Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 6 (4): 785-788, 2006
ISSN 1812-5654

A Logistic Regression Analysis of the Ischemic Heart Disease Risk

[1]Irfana P. Bhatti, [1]Heman D. Lohano, [2]Zafar A. Pirzado and [1]Imran A. Jafri,
[1]Sindh Agriculture University, Tando Jam, Pakistan
[2]Chandka Medical College, Larkana, Pakistan

Abstract: The main objective of the present study is to investigate factors that contribute significantly to enhancing the risk of ischemic 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 is applied for exploring the factors affecting the disease. The result of the study show the factors that contribute significantly to enhancing the risk of ischemic heart disease are the use of banaspati ghee, living in urban area, high cholesterol level, age group of 51 to 60 years. Other significant factors are Apo Protein A, Apo Protein B, cholesterol level, high density Lipo protein, low density Lipo protein, phospholipids, total lipid and uric acid.

Key words: IHD, heart, logistic regression, factors

INTRODUCTION

Ischemic heart disease is the most common form of heart disease. Ischemic Heart Disease (IHD) is a serious medical problem that causes illness and death and involves high private and public health care costs in Pakistan. At present, the prevalence of IHD in Pakistan is 6.8% (Adviware, 2005). IHD is common heart disease in the United States of America, where its prevalence rate is also 6.8% (Adviware, 2005).

For reducing the prevalence of ischemic heart disease, there is a need of exploring the factors that are responsible to enhancing the risk of this disease. There have been some efforts in previous studies on the topic. Sarwar *et al.* (2004) conducted a survey in the rural areas of Peshawar district of Pakistan and found that major causes of ischemic heart disease are excessive consumption of fatty food, sedentary lifestyle, lack of regular exercise and stressful pattern of life. However, the results of this study were based on computing simple averages of variables and did not include the detailed diet pattern, protein types and other major factors that may be responsible for enhancing the risk of ischemic heart disease.

The present study uses logistic regression model to investigate factors that contribute significantly to enhancing the risk of ischemic heart disease. For analyzing this problem, we observe whether a person has or does not have ischemic heart disease and investigate twenty two independent variables that may be the factors affecting the ischemic heart disease risk. Logistic regression analysis allows one to predict probability of a binary dependent variable from a set of independent variables that may be continuous, discrete, or a mix of them. Logistic regression method is a powerful technique because it is relatively free of restrictions and it allows analyzing a mix of all types of predictors. The next section presents methodology. The last section presents the results and discussion.

MATERIALS AND METHODS

Logistic regression model: The logistic regression model can be written as:

$$\text{prob}(Y = 1) = \frac{e^Z}{1 + e^Z} \quad (1)$$

where Y is binary dependent variable (Y = 1 if event occurs; Y = 0 otherwise), e is the base of the natural logarithms and Z is:

$$Z=\beta_0+\beta_1X_1+\beta_2X_2+ \ldots\ldots\ldots\ldots+\beta_PX_P$$

with constant β_0, coefficients β_j and predictors X_j, for p predictors (j=1,2,3,.....,p).

Data: Data for the study were collected from Chandka Medical College Hospital, Larkana city in Pakistan during the year 1998. In this hospital, patients visit from various districts of Northern part of Sindh province. There were 585 observations in the data set. The data set comprises one dependent variable (Diagnosis) and twenty two independent variables as given in Table 1. In the data set, there were 101 patients without IHD (control) and 484

Corresponding Author: Irfana P. Bhatti, Sindh Agriculture University, Tando Jam, Pakistan

patients with IHD. The collected data were analyzed using SPSS PC+ (version 10.0). Before the analysis, the variables were created, labeled and categorized using indicator variable coding scheme (SPSS, 1997).

Estimation method: Parameters of the model are estimated using the maximum likelihood method. By this method, the estimates of coefficients are the values that maximize the probability of drawing the sample actually obtained (Kennedy, 2003). Backward stepwise elimination method was applied to select significant factors. Backward elimination starts with all of the variables in the model. Then, at each step, variables are evaluated for entry and removal. The score statistic is always used for determining whether variable should be added to the model. Just as in forward selection, the Wald Statistic, the likelihood ratio statistic, or the conditional statistic can be used to select for removal.

