International Journal of Research
Available at https://edupediapublications.org/journals

p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 03 Issue 14
October 2016

Heart Disease Prediction (Empirically) Using Logistic Regression Analysis by Using SAS Software

Abdullah Mohammed Rashid

M.Sc (Applied Statistic's), Department of Statistic's,

University college of Science, Osmania University, India

Email Id: abdxxxx993@gmail.com

Abstract

The main objective of the present study is to investigate factors that contribute significantly to enhancing the risk of heart disease as well as accurately predict the overall risk. The dependent variable of the study is to diagnosis whether the patient has the disease or does not have the disease. Logistic regression analysis is applied for exploring the factors affecting the disease. The early prediction of cardiac diseases can aid in making decisions to lifestyle changes in high risk patients and in turn reduce their complications. This is a technique of using historical information on a certain attribute or event to identify patterns which will assist in predicting a future value of the same with a certain probability attached to it. Its application is invaluable in the field of medical sciences. This paper presents the steps involved in developing a Logistic Regression model based on patient's heart disease risk. The power of SAS in analyzing data patterns and developing such models is also demonstrated where appropriate and relevant portions of SAS code are included where ever possible.

Introduction:

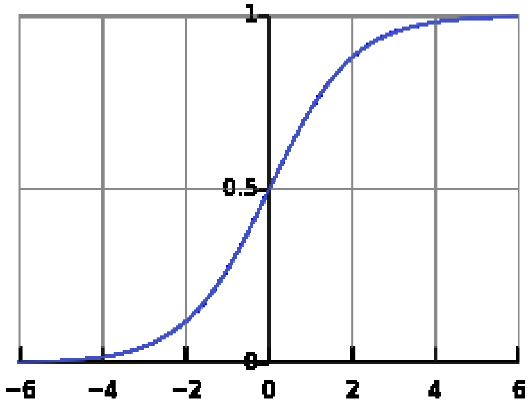
Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function, which is the cumulative logistic distribution. Thus, it treats the same set of problems as probit regression using similar techniques, with the latter using a cumulative normal distribution curve instead. Equivalently, in the latent variable interpretations of these two methods, the first assumes a standard logistic distribution of errors and the second a standard normal distribution of errors. Logistic regression can be seen as a special case of generalized linear model and thus analogous to linear regression. The model of logistic regression, however, is based on quite different assumptions (about the relationship between dependent and independent variables) from those of linear regression. In particular the key differences of these two models can be seen in the following two features of logistic regression. First, the conditional distribution y | x is a Bernoulli distribution rather than a Gaussian distribution, because the dependent variable is binary. Second, the predicted values are probabilities and are therefore restricted to (0, 1) through the logistic distribution function because logistic regression predicts the probability of particular outcomes.

Logistic regression is an alternative to Fisher's 1936 method, linear discriminant analysis [5]. If the assumptions of linear discriminant analysis hold, application of Bayes rule to reverse the conditioning results in the logistic model, so if linear discriminant assumptions are true, logistic regression assumptions must hold. The converse is not true, so the logistic model has fewer assumptions than discriminant analysis and makes no assumption on the distribution of the independent variables. Logistic Regression is a predictive analysis, like Linear Regression, but Logistic Regression involves predicting a dichotomous dependent variable. The predictors in a regression analysis can be continuous or dichotomous, but Ordinary Least Squares (OLS) Regression is not appropriate if the outcome is dichotomous. The OLS Regression uses a normal probability theory; whereas, Logistic Regression uses a binomial probability theory. The binomial probability theory makes this analysis a bit more complicated mathematically.

The logistic function:

The logistic function is useful because it can take an input with any value from negative to positive infinity, whereas the output always takes values between zero and one [6] and hence is interpretable as a probability. The logistic function $\sigma(t)$ is defined as below,

$$\sigma(t) = \frac{e^t}{e^t+1} = \frac{1}{1+e^{-t}} \qquad (9)$$



The standard logistic function $\sigma(t)$, $\sigma(t) \in (0,1)$ for all t.

The above graph is a logistic on the t-interval (-6, 6).

