Open Access Scientific Reports

Research Article Open Access

Estimating Odds Ratios in Logistic Regression of Dichotomous Data

Oyeka ICA[1] and Okeh UM[2]*

[1]Department of Applied Statistics, Nnamdi Azikiwe University, Awka Nigeria
[2]Department of Industrial Mathematics and Applied Statistics, Ebonyi State University Abakaliki, Nigeria

Abstract

This paper proposes an odds ratio type measure of strength of association between screening test results and state of nature or condition in a population from diagnostic screening tests based on logistic regression analysis of dichotomous outcomes. The proposed method unlike in the analysis of data from most screening tests requires that the response to the condition of interest is dichotomous, assuming one of two possible values. The predisposing factors in this study are categorical variables. This would enable the fitting of a logistic regression model to help in the estimation of desired probabilities, odds and odds ratios of positive responses. A test statistic to assess the statistical significance of the proposed measure based on the logistic regression is developed. The proposed method is illustrated with some sample data and the results are shown to compare favourably with what is obtained using the usual expression for the odds ratio.

Keywords: Gestational diabetes mellitus; Odds; Odds ratio; Logistic regression; Dichotomous

Introduction

Often a candidate for an examination or a job interview may wish to estimate the probability of his success given some predisposing factors such as the number of hours he studied per day or per week, the nature, type and duration of the examination, the condition; prior qualifications his age ,gender, ethnic group, state of origin etc. A clinician conducting a diagnostic test or drug trials for a certain condition may wish to know the odds that his subjects or patients respond positive given their various characteristics such as age, gender, body weight, family history [1] etc. A gynecologist or a pediatrician may wish to estimate the odds that a new-born baby is under-weight or has more than normal gestation period given the mothers age, parity, body weight and child's gender [2] etc. In all the situations the response to the condition of interest is dichotomous, assuming one of two possible values. The predisposing factors may be either categorical or continuous variables [3]. This would enable the fitting of a logistic regression model to help in the estimation of desired probabilities, odds and odds ratios of positive responses as discussed below[4-6].

The proposed model

Let y_i be the response of the ith randomly selected subject to the condition of interest assuming values of either 1(positive response) or 0 (negative response) for i=1,2,…,n. Let x_{i1}, x_{i2},...., x_{ik} be the score by the ith subject on the independent explanation, or predetermined variables X_1, X_2,...., X_K respectively,

$$Let\ y_i = \begin{cases} 1, if\ \ the\ ith\ subject\ responds\ \ positive\ \ to\ the\ condition\ of\ \mathrm{int}\, erest. \\ 0, if\ \ the\ ith\ subject\ responds\ negative\ ....,,... \end{cases} \tag{1}$$

for i=1, 2,...., n

Then a logistic model regressing the probability that the ith subject responds positive to the condition under study on the independent variables x_{i1}, x_{i2},...., x_{ik} is

$$P(y_i = 1/x_{i1}, x_{i2}, ..., x_{ik}) = P_i(1/x) = \frac{1}{1+e^{-(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + e_i)}} \tag{2}$$

The corresponding logistic regression model of the probability of non-response is

$$P_i(0/x) = 1 - P_i(1/x) = \frac{e^{-(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + e_i)}}{1+e^{-(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + e_i)}} \tag{3}$$

where the β_{js} are partial regression coefficients and e_i are error terms not correlated with x_{ijs}, for i=1,2,...,n and j= 1,2,...k.

Applying the method of least squares to Equation 2 enables us obtain unbiased estimates of β_j as $\hat{\beta}_j = b_j$ for j=0,1, 2...,k there by yielding the estimated regression Equation for $P_i(1/x)$ as

$$\hat{P}_i(1/x) = \frac{1}{1+e^{-(b_0 + \sum_{j=1}^{k} b_j x_{ij})}} \tag{4}$$

for i=1, 2, ...,n; j=1, 2,....,k

The corresponding estimated regression Equation for $P_i(0/x)$ as

$$\hat{P}_i(0/x) = 1 - \hat{P}_i(1/x) = \frac{e^{-(b_0 + \sum_{j=1}^{k} b_j x_{ij})}}{1+e^{-(b_0 + \sum_{j=1}^{k} b_j x_{ij})}} \tag{5}$$

The following analysis of variance (ANOVA), Table 1 is used to test the adequacy of Equations 4 and 5 based on the F-test statistic:

Where $\underline{u}' = (u_1, u_2, ..., u_n), u_i = In\left(\frac{P_i(1)}{1-P_i(1)}\right); \underline{b}' = (b_0, b_1, ..., b_k)$ and X is an n×(k+1) matrix of regressors. The null hypothesis to be tested for the adequacy of Equation 4 using the results of Table 1 is

$$H_0: \beta_1 = \beta_2 = ... = \beta_k = 0 \text{ vs } H_1: \beta_j \neq 0 \text{ for some } j \tag{6}$$

j=1, 2,....,k

***Corresponding author:** Okeh UM, Department of Industrial Mathematics and Applied Statistics, Ebonyi State University Abakaliki, Nigeria, E-mail: uzomaokey@ymail.com

Received January 03, 2013; **Published** January 08, 2013

Citation: Oyeka ICA, Okeh UM (2013) Estimating Odds Ratios in Logistic Regression of Dichotomous Data. 2:608 doi:10.4172/scientificreports.608

Screening Test Result	GDM Present (B)	GDM Absent ($\overline{B}$)	Total
GDM positive (A)	n_{11}=18	n_{12}=18	$n_{1.}$=36
GDM Negative ($\overline{A}$)	n_{21}=35	n_{22}=230	$n_{2.}$=265
Total	$n_{.1}$=53	$n_{.2}$=248	$n_{..}$=301

Table 1: Four-fold table for the screening test results and gold standard of risk pregnant women for GDM Gold Standard.

H_0 is rejected at the α level of significance if

$$F \geq F_{1-\alpha;\, k,\, n-k-1} \tag{7}$$

Otherwise H_0 is accepted where $F_{1-\alpha;\, k,\, n-k-1}$ is the critical value of the F-distribution with k and n-k-1 degrees of freedom for a specified α level. If the model fits, that is if H_0 is rejected so that not all the β_{js} are zero, then we may proceed to estimate the required probabilities, odds and odds ratios of positive responses to the condition of interest. Thus assuming that H_0 is rejected then we estimate from Equations 4 and 5 the odds that the i^th subject responds positive to the condition under study given the independent variables $X_1=x_{i1}$, $X_2=x_{i2}$,.... $X_k=x_{ik}$ as

$$\hat{\Omega}_{i\,(X_1=x_{i1},X_2=x_{i2},\ldots,X_k=x_{ik})}$$

$$\hat{\Omega}_{ix} = \frac{P_i(1/x)}{P_i(0/x)},$$

That is

$$\hat{\Omega}_{ix} = e^{b_0+\sum_{j=1}^{k} b_j x_{ij}} \tag{8}$$

Now suppose X_i is increased by α units and X_s is decreased Υ units, while holding all other independent variable at constant levels, that is, suppose

$$X_i' = x_{il}' = x_{il} + \alpha \text{ and } X_s' = x_{is}' = x_{is} - \Upsilon,$$

with other independent variables held constant, then the resulting odds using Equation 8 is

$$\hat{\Omega}_{i(X_1=x_{i1},X_2=x_{i2},\ldots.X_l'=x_{il}+\alpha,\ldots.X's=x_{is}-\Upsilon;\ldots X_k=x_{ik})} = \hat{\Omega}_{ix'}$$

That is

$$\hat{\Omega}_{ix'} = e^{b_0+b_l(x_{il}+\alpha)+bs(x_{is}-\Upsilon)+\sum_{J\neq bs}^{k} b_j . x_{ij}}$$

$$= e^{(b_l\alpha-b_s\Upsilon)+(b_0+\sum_{j\neq ls}^{k} b_j x_{ij})} = e^{(b_l\alpha-b_s\Upsilon)} e^{(b_0+\sum_{j\neq ls}^{k} b_j x_{ij})}$$

That is

$$\hat{\Omega}_{ix'} = \frac{\hat{P}_i(1/x')}{\hat{P}_i(0/x')} = e^{(b_l . \alpha_l - b_s . \Upsilon)} e^{b_0+\sum_{j\neq ls}^{k} b_j . x_{ij}}$$

That is

$$\hat{\Omega}_{ix'} = e^{(b_l . \alpha - b_s . \Upsilon)} \hat{\Omega}_{ix} \tag{9}$$

In other words, the odds of positive response by the i^th subject when the levels of some independent variables are adjusted is a function of the odds of positive response when the levels of all the independent variables are held constant. Hence the estimated odds ratio of positive responses by the i^th subject under these conditions is from Equation 9

$$\hat{w}_i = o = \frac{\hat{\Omega}_{ix'}}{\hat{\Omega}_{ix}} = e^{b_l\alpha - b_s\Upsilon} \tag{10}$$

for l=1, 2,....,k; s=1, 2,...,k; $l \neq s$

Estimation odds ratio from logistic regression of data

Note that since the right hand side of Equation 10 is independent of i, i=1, 2,…,n. Equation 10 may be interpreted as the estimated odds ratio of positive responses by any randomly selected subjects under the specified conditions. In obtaining the odds ratio of Equation 10 it is assumed that some independent variables are increased or decreased by some constant. It is however also possible that some of these independent variables are increased or decreased proportionately, that is by some percentage or proportion of the independent variables themselves. Thus suppose $X_l' = (1+\alpha).X_{il}$ assuming a value of (1+α) x_{il} and $X_s' = (1-\Upsilon).X_{is}$ assuming a value of (1-Υ)x_{is}, holding other independent variables constant. Then the resulting odds of positive response by the i^th randomly selected subject is,

$$\hat{\Omega}_{ix'} = e^{b_0+b_l(1+\alpha).x_{il}+b_s(1-\Upsilon).x_{is}+\sum_{j=1}^{k} b_j x_{ij}}$$

Where $x_{il} + \alpha = X_l' = x_{il}' = (1+\alpha)x_{il}$ and

$$x_{il} - \Upsilon = X_s' = x_{is}' = (1-\Upsilon)x_{is}$$

$$= e^{b_l\alpha x_{il} - b_s . \Upsilon . x_{is} + (b_0+\sum_{j=1}^{k} b_j x_{ij})}$$

$$= e^{b_l\alpha x_{il} - b_s . \Upsilon . x_{is}} e^{(b_0+\sum_{j=1}^{k} b_j x_{ij})}$$

That is

$$\hat{\Omega}_{ix'} = e^{b_l\alpha x_{il} - b_s . \Upsilon . x_{is}} \hat{\Omega}_{ix} \tag{11}$$

Thus, the resulting estimated odds ratio of positive response by the i^th subject when some independent variables are proportionately increased or decreased while holding others at constant levels is from Equation 11.

$$\hat{w}_i = o = \frac{\hat{\Omega}_{ix'}}{\hat{\Omega}_{ix}} = e^{b_l . \alpha . x_{il} - b_s . \Upsilon . x_{is}} \tag{12}$$

for l=1, 2,....,k and s=1, 2,...,k, $l \neq s$ and i=1, 2,...n

llustrative example

Table 2 shows the data obtained from a collection of hospitals in Ebonyi State covering from January 2010 to December 2011 particularly from the medical record unit of these hospitals. It was the result of a retrospective study on the effect of four independent risk factors (variables) in the development of gestational diabetes mellitus (GDM). A sample of 301 risk pregnant women who satisfied the inclusion criteria based on WHO, 1999 [7] standard were considered. All the risk factors (family history-FH, obesity, age, and previous fetal weight) considered in this work are dichotomous in which case; it has been coded for use in estimating the odds ratio in logistic regression. The dependent variable is GDM. We here present sample data obtained in a diagnostic screening test to confirm the presence or absence of GDM among the sampled subjects from a certain population. The proposed method is illustrated using the sample data of table 1.

Analysis showed the following results:

	B	S.E.	Wald	df	Sig.	Exp(B)	
Step 1(a)	OBESITY	-.082	.255	.104	1	.747	.921
	AGE	-.020	.172	.013	1	.909	.981
	FH(1)	-.211	1.145	.034	1	.854	.810
	PFW	-.125	1.098	.013	1	.909	.882
	Constant	3.503	7.924	.195	1	.658	33.209

(a) Variable (s) entered on step 1: OBESITY, AGE, FH, PFW.

Table 2: Variables in the Equation.

$$Se = P(A|B) = \frac{18}{53}; Sp = P(\bar{A}|\bar{B}) = \frac{230}{248}; \Omega_A = \frac{P(B|A)}{P(\bar{B}|A)}; \Omega_{\bar{A}} = \frac{P(B|\bar{A})}{P(\bar{B}|\bar{A})}$$

$$w = \frac{\Omega_A}{\Omega_{\bar{A}}} = \frac{SeSp}{(1-Se)(1-Sp)} = \frac{\frac{18}{53} \times \frac{230}{248}}{\left(1-\frac{18}{53}\right)\left(1-\frac{230}{248}\right)} = \frac{0.3396 \times 0.9274}{0.6604 \times 0.0726}$$

$$w = 6.5741 = \hat{w}_i = o = \frac{\hat{\Omega}_{ix'}}{\hat{\Omega}_{ix}}$$

This result is similar to the one obtained using SPSS version 17.

Testing the adequacy of model

Regression analysis showed the following results. Now from Equation 6, where $\beta_j \neq 0$ since we have from analysis that obesity=-0.082, Age=-0.020, Family History(FH)= -0.211, PFW=-0.125 and Constant=3.503.

We here reject H_0 and conclude that the risk factors have significant relationship. Odds ratio values for the risk factors showed Obesity=0.921, Age=0.981, FH=0.810, PFW=0.88 and constant=33.209. Significant values of risk factors are Obesity=0.747, Age=0.981, FH=0.810, PFW=0.882 and constant=33.209. These indicates high significance for their relationship. It also shows the effects of these risk factors on the occurrence of GDM.

Summary and Conclusion

This paper has proposed a statistical method for measuring the probability of success based on some predisposing factors from the association between diagnostic screening test results and state of nature or condition in a population based on the probabilities, odds, odds ratio estimated from logistic regression. A test statistic for the statistical significance of the proposed measure of association are developed based on the estimation of logistic regression of the dichotomous dependent variable and some covariates variables. The proposed method is illustrated with sample data and shown to compare favourably with results that would have been obtained using the traditional expression for the odds ratio test and other statistic for measure of association.

References

1. Hall GH, Round AP (1994) Logistic regression-explanation and use. J R Coll Physicians Lond 28: 242-246.
2. Lee J (1986) An insight on the use of multiple logistic regression analysis to estimate association between risk factors and disease occurrence. Int J Epidemiol 15; 22-29.
3. Fleiss JL, Bruce L, Paik MC (2003) Statistical Methods for Rates and Proportions. (3rdedn), Wiley, New York.
4. Fienberg SE (1980) The Analysis of Cross-Classified Categorical Data (2ndedn). The MIT Press, Cambridge.
5. Hosmer DW, Lemeshow S (1989) Applied Logistic Regression. Wiley, New York.
6. Van Houwelingen JC, le Cessie S (1988) Logistic regression: a review. Stat Neerl 42: 215-232.
7. World Health Organization (1999) Definition, diagnosis and classification of diabetes mellitus and its complications, Part 1. WHO Department of Non communicable disease surveillance, Geneva.