Prediction, analysis and estimation of factors affecting cardiac patients using regression models

Suhair Jameel Mushrif Al-Neyazy

University of Baghdad

Abstract: The heart disease, myocardial infarction, heart failure, and stroke, in addition to high blood pressure, are among the most dangerous diseases for human life, as it may result in many complications that lead to death. Therefore, the study aimed to estimate a logistic regression model and predict the probability of heart disease, in addition to identifying the most important statistical methods in analyzing data of heart patients depending on the factors affecting their incidence (patient's gender, patient age, smoking, blood pressure).

Keywords: regression, unbiased estimator, multi-linear.

Introduction

High blood pressure is a pathological condition that occurs as a result of the increase in the pressure force of the blood on the inner walls of the blood vessels. This rise in the long run leads to many complications in the body, including cardiovascular disease, myocardial infarction, heart failure and stroke, in addition to kidney failure. High blood pressure is a problem. It is a major public health problem, as it causes the death of (7.5) million people annually, which is more than ((1.12 of the total deaths in the world. It is estimated that high blood pressure affects the Eastern Mediterranean region about 40%)) of the adult population over the age of 25 (years and above, about 60%) does not know a clear and specific cause of infection with the disease, but there are many risk factors that can raise the incidence of it, and this type is called “high blood pressure, which leads to heart disease.” In this research, The use of independent variables (influencing) (patient's gender, patient's age, smoking, blood pressure,) to predict the infection of these patients through the patient database to know which of the factors affecting the prediction of the probability of infection so that the probability of heart disease can be estimated. For the purpose of studying the effect of these factors, we analyzed them using the multiple linear regression method. We used a sample of 300 views and chose a specific age group from (40-90) years.

Search objective:

This research aims to determine the factors affecting heart disease through the use of multiple linear regression and to indicate the importance of each variable in relation to the other variables.

The study problem:

Affecting factors (patient's gender, patient's age, smoking, blood pressure) and their impact on heart disease. Factor analysis, logistic model estimation, and prediction of the probability of infection in the dependent variable $y_i$and the independent variables x_i.

The importance of studying:

The importance of this study comes to the problem of heart disease by identifying the factors affecting it using the logistic regression model and identifying them with the injury variable and the model's classification of patients to injured or uninjured and how to implement the prediction of the logistic regression model.

Regression analysis (linear regression):

Regression analysis is a statistical tool that builds a statistical model in order to estimate the relationship between one quantitative variable and the dependent or dependent variable and another quantitative variable or several quantitative variables which are the independent variables so that it produces a statistical equation that shows the relationship between the variables.

This equation can be used to find out the type of relationship between the variables and to estimate the dependent variable using other variables.

Also when the relationship in the statistical model is between a dependent variable (dependent) and one independent variable, then this model is the simplest regression model, and the model is called simple liner regression, and when several independent variables are more than one quantitative variable, the model is called multiple regression. Multiple linear regression will be used to fit the data used in the research.

Multiple linear regression is not just a single method, but rather a set of methods that can be used to find out the relationship between a continuous dependent variable and a number of independent variables that are usually continuous. The linear equation in multiple linear regression is:

Yi=B0+B1xi1+B2xi2+….+Bkxik+ui

Whereas:

Yi: represents the dependent variable.

B0: a constant.

B1: the slope of the regression Y on the first independent variable.

B2: the slope of the regression y on the second independent variable.

Bk: the slope of the regression y on the last independent variable k. (x1,x2,…….,xk) independent variables.

Ui: random error.

Multiple linear regression can be used if the following conditions are met:

1- The relationship between the independent variables and the dependent variable must be linear.

2- The data should be normally distributed for the independent variables and the dependent variable.

3- The dependent variable values must be of at least ordinal level.