Now assume that t is a linear function of a single explanatory variable x (the case where t is a linear combination of multiple explanatory variables is treated similarly). We can then express t as below,

$$t = \beta_0 + \beta_1 x \qquad (10)$$

And the logistic function now can be written as,

$$F(x) = \frac{1}{1+ e^{-(\beta 0+ \beta 1x)}} \qquad (11)$$

F(x) is interpreted as the probability of the dependent variable equaling a 'success' or 'case' rather than a failure or non-case. It is clear that the response variables Y_i are not identically distributed: P (Y_i = 1 | X) differs from one data point X_i to another, though they are independent given design matrix X and shared with parameters β.

Odds Ratio:

The odds of the dependent variable equaling a case (given some linear combination x of the predictors) is equivalent to the exponential function of the linear regression expression. This illustrates how the logit serves as a link function between the probability and the linear regression expression. Given that the logit ranges between negative and positive infinity, it provides an adequate criterion upon which to conduct linear regression and the logit is easily converted back into the odds. [6]

So we define odds of the dependent variable equaling a case (given some linear combination x of the predictors) as below,

$$\text{Odds} = e^{\beta 0+ \beta 1x} \qquad (12)$$

For a continuous independent variable the odds ratio can be defined as,

$$\text{OR} = \frac{odds\,(x+1)}{odds\,(x)} = \frac{\frac{F(x+1)}{1-F(x+1)}}{\frac{F(x)}{1-F(x)}} = \frac{e^{\beta 0+ \beta 1(x+1)}}{e^{\beta 0+ \beta 1x}}$$

$$= e^{\beta 1} \qquad (13)$$

This exponential relationship provides an interpretation for β_1: The odds multiply by $e^{\beta 1}$ for every 1-unit increase in x.

For a binary independent variable the odds ratio is defined as ad/bc where a, b, c and d are cells in a 2x2 contingency table. [7]

Assumptions of logistic Regression:

- Logistic regression does not assume a linear relationship between the dependent and independent variables.
- The dependent variable must be a dichotomy (2 categories).
- The independent variables need not be interval, non-normally distributed, non-linearly related, nor of equal variance within each group.
- The categories (groups) must be mutually exclusive and exhaustive; a case can only be in one group and every case must be a member of one of the groups.
- Larger samples are needed than for linear regression because maximum likelihood coefficients are large sample estimates. A

minimum of 50 cases per predictor is recommended.

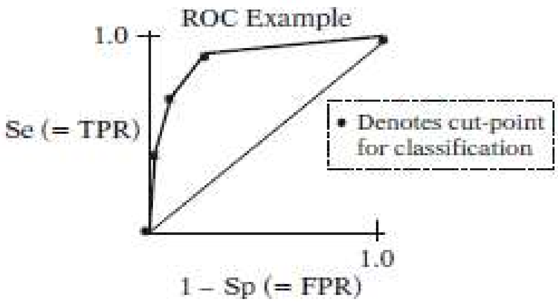
Application of Logit model:

1. It can be used to identify the factors that affect the adoption of a particular technology say, use of new varieties, fertilizers, pesticides etc., on a farm.
2. In the field of marketing, it can be used to test the brand preference and brand loyalty for any product.
3. Gender studies can use logit analysis to find the factors which affect the decision making status of men/women in a family.
4. We want to model the probabilities of a response variable as a function of some explanatory variables, e.g. "success" of admission as a function of gender.
5. we want to perform descriptive discriminate analyses such as describing the differences between individuals in separate groups as a function of explanatory variables, e.g. student admitted and rejected as a function of gender
6. we want to predict probabilities that individuals fall into two categories of the binary response as a function of some explanatory variables, e.g. what is the probability that a student is admitted given she is a female
7. We want to classify individuals into two categories based on explanatory variables, e.g. classify new students into "admitted" or "rejected" group depending on their gender.

ROC Curve:

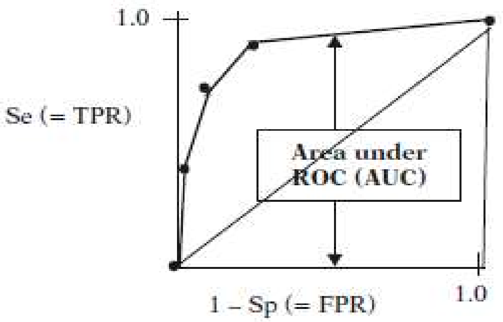
A Receiver Operating Curve (ROC) is a plot of sensitivity (Se) by 1 – specificity (1 – Sp) values derived from several classification tables corresponding to different cut-points used to classify subjects into one of two groups, e.g., predicted cases and non-cases of a disease.

