International Journal of Multidisciplinary and Current Research ISSN: 2321-3124
Research Article Available at: http://ijmcr.com

Modeling of Binary Logistic Regression for Obesity among Students in Rural Area of Visakhapatnam

Nagendra Kumar.K[1],Muniswamy.B[2] D.BVN.Suresh[3],Sreelatha.Ch[4] and Jeevan Kumar.D[2]

[1]Department of Statistics, Andhra University, Visakhapatnam-530 003, Andhra Pradesh. India
[2]Department of Engineering Mathematics, Andhra University,Visakhapatnam-530003, Andhra Pradesh, India

Accepted 01 April 2016, Available online 06 April 2016, **Vol.4 (March/April 2016 issue)**

Abstract

Logistic regression analysis examines the impact of different factors on dichotomous outcomes by way of estimating the chance of the event occurrence logistic regression, additionally called a logistic model, is a statistical procedure used to model dichotomous effects. In the logistic model the log odds of the dichotomous outcome is modeled as a linear combination of the predictor variables. The log odds quantitative relation in logistic regression provides an outline of the probabilistic relationship of the variable and also the outcomes. In conducting logistical regression, choice procedures are utilized in choosing vital predictor variables diagnostics are used to test that assumptions are legitimate which include independent of errors, linearity is the logit for continuous variability absence of multicolinearity, and shortage of strongly influential outliers and take a look at data point is calculated to see the aptnes of the model. This study used the binary logistic regression model to analyze obese (overweight) and obesity among rural area in Visakhapatnam, distance, AP, India. The idea of their demographics profile, records, weight loss program and lifestyle.The results indicate that overweight and obesity of peoples are influenced via obesity in family and also the interaction between a Students quality and routine meals intake.

Keywords: *Binary Logistic Regression, logit model, Odds Ratio, Model validation, Hosmer and Lemeshow Test.*

1. Introduction

Logistic regression is a sort of regression analysis used to predict the outcome of a categorical response variables based on one or more predictor variables. Binary logistic regression is one of the logistic regression analysis techniques where by the response variable is dichotomous having two categories, coded as success '1' or failure '0'. The final results is not a prediction of a numerical cost, as in linear regression, however a probability of belonging to certainly one of conditions, that may take on any values between 0 and 1. Each theoretical and empirical outcome suggests that when the response variable is binary, the form of the response function is curvilinear, that resembles either as a fitted S or as a reverse tilted S.

This sort of shape is frequently known as sigmoidal or S-shaped. The S-shaped curves are fitted using the logistic characteristic. When the response is a proportion as a response, a logit transformation is used to link the response variable the set of predictor variables. In a logistic regression model the log odds of the dichotomous outcome is modeled as a linear combination of the predictor variables.

Logistic regression calculates the probability of success over the probability of failure in the form of an odds ratio. The odds ratio is a measure of effect size, describing the strength of non-independent association between two binary information values. Logistic regression has been specifically famous with clinical studies in which the dependent variable is whether or not a patient has a disease.

Logistic regression forms a best fitting equation or function using the maximum likelihood method, which maximizes the probability of classifying the observed data into the appropriate category given the regression coefficients. Logistic regression model building manner includes choice of predictor variables and the use of diagnostic strategies.

2. Methodology

Logistic regression model assumes that the log–odds of an observation can be expressed as a linear function of the s input variables w.

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 as follows

$$L = \ln\left(\frac{p(w)}{1-p(w)}\right) = \sum_{h=0}^{s} \beta_h w_h \quad (2.1)$$

In equation (2.1), the constant term β_h is added by setting w_0=1 this produce s+1 parameters.

The left hand side of equation (2.1) is called the logit of p, which implies that logistic regression.

Now apply exponent on both sides of equation (2.1), we have

$$\frac{p(w)}{1-p(w)} = \exp\left(\sum_{h=0}^{s} \beta_h w_h\right) \quad (2.2)$$

$$= \prod_{h=1}^{s} \exp\left(\beta_h w_h\right) \quad (2.3)$$

Equation (2.3) tells us that logistic model are multiplicative in their inputs, and it produce a way to interpret the coefficients.

The logit equation can also invented to get a new expression for

$$p(w) = \frac{\exp(\pi)}{1+\exp(\pi)} \quad (2.4)$$

Where $\pi = \sum_{h=0}^{s} \beta_h w_h$ (2.5)

The right hand side of equation (2.5) is the sigmoid of ? which maps the real line to the interval (0, 1) and is approximately linear near to origin. An important fact about p(w) is that the derivative

$$p'(w) = p(w)\left[(1-p(w))\right]$$

The derivation goes as follows

$$p(w) = \frac{\exp(\pi)}{1-\exp(\pi)}$$
$$= \exp(\pi)\left(1+\exp(\pi)\right)^{-1} \quad (2.6)$$

By using the product rule to solve the equation (2.6), we have

$$p'(w) = \exp(\pi)(-(1+\exp(\pi))^{-2}\exp(\pi) + \exp(\pi)(1+\exp(\pi))^{-1}$$
$$= -\exp(\pi)(1+\exp(\pi))^{-2}\exp(\pi) + \exp(\pi)(1+\exp(\pi))^{-1}$$
$$= -(\exp(\pi))^2(1+\exp(\pi))^{-2} + \exp(\pi)(1+\exp(\pi))^{-1}$$
$$= \exp(\pi)(1+\exp(\pi))^{-1} - (\exp(\pi))^2(1+\exp(\pi))^{-2}$$
$$= \frac{\exp(\pi)}{(1+\exp(\pi))} - \frac{(\exp(\pi))^2}{(1+\exp(\pi))}$$

$$= \frac{\exp(\pi)(1+\exp(\pi)) - (\exp(\pi))^2}{(1+\exp(\pi))^2}$$

$$= \frac{\exp(\pi)}{(1+\exp(\pi))^2}$$

$$= \frac{\exp(\pi)}{(1+\exp(\pi))}\left(\frac{1}{1+\exp(\pi)}\right) \quad (2.7)$$

$$\therefore \quad P'(w)=P(w)\ (1-P(w)) \quad (2.8)$$

The solution of a Logistic Regression problem is the set of parameters β that maximizes the likelihood of the data, which is expressed as the product of the predicted probabilities of the N individual observations.

$$L(W/P) = \prod_{k=1}^{N}\prod_{z_k}^{N} P(w_k) \prod_{k=0}^{N}\prod_{z_k}^{N}[1-P(w_k)] \quad (2.9)$$

(W,z) is the set of observation W is a k+1 by N matrix of inputs, where each column corresponds to an observation and the first row is 1,z is an N-dimensional vector of responses and (w_k,z_k) are the individual observations.

Taking the log of equation(2.9) to get

$$L(W/P) = \sum_{k=1}^{N}\sum_{z_k=1}^{N} \log P(w_k) + \sum_{k=0}^{N}\sum_{z_k=0}^{N} \log[1-P(w_k)] \quad (2.10)$$

Maximizing the log-likelihood will maximize the likelihood. The quantity -2*log-likelihood is called the deviance of the model. It is analogous to the Residual Sum of Squares(RSS) of a linear model. Ordinary least square minimizes RSS; logistic regression minimizes deviance.

$$Psuedo - R^2 = 1 - \frac{deviance}{null\ \text{deviance}} \quad (2.11)$$

To maximize the log-likelihood, we take its gradient with respect to β;

$$\frac{\partial}{\partial\beta}\log L = \sum_{k=1}^{N}\sum_{z_k=1}^{N} \frac{P_k'}{P_k(w_k)} w_k - \sum_{k=0}^{N}\sum_{z_k=0}^{N} \frac{P_k'}{1-P_k(w_k)} w_k \quad (2.12)$$

The maximum occurs where the gradient is zero. From the above function $P'(w_k)=P(w_k)\ [1-P(w_k)]$, therefore equation (2.12) becomes

$$\frac{\partial}{\partial\beta}\log L = \sum_{k=1}^{N}\sum_{z_k=1}^{N} \frac{P(w_k)(1-P(w_k))}{P(w_k)} w_k - \sum_{k=0}^{N}\sum_{z_k=0}^{N} \frac{P(w_k)(1-P(w_k))}{1-P(w_k)} w_k$$

$$\frac{\partial}{\partial\beta}\log L = \sum_{k=1}^{N}\sum_{z_k=1}^{N} (1-P(w_k))w_k - \sum_{k=1}^{N}\sum_{z_k=0}^{N} P(w_k)$$

$$= \sum_{k=1}^{N} z_k(1-P(w_k))w_k - \sum_{k=1}^{N}(1-z_k)P(w_k)w_k$$

$$= \sum_{k=1}^{N} [z_k(1-P(w_k)) - (1-z_k)P(w_k)]w_k \quad (2.13)$$

Equation (2.13) merges the cases (z_k=1and z_k=0) into a single sum. This produces the set of simultaneous equations that are true at the optimum;

$$\sum_{k=1}^{N} z_k w_k - P(w_k)w_k = 0 \quad (2.14)$$

It should be noted that from Equation (2.14), the sum of probability mass across each coordinate of the w_k vector is equal to the count of observations with that coordinate value for which the response was true.

The coefficients β can be solved by using the Newton's method.

Assuming that we start with an initial guess β_0, we can take the Taylor expansion of f around β_0:

$$F(\beta_0 + \Delta) \approx f(\beta_0) + f'(\beta_0)\Delta \quad (2.15)$$

f is a matrix and it is the Jacobean of first derivatives of f with respect to β. Setting the left hand side to zero, then solve for Δ as follows;

$$\Delta_0 = -[f'(\beta_0)] - f(\beta_0) \quad (2.16)$$

Then we update our estimate for β;

$$\beta_1 = \beta_0 + \Delta_0 \quad (2.17)$$

In this paper, f is the gradient of the log-likelihood, and its Jacobean is the Hessian of the log-likelihood function.

$$H = \frac{\partial^2}{\partial \beta^2} \log L = \sum_{k=1}^{N} w_k P(w_k)(1-P(w_k))w_k^T \quad (2.18)$$

$$= WZW^T \quad (2.19)$$

3. Data Analysis

The binary logistic model was conducted to get a relation between being overweight and obese with student's characteristics. The response variable is dichotomous with two possible outcomes.

Here, the strategies indicated that seven variables which are Gender, Age, Body Mass Index, obesity among families, Systolic Blood Pressure, Diastolic Blood Pressure, and taking routine meals.

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 the p-value =0.302 is obtained is greater than 0.05, then the model is fit to the data well.

Table3.1:-Hosmer and Lemeshow Test

Step	Chi-square	Df	Sig.
1	9.495	8	.302

Model Validation

The classification table from the SPSS output result summarizes the observed group and the predicted group classification. The overall correctly specified group percentage is 68.6.

Logistic Regression

Table3.2: Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	397.668[a]	.570	.693

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

Table 3.3: Classification Table[a]

	Observed		Predicted		
			Obesity		Percentage Correct
			0	1	
St ep 1	Obesity	0	93	54	68.8
		1	67	86	59.2
	Overall Percentage				68.7

(b) The cut value is .500

The logistic regression output from SPSS for the Obesity patient's data with Gender, age, body mass index, SBP,DBP, Obesity among family and routine meals are the explanatory variables.

The fitted model is

$$\ln\left(\frac{p(w)}{1-p(w)}\right) = -2.715 + 0.132age + 0.193BMI + 0.026OAF + 0.325RM$$

The Binary logistic regression analysis, there are four factors are significant out off six factors were tested and identified as having influence significantly the performance of obesity. These factors are Age, BMI, Obesity among family and Regular Meals. From the SPSS output table (6.4), When the age of the respondents where increased by one year the chance to have Obese will be increased by a factor of 1.052 when other factors remaining constant (95% CI: 1.014 to 1.075, p-value<0.05).

Table3.4: 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	-0.418	0.327	0.572	1	0.415	0.316	0.337	0.586
	Age	0.132	.0.02	5.867	1	0.017	1.052	1.014	1.075
	BMI	0.193	0.043	5.127	1	0.004	1.985	0.725	1.056
	SBP	0.027	0.016	0.056	1	0.813	1.004	0.913	1.064
	DBP	0.021	0.127	0.871	1	0.512	0.354	0.378	1.037
	Obesity among family								
		0.064	0.026	6.431	1	0.002	1.098	0.352	1.072
	Regular Meals								
		0.325	0.473	3.107	1	0.041	1.32	0.632	1.035
	Constant	-2.715	1.416	3.261	1	0.049	21.57		

(c) Variable(s) entered on step 1: Gender, Age, BMI, SBC, DBP, Obesity among family and Routine Meals

When the BMI of the respondents where increased by the chance to obesity will be increased by a factor of 1.985 when other factors remaining constant (95% CI: 0.725 to 1.056, p-value<0.01). When the obesity among family of the respondents where increased by the chance to obesity will be increased by a factor of 1.098 when other factors remaining constant (95% CI: 0.352 to 1.072, p-value<0.01). When the regular meals of the respondents where increased by the chance to obesity will be increased by a factor of 1.320 when other factors remaining constant (95% CI: 0.632 to 1.035, p-value<0.01).As a conclusion, the four of these factors can influence the performance of Overweight at risk for Obesity.

Conclusion

Binary logistical regression is applicable once the response variable is binary either success of failure. Through this model, likelihood for getting outcome for a particular factor is determined by exploitation odds quantitative relation. The results of this study indicated that being overweight and obese among rural area students are influenced by genetic factor which is obesity in their family relations and their lifestyle that is dietary intake. The risk of a rural student being overweight and obese increased with having obese family members. The risk of a rural student's takes routine meals being overweight and obese is higher than who frequently takes routine meals. This is due to different dietary intakes among different quality.

References

[1]. Agresti A.(1996), An Introduction to Categorical Data Analysis , wiley.
[2]. David W. Hosmer and Stanley Lemeshow, Applied Logistic Regression, Second Edition.
[3]. Ogden CL, Carroll MD, Curtin LR, Lamb MM. & Flegal KM.(2010) Prevalence of high body mass index in U.S. children and adolescents, 2007-2008. JAMA, 303(3), 242-249.
[4]. Sheperd, T. M. (2003). Effective management of obesity - Research findings that are changing clinical practice. Journal of family practice, 52(1), 34-42
[5]. Sundquist, J., & Johansson, S. (1998). The influence of socioeconomic status, ethnicity, and lifestyle on body mass index in a longitudinal study. International journal of epidemiology, 27, 57-63.
[6]. World Health Organization (WHO). Obesity: preventing and managing the global epidemic. WHO Obesity Technical Report Series, No. 894. Geneva: World Health Organization, 2000.
[7]. Jewell NP. (2004). Statistics for Epidemiology.New Yark, Chapman & Hall/CRC.
[8]. Kutner M.H; Nachtsheim C.J and Neter J. (2004),Applied linear regression models (fourth edition)
[9]. Mitchell C. And Dayton (1992), Logistic Regression Analysis,University of Maryland.
[10]. Kleinbaum DG,Klein M.(2002). Logistic Regerssion: A Self-Learning Text. 2nd Ed.New York,Springer-Verlag.