Songklanakarin J. Sci. Technol.
38 (2), 171-176, Mar. - Apr. 2016

http://www.sjst.psu.ac.th

Original Article

An alternative method for logistic regression on contingency tables with zero cell counts

Nurin Dureh*, Chamnein Choonpradub, and Phattrawan Tongkumchum

Department of Mathematics and Computer Sciences, Faculty of Science and Technology, Prince of Songkla University, Pattani Campus, Mueang, Pattani, 94000 Thailand.

Received: 20 June 2014; Accepted: 4 October 2015

Abstract

This paper introduces an alternative method for solving a problem of non-convergence in logistic regression. The method does not require any special software to be developed. It simply involves modifying the data by replacing the zero count by 1 and doubling a corresponding non-zero count. The method is compared with that based on penalized likelihood suggested by Firth. Results show that the data modification method provides statistical significance of associations similar to Firth's method while using standard logistic regression output.

Keywords: zero cell count, logistic regression, data modification

1. Introduction

The method we propose extends results given in Dureh *et al.* (2015), where several methods for testing association in two-by-two tables containing at least one small count (possibly zero) were compared. The result show that the Conditional Binomial Exact Test (Rice, 1988), Lancaster's mid-P test (Biddle and Moris, 2011) and the penalized maximum likelihood (Firth, 1993) have similar power in testing association in tables with small marginal totals. In this study, we consider more general situations with a binary outcome and one or more determinants, each of which is a factor with two or more levels. With such data, grouping into a contingency table of counts and logistic regression is commonly used to fit a model. However, when the contingency table has at least one cell containing a zero count, the method may fail to converge (Albert and Anderson, 1984; Aitkin and Chadwick, 2003; Bester and Hansen, 2005; Eyduran, 2008).

A penalized likelihood (PL) procedure to solve this problem for generalized linear models was proposed by Firth (1993) and further studied by Heinze (2006, 2009) and Heinze and Shemper (2002) in logistic regression. Since this method requires special software we considered the possibility of simply modifying the data rather than the method. Lunn and McNeil (1995) used a similar approach for modeling competing risks in survival analysis. Agresti (2002) and Clogg *et al.* (1991) also recommend data modification in preference to new methodology when cell counts are small or data incomplete.

* Corresponding author.
Email address: dnurin@gmail.com

2. Methodology

2.1 Logistic regression model

Suppose Y is a binary response variable where Y=1 denotes an outcome successes (e.g. present of disease) and Y=0 otherwise (absent of disease). We also have a set of covariates $X=(x_1,x_2,\ldots x_{p)}$, which can be discrete, continuous or a combination. If p is the probability of a successful outcome, Pr(Y=1|X), the logistic regression model is given by:

$$\pi = P(Y=1 \mid X) = \frac{1}{1+\exp(-(\beta_0+\beta_1 x_1+\ldots+\beta_p x_p))}$$

or $$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = b_0 + b_1 x_1 + \ldots + b_p x_p$$

In this study we demonstrate the use of DM method for logistic regression with the categorical covariates and extend results for 2x2 tables to $2x2^p$, and similar tables of summary counts.

2.2 Data modification

The data modification method (DM) is improved from the standard approach suggested by Agresti (2002). In a 2 by 2 table with counts *a, b, c* and *d* as in Table 1A, the sample odd ratio $\hat{\theta} = ad/bc$ equals 0 or ∞ if any count is 0, then Agresti's estimator of the OR is

$$\hat{\theta} = \frac{(a + 0.5)(d + 0.5)}{(b + 0.5)(c + 0.5)}$$

To deal with such kind of problem in logistic regression, we introduce a new simple method for which the statistical significance determined by Wald's test from logistic regression aligns closely with Firth's method. The Firth procedure is the current method of choice for logistic regression in tables with zero cell counts (Heinze, 2009); it removes the $O(n^{-1})$ asymptotic bias of the maximum likelihood estimator of the $\log(\hat{\theta})$. Coverage rates of its confidence intervals are shown to be close to nominal values.