Testing hypothesis about the coefficients: The statistical significance of each of the coefficients is evaluated using the Wald test. The Wald Statistic is defined as:

$$W_j = \left(\frac{\beta_j}{S.E_{\beta_j}}\right)^2 \quad \text{where } j=1,2,3,.....,P$$

Partial correlation: A statistic that is used to look at the partial correlation between the dependent and each of the independent variables is the R statistic. R can range in value from -1 to +1. A positive value indicates that as the variable increases in value, so does the likelihood of the event occurring. If R is negative, the opposite is true. Small values for R indicate that the variable has a small contribution to the model. The R statistic can be defined as:

$$R = \pm\sqrt{\left(\frac{\text{Wald Statistic - 2K}}{-2LL_{(O)}}\right)}$$

where K is the degrees of freedom for the variable. The denominator is -2 times the log-likelihood of a base model that contains only the intercept, or a model with no variable if there is no intercept. The sign of the corresponding coefficient is attached to R. The value of 2K in Eq. 5 is an adjustment for the number of parameters estimated. If the Wald statistic is less than 2K, R is set to 0.

Interpretation of coefficients using odds: For interpretation of the logistic coefficients, denote $\hat{Y} = prob\ (Y = 1)$ in Eq. 1:

$$\hat{Y}_i = \frac{e^Z}{1+e^Z}$$

The logistic regression model can be written in terms of the odds of an event occurring. The odds of an event occurring are defined as the ratio of the probability that it will occur to the probability that it will not. The logistic model in terms of log of the odds can be defined as:

$$\log\left(\frac{\hat{Y}_i}{1-\hat{Y}_i}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_P X_P$$

Since it is easier to think of odds rather than log odds, the logistic regression equation can be written in terms of odds as:

$$\frac{\hat{Y}_i}{1-\hat{Y}_i} = e^{\beta_0+\beta_1 X_1+\beta_2 X_2+.....+\beta_P X_P} = e^{\beta_0} e^{\beta_1 X_1} e^{\beta_2 X_2} .....e^{\beta_P X_P}$$

Assessing the goodness of fit of the model: Whenever a model is fitted to the data, the main objective is to know how well the model fits not only the sample of data from which it is derived, but also the population from which the sample data were selected. Define the log-likelihood function as:

$$\text{Log-likelihood} = \sum_{i=1}^{n}\left[Y_i \ln\left(\hat{Y}_i\right) + \left(1-Y_i\right)\ln\left(1-\hat{Y}_i\right)\right]$$

where, $Y_{i's}$ are actual outcomes and $\hat{Y}_{i's}$ are the predicted probabilities of event occurring. Some of the the goodness-of-fit statistics used for the model are Cox and Snell R^2 and $\tilde{R}^2$. These statistics attempt to quantify the proportion of "explained variation" in the logistic regression model. The Cox and Snell R^2 can be defined as

$$R^2 = 1 - \left[\frac{L(0)}{L(\beta)}\right]^{2/N}$$

where, L(0) is the likelihood for the model with only a constant, L(β) is the likelihood for the model under consideration and N is the sample size. The problem with this measure for logistic regression is that it can not achieve a maximum value of 1. The Cox and Snell R^2 so that the value of 1 could be achieved. The Nagelkerke $\tilde{R}^2$ can be defined as,

$$\tilde{R}^2 = \frac{R^2}{R^2_{MAX}}$$

where, $R^2_{MAX} = 1-[L(0)]^{2/N}$. Nagelkerke $\tilde{R}^2$ reveals about the variation in the outcome variable which is explained by the logistic regression model.

Another approach for testing goodness of fit is Chi-square test. Define Chi-square statistic as:

$$\chi^2 = 2\left[\begin{matrix}(\text{log likelihood of bigger model})\\ -(\text{log likelihood for smaller model})\end{matrix}\right]$$

