUCTION

Heart disease is now the world's biggest killer of both men and women. Half of the deaths in the United States and other developed countries occur due to Cardio Vascular Disease(CVD). *WORLD HEART DAY* is celebrated every year on 29th September with the intent of raising awareness about cardiovascular disease.

A. Cardio Vascular Disease (CVD)

The Cardiovascular disease is made up of the heart and blood vessels. CVD is defined as any serious abnormal condition of the heart or blood vessels (arteries and veins). CVD includes Coronary Heart Disease (CHD), Congential heart disease and many other conditions.
Risk factors are variables that predict who is most likely to develop CVD. Most of the risk factors for CVD and stroke are modifiable or entirely preventable.
By modifying risk factors, you decrease the chances of getting disease. The more risk factors one has, the higher the risk of developing heart disease.

B. Modifiable Risk Factors

Tobacco use, high blood pressure, physical inactivity, high blood cholesterol, obesity, heavy alcohol and poor nutrition

C. Non-Modifiable Risk factors

Family history (hereditary) and age factor
Medical diagnosis plays a vital role and yet complicated task that needs to be executed efficiently and accurately. Learning of the risk components connected with heart disease helps medical services experts to recognize patients at high risks of having heart disease.

1) *Prevention:* A healthy life style can help prevent heart disease and slow its progress. A heart-healthy includes maintaining a healthy diet, regular exercise, weight maintained, no smoking, moderate drinking, controlling hypertension and managing stress.

II. DATA DESCRIPTION

A. Data Set Information

This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to this date. The "goal" field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4. Experiments with the Cleveland database have concentrated on simply attempting to distinguish presence (values 1,2,3,4) from absence (value 0). The dataset consists of 297 values, in which 201 are males and 96 females.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue XI, Nov 2018- Available at www.ijraset.com

DATA SPECIMEN

Age	Sex	cp	trestbps	chol	fbs	restecg	thalach	exang	old peak	slope	ca	thal
63	1	1	145	233	1	2	150	0	2.3	3	0	6
67	1	4	160	286	0	2	108	1	1.5	2	3	3
67	1	4	120	229	0	2	129	1	2.6	2	2	7
37	1	3	130	250	0	0	187	0	3.5	3	0	3
41	0	2	130	204	0	2	172	0	1.4	1	0	3
56	1	2	120	236	0	0	178	0	0.8	1	0	3
62	0	4	140	268	0	2	160	0	3.6	3	2	3
57	0	4	120	354	0	0	163	1	0.6	1	0	3
63	1	4	130	254	0	2	147	0	1.4	2	1	7
53	1	4	140	203	1	2	155	1	3.1	3	0	7
57	1	4	140	192	0	0	148	0	0.4	2	0	6
56	0	2	140	294	0	2	153	0	1.3	2	0	3
56	1	3	130	256	1	2	142	1	0.6	2	1	6
44	1	2	120	263	0	0	173	0	0	1	0	7
52	1	3	172	199	1	0	162	0	0.5	1	0	7
57	1	3	150	168	0	0	174	0	1.6	1	0	3
48	1	2	110	229	0	0	168	0	1	3	0	7
54	1	4	140	239	0	0	160	0	1.2	1	0	3
48	0	3	130	275	0	0	139	0	0.2	1	0	3

Age – in years
Gender – 1 = male; 0 = female.
CP – Chest pain type
1 = typical angina
2 = atypical angina
3 = non – angina pain
4 = asymptomatic.
Trestbps – resting blood pressure (in mm).
restecg – resting electrocardiographic results
0 = normal
1 = having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.005 mv) ;
2 = showing probable or definite left ventricular hypertrophy by Ester's criteria.
Chol – serum cholesterol in mg/dl
Fbs – fasting blood sugar > 120 mg/dl
1 = true ; 0 = false.
thalach – maximum heart rate achieved.
Exang – exercise induced angina
1 = yes ; 0 = no.
Old peak – ST depression induced by exercise relative to rest
Slope – the slope of the peak exercise ST segment
1 = up sloping
2 = flat
3 = down sloping.
Ca – number of major vessels (0-3) coloured by flourosopy
Thal - 3 = normal; 6 = fixed defect; 7 = reversible defect.