Equivalently, the ROC is a plot of the true positive rate (TPR = Se) by the false positive rate (FPR = 1 - Sp).



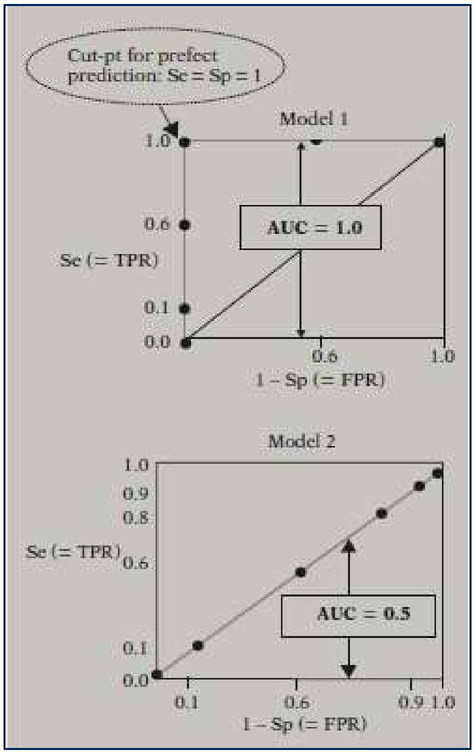
The ROC was first developed by electrical engineers and radar engineers during World War II for detecting enemy objects in battle fields, also known as the signal detection theory; in this situation, a signal represents the predicted probability that a given object is an enemy weapon. ROC analysis is now widely used in medicine, radiology, psychology and, more recently in the areas of machine learning and data mining.

More specifically, an ROC provides an appropriate answer to the question we previously asked when we compared classification tables for two models: How often will a randomly chosen (true) case have a higher probability of being predicted to be a case than a randomly chosen true non-case?



Moreover, we will see that the answer to this question can be quantified by obtaining the area under an ROC curve (AUC): the larger the area, the better the discrimination. First, we provide the two ROCs derived from hypothetical Models 1 and 2 that we considered in the previous section. Notice that the ROC for each model is determined by connecting the dots that plot pairs of Se and 1 – Sp values obtained for several classification cut-points.

For Model 1, the area under the ROC is 1.0.



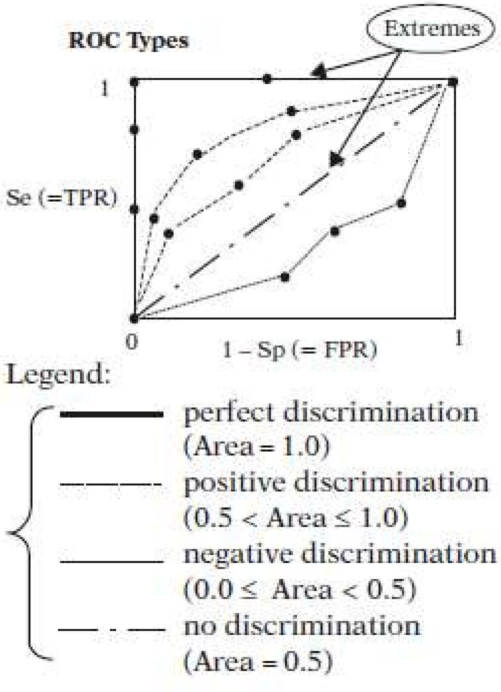
In contrast, for Model 2, the area under the ROC is 0.5.

Since the area under the ROC for Model 1 is twice that for Model 2, we would conclude that Model 1 has better discriminatory performance than Model 2. The AUC measures discrimination, that is, the ability of the model to correctly classify those with and without the disease. We would expect a model that provides good discrimination to have the property that true cases have a higher predicted probability (of being classified as a case) than true non-cases. In other words, we would expect the true positive rate (TPR = Se) to be higher than the false positive rate (FPR = 1 - Sp) for all cut-points.

Observing the above ROCs, we see that, for Model 1, TPR (i.e., Se) is consistently higher than its corresponding FPR (i.e., 1 - Sp); so, this indicates that Model 1 does well in differentiating the true cases from the true non-cases.

In contrast, for Model 2 corresponding true positive and false positive rates are always equal, which indicates that Model 2 fails to differentiate true cases from true non-cases.