Table 1: Description of dependent and independent variables of the Model

Variable label	Variable name	Level	Value label
Dependent variable			
Diagnosis	Ischemic heart disease	0	Control
		1	Disease
Independent variables			
Age	Age in groups	1	20 to 40 years
		2	41 to 50 years
		3	51 to 60 years
		4	61 to 80 years
Cooking oil	Cooking oil	1	Desi Ghee
		2	Banaspati Ghee
		3	Desi and Banaspati ghee
		4	Vegetable oil
Profession	Profession of the patient	1	Labour
		2	Service
		3	Business
		4	Farm owner
		5	House wife
Locality	Locality of the patient	0	Urban
		1	Rural
Sex	Sex of the patient	0	Male
		1	Female
Education	Literacy level of the patient	0	No
		1	Yes
BP	Blood Pressure	0	Normotension
		1	Hypertension
Smoke	Smoking	0	Non Smoker
		1	Smoker
Nuser	Narcotic user	0	No
		1	Yes
DM	Diabetes miletus	0	No
		1	Yes
IHD	History of IHD	0	No
		1	Yes
Apo-A	Apo protein A		
Apo-B	Apo protein B		
Chol	Cholesterol level		
HDL	High density lipo protein		
LDL	Low density lipo protein		
PL	Phospholipids		
TL	Total lipid		
UA	Uric acid		
Weight	Weight of the patient in kg		
Glucose	Sugar level of the patient		
TG	Triglyceride		

Degrees of freedom (df) are the difference between degrees of freedom for the bigger model and smaller models. The constant only model has 1 df (for the constant) and full model has 1 df for each individual effect and one for the constant.

RESULTS AND DISCUSSION

Estimation of the logistic regression model is presented in Table 2. This table reports the estimated coefficients, standard error for coefficients, Wald Statistic, degrees of freedom and level of significance for Wald Statistic and partial correlation of the logistic regression model. The significant independent variables are age in groups (20 to 40 years, 41 to 50 years, 51 to 60 years, 61 to 80 years), cooking oil (desi ghee, banaspati ghee, desi and banaspati ghee, vegetable oil), Locality (Urban, Rural), Apo Protein A, Apo Protein B, cholesterol level, high density Lipo protein, low density Lipo protein, Phospholipids, total lipid and uric acid. Besides main effects, interactions and square of the quantitative variable were tested. None of the interaction was significant and square of the variables were found significant except Chol (Cholestrol) that has quadratic relationship with the risk of IHD.

Testing hypothesis about the coefficients: To test the null hypothesis for possible rejection, the Wald Statistic and its corresponding significance level were calculated as shown in Table 2. The probabilities of Wald Statistic reveal that in comparison with Age 4 (61-80 years), Age 2 (20-40) has significantly less chances of prevalence of the disease since the coefficient of Age 2 is negative and its significance level is less than 0.05, already established level of significance. Age 3 (51-60) is under the risk of IHD, its significance value is pretty close to 0.05. In contrary to vegetable oil, banaspati ghee users are under the risk of IHD. Locality has negative coefficient that unveils that urban dwellers are more sufferers of the disease than villager are because categorical value 0 was used for urban and 1 was used for rural as given in Table 1. Apo A (Apo Protein A) and HDL (High Density Lipo Protein) have negative coefficients that show that the less value of Apo-A and/or DHL, the higher risk of the disease. Unlike Apo A and HDL, Apo B (Apo-Protein B), LDL (Low Density Lipo Protein), PL (Phospholipids), TL (Total Lipid), UA (Uric Acid) have positive coefficients that indicate that the disease risk increases with the increasing values of these factors. Factor Chol (Cholesterol) has quadratic relationship with the disease risk since $Chol^2$ is also significant. The negative coefficient of Chol and positive coefficient of $Chol^2$ reveals that both extreme cholesterol levels (Very high and very low) are risky.

Partial correlation: The partial correlation between the dependent and each of the independent variable is shown in the table, column labeled R. For instance, the partial correlation between IHD (DV) and Apo-A (IV) is -0.1615. The negative sign shows the negative relationship between the independent variable and the dependent variable. As discussed earlier, the levels of Apo (A), HDL decreases, the likelihood of IHD increases. On the contrary, positive sign shows the increasing risk of IHD with higher levels of factors, namely, Apo B (Apo-Protein B), LDL (Low Density Lipo Protein), PL (Phospholipids), TL (Total Lipid), UA (Uric Acid). R ranges from -1 to +1. Thus, these small values of R indicate that the variable