Our DM adjustment is similar to Agresti's approach. The modified table is that shown in table 1B after replacing the original cell entries *a, c* by *a** and *c**, while *b* and *d* remain the same as indicated in table 1. Hence $\hat{\theta} = a^*d/bc^*$ with $a^*=2a$ and $c^*=1$ for the data of Table 1A.

The p-values for testing no association between outcome and explanatory variables with the DM method is then calculated by logistic regression, testing a null hypothesis, H_0: $\beta = 0$, where $\beta = \log(\theta)$ is the log(OR). Then $\hat{\beta} = \log\left(\frac{a^* \times d}{b \times c^*}\right)$. Using the Mantel Haenszel test (McNeil, 1996), the standard error $SE(\hat{\beta}) = \sqrt{\frac{1}{a^*} + \frac{1}{b} + \frac{1}{c^*} + \frac{1}{d}}$.

Then the Wald's test statistic is $z = \log\left(\frac{a^* \times d}{b \times c^*}\right) / SE = \frac{\hat{\beta}}{SE(\hat{\beta})}$

However, the standard errors of the log OR from the DM method give incorrect confidence intervals as a consequence of the increased sample sizes. To avoid such bias, we adjust the $SE(\hat{\beta})$ by using the expected counts of *a, b, c, d* (namely, $\hat{a}$, $\hat{b}$, $\hat{c}$ *and* $\hat{d}$), which can be calculated as $n_i{}^*p_i$ where n_i is the total number for each group of independent variables ($n_1=a+c$, $n_2=b+d$) and p_i is the fitted probability of the successful outcome Y=1 for a modified data table. The new standard error is then calculated as $SE(\hat{\beta}) = \sqrt{\frac{1}{\hat{a}} + \frac{1}{\hat{b}} + \frac{1}{\hat{c}} + \frac{1}{\hat{d}}}$

This method generalizes readily to logistic regression models which test the association of categorical explanatory variables with a binary outcome (termed "positive" or "negative") where a zero "positive" count has occurred for some cell within the covariate cross-classification, so that a complete separation of outcomes can be achieved and logistic regression fails to converge (Heinze, 2009). In such cases DM replaces the zero count by 1, and doubles all other cell counts with negative outcomes for the same explanatory variables that correspond to the zero. Then the output from logistic regression of the modified data is used for inference.

3. Results

Example 1: Constructed data set

To illustrate this procedure in 2 by 2 tables, we construct a zero count data set (1A) and a modified data set (1B). The constructed data set consisted of 36 two-by-two tables with n1>4 and two properties: at least one cell contains a zero count; and, the p-value from Firth's method was close to 0.05 (between 0.01 and 0.10).

Each table contains a zero cell and other small counts. These tables fail to satisfy the assumption in Pearson's chi-squared test and also give infinite parameter estimates when using logistic regression. We applied our proposed method to these data and then compared the results with other commonly used tests of associations, including, Fisher's exact test, (Seneta and Phipps, 2001), Lancaster's mid-P test, Agresti's method adding 0.5 to each cell and Firth's method.

P-value for test association in two-by-two tables with zero cell counts

Figure 1 shows p-values given by (a) Firth's method, (b) logistic regression using the DM method, (c) Fisher's

Table 1. General counts of a two-by-two table with a zero count (1A) and modified table (1B).

	1A	
response (y)	group (x)	
	1	0
1(negative)	a	b
0 (positive)	c=0	d

→

	1B	
response (y)	group (x)	
	1	0
1(negative)	a*= a+a	b
0 (positive)	c*=1	d

exact test, (d) Lancaster's mid-P test and (e) the method suggested by Agresti. Logistic regression with the DM method usually agrees closely in p-values with Firth's method and tends to track the p-values of Firth's method. In comparison, the method suggested by Agresti, the Fisher's exact test and Lancaster's mid-P test have higher P-values, consistent with them being more conservative tests of association in 2 by 2 tables (see Seneta and Phipps, 2001). Our findings suggest that the DM method is an appropriate alternative to Firth's method for judging statistical significance of associations in more general logistic regression when zero counts occur in the response variable.

