INVESTIGATE RISK FACTORS IN THE HIGH BLOOD PRESSURE USING MULTIPLE LOGISTIC REGRESSION

Nagendra Kumar.K*

Muniswamy.B*

D.B.V.N. Suresh*

Ch.Sreelatha*

ABSTRACT:

This paper tries to investigate predictors of incident in hypertension using Multiple Logistic Regression. A random sample of 250 patients was selected from King George Hospital (KGH) Andhra Pradesh, Visakhapatnam, India for the study. The information consists of five instructive variables (Age, gender, BMI, SBP and DBP) and a response variable (hypertension is binary variable). Hypertension is one among the vital public health challenges in worldwide owing to its high frequency and concomitant risks of obesity, kidney and cardiovascular disease. A hypertension model is made to check the interaction and significance between the risk factors. In this study, we have a tendency to analyze the results from multiple logistic regression and to model the relationship between the ordinal outcome variable. The significance variable is chosen supported the P-value, the level of significant model is associated with α =0.05.There are two risk factors are significant out of five explanatory variables are identified. These risk factors are influence significantly in the performance of human blood pressure. The result of this analysis further revealed that age (p-value<0.05) and Body Mass Index (p-value<0.05) are significant, while Gender, SBP and DBP are insignificant. The Logistic regression was expressly mentioned.

KEYWORDS: Multiple Logistic Regression, Odds Ratio, Model Validation, Hypertension.

* Department of Statistics, Andhra University, Visakhapatnam-530 003, Andhra Pradesh

1. INTRODUCTION

Hypertension is one among the necessary public health challenges worldwide as a result of high frequency concomitant risks are obesity and cardiovascular disease. It has been known as a number one risk issue for mortality and ranked third as a cause of disability- adjusted life –years (Ezzati et al., 2002). The fast epidemic of cardiovascular disease in India was documented by studies done at numerous places across the country. The National Nutrition Monitoring Bureau (NNMB), which monitors the organic process standing of the population in nine states of India has estimated the prevalence of cardiovascular disease among the agricultural adult(aged eighteen and above) population of India to be 25% throughout 2004-2005(NNMB,2006).

Various risk factors are related to hypertension as well as age, gender, race, physical activity and socioeconomic category. Compared with the year 2000, the number of adults with hypertension is predicted to increase by 60 percent to a total 1.56 billion by the year 2025. The growing prevalence of obesity is increasingly recognised as one of the most important risk factors for the development of hypertension (Sing et al., 1999). Population studies have additionally shown that blood pressure correlated with body mass index (BMI) and different anthropometrical indices of obesity like waist-hip ratio. In the Framingham study,70 percent of latest cases of hypertension were associated with excess body fat(Kotsis et al.,2005).

2. REVIEW LITERATURE

Jewel (2004) admits that linear regression can be used to model risk difference. However, he immediately follows this with a discussion of the potential for predictions outside of the 0-1 range and then goes on to present logistic regression as the model of choice in general.

Kleinbaum and Klein(2002) make the following argument in favour of the logistic function being a reasonable formula for risk. The S-shape of f(z) has a nice epidemiological property. If we let f(z) represents risk of disease, and let z represent a combination of risk factors, than an individual's risk is minimal for low z's until some threshold is reached. The risk then rises rapidly over a certain range of intermediate z values, and then remains extremely high once z gets large enough. This threshold idea is thought by epidemiologists to apply to a variety of disease conditions, and so f(z) is widely applicable for considering the multivariable nature of epidemiologic research questions. Kleinbbaum and Klein's argument provides a level of credibility to the logistic function, since logistic regression predicts risk. That is, logistic regression predicts the proportion of 1's in the outcome variable, while forcing the logistic function shape on the predicted values.

Soudarssanane et al (2006) carried out titled "A key predictors of high blood pressure and hypertension among adolescents: a simple prescription for prevention". A sample of 673 adolescents (351 males and 322 females) in the 15-19 years age group was used for the study. The univariate analysis followed that mean Systolic Blood Pressure(SBP) and mean Diastolic Blood Pressure (DBP) were 113.6 and 74.3 mm Hg respectively(114.1 and 74.6 in males, 113.1 and 74.1 in females). Mean Blood Pressure(MBP) showed significant correlation with age.MBP and prevalence of hypertension, weight, height and Body Mass Index. Of these, BMI and higher salt intake emerged as independent predictors by multivariate analysis. Findings were confirmed by the case control study, and the major risk factors for hypertension among adolescents are BMI and higher salt intake.