The two ROCs we have shown actually represent two extremes of what typically results for such plots. Model 1 gives perfect discrimination whereas Model 2 gives no discrimination.



We show in the figure at the left several different types of ROCs that may occur. Typically, as shown by the two dashed curves, the ROC plot will lie above the central diagonal (45^0) line that corresponds to Se = 1 - Sp; for such curves, the AUC is at least 0.5. It is also possible that the ROC may lie completely below the diagonal line, as shown by the dotted curve near the bottom of the figure, in which case the AUC is less than 0.5. This situation indicates negative discrimination, i.e., the model predicts true non-cases better (i.e., higher predicted probability) than it predicts true cases. An AUC of exactly 0.5 indicates that the model provides no discrimination, i.e., predicting the case/non-case status of a randomly selected subject is equivalent to flipping a fair coin.

A rough guide for grading the discriminatory performance indicated by the AUC follows the traditional academic point system, as shown below [29].

0.90–1.0 = excellent discrimination (A)
0.80–0.90= good discrimination (B)
0.70–0.80 = fair discrimination (C)
0.60–0.70 = poor discrimination (D)
0.50–0.60 = failed discrimination (F)

However, that it is typically unusual to obtain an AUC as high as 0.90, and if so, almost all exposed subjects are cases and almost all unexposed subjects are non-cases (i.e., there is nearly complete separation of data points). When there is such “complete separation,” it is impossible as well as unnecessary to fit a logistic model to the data.

Results:

Data and Variables:

The data was collected for 500 patients from KIMS hospital, Hyderabad. Performed a study to investigate factors that contribute significantly to enhancing the risk of heart disease. The dependent variable of the study is diagnosis whether the patient has the disease or does not have the disease. Logistic regression analysis was applied for exploring the factors affecting the disease. The entire analysis was done using SAS 9.2 software [20].

The independent variables used in this study were – Age, Chest Pain, RestBP, BP, Rest Electro, Max Heart rate and Gender. Did some codification to these independent variables in order to make use in modelling process.

The SAS System

The FREQ Procedure

Age	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	110	22.00	110	22.00
2	143	28.60	253	50.60
3	181	36.20	434	86.80
4	66	13.20	500	100.00

Variable: Chest_Pain ; The variable Chest_Pain was coded as below.

Chest Pain	Code
asympt	1
atyp_angina	2
non_anginal	3
typ_angina	4

Chest_pain	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	153	30.60	153	30.60
2	143	28.60	296	59.20
3	107	21.40	403	80.60
4	97	19.40	500	100.00

Variable: Rest_bpress ; Variable Rest_bpress was codified as below.

Rest_bpress	Code
98-130	1
131-160	2
161-200	3

Rest_bpress	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	165	33.00	165	33.00
2	157	31.40	322	64.40
3	178	35.60	500	100.00

Variable: Blood_Sugar

Blood_Sugar	Code
TRUE	1
FALSE	0

lood_Sugar	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	300	60.00	300	60.00
1	200	40.00	500	100.00

Variable: Rest_electro

Rest_Electro	Code
Normal	1
Left_Vent_Hyper	2
St_t_Wave_abnormality	3

Rest_electro	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	213	42.60	213	42.60
2	143	28.60	356	71.20
3	144	28.80	500	100.00

Variable: Max_heart_rate

Max_Heat_Rate	Code
87-120	1
121-150	2
151-188	3

Max_heart_rate	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	159	31.80	159	31.80
2	202	40.40	361	72.20
3	139	27.80	500	100.00

Variable: Exercise_angina

Exercise	Code
Yes	1
No	0

Exercise_angina	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	265	53.00	265	53.00
1	235	47.00	500	100.00

Variable: Disease

Disease	Code
Positive	1
Negative	0

Disease	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	432	86.40	432	86.40
1	68	13.60	500	100.00

Variable: Height

Hieght	Code
5.2-5.3	1
5.4-5.5	2
5.6-5.8	3

Height	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	121	24.20	121	24.20
2	170	34.00	291	58.20
3	209	41.80	500	100.00

Variable: Gender

Gender	Code
Male	1
Female	0

Gender	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	258	51.60	258	51.60
1	242	48.40	500	100.00