Comparison of standard errors

Standard errors of the log odds ratio are used to compare the accuracy of methods as shown in Figure 2. The standard errors for the DM and Agresti's method are a little smaller than those for the Firth's procedure. The small standard errors provide narrower limits for confidence intervals. Corresponding results were found in the study of Gart and Thomas (1972), which concluded that confidence interval for log odds ratio in logistic regression are generally too narrow, especially when the sample sizes are small.

Example 2: Comparison of p-values using a simulation data set

Data for 2 by 2 table frequencies were simulated using the Poisson and Binomial distributions. In the first case, counts *a, b,* and d are generated from independent Poisson distribution with specific means equal to N*(1-π, 1-π, π) for N=10, 25, 50 and λ=N π =3. However, *c* was forced to be a zero count since our purpose is to study the use of the DM

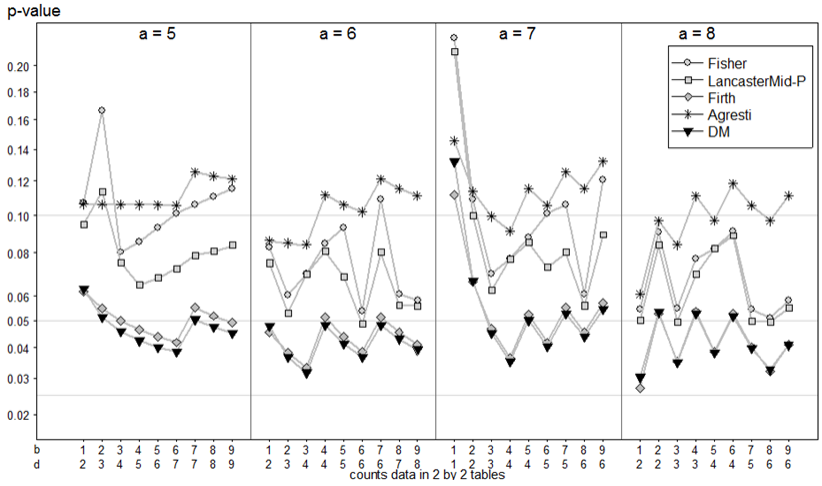

Figure 1. P-values of test for independence in two-by-two tables with a zero count for 36 tables with specified values of the counts (a, b, c, d).

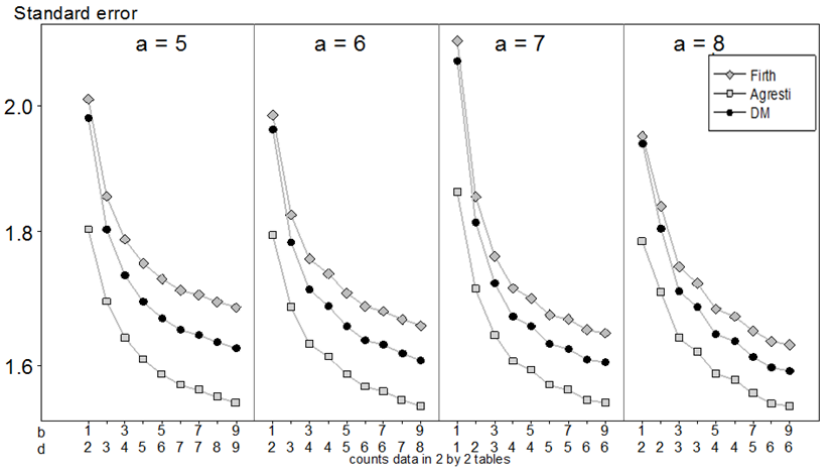

Figure 2. Standard error of log odds ratio of test for independence in two-by-two tables with a zero count for 36 tables with specified values of the counts (a, b, c, d).

method. In addition, we also simulated the data table using the Binomial distribution with the same expected values for counts *a, b, d*, and sample sizes.

The choice $\lambda=N\ \pi=3$ provides tables in which group 1 has an expected 3 cases with positive outcomes. Hence, outcome *d* has corresponding expected value 3 in all simulations. In both groups, the outcomes were generated with corresponding rates of negative results (i.e. 70% probability). We conditioned on the final cell count *c* being 0. The expected number 3 is towards the upper limit for a confidence interval for the cell mean given that 0 counts have occurred.