Obesity is a well-established risk factor for hypertension(Rankinen et at., 2007). In this study the prevalence of high blood pressure accrued with BMI. In the Ansan study conducted in Korea, BMI and abdominal circumference was found to be a risk factor for hypertension(Jo et al.,2001). Elsewhere in Asia, the

prevalence of overweight and hypertension was most common in japan(Singh et al., 2010)

Risk factors known weren't a similar altogether the studies conducted in numerous places and it stressed the requirement for identification of risk factors within the specific area for better prevention and management of high blood pressure and its consequences.

3.METHODOLOGY

A form (questionnaire) specially ready for this study was used for data collection. Data on gender, age, height, weight, body mass index (BMI), Systolic Blood Pressure and Diastolic Blood Pressure was collected. Positive designation of high blood pressure disease was created once the beat blood pressure was greater than 140 mm Hg and/or heart beet pressure is greater than or equal to 90 mm Hg

Therefore BMI was calculated employing a straightforward equation (body weight in unit is divided by (height m)^{2}.

The fundamental model for any multiple regression analysis assumes that the outcome variable is a linear combination of a set of predictors, and this is represented as:

$$Y = \beta_0 + \beta_1 w_1 + \ldots\ldots\ldots\ldots + \beta_k w_{ik} + \varepsilon$$

$$= \sum_{k=0}^{p} Y = \beta_k w_{ik} + \varepsilon \qquad \text{where i=1,2,.....,N} \qquad \ldots\ldots \qquad (3.1)$$

Where β_0 is the Intercept of the model expected value of Y, where the w's are set to zero, β_k is the regression coefficient for each corresponding predictor variables w_{ik} . ε is the error term.

The binary logistic model is based on a linear relationship between the natural logarithm (ln) of the odds of an event, and a numerical independent variable. The form of this relationship is as follows:

$$L=\ln\left(\frac{\pi_i}{1-\pi_i}\right)=\sum_{k=0}^{p}Y=\beta_k w_{ik}+\varepsilon \quad \ldots\ldots \quad (3.2)$$

The logistic regression model indirectly models the response variable based on probabilities associated with the values of Y.

Let π_i be the probability that Y=1 and

$1-\pi_i$ be the probability that Y=0

These probabilities are represented as

$$\left.\begin{array}{l}\pi_i = P(Y=1/w_1,w_2,\ldots\ldots,w_{ik}) \\ \text{or} \\ =P(Y=0/w_1,w_2,\ldots\ldots,w_{ik})\end{array}\right\} \quad \ldots\ldots \quad (3.3)$$

$$\log it\ \pi_i=\ln\left(\frac{\pi_i}{1-\pi_i}\right)=e^{\sum_{k=0}^{p}\beta_k w_{ik}} \quad \ldots\ldots \quad (3.4)$$

But the general form of logistic model is given by

$$\log it\ \pi_i=\ln\left(\frac{\pi_i}{1-\pi_i}\right) \quad \ldots\ldots \quad (3.5)$$

Or $$=\ln\left(\frac{\pi_i}{1-\pi_i}\right)=e^{\sum_{k=0}^{p}\beta_k w_{ik}}, \quad where \quad i=1,2,\ldots\ldots N$$

And $\frac{\pi_i}{1-\pi_i}$ are the odd of developing any disease for a subject with risk factors.

