Reporting dichotomous data using Logistic Regression in Medical Research: The scenario in developing countries

Sathian B[1]

[1]Assistant Professor, Department of Community Medicine, Manipal College of Medical Sciences, Pokhara, Nepal

Editorial

Corresponding Author:
Dr. Brijesh Sathian, Assistant Professor, Department of Community Medicine, Manipal College of Medical Sciences, Pokhara, Nepal.
Email: drsathian@gmail.com

Odds ratio and relative risk have been widely applied in public health and medical research. Clinicians often comment that they are more interested in finding out the risk factors for the diseases they treats in their own country. Many medical research problems call for the analysis and prediction of a dichotomous outcome: whether smokers will have a chance of developing lung cancer, hyperuricemia patients have the risk of getting cardio vascular disease. Traditionally, these research questions were addressed by either ordinary least squares (OLS) regression or linear discriminant function analysis. Both techniques were subsequently found to be less than ideal for handling dichotomous outcomes due to their strict statistical assumptions, i.e. linearity, normality, and continuity for OLS regression and multivariate normality with equal variances and covariances for discriminant analysis[1-4]. Logistic regression was proposed as an alternative in the late 1960s and early 1970s[1], and it became routinely available in statistical packages in the early 1980s. Since that time, the use of logistic regression has increased in all science disciplines. The current wide availability of statistical software applications and good statisticians have resulted in the escalated use of logistic regression. But in developing and under developing countries, its use is low. There are several reasons for this, one being limited knowledge regarding what to expect in an article that uses logistic regression techniques, how to understand the methodology and how to report the results. These destroy the applicability of the research data. Some very good clinical research studies from developing countries might not be using logistic regression and they will not explore the data in the manner of good research methodology. This probably results in lowered quality of reporting and the research article might not be accepted in popular indexed journals. In binary logistic regression, we use dichotomous variables.

Two variables used in the logistic regression equation are dependent variables (disease) denoted as Y and independent variable (risk factor) denoted as X. Dichotomous variable is a special case of categorical variable with two outcomes only. Examples of dichotomous variables in Medical fields are in cohort and Clinical trials Y = Cure / no cure, X =Therapy, Other Pt. Variables. In case control studies Y = Case / Control (cancer / non-cancer), X = Risk factors [Age, Sex, Smoking, Occupation]. In cohort studies, Y = MI / No MI, X = Risk factors [Age, Sex, family history etc.]. In the case of looking for a dependence structure, with a dependent variable and a set of explanatory variables (one or more), we can use the logistic regression method. Multiple linear regression may be used to investigate the relationship between a continuous (interval scale) dependent variable, such as Height, Weight, Creatinine, Uric acid and lipid profiles levels. However, socio-demographic and economic variables are very often categorical, rather than interval scale. In many cases, research focuses on models where the dependent variable is categorical. For example, the dependent variable might be Diseased or not otherwise clotting time ≤6 minutes coded as 0 and clotting time >6 minutes coded as 1 (as we saw in Exercise 1) , and we could be interested in how this variable is related to gender, country, blood group, etc. In this case we could not carry out a multiple linear regression as many of the assumptions of this technique will not be met. Instead we would carry out a logistic regression.

Nepal Journal of Epidemiology 2011;1(4):111-113

Published online by NepJOL-INASP
www.nepjol.info/index.php/NJE

Table 1: Cross tabulation of Gender and Clotting time of Nepalese Students[5]

Clotting Time	Gender		
	Male	Female	Total
<6mins	60(93.75)	52(81.25)	112(87.5)
>6mins	4(6.25)	12(18.75)	16(12.5)
Total	64(50)	64(50)	128

Table 1 is a cross tabulation of two binary variables for a sample of 128 students blood clotting time.

- Whether or not the student is perceived to have a clotting time >6 minutes (which we will later model as the response).
- Gender (which we will later model as the explanatory variable).

We can see that the majority of the students (87.5%) are not perceived to have a clotting time >6 minutes and 50% of them are female.

The conditional probabilities of having a clotting time >6 minutes, given gender are shown in square brackets after each of the cell frequencies. For example the probability of being perceived to have a clotting time >6 minutes for male is 0.06, and for female is 0.19.

Odds and Relative Odds

A useful way of using the information in cross tabulations where one dimension of the table is an outcome of interest (whether 2x2 tables or more complicated ones), is to calculate odds and relative odds (odds ratios).

Odds

In the above table, the odds of a male being seen to have a clotting time >6 minutes are 4/60 = 0.0667 or 0.0667 to 1. In betting terms that is about 14.9: 1. For female, the corresponding odds are 12/52= 0.2307, or 0.2307 to 1. It is equivalent to 2.3 to 10. In betting terms that is about 4.3: 1. Note that odds are not the same as probabilities – they are not restricted to the range 0 to 1.

Relative odds

We can also consider the information in the table in terms of relative odds. The relative odds of a female compared with a male being seen as having a clotting time >6 minutes are 0.2307/ 0.0667 or 3.46 to 1. In other words a female is 3.46 times more likely than a male to be seen as having a clotting time >6 minutes.

Equally, students perceived to have clotting time >6 minutes are 3.46 times more likely to be female rather than male, compared with students without perceived clotting time >6 minutes. Relative odds are symmetrical in that sense; like correlation, we do not think of this measure in terms of a dependent variable and an explanatory variable. We just think in terms of the association between two variables.