Figure 3 shows the level of agreement of the p-values from DM method, Fisher's exact test, and Lancaster's mid-p test are compared to p-values for Firth's method. The upper panel graphs provide the results for the data tables simulated from the Poisson distribution, and the lower panel graphs are the results for the data tables simulated from the Binomial distribution. For either distribution, the majority of p-values from DM method fall close to the line of identity with Firth's p-values, for which the two p-values agree exactly. In comparison the other two methods, Fisher's exact test and Lancaster's mid-P test tends to have a larger p-values compare to Firth's. This is consistent with Fisher's test being more conservative than Firth's test.

Example 3: Condom use and first-time urinary tract infection study

The case-control study of Foxman *et al.* (1997) examines urinary tract infection related to age and contraceptive use. The data set consists of 130 college women with urinary tract infections and 109 uninfected controls, and includes binary covariates age (*age*), oral contraceptive use (*oc*), condom use (*vic*), lubricated condom use (*vicl*), spermicide use (*vis*) and diaphragm use (*dia*). There are no cases of women with the uninfected urinary tract and use of diaphragm. This is an example of an aggregated data set where one cell has a zero count. The data are available in the package logistf of the R program (Heinze and Ploner, 2004). Comparing logistic regression results with DM and Firth's method gives results as shown in Table 2.

The two methods give similar results. Factors *age*, *vic*, *vicl* and *dia* are associated with urinary tract infection with p-values less than 0.05. However, when the standard errors of the log odds ratio in the model are considered, the DM method gives smaller estimates of effects and standard errors and correspondingly shorter 95% confidence intervals than those for Firth's method.

Example 4: Child Deaths from External cause in Thailand

The data here are based on the Thai 2005 Verbal Autopsy (VA) study (Rao *et al.*, 2010) for correcting misreported cause of death for children under five. The data consists of one determinant, DR.hGrp, which is the combined variable of reported cause of death and place of death (inside/outside hospital). The binary outcome is whether the child died from perinatal (ICD chapter P) or congenital (chapter Q) causes versus other causes. These data are listed in the left panels of Table 3 with modified data for using the DM method asterisked in the right panel.

DM and Firth's method return similar results for coefficients, standard errors of log odds ratios and p-values as shown in Table 4.

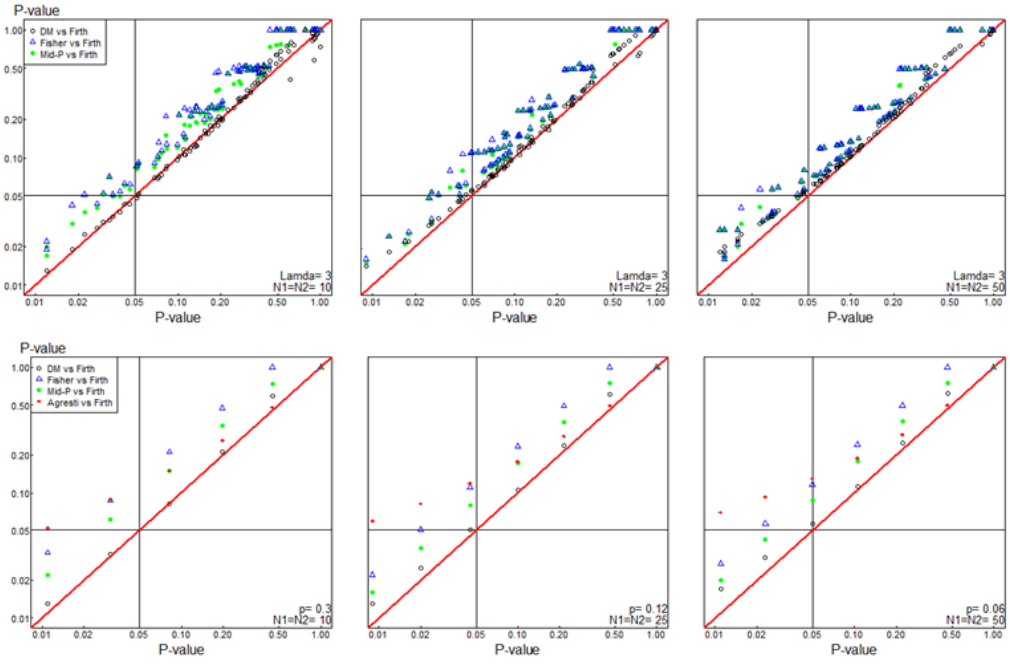

Figure 3. P-values from Fisher's exact test, Lancaster's mid-p test and DM method compared with Firth's method.