By logit transformation of the inverse of log odds to favours Y=1,we obtain the linear component as

$$\frac{\pi_i}{1-\pi_i} = e^{\sum_{k=0}^{p} \beta_k w_{ik}}$$

Using the inverse of logit transformation of the natural logarithm of the odds (log odds) to favour Y=1, we equals to the linear component to have

$$\log it\ \pi_i = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \sum_{k=0}^{p} \beta_k w_{ik} \quad \text{where i=1,2,.....,N} \qquad \ldots\ldots\ldots \qquad (3.6)$$

Therefore, from the equation (6), then we get

$$= e^{\sum_{k=0}^{p} \beta_k w_{ik}}$$

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = e^{\sum_{k=0}^{p} \beta_k w_{ik}} \qquad \ldots\ldots\ldots \qquad (3.7)$$

4.MAXIMUM LIKELIHOOD ESTIMATION FOR LOGISTIC REGRESSION:

We are estimate the P+1 unknown parameters β in equation (3.6). With MLE, by finding the set of parameters for which the probability of the observed is maximum. Therefore each y_i represents a binomial count in the i^{th} population, the joint probability density function of Y is

$$f(Y/\beta) = \prod_{i=1}^{N} \frac{n_i!}{y_i!(n_i - y_i)} \pi_i^{y_i} (1-\pi_i)^{n_i - y_i} \qquad \ldots\ldots\ldots \qquad (4.1)$$

Where β is form π_i in equation (3.3). For every population there are $\binom{n_i}{y_i}$ different ways to arrange y_i success from among n_i trials π_i , then the probability of success of y_i is $\pi_i^{y_i}$ and also the probability of failure is

$(1-\pi_i)^{n_i-y_i}$. The likelihood function is same form as the probability density function, except that the parameters of the function are reversed. The likelihood function express the values of β, in terms of known fixed values for Y. Thus,

$$L(\beta/Y) = \prod_{i=1}^{N} \frac{n_i!}{y_i!(n_i-y_i)} \pi_i^{y_i}(1-\pi_i)^{n_i-y_i} \quad \ldots\ldots \quad (4.2)$$

The maximum likelihood estimates are the values for β that maximize the likelihood function in Equation (4.2). Thus, finding the maximum likelihood estimates requires computing the first and second derivatives of the likelihood function. Since the factorial terms do not contain any of the π_i, they are essentially constants that can be ignored. Therefore, maximizing the equation without the factorial terms will come to the same result, as if they were included. By rearranging the terms, the equation to be maximized which is the conditional likelihood can be written as:

$$L\left(\beta/Y\right) = \prod_{i=1}^{N} \pi_i^{y_i}(1-\pi_i)^{n_i-y_i}$$

$$L\left(\beta/Y\right) = \prod_{i=1}^{N} \left(\frac{\pi_i}{1-\pi_i}\right)^{y_i} \left(1-\pi_i\right)^{n_i} \quad \ldots\ldots \quad (4.3)$$

Consider equation (3.7) then we get..

$$\frac{\pi_i}{1-\pi_i} = e^{\sum_{k=0}^{p} \beta_k w_{ik}} \quad \ldots\ldots \quad (4.4)$$

This, after solving for π_i (The same thing as the result of equation (3.3)) becomes

$$\pi_i = \frac{e^{\sum_{k=0}^{p} \beta_k w_{ik}}}{1+e^{\sum_{k=0}^{p} \beta_k w_{ik}}} \quad \ldots\ldots \quad (4.5)$$

Substituting equation (4.4) for the first term and equation (4.5) for the second term, in equation (4.3) becomes..

$$L\left(\beta/Y\right)=\prod_{i=1}^{N}\left(e^{\sum_{k=0}^{p}w_{ik}\beta_k}\right)^{y_i}\left(1-\frac{e^{\sum_{k=0}^{p}w_{ik}\beta_k}}{1+e^{\sum_{k=0}^{p}w_{ik}\beta_k}}\right)^{n_i} \quad \text{.....} \quad (4.6)$$

$$L\left(\beta/Y\right)=\prod_{i=1}^{N}\left(e^{\beta_0 y_i+y_i\sum_{k=1}^{p}w_{ik}\beta_k}\right)\left(1+e^{\beta_0+\sum_{k=1}^{p}w_{ik}\beta_k}\right)^{-n_i} \quad \text{.....} \quad (4.7)$$

This is the kernel of the likelihood function to maximize. We have simplified further by taking its log. Since the logarithm is a monotonic function any maximum of the likelihood function will also be a maximum.