After obtaining the results of the regression equation, we must show whether these coefficients are statistically acceptable, that is, statistically significant, noting that the significance is for each coefficient separately. In order to judge the significance of the regression coefficients, we use the T-test and the corresponding probability level, of course.

spss will automatically extract the T-test and its level of probability. It will also obtain statistics used to know the significance, including R is the simple correlation coefficient, which measures the strength of the relationship between one independent variable with one dependent variable, and r^2 is used to interpret the explanatory power of the multiple linear regression model because it takes into account independent variables.

We also use the f-statistic to judge the significance of the estimated model as a whole at a significant level.

Use the estimated least squares method to estimate parameters $B_0, B_1, B_2, \ldots\ldots, B_k$

Y=XB+U

U=Y-XB

$$\sum u_{i^2} \approx \min$$

$$\sum u_{i^2} = u^2 1 + u^2 2 + \ldots\ldots + u^2 n$$

$$= (\mathrm{U}_1 U_2 U_3 \ldots\ldots \mathrm{U}_n)\begin{bmatrix} U_1 \\ U_2 \\ U_3 \end{bmatrix}$$

$$U'U = (Y - XB)'(\mathrm{Y} - \mathrm{XB})$$

$$= (\mathrm{Y}' - \mathrm{X}'\mathrm{B}')(\mathrm{Y} - \mathrm{XB})$$

$$= (\mathrm{Y}'\mathrm{Y} - \mathrm{Y}'\mathrm{XB} - \mathrm{B}'\mathrm{X}'\mathrm{Y} + \mathrm{B}'\mathrm{X}'\mathrm{Y} + \mathrm{B}'\mathrm{X}'\mathrm{XB})$$

$$=$$

It is noticed from the above formula that the second and third terms are of a specific value and have nothing to do with the matrix and therefore they can be summed

$$=Y\,Y' - 2B'X\,X'B$$

$$\frac{\partial y}{\partial x} = -2x\,y' + 2x\,'xb = 0$$

$$x\,'xb = x\,y'$$

$$b = (x\,'x\,)^{-1}x\,y'\,\ldots\ldots..$$

The covariance and covariance matrix of the estimated parameters

As:

Y=xb+u

Substitute into the above equation

b=(x'x)$^{-1}$*x'y*

Produce

$\bar{B}$ *= B + =(x'x)*$^{-1}$*x'y*

E($\bar{B}$*) = B*

Because

Eu = B

The above result means that (b) is nothing but an 'unbiased estimator'

Rewriting the above formula, we get

$$\mathrm{var-cov}(b)\,\mathrm{var-cov}(B)$$

$$= E\,[(b-B)(b-B)']$$

$$= E\begin{bmatrix} b^{0} \\ b_1 \\ b_k \end{bmatrix}\begin{bmatrix} B_0 \\ B_1 \\ B_K \end{bmatrix}[(\mathrm{b}_0 - B_0)(\mathrm{b}_1 - B_1)\ldots\ldots\ldots..(b_k - B_K)]$$

$$\begin{pmatrix} E\,(b_0 - B_0)^2 \ldots.\mathrm{E}(\mathrm{b}_0 - B_0)(\mathrm{b}_1 - B_1)\ldots\ldots...\mathrm{E}(\mathrm{b}_0 - \mathrm{B}_0)(\mathrm{b}_k - B_K) \\ \vdots \\ \vdots \\ E\,(b_k - B_k)(b_0 - B_0)\ldots..E\,(\mathrm{b}_k - B_K)(b_1 - B_1)\ldots\ldots..E\,(b_k - B_k)^2 \end{pmatrix}$$

This matrix is known as the covariance and covariance matrix for the estimated features, whereby the locations of each of the variances for the estimated features represented by the matrix diameter can be determined.

As for the elements outside the diameter range, they represent the common variance between any two of these estimated parameters.