The objectives of the study are

To study the independence among the males and females with respect to

1) Exang(exercise induced angina) vs slope
2) age vs cp(chest pain)
3) age vs restbp(resting bloodpressure)
4) age vs chol(cholesterol)
5) age vs thalach(maximum heart rate).

B. Correlation Structure

Table 1: Correlations

		age	restbp	chol	thalach	oldpeak
Age	Pearson Correlation	1	.290**	.203**	-.395**	.197**
	Sig. (2-tailed)		.000	.000	.000	.001
	N	297	297	297	297	297
Restbp	Pearson Correlation	.290**	1	.132*	-.049	.191**
	Sig. (2-tailed)	.000		.023	.399	.001
	N	297	297	297	297	297
Chol	Pearson Correlation	.203**	.132*	1	.000	.039
	Sig. (2-tailed)	.000	.023		.999	.508
	N	297	297	297	297	297
Thalach	Pearson Correlation	-.395**	-.049	.000	1	-.348**
	Sig. (2-tailed)	.000	.399	.999		.000
	N	297	297	297	297	297
Oldpeak	Pearson Correlation	.197**	.191**	.039	-.348**	1
	Sig. (2-tailed)	.001	.001	.508	.000	
	N	297	297	297	297	297

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

Interpretation

1) Here the correlation between age and maximum heart rate achieved shows negative correlation(i.e., Increase in one factor results in the decrease in another factor)
2) The factors maximum heart rate achieved and old peak also shows negative correlation.

CHI - SQUARED TEST

The Chi-Squared test is used to determined independence of two factors.

The Chi-Squared statistic is defined as

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

where O_i is the observed frequency, and E_i is the expected frequency.

The hypothesis is rejected at the chosen significance level(α) if the test statistic is greater than the critical value defined as

$$\chi^2_{1-\alpha,k-1}$$

(I) Slope versus Exercise induced angina in males and females.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue XI, Nov 2018- Available at www.ijraset.com

a) H_0 : Slope is independent of exercise induced angina in females
Vs H_1 : Slope is dependent of exercise induced angina in females
where Slope – the slope of the peak exercise ST segment
1 = up sloping
2 = flat
3 = down sloping

Table 2

		slope			Total
		upsloping	flat	downsloping	
exang	no	41	30	3	74
	yes	5	15	2	22
Total		46	45	5	96

Table3: Chi-Square Test (Females)

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	7.369[a]	2	.025
Likelihood Ratio	7.703	2	.021
Linear-by-Linear Association	6.843	1	.009
N of Valid Cases	96		

1) Inference: From the table we reject the null hypothesis since p<0.05 and conclude that the up slopping and flat slopping are dependent on exercise induced angina in females.

(I).b) H_0 : Slope is independent of exercise induced angina in males
Vs H_1 : Slope is dependent on exercise induced angina in males.

Table 4

		slope			Total
		upslopping	flat	downsloping	
exang	No	72	45	9	126
	Yes	21	47	7	75
Total		93	92	16	201

Table5: Chi-Square Test(Males)

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	16.375[a]	2	.000
Likelihood Ratio	16.783	2	.000
Linear-by-Linear Association	11.612	1	.001
N of Valid Cases	201		

2) Inference: From the table we reject the null hypothesis since p<0.05 and conclude that slopping is dependent on exercise induced angina in males.

(II) Age versus chest pain levels in males and females.

H_0^1 : Age is independent of chest pain levels in males

Vs H_1^1 : Age is dependent on chest pain levels in males.

H_0^2 : Age is independent of chest pain levels in females

Vs H_1^2 : Age is dependent on chest pain levels in females.

Table6: Chi-Square Tests(Females)

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	117.278[a]	102	.143
Likelihood Ratio	107.101	102	.345
Linear-by-Linear Association	.801	1	.371
N of Valid Cases	96		

Table7: Chi-Square Tests(Males)

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	123.394[a]	111	.198
Likelihood Ratio	128.444	111	.123
Linear-by-Linear Association	2.940	1	.086
N of Valid Cases	201		

3) Inference: From the above tables we reject the null hypothesis since p<0.05 and conclude that age is dependent on chest pain levels in both males and females.