$$L\left(\beta/Y\right)=\prod_{i=1}^{N}\left(e^{\beta_0 y_i+y_i\sum_{k=1}^{p}w_{ik}\beta_k}\right)\left(1+e^{\beta_0+\sum_{k=1}^{p}w_{ik}\beta_k}\right)^{-n_i}$$

$$\text{Log L}=\sum_{i=1}^{N}\log\left\{\left(e^{\beta_0 y_i+y_i\sum_{k=1}^{p}w_{ik}\beta_k}\right)\left(1+e^{\beta_0+\sum_{k=1}^{p}w_{ik}\beta_k}\right)^{-n_i}\right\}$$

$$=\sum_{i=1}^{N}\left\{\log\left(e^{\beta_0 y_i+y_i\sum_{k=1}^{p}w_{ik}\beta_k}\right)+\log\left(1+e^{\beta_0+\sum_{k=1}^{p}w_{ik}\beta_k}\right)^{-n_i}\right\}$$

$$=\sum_{i=1}^{N}\left\{\left(\beta_0 y_i+y_i\sum_{k=1}^{p}w_{ik}\beta_k\right)-n_i\log\left(1+e^{\beta_0+\sum_{k=1}^{p}w_{ik}\beta_k}\right)\right\}$$

$$= \sum_{i=1}^{N} \left\{ \left(y_i \sum_{k=0}^{p} w_{ik} \beta_k \right) - n_i \log \left(1 + e^{\sum_{k=0}^{p} w_{ik} \beta_k} \right) \right\} \quad \ldots\ldots \quad (4.8)$$

Differentiating equation (4.7) with respect to β_k,

$$\frac{\partial}{\partial \beta_k} \log L = \frac{\partial}{\partial \beta_k} \left\{ \sum_{i=1}^{N} \left\{ \left(y_i \sum_{k=0}^{p} w_{ik} \beta_k \right) - n_i \log \left(1 + e^{\sum_{k=0}^{p} w_{ik} \beta_k} \right) \right\} \right\}$$

$$\frac{\partial}{\partial \beta_k} \log L = \sum_{i=1}^{N} y_i \frac{\partial}{\partial \beta_k} (\beta_0 + \beta_1 w_{i1} + \beta_2 w_{i2} + \ldots\ldots + \beta_k w_{ik})$$

$$= \sum_{i=1}^{N} y_i w_{ik}$$

Consider the second term

$$= -n_i \frac{\partial}{\partial \beta_k} \left(\log(1 + e^{\sum_{k=0}^{p} w_{ik} \beta_k}) \right)$$

$$= -n_i \frac{1}{1 + e^{\sum_{k=0}^{p} w_{ik} \beta_k}} \frac{\partial}{\partial \beta_k} \left(1 + e^{\sum_{k=0}^{p} w_{ik} \beta_k} \right)$$

$$= -n_i \frac{1}{1 + e^{\sum_{k=0}^{p} w_{ik} \beta_k}} e^{\sum_{k=0}^{p} w_{ik} \beta_k} . w_{ik}$$

$$= -n_i w_{ik} \frac{e^{\sum_{k=0}^{p} w_{ik} \beta_k}}{1 + e^{\sum_{k=0}^{p} w_{ik} \beta_k}}$$

$$= -n_i w_{ik} \pi_i \qquad \ldots\ldots \qquad (4.9)$$

$$where\ \pi_i = \frac{e^{\sum_{k=0}^{p} w_{ik}\beta_k}}{1+e^{\sum_{k=0}^{p} w_{ik}\beta_k}}$$

$$\frac{\partial}{\partial \beta_k}\log L = \sum_{i=1}^{N} y_i w_{ik} - n_i \pi_i w_{ik} \qquad \ldots\ldots \qquad (4.10)$$

$$= \sum_{i=1}^{N} w_{ik}(y_i - n_i \pi_i)$$

$$\frac{\partial}{\partial \beta_k}\log L = \sum_{i=1}^{N} w_{ik}(y_i - \phi_i) \qquad \text{where } n_i \pi_i = \phi_i$$

So that, the log likelihood to write a matrix form is given by

$$\frac{\partial}{\partial \beta_k}\log L = W^T(Y - \phi) \qquad \ldots\ldots \qquad (4.11)$$

Which is a column vector of length P+1, then the elements are $\frac{\partial}{\partial \beta_k}\log L$

Let ϕ be a column vector of length N, with elements $\phi_i = n_i \pi_i$. The maximum likelihood estimates for β can be found by setting each of the P+1 equations in Equation (4.10) equal to zero, and solving each β_k.

The general form of the matrix of second partial derivatives are

$$\frac{\partial^2}{\partial \beta_k \partial \beta_{k^1}}\log L = \frac{\partial}{\partial \beta_{k^1}}\left\{\frac{\partial}{\partial \beta_k}\log L(\beta)\right\} \qquad \ldots\ldots \qquad (4.12)$$

$$= \frac{\partial}{\partial \beta_{k^1}}\left\{\sum_{i=1}^{N}\left(y_i w_{ik} - n_i \pi_i w_{ik}\right)\right\}$$

$$= \frac{\partial}{\partial \beta_{k^1}}\sum_{i=1}^{N} y_i w_{ik} - \frac{\partial}{\partial \beta_{k^1}}\sum_{i=1}^{N} n_i \pi_i w_{ik}$$

$$now\ \text{consider}\ -\frac{\partial}{\partial \beta_{k^1}}\sum_{i=1}^{N} n_i \pi_i w_{ik}$$

$$=-\sum_{i=1}^{N} n_i w_{ik} \frac{\partial}{\partial \beta_{k^1}} \left(\frac{e^{\sum_{k=0}^{N} \beta_k w_{ik}}}{1+e^{\sum_{k=0}^{N} \beta_k w_{ik}}} \right) \quad \ldots\ldots \quad (4.13)$$

Now

$$= \frac{\partial}{\partial \beta_{k^1}} \left(\frac{e^{\sum_{k=0}^{p} \beta_k w_{ik}}}{1+e^{\sum_{k=0}^{p} \beta_k w_{ik}}} \right)$$

$$= \frac{\partial}{\partial \beta_{k^1}} \left(\frac{1}{1+e^{-\sum_{k=0}^{p} \beta_k w_{ik}}} \right)$$

$$= \left\{ \frac{(1+e^{-\sum_{k=0}^{p} \beta_k w_{ik}}).0-1.(e^{-\sum_{k=0}^{p} \beta_k w_{ik}}) \text{-}(w_{ik^1})}{\left(1+e^{-\sum_{k=0}^{p} \beta_k w_{ik}}\right)^2} \right\}$$

$$= \frac{e^{-\sum_{k=0}^{p} \beta_k w_{ik}}}{\left(1+e^{-\sum_{k=0}^{p} \beta_k w_{ik}}\right)^2} . w_{ik^1}$$

It can be written as

$$= -\sum_{i=1}^{N} n_i w_{ik} \pi_i (1-\pi_i) \; w_{ik^1}$$

$$\frac{\partial^2}{\partial \beta_k \partial \beta_{k^1}} logL(\beta) = -\sum_{i=1}^{N} n_i w_{ik} \pi_i (1-\pi_i) \; w_{ik^1} \quad \ldots\ldots \quad (4.14)$$

Now we can write matrix form

$$\frac{\partial^2}{\partial\beta_k\partial\beta_{k^1}} logL(\beta) = - W^T Z\ W \quad \text{.....} \quad (4.15)$$

Where Z is a square matrix of order N, with elements $n_i\pi_i(1-\pi_i)$ on the diagonal.

5. DATA ANALYSIS

The logistic regression output from SPSS for the hypertension patient's data with gender, age, Body Mass Index, Systolic Blood Pressure and Diastolic Blood Pressure are the explanatory variables.

The fitted model is

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = -3.927 + 0.30age + 0.186BMI$$

We shall first examine the hypothesis that all of the regression coefficients for the explanatory variables are zero, for logistic regression. $H_0 = \beta_1 = \beta_2 = 0$

This hypothesis is tested by a chi-square statistic with 5 decrees of freedom. This is given in the SPSS output (table: 6.1) as Chi-square calculated value is 23.615 and it significant (the P-value is 0.000). We reject H_0 and conclude that one or more of the explanatory variables can be used to predict the hypertension in the patients. The P-value for age is (p-value=0.011) is less than 0.05. And the P-value of BMI(0.003)is less than 0.01 respectively.

HOSMER AND LEMESHOW TEST

Hosmer and Lemeshow test is based on grouping cases in to deciles of risk. It compares the observed probability with the expected probability within each deciles. The P-value is greater than (>) 0.05, there is no significant difference between the observed probability and the expected probability. From the given bellow table(5.1) the p-value =0.132 is obtained is greater than 0.05, then the model is fit to the data well.

Table: 5.1-Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	17.942	8	.132

MODEL VALIDATION

The classification table(6.3) from the SPSS output result summarizes the observed group and the predicted group classification. The overall correctly specified group percentage is 71.6

CONCLUSION:

The multiple logistic regression analysis, there are two factors are significant out off five factors were tested and identified as having influence significantly the performance of human blood pressure(hypertension). These factors are age and Body Mass index(BMI).From the SPSS output table (6.4), When the age of the respondents where increased by one year the chance to have high blood pressure will be increased by a factor of 1.031 when other factors remaining constant (95% CI: 1.007 to 1.056, p-value<0.05). When the BMI of the respondents where increased by the chance to have high blood pressure will be increased by a factor of 1.985 when other factors remaining constant (95%

CI: 0.934 to 1.038, p-value<0.01).As a conclusion, the two of these factors can influence the performance of blood pressure at risk for hypertension.

6. Logistic Regression

Block 1: Method = Enter

Table:-6.1

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	23.615	5	.000
	Block	23.615	5	.000
	Model	23.615	5	.000

Table6.2 :Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	304.227[a]	.590	.653

a) Estimation terminated at iteration number 4 because parameter estimates changed by less than 0.001

Table 6.3 : Classification Table[a]

Observed			Predicted		
			Hypertension		Percentage Correct
			0	1	
Step 1	Hypertension	0	146	13	91.8
		1	58	33	36.3
	Overall Percentage				71.6

a) The cut value is .500

Table 6.4 : Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)

								Lower	Upper
Step 1[a]	Gender	-.568	.278	.989	1	.635	.567	.329	.976
	Age	.030	.012	6.388	1	.011	1.031	1.007	1.056
	BMI	.186	.057	5.334	1	.003	1.985	.934	1.038
	SBP	.004	.016	.056	1	.813	1.004	.973	1.035
	DBP	.015	.026	.343	1	.558	.985	.935	1.037
	Constant	-3.927	1.517	3.726	1	.034	18.680		

a) Variable(s) entered on step 1: Gender, Age, BMI, SBC, and DBP.

References:

Agresti A.(1996), "An Introduction to Categorical Data Analysis",Wiley

A Rashid K,KA Narayan, Azizah Hj Ab Manan. The Prevalence of Hypertension among the Elderly in Fourteen Villages in Kedah, Malaysia.Malaysian Journal of Medicine and Health Sciences Vol. 4(2) June 2008:33-39

Brundtland GH. 2002. From the World Health Organization. Reducing risks to health, promoting healthy life. Jama. 288:1974.

Cifkova R. Et al (2004),"Prevalence, awareness,treatment and control of hypertension in the Czech Republic" Journal of human hypertension.

D.W. Hosmer and S. Lemeshow, Applied logistic regression,second edition John Wiley and Sons,2000.

Jewell NP. (2004). Statistics for Epidemiology.New Yark, Chapman & Hall/CRC.

Kutner M.H; Nachtsheim C.J and Neter J. (2004),"Applied linear regression models (fourth edition)"

Mitchell C. And Dayton (1992)," Logistic Regression Analysis",University of Maryland.

Kleinbaum DG,Klein M.(2002). Logistic Regerssion: A Self-Learning Text. 2nd Ed.New York,Springer-Verlag.

Soudarssanane M.B.,Karthigeyan S., Stephen A. and Sahai,A.(2006): Key Predictors of High Blood Pressure and Hypertension among Adoleicen