As for calculating the values of each of the variances of these parameters and the common variance between them, it can be reached after substituting the previous formula with its equivalent, as follows:

$$\text{var}-\text{cov}(b) = E[(\text{x}'\text{x})^{-1}x'u][(x'x)^{-1}x'u]'$$
$$= E[(x'x)^{-1}x'uu'x(x'x)^{-1}$$
$$= (6\ln x'x)^{-1}x'x(x'x)^{-1}$$
$$\text{var}-\text{cov}(b) = \sigma^2\ln(x'x)^{-1}x'x(x'x)^{-1}$$
$$\text{var}-\text{cov}(b) = \sigma^2(x^{-1}x)^{-1}$$

Analysis of deviations in the case of a multi-linear model

The sum of the squares of the deviations is $E(y_i - \hat{y})^2$

The first is the sum of the squares of the deviations shown $E(\hat{y}_i - \bar{y})^2$ The second part represents the sum of the squares of the unexplained deviations (the sum of the residuals).

Meaning:

$$\sum_{i=1}^{n}(\text{y}_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{\text{y}}_i - \bar{y})^2 + \sum_{i=1}^{n}ui^2 \quad \text{…………………} (1)$$

Since the coefficient of determination R^2 is considered a key influence in evaluating the extent of the significance of the assumed relationship between the dependent variable and the independent variable, the same role takes it in the case of the general linear model, where it is called the multiple determination coefficient and is symbolized by the symbol (RY 1,2,……….K), where The sequence (1,2,…………K) refers to the independent variables

Using matrices to express all sources of deviations in the multiple model, as follows:

$$e'e = (\text{Y}-\text{XB})'(\text{Y}-\text{Xb})$$
$$\text{e}'\text{e} = (\text{Y}'\text{Y}-2\text{B}'\text{X}'\text{Y}+B'X'XB)$$

Substituting in the value of (B) we get:

$$e'e = Y'Y - 2B'X'Y + B'X'X(X'X)^{-1}X'Y$$
$$e'e = Y'Y - 2B'X'Y + B'X'Y$$
$$e'e = Y'Y - B'X'Y$$

Compared with formula No. (1), the sum of the total deviations can be rewritten in terms of matrices as follows:

$$Y'Y = B'X'Y + e'e ....(2)$$

Relationship No. (2) above shows the main sources of deviations of the wave observations of the dependent variable (y).

whereas :

$Y'Y$: represents the total deviations SST

$B'X'Y = \widehat{Y}'\widehat{Y}$ represent the deviations shown SSR

$e'e$ = represent the deviations not shown

Since the

$$R^2 = \frac{SSR}{SST}$$

$$R^2 = \frac{B'X'Y}{Y'Y} = \frac{B'X'Y}{\sum Yi^2} ....(3)$$

Accordingly, the three sources of deviation can be converted in terms of the coefficient of determination as follows:

From formula No. (3) we have

$$(\mathrm{Y'Y})\,\mathrm{R}^2 = B'X'Y ....(4)$$

Or

$$R^2 \sum Y_i^2 = B'X'Y$$

From formula No. (4) represents the deviations shown in terms of the multiple determination coefficient from formula No. (2) after compensation and rearranging we get

$$e'e = Y'Y - Y'YR^2$$

$$e'e = Y'Y\,(1 - R^2)........(5)$$

Or

$$\sum ei^2 = (1-R^2)\sum Yi^2$$

Formula No. (5) above represents the unexplained deviations in terms of the multiple determination coefficient, which are formulas (1) and (2) the cornerstone in building the analysis of variance table shown below for (k) variables

calculated F	**Average sum of squares**	**Sum of squares**	**Degree** freedom	**S.O.V**
$F_0 = \dfrac{R^2 / K}{(1-R^2) / n-K-1}$	$R^2 Y'Y / K$	$B'X'Y = R^2 Y'Y$	**K**	**Deviations shown SSR**
	$(1-R^2)Y'Y / n-k-1$	$e'e = (1-R^2)Y'Y$	**n-k-1**	**Deviations not shown SSE**
		$Y'Y$	**n-1**	**Total deviations SST**

Thus, the value of (F0) the calculated operation from the table above can be compared for the degree of freedom (k), (nk-1) for a certain level of significance with its tabular counterpart. There is no effect of any of the independent variables (X1,X2,……..XK) on the dependent variable (Yi).