The below is the SAS code to fit logistic regression model.

```
#@ Modified Variables: AgeChest_pain  Rest_bpress Blood_Sugar Rest_electro
      Max_heart_rate    Exercise_angina   Disease     Hight Gender ;
#Raw variables:  Age    Chest_pain  Rest_bpress_OLD   Blood_Sugar
      Rest_electroMax_heart_rate_OLDExercise_angina   Disease     Hight_OLD
      Gender;


/* Fitting Logistic Regression */
proc logistic data=Disease_Data Desc;
model Disease = AgeChest_pain  Rest_bpress Blood_Sugar Rest_electro
      Max_heart_rate    Exercise_angina   Hight Gender/lackfit
outroc=rocdata;
output out = logmodel p=pred;
run;


/* ROC Plot */
proc gplot data=rocdata;
plot _SENSIT_*_1MSPEC_;
run;
quit;
```

The below is output for Logistic Regression analysis

The SAS System

The LOGISTIC Procedure

Model Information	
Data Set	WORK.Disease_Data
Response Variable	Disease
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	500
Number of Observations Used	500

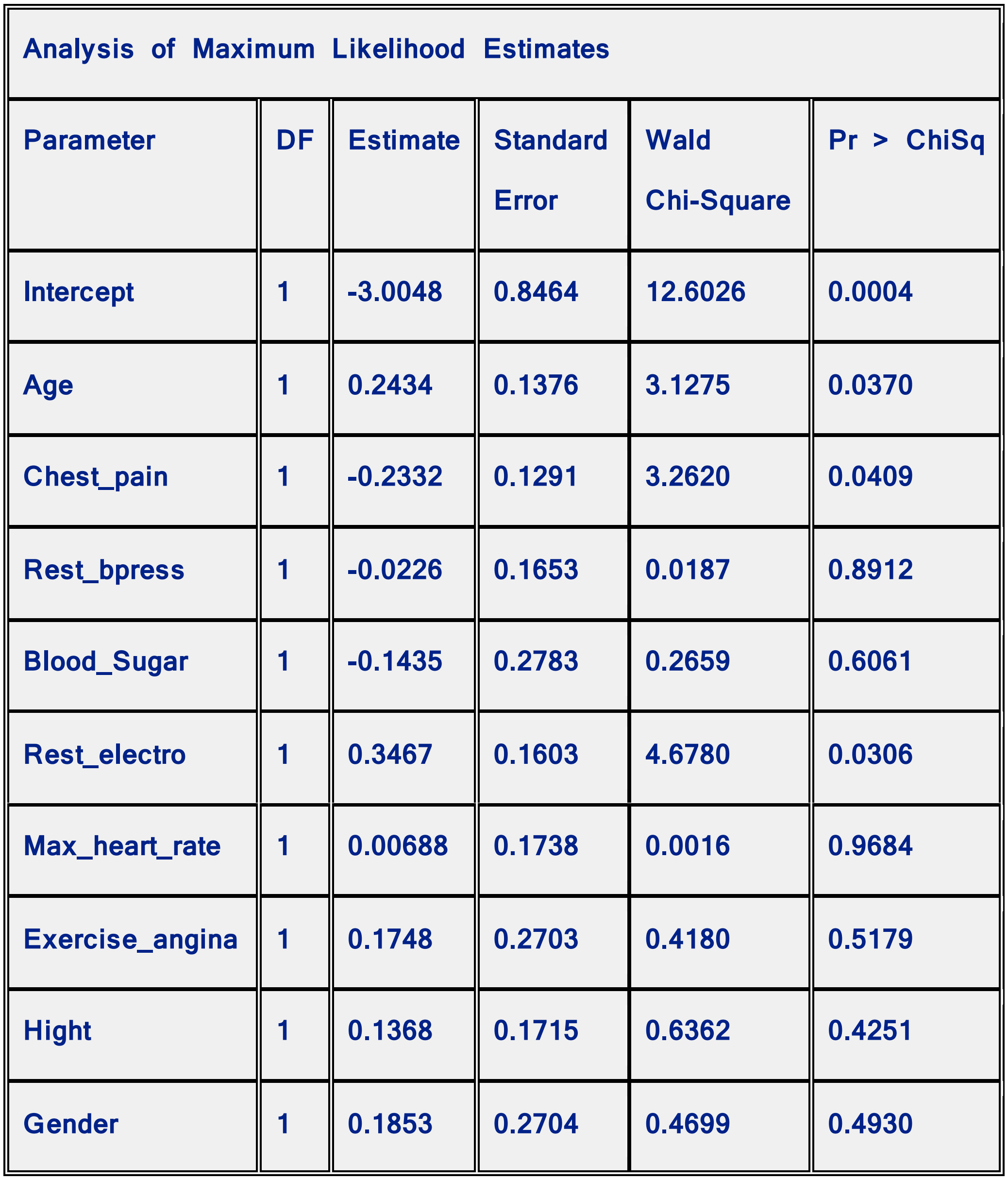
Response Profile		
Ordered Value	Disease	Total Frequency
1	1	68
2	0	432

Probability modeled is Disease=1.