(III) Age versus resting blood sugar levels in males and females.

H_0^1 : Age is independent of resting blood sugar levels in males

Vs H_1^1 : Age is dependent on resting blood sugar levels in males.

H_0^2 : Age is independent of resting blood sugar levels in females

Vs H_1^2 : Age is dependent on resting blood sugar levels in females.

Table8: Chi-Square Tests(Females)

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	1130.199[a]	1156	.701
Likelihood Ratio	396.110	1156	1.000
Linear-by-Linear Association	9.515	1	.002
N of Valid Cases	96		

Table9: Chi-Square Tests(Males)

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	1433.934[a]	1591	.998
Likelihood Ratio	625.593	1591	1.000
Linear-by-Linear Association	14.456	1	.000
N of Valid Cases	201		

4) Inference: From the above tables we reject the null hypothesis since p<0.05 and conclude that age is dependent on resting blood pressure levels in both males and females.

(IV) Age versus cholesterol levels in males and females.

H_0^1 : Age is independent of cholesterol levels in males

Vs H_1^1 : Age is dependent on cholesterol levels in males.

H_0^2 : Age is independent of cholesterol levels in females

Vs H_1^2 : Age is dependent on cholesterol levels in females.

Table10: Chi-Square Tests(Females)

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	2709.505[a]	2652	.214
Likelihood Ratio	600.781	2652	1.000
Linear-by-Linear Association	6.496	1	.011
N of Valid Cases	96		

Table11: Chi-Square Tests(Males)

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	4417.332[a]	4366	.290
Likelihood Ratio	1129.595	4366	1.000
Linear-by-Linear Association	3.819	1	.051
N of Valid Cases	201		

5) Inference: From the above tables, since p<0.05 we reject the null hypothesis and conclude that age is dependent on cholesterol levels in females. But in males, p>0.05 we accept the null hypothesis and conclude that age is independent of cholesterol levels in males.

(V) Age versus maximum heart rate achieved (thalach) in males and females.

H_0^1 : Age is independent of maximum heart rate achieved in males

Vs H_1^1 : Age is dependent on maximum heart rate achieved in males.

H_0^2 : Age is independent of maximum heart rate achieved in females

Vs H_1^2 : Age is dependent on maximum heart rate achieved in females.

Table12: Chi-Square Tests(Females)

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	1802.042[a]	1632	.002
Likelihood Ratio	493.540	1632	1.000
Linear-by-Linear Association	16.367	1	.000
N of Valid Cases	96		

Table 13: Chi-Square Tests(Males)

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	2980.651[a]	2997	.580
Likelihood Ratio	956.098	2997	1.000
Linear-by-Linear Association	32.300	1	.000
N of Valid Cases	201		

6) Inference: From the above tables we reject the null hypothesis since p<0.05 and conclude that age is dependent on maximum heart rate achieved (thalach) in both males and females.

III. CONCLUSIONS

A. The up slopping are flat slopping are dependent on exercise induced angina in females and males.
B. Age is dependent on chest pain levels in both males and females.
C. Age is dependent on resting blood pressure levels in both males and females.
D. Age is dependent on cholesterol levels in females. But, age is independent of cholesterol levels in males.
E. Age is dependent on maximum heart rate achieved (thalach) in both males and females.

REFERENCES

[1] The data was collected from the Cleveland Clinic Foundation (cleveland.data) Website: https://archive.ics.uci.edu/ml/datasets/Heart+Disease
[2] Montgomery, D.C. and Peck, E.A. and Vining, G.G.(2003), Introduction to Linear Regression, John Wiley, New York.
[3] Anderson, T.W. (2004), An Introduction to Multivariate Analysis, John Wiley, New York.
[4] Heart disease prediction using machine learning and data mining technique. (IJCSC) vol. 7 no.1 September 2015 – march 2016.