Table 2. Logistic regression analysis of condom use and first-time urinary infection study.

Variable	DM				Firth's method			
	coef (coef)	SE (95% CI)	OR	p-value	coef (coef)	SE (95% CI)	OR	p-value
age	-1.07	0.39	0.34 (0.16,0.75)	0.007	-1.11	0.42 (0.14,0.76)	0.33	0.006
oc	-0.15	0.43	0.86 (0.37,2.02)	0.731	-0.07	0.44 (0.39,2.23)	0.93	0.875
vic	2.04	0.51	7.72 (2.85,20.94)	<0.001	2.27	0.55 (3.30,28.33)	9.67	<0.001
vicl	-1.92	0.50	0.15 (0.06,0.39)	<0.001	-2.11	0.54 (0.04,0.35)	0.12	<0.001
vis	-0.81	0.41	0.45 (0.20,1.00)	0.048	-0.79	0.42 (0.20,1.03)	0.45	0.054
dia	1.16	1.04	3.18 (0.41,24.54)	0.052	3.10	1.67 (0.83,589.36)	22.11	0.005

Table 3. Number of child deaths from congenital and other causes.

DR.hGrp	Cause of deaths		Cause of deaths*	
	Other	Congenital	Other	Congenital
Perinatal inside hospital	9	3	9	3
Congenital inside hospital	3	0	6*	1*
External+ inside hospital	18	25	18	25
All causes outside hospital	21	24	21	24

*Modified data using DM method.

Table 4. Logistic regression analysis of number of child deaths from congenital and other causes.

Variable	DM			Firth's method		
	coef	se(coef)	p-value	coef	se(coef)	p-value
Intercept	-1.099	0.667	0.099	-0.999	0.651	0.089
Perinatal inside hospital	0	-	-	-	-	-
Congenital inside hospital	-0.693	1.792	0.585	-0.947	1.863	0.531
External+inside hospital	1.427	0.735	0.052	1.319	0.720	0.046
All causes outside hospital	1.232	0.731	0.092	1.129	0.716	0.087

In this analysis the p-values are based on contrasts between the omitted level for the factor (perinatal inside hospital) and each other level, and we see that only one of these differences (perinatal versus external+) is statistically significant at the 5% level. A p-value for testing the hypothesis that there is no mortality difference between the three cause groups is provided by an Anova test, which has p-value 0.038 for these data based on the DM method. While p-values for LR test and Wald test given by Firth's method are 0.081 and 0.193, respectively.

4. Discussion and Conclusions

This study provides an alternative method for solving the problem of non-convergence in logistic regression. Firth's method has previously been recommended for analysis data with such a problem (Heinze and Schemper, 2002; Eyduran, 2008), but in this study it was found that the data modification (DM) method generally provides smaller p-values to those from Firth's method. However, in 2 by 2 tables, with small total counts, we have consistent evidence that the

results of DM and Firth's method align closely. While Agresti's method is used for the zero count problems, especially in two-by-two tables, the DM method gives closer result to Firth's method. We have demonstrated that the DM method can be used as an alternative to Firth's method in more general logistic regression when zero counts occur in the response variable and observed the same close correspondence in results. The DM method uses logistic regression methods for maximum likelihood estimation. Logistic regression methods are well known and have the advantage of not requiring more specialized statistical software. The DM method might also be applicable with continuous covariates, but this possibility needs to be considered in further study comparing methods.

The user should be aware too of the potential bias of DM as an estimator of the log-OR and its standard error (underestimated). This bias occurs in tables of small cell counts (e.g. in Table 2 for the factor dia), including the situation of separation. It is known that the Wald test and confidence interval become unsuitable (Heinze and Shemper, 2002). However, the DM estimator holds the correct level of significance in the association