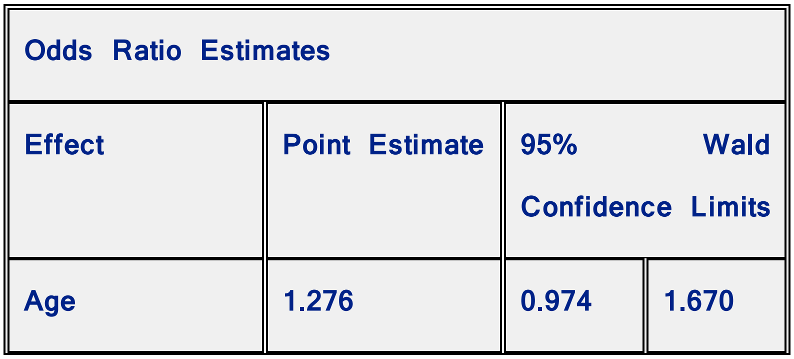
Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	399.635	404.413
SC	403.850	446.559
-2 Log L	397.635	384.413

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	13.2228	9	0.1528
Score	13.3960	9	0.1455
Wald	12.9345	9	0.1656



Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-3.0048	0.8464	12.6026	0.0004
Age	1	0.2434	0.1376	3.1275	0.0370
Chest_pain	1	-0.2332	0.1291	3.2620	0.0409
Rest_bpress	1	-0.0226	0.1653	0.0187	0.8912
Blood_Sugar	1	-0.1435	0.2783	0.2659	0.6061
Rest_electro	1	0.3467	0.1603	4.6780	0.0306
Max_heart_rate	1	0.00688	0.1738	0.0016	0.9684
Exercise_angina	1	0.1748	0.2703	0.4180	0.5179
Hight	1	0.1368	0.1715	0.6362	0.4251
Gender	1	0.1853	0.2704	0.4699	0.4930

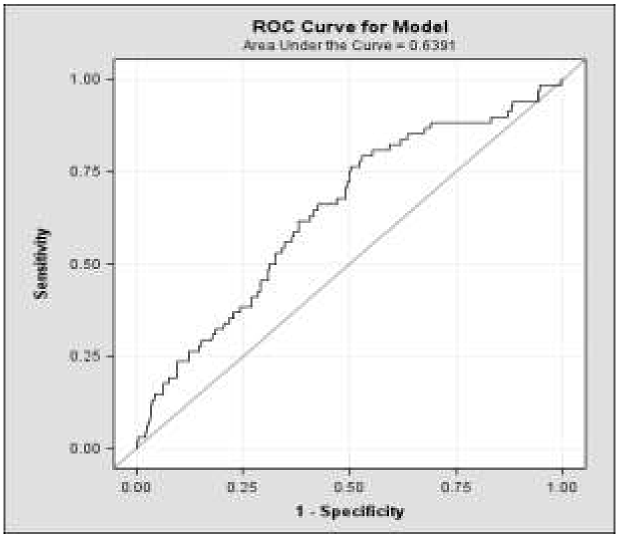


Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Age	1.276	0.974	1.670

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Chest_pain	0.792	0.615	1.020
Rest_bpress	0.978	0.707	1.352
Blood_Sugar	0.866	0.502	1.495
Rest_electro	1.414	1.033	1.936
Max_heart_rate	1.007	0.716	1.416
Exercise_angina	1.191	0.701	2.023
Hight	1.147	0.819	1.605
Gender	1.204	0.709	2.045

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	63.9	Somers' D	0.278
Percent Discordant	36.1	Gamma	0.278
Percent Tied	0.0	Tau-a	0.066
Pairs	29376	C	0.639

ROC Plot:



Hosmer-Lemeshow Test:

Partition for the Hosmer and Lemeshow Test					
Group	Total	Disease = 1		Disease = 0	
		Observed	Expected	Observed	Expected
1	50	4	3.13	46	46.87
2	50	4	4.16	46	45.84
3	51	2	4.86	49	46.14
4	50	3	5.28	47	44.72
5	50	9	5.82	41	44.18

Partition for the Hosmer and Lemeshow Test					
Group	Total	Disease = 1		Disease = 0	
		Observed	Expected	Observed	Expected
6	50	6	6.47	44	43.53
7	50	12	7.31	38	42.69
8	50	7	8.30	43	41.70
9	50	8	9.90	42	40.10
10	49	13	12.77	36	36.23

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
9.4412	8	0.3065

Conclusions :

The main objective of this study is to investigate factors that contribute significantly to enhancing the risk of heart disease. The dependent variable of the study is diagnosis whether the patient has the disease or does not have the disease. From the coefficients of logistic regression model it is observed that the variables- Age, Chest_pain and Rest_eletro are turned as very significant. It means that there is a highly important factor in order to predict the risk of heart disease. From the HL Test statistic, the P value is 0.306, which is greater than 0.05. So there is an evidence not to reject the null hypothesis. This means that, there is similarity between observed and predicted values. Since the percentage of Concordance is 63.9, it is reasonably good; and from ROC plot the AUC value is moderately good (0.6391) hence we can say that our model will perform well to predict the risk of heart disease for a particular patient.

References:

[i.] Cramer, J. S. (2002). The origins of logistic regression.

[ii.] Vasisht, A. K. (2007). Logit and probit analysis. *From< ww. iasri. res. in/.../5-Logit% 20and% Probit, 20.*

[iii.] Walker, S. H., & Duncan, D. B. (1967). Estimation of the probability of an event as a function of several independent variables. *Biometrika*, *54*(1-2), 167-179.

[iv.] Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, 215-242.

[v.] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 6). New York: springer.

[vi.] Hosmer Jr, D. W., & Lemeshow, S. (2004). *Applied logistic regression.* John Wiley & Sons.

[vii.] Mitchell, L. (2007). The cambridge dictionary of statistics Everitt BS (ed.)(2006) ISBN: 0521690277; 432 pages;£ 17.99, $40.50 Cambridge University Press; http://www. cambridge. org.

[viii.] Menard, S. (2002). *Applied logistic regression analysis* (No. 106). Sage.

[ix.] Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of clinical epidemiology*, *49*(12), 1373-1379.

[x.] Park, H. (2013). An introduction to logistic regression: from basic concepts to interpretation with particular attention to nursing domain. *Journal of Korean Academy of Nursing*, *43*(2), 154-164.

[xi.] Hosmer, D. W., Hosmer, T., Le Cessie, S., & Lemeshow, S. (1997). A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in medicine*, *16*(9), 965-980.

[xii.] Menard, S. (2002). *Applied logistic regression analysis* (No. 106). Sage.

[xiii.] Bewick, V., Cheek, L., & Ball, J. (2005). Statistics review 14: Logistic regression. *Critical Care*, *9*(1), 1.

[xiv.] Katz, M. H. (2011). *Multivariable analysis: a practical guide for clinicians and public health researchers*. Cambridge university press.

[xv.] Morris, J. A., & Gardner, M. J. (1988). Statistics in Medicine: Calculating confidence intervals for relative risks (odds ratios) and standardised ratios and rates. *British medical journal (Clinical research ed.)*, *296*(6632), 1313.

[xvi.] Morris, J. A., & Gardner, M. J. (1988). Statistics in Medicine: Calculating confidence intervals for relative risks (odds ratios) and standardised ratios and rates. *British medical journal (Clinical research ed.)*, *296*(6632), 1313.

[xvii.] Tjur, T. (2009). Coefficients of determination in logistic regression models—A new proposal: The coefficient of discrimination. *The American Statistician*, *63*(4), 366-372.

[xviii.] Bewick, V., Cheek, L., & Ball, J. (2004). Statistics review 13: receiver operating characteristic curves. *Critical care*, *8*(6), 1.

[xix.] Allison, P. D. (2012). *Logistic regression using SAS: Theory and application*. SAS Institute.

[xx.] Hilbe, J. M. (2009). *Logistic regression models*. CRC press.