Either if the value of (F0) the calculated process is greater than its tabular value, then this means that the studied linear relationship is significant and there is an effect and its relationship is close between the independent variable and the dependent variable.

null hypothesis:

H_0:B_1=B_2=………B_k=0

Alternative Hypothesis

$$H_1 : B_1 \neq B_2 \neq ....B_k \neq 0$$

Selection factor (R^2) Adjusted

The characteristic of the coefficient of determination R^2 is that if an independent variable is added to the model, its value will rise even if the added variable is not of the importance it deserves with it being included in the model, so for the purpose of obtaining a better criterion to measure the extent to which different groups of variables are capable of analyzing the relationship under study and at the same time takes into account the number of included variables is calculated by the so-called corrected coefficient of determination, R'^2, which is calculated according to the following formula:

$$R'^2 = 1 - \frac{n-1}{n-k-1}(1-R^2)$$

It is noted that the value of R^2 will decrease when adding an independent variable if this addition does not lead to a reduction (1-R^2) to compensate for the increase that occurs in (n-1/nk-1) as a result of the increase in the value of k. In other words, it is better not to add a variable to the model. If adding it causes a devaluation of R^2.

Partial correlation coefficient:

The partial correlation coefficient represents the net correlation between the dependent variable and the independent variable after deleting the combined effect of the rest of the independent variables on both the dependent variable and the independent variable.

That is, after fixing the other variables in the model, for example *($PX_1Y.X_2$) or ($P1Y._2$),* it means the partial correlation coefficient between *X1* and Y after deleting the effect of *X_2* on both *Y* and the multiple, knowing that its value is limited between (-1) and (+1) and takes the sign of the corresponding parameter is calculated according to the following formula:

$$P1Y\ .2 = \frac{p_1y - p_{12}p_2y}{\sqrt{(1-p^2{}_{12})(1-p_2y^2)}}$$

Either the other partial correlations are calculated as follows:

$$p_1y\cdot_{23} = \frac{p_1y\ .3y - p_{12}.3p_2y\cdot_3}{\sqrt{(1-p_{12.3})^2(1-p_2y\cdot_3)^2}}$$

$$p_1y\cdot_{234} = \frac{p_1y\ .3y - p_{12.34} - p_2y\cdot_{34}}{\sqrt{(1-p_{12.34})^2(1-p_2y\cdot_{34})^2}}$$

Thus, we can conclude for other cases.

Linear model problems

In the previous study of regression models, basic assumptions were relied upon. Therefore, the accuracy of estimating the parameters of the model in applied reality depends on the validity of these hypotheses. If some of these hypotheses are inaccurate, we have some problems.

1- The problem of heterogeneity of variance

In general, we face the problem of heterogeneity of variance in the case of estimating the parameters of multiple models on cross-sectional data, where there is a large variance in the values of their variables. As a result, the hypothesis takes the following form:

$$E(ei^2) = \sigma^2 ei$$
$$E(e_i e_j) = 0 ....\forall i \neq j$$

In the case of the general linear model, the above hypothesis takes the following form:

$$E(ee')\begin{pmatrix} \sigma_1^2 & 0.......... & 0 \\ 0 & \sigma_2^2......... & 0 \\ 0 & 0........... & \sigma_n^2 \end{pmatrix}$$

Whereas:

$$\sigma_1^2 \neq \sigma_2^2 \neq .... \neq \sigma_n^2$$

Under the hypothesis of heterogeneity of variance above, the use of the least squares (OLS) method to estimate the parameters of the model is not feasible, as the parameters will not be the best unbiased linear estimate, in other words, the parameters estimated by this method do not have the least variance property.