Table 2: Estimates of the Logistic Regression Model

Variable	β	SE	Wald	df	Sig	R
Age			13.7997	3	0.0032	0.1204
Age1	-2.3978	2.4514	0.9568	1	0.3280	0.0000
Age2	-6.2375	2.2203	7.8921	1	0.0050	-0.1046
Age3	3.4259	1.7653	3.7663	1	0.0523	0.0573
Cooking oil			6.4537	3	0.0915	0.0290
Oil 1	9.0297	95.5912	0.0089	1	0.9247	0.0000
Oil 2	2.8424	1.1218	6.4196	1	0.0113	0.0906
Oil 3	1.2822	1.8160	0.4985	1	0.4801	0.0000
APO-A	-0.2814	0.0703	16.0337	1	0.0001	-0.1615
APO-B	0.2503	0.0713	12.3115	1	0.0005	0.1384
CHOL	-0.9672	0.3134	9.5245	1	0.0020	-0.1182
$CHOL^2$	0.0019	0.0007	8.1320	1	0.0043	0.1067
HDL	-0.2855	0.0886	10.3743	1	0.0013	-0.1247
LDL	0.1177	0.0376	9.7813	1	0.0018	0.1202
Locality	-3.2968	1.6541	3.9726	1	0.0462	-0.0605
PL	0.0625	0.0202	9.5558	1	0.0020	0.1185
TL	0.0273	0.0082	11.1455	1	0.0008	0.1303
UA	5.0924	1.3200	14.8843	1	0.0001	0.1547
Constant	56.4146	25.4441	4.9160	1	0.0266	

has a small partial contribution to the model. If the Wald Statistic is less than 2K, as in the case of variable Age 1, Oil 1 and Oil 3, R is set to 0 as shown in Table 2.

Interpretation of the regression coefficients (using odds): The logistic regression equation model can be rewritten in terms of the odds:

$$\frac{\Pr(\text{disease})}{\Pr(\text{non}-\text{disease})} = e^{56.4146 - 2.3978(\text{Age1}) + \ldots\ldots + 5.0924(\text{UA})}$$

Then *e* raised to the power β_i is the factor by which the odds change when the *ith* independent variable increases by one-unit. For instance, when the value for locality changes from 0 to 1, the odds are decreased by a factor of 0.0370, which is Exp (β). Likewise, when Age 1 (20-40 years) is compared to the Age 4 (61-80 years), the odds are decreased by a factor of 0.0909, Age 2 (41-50 years) is compared to Age 4, the odds are decreased by 0.0020 and when Age 3 (51-60 years) is compared to Age4, the odds are increased by a factor of 30.7507.

Goodness of fit of the model: We test Goodness-of-Fit Statistics for the model with all independent variables. The value of Cox and Snell R^2 is 0.571, which reveals that about 57% of the variation in the outcome variable is explained by the model. The Nagelkerke R^2 is 0.95, which indicates that about 95% of the variation in the outcome variable is explained by the logistic regression model. Chi-square statistic is 495.751 with degree of freedom 16, where the significance level is 0.0000. The chi-square is highly significant and indicates that the model perfectly fits the data under study.

Table 3: Classification of Cases Predicted

		Non-Diseased	Diseased	Accuraey (%)
Observed	Non-Diseased	96	5	95.05
	Diseased	3	481	99.38
			Overall	98.63

Classification of cases: One method of assessing the success of a model is to evaluate its ability to predict correctly the outcome category for cases for whom outcome is known. If a case has diagnosed diseased, for instance, it can be seen if the case is correctly classified as diseased on the basis of other predictor variables.

Results in Table 3 show that 96 individual not having IHD where correctly predicted by the model not to have IHD. It means that 95.05% of the individuals correctly classified without IHD. Similarly, 481 individuals with IHD were correctly predicted to have IHD i.e., 99.38% of the individuals were classified correctly with IHD. The off-diagonal entries tell that how many individuals were incorrectly classified. It is also obvious that 5 individuals without IHD were predicted incorrectly having IHD or it is called the type-I error in hypothesis testing. Similarly, 3 individuals having IHD were incorrectly predicted without IHD. A total of 8 individuals were misclassified in the data set. Overall 98.63% of the 585 cases were correctly classified, which tells that the model fits the data perfectly.

REFERENCES

Adviware, 2005. Statistics by Country for Ischemic Heart Disease. Adviware pty Ltd. Available at http://www.wrongdiagnosis.com/i/ischemic-heart-disease/stats-country.htm, 2005).

Kennedy, P., 2003. Guide to Econometrics. 5th Edn., Cambridge MT: The MIT Press, 1992.

Sarwar, G., I. Khan, R. Iqbal, A.K. Afridi, A. Khan and R. Sarwar, 2004. Risk Factors of Ischemic Heart Diseases in Peshawar. The Journal of Post Graduate Medical Insititute 18: 685-88.

SPSS (Statistical Package for Social Sciences), 1997. SPSS Professional Statistics, 7.5.