Application of Logistic regression in Table 1

We can fit a logistic regression model Logit P = $\beta_0+\beta_1X$ to the data in Table 1. We get:

Logit P = -2.708+1.242Gender

Which we can interpret as the log odds of a male student (Gender=0) seen as having a clotting time >6 minutes being equal to –2.708, hence the odds of a male student having a clotting time >6 minutes are: exp (-2.708) = 0.062. The log odds of a female student (Gender=1) having a perceived clotting time >6 minutes are -2.708+1.242= -1.466. Hence the odds of a female student having a clotting time >6 minutes are exp (-1.466) = 2.31. Alternatively, we can say that the odds for female students are exp (1.242) = 3.46 times as high as they are for male. That is, the relative odds of perceiving a female student to have clotting time >6 minutes compared with a male student are 3.46.

More simplified interpretation of Coefficients in Logistic regression

Because of these complicated algebraic translations, our regression coefficients are not as easy to interpret. B represents "the change in Y with one unit change in X" is no longer applicable. Instead, we have to translate using the exponent function. And, as it turns out, when we do that we have a type of "coefficient" that is pretty useful. This coefficient is called the odds ratio. The odds ratio is equal to exp (B), or sometimes written e^B. An odds ratio 2 means that the probability that Y equals 1 is twice as likely as the value of X is increased one unit. An odds ratio of .5 indicates that Y=1 is half as likely with an increase of X by one unit (so there is a negative relationship between X and Y). An odds ratio of 1.0 indicates there is no relationship between X and Y.

Relevance of Sample size determination in the studies with logistic regression analysis

The sample size of any given study should be an optimal value to get valid results from logistic regression analysis of dichotomus data. Logistic regression analysis involves complicated sample size determination formulae. One method is to find out sample size determination formulae from reliable and relevant publications in good medical research journals to calculate the sample size[6,7]. The alternative method is to use existing sample-size calculation software[7].

In Summary, this editorial is a brief introduction to the importance of logistic regression and not revealing the all applications of logistic regression and its uses. Researchers should read good medical statistics and research methodology books and relevant research articles published in reputed medical journals[1-12]. SPSS and other statistical softwares can be used to analyse the Data[13].

Conflict of Interests

The author has no conflict of interest arising from the study.

References

1. Cabrera AF. Logistic regression analysis in higher education: An applied perspective. Higher Education: Handbook of Theory and Research 1994;10:225–56.
2. Cleary PD, Angel R. The analysis of relationships involving dichotomous dependent variables. Journal of Health and Social Behavior 1984;25:334–48.
3. Efron B. The efficiency of logistic regression compared to normal discriminant analysis. Journal of the American Statistical Association 1975;70:892–98.
4. Press SJ, Wilson S. Choosing between logistic regression and discriminant analysis. Journal of the American Statistical Association 1978;73:699–705.
5. Roy B, Banerjee I, Sathian B, Mondal M, Saha CG. Blood Group Distribution and Its Relationship with Bleeding Time and Clotting Time: A Medical School Based Observational Study among Nepali, Indian and Srilankan Students. Nepal Journal of Epidemiology 2011;1(4):135-40.
6. Sathian B, Sreedharan J, Baboo NS, Sharan K, Abhilash E S, Rajesh E. Relevance of Sample Size Determination in Medical Research. Nepal Journal of Epidemiology 2010; 1(1): 4-10.
7. Hsieh FY, Bloch DA, Larsen MD. A simple method of sample size calculation for linear and logistic regression. Stat Med. 1998;17(14):1623-34.
8. Mittal A, Sathian B, Kumar A, Chandrasekharan N, Farooqui MS, Singh S, Yadav KS. Hyperuricemia as an Additional Risk Factor for Coronary Artery Disease: A Hospital Based Case Control Study in Western Region of Nepal. Nepal Journal of Epidemiology 2011;1(3): 81-85.
9. Scott KG, Mason CA, Chapman DA. The use of epidemiological methodology as a means of influencing public policy. Child Development 1999;70(5):1263–72.
10. Banerjee I, Roy B, Sathian B, Banerjee I, Kumar SS, Saha A. Medications for Anxiety: A Drug utilization study in Psychiatry Inpatients from a Tertiary Care Centre of Western Nepal. Nepal Journal of Epidemiology 2010; 1(4):119-25.
11. Sreeramareddy CT, Ramakrishnareddy N, Harsha Kumar HN, Sathian B, Arokiasamy JT. Prevalence, distribution and predictors of tobacco smoking and chewing in Nepal: a secondary data analysis of Nepal Demographic and Health Survey-2006. Substance Abuse Treatment, Prevention, and Policy 2011;6:33.
12. Roy B, Banerjee I, Sathian B, Mondal M, Kumar SS, Saha CG. Attitude of Basic Science Medical Students towards Medicine and Surgery Post Graduation: A Questionnaire based Cross-sectional Study from Western Region of Nepal. Nepal Journal of Epidemiology 2010; 1(4):126-34.
13. SPSS Regression 17. [online] 2007 [cited 2011 December 15]. Available from: http://www.helsinki.fi/~komulain/Tilastokirjat/IBM-SPSS-Spec-Regression.pdf.

Additional Notes: Cover design and journal layout by NJE Associate Managing Editor, Dr. Nishida Chandrasekharan.

Article Information	
Article history	
Received	20 September 2011
Received in revised form	22 September 2011
Accepted	25 September 2011

Published online by NepJOL-INASP
www.nepjol.info/index.php/NJE