2- Testing for the existence of a heterogeneity problem

Several tests have been developed to detect the problem of error heterogeneity, and the simplest of these tests is the Spearman Rank Correlation Test (Spearman), which depends on the absolute values of errors resulting from the difference between the true value and the estimated values of the dependent variable and the values of the independent variable. Calculating this indicator requires estimating the parameters of the model Linearity using the (OLS) method first, from which the deviations are calculated, then the Spearman coefficient of rank correlation is extracted according to the following formula:

$$r_{ex} = 1 - \frac{6\sum_{i=1}^{n} Di^2}{n(n^2-1)}$$

whereas:

n: is the sample size.

Di: represents the difference between the rank of the absolute values of the deviations and the rank of the independent variable under research, and whenever the value of the rank correlation coefficient is high and close to the correct one, this indicates a strong relationship between the deviations and the independent variable, and thus the existence of the problem of heterogeneity of error variance.

In addition, we can test the significance of the homogeneity of the error variance in the sample under research, where in such a case the value of the standard deviation of the Spearman coefficient of rank correlation must be calculated according to the following formula:

$$S_{(r_{ex})} = \frac{1}{\sqrt{n-1}}$$

3- The problem of autocorrelation

The autocorrelation problem appears in the case of a relationship between the values of random errors in the linear model, especially when using time series data in measuring model variables, as the random error in each time period depends on the errors of the previous time periods. In addition, such a problem may appear as a result of the procedure a modification in the data used for the purposes of estimating the parameters of the studied model, such as resorting to estimating the values of some observations of the model variables, and since this estimation process usually depends on taking the averages of the values of successive observations, and this in turn leads to creating a relationship between the values of random errors of the linear model. Therefore, it is necessary to reconsider the hypothesis of the lack of common variance between random errors when estimating the parameters of the linear model, whether it is a simple or multiple model, in other words, reconsidering the following hypothesis:

$$E\,(e_i e_j) = 0 \ldots\ldots\ldots \forall\, \mathrm{i} \neq j$$

Which is correct in light of the above as follows:

$$E\,(e_i e_j) \neq 0 \ldots\ldots\ldots\ldots \forall i \neq j$$

That is, there is an autocorrelation between the values of random errors. One of the simplest and most common types of autocorrelation is first-degree autocorrelation, according to which the random error for each period depends linearly on the random error of the previous periods.

$$e_i = pe_i + v_i$$

Since the random error in the above model takes the same basic assumptions when applying Ordinary Least Squares (OLS) in other words:

$$v_i \;\square\; N\,(0, \sigma^2) \ldots\ldots \forall i \neq j$$

$$E\,(v_i v_j) = 0$$

Test for the existence of the autocorrelation problem:

There are several tests to detect the existence of the autocorrelation problem, the most important and most widely used is what is known as the Durbin Watson (D-W) test, as it is suitable for testing the existence of the first order autocorrelation problem, whereby the null hypothesis is placed against the alternative hypothesis:

$$H_0 : \Omega = 0$$

$$H_1 : \Omega \neq 0$$

The estimated random errors are used in calculating the (D-W) formula as follows:

$$D - W = \frac{\sum_{i=1}^{n} (e_i - e_i - 1)^2}{\sum_{i=1}^{n} e_i^{\,2}}$$

Whereas:

$$e_i = Y_i - \widehat{Y}_i$$

Replacing the formula in large operations, we get:

$$\sum_{i=2}^{n} e_i^{\,2} \approx \sum_{i=1}^{n} e_i^{\,2} \approx \sum e_i^{\,2} - 1$$

$$D - W = 1 - \frac{2\sum_{i=1}^{n} e_i e_i - 1}{\sum_{i=1}^{n} e_i^{\,2}} + 1$$

$$IFp = -1 \rightarrow D - W = 4$$

$$p = 0 \rightarrow D - W = 2$$

$$p = 1 \rightarrow D - W = 0$$

Therefore, whenever the value of DW is close to zero, this indicates the existence of positive autocorrelation, while whenever this value is close to the 4, the more this indicates the presence of negative autocorrelation, and finally, whenever that value approaches 2, this indicates the absence of autocorrelation. The test must first calculate the value of the DW statistic

2(1-p^) Then we extract from the tables for the DW statistics the minimum value (DL) and the highest value (DU) with a degree of freedom equal to the sample size n and the number of estimated parameters k for a specific significance level. :

if it was

Having a positive autocorrelation *1-D-W<DL*

No autocorrelation *2-DU<D-W<DL*

The presence of a negative autocorrelation *3-D-W>4-DL*

Otherwise, the test fails.

Application side

The formula of the logistic model that was used in the analysis, estimation and prediction of the probability of injury takes the following form:

$$P(Y)= \frac{1}{1+e^{-(\beta_0+\beta_1 X_{1i}+\beta_2 X_{2j}+\beta_3 X_{3k}+\beta_4 X_{4l})}}$$

Whereas:

Y: a dependent variable that takes (1.0) infected or uninfected

independent variables:

x_{1i} :Patient's gender variable (male, female)

x_{2j}:patient age variable

x_{3k}:smoking variant

x_{4l}:variable blood pressure

The method of entering variables into the model: where a variable was entered and then a new variable was added to the model to see the extent to which the model's prediction improved until the best prediction of the regression model for the four variables affecting our research is reached.

Where the wald statistic was used as in the equation

The method of entering variables into the model: where a variable was entered and then a new variable was added to the model to see the extent to which the model's prediction improved until the best prediction of the regression model for the four variables affecting our research is reached.

Where the wald statistic was used as in the equation

Wald=$(\frac{\beta}{SE_{\beta}})^2$

To find out the effect of each of the factors

where: β: the parameter value of the predictive variable and SE_{β}standard error

Table of variables in the final model equation

Exp(β) confidence		Exp(β) value	Sig value	Degree freedom	Wald value	error S.E	β value	variable
maximum	minimum							
76.895	8.071	24.912	0.000	1	31.262	.575	3.215	X_{1i}
1.106	1.045	1.075	0.000	1	25.357	.014	0.073	X_{2j}
0.728	0.086	.251	0.011	1	6.462	.544	-1.384	X_{3k}
0.032	0.004	.011	0.000	1	66.730	.555	-4.532	X_{4l}
		1.039	0.965	1	0.002	.873	0.038	Constant

We note that the variables X_{1i} the patient's gender, X_{2j} the patient's age

Two variables are important for classification and prediction. When changing by one unit, the likelihood of developing the disease increases. From this table, the value of Exp(β) for variable X_{1i} indicates that when changing from the value zero (female) to the value (1) male, the probability of heart disease is thus the relationship between The variable of the patient's gender and the incidence of heart disease is a valid relationship, meaning that males are affected more than females.

If the value of the variable X_{3k} increases by one unit, that is, there is a change from smoking to not smoking, the probability of heart disease decreases because the value of Exp(β) for X_{3k} is less than one, meaning the incidence decreases by 0.251 times.

If the value of the variable X4l increases by one unit in blood pressure from high to normal, the probability of infection decreases because the value of Exp(β) is less than (1) and is equal to 0.011.

When we substitute equation No. (1) of the model with the values of (β) from the table, we get the estimated equation of the logistic model, which is used for forecasting.

$$P(Y) = \frac{1}{1+e^{-(0.038+3.215X_{1i}+0.073X_{2j}-1.384X_{3k}-4.532X_{4l})}} \quad \text{……….. (5)}$$

This model is ideal and is able to classify patients into infected and uninfected by 114%.

* Likelihood to predict injury

$$P(Y) = \frac{1}{1+e^{-(0.038+3.215(1)+0.073(70)-1.384(0)-4.532(1))}}$$

$$P(Y) = \frac{1}{1+e^{-(3.831)}} = \frac{1}{1+0.027} = 0.974$$

There is a chance of 0.974 that a 70-year-old male patient who does not smoke and whose blood pressure is high will develop.

$$P(Y) = \frac{1}{1+e^{-(0.038+3.215(1)+0.073(70)-1.384(0)-4.532(0))}}$$

$$P(Y) = \frac{1}{1+e-(8.363)} = \frac{1}{1+0.0002} = 0.999$$

He does not have this disease if his blood pressure is normal, with a probability of (0.999):

$$P(Y) = \frac{1}{1+e^{-(0.038+3.215(0)+0.073(70)-1.384(0)-4.532(1))}}$$

$$P(Y) = \frac{1}{1+e-(0.616)} = \frac{1}{1+0.54} = 0.649$$

A 70-year-old female patient with heart disease who does not smoke and has high blood pressure has a probability of (0.649):

$$P(Y) = \frac{1}{1+e^{-(0.038+3.215(0)+0.073(70)-1.384(0)-4.532(0))}}$$

$$P(Y) = \frac{1}{1+e-(5.148)} = \frac{1}{1+0.006} = 0.994$$

A 70-year-old patient who does not smoke and whose blood pressure is normal does not develop heart disease, with a probability of (0.994)

Recommendations

- Using the logistic regression model in the field of economic and social sciences, as well as limiting it to medical and educational studies.

- Introducing more economic and social variables affecting the odds of heart disease and its predictions in determining the incidence of the disease.

- Establishing a database in hospitals using statistical programs to collect and classify examinations and save them using electronic archiving programs.

- Establishing health care centers in all governorates of Iraq for early detection and identification of heart disease.

Conclusion:

Applying the multiple logistic regression analysis method that blood pressure is one of the most important factors affecting the incidence of heart disease, as the higher the age with high blood pressure, the higher the incidence, and the estimated model is good in classifying patients into injured and uninfected.

References:

1. Batarseh, Salih Rashid, 2009, Book of Statistics and Probabilities, Osama House for Publishing and Distribution - Jordan - Amman.

2. Babtain, Adel Ahmed, 2010. Logistic regression and how to use it in building prediction models for data with two-valued dependent variables. Umm Al-Qura University: College of Education.

3. Youssef, Kholoud Youssef Khamo, Comparison of the modified least chi-square method with other methods in analyzing classified data, Master’s thesis in Statistics, College of Administration and Economics, University of Baghdad 1993.

4. osmer, David W. & Lemeshow, Stanely (2000). Applied Logistic Regression. 2nd edition. New York: Johnson Wiley & Sons, Inc.

5.Dallal,Gerard E. (2001). Logistic Regression. Available at:www.tufts.edu/~gdallal/logistic.htm

6. Agresti, Alan (2007). Categorical Data Analysis. Second edition. New York: Johnson Wiley & Sons, Inc.

7. Murray GD. Assessing the clinical impact of a predictive sytstem in severe head injury. *Med Inform (Lond)* 1990;15:269–73.

8. Murray LS, Teasdale GM, Murray GD, Jennett B, Miller JD, Pickard JD, et al. Does prediction of outcome alter patient management? *Lancet.* 1993;341:1487–91.

9. Kang HY, Ko SK, Liew D. Results of a Markov model analysis to assess the cost-effectiveness of statin therapy for the primary prevention of cardiovascular disease in Korea: The Korean Individual-Microsimulation Model for Cardiovascular Health Interventions. *Clin Ther.* 2009;31:2919–30.

10. Cooper GF, Aliferis CF, Ambrosino R, Aronis J, Buchanan BG, Caruana R, et al. An evaluation of machine-learning methods for predicting pneumonia mortality. *Artif Intell Med.* 1997;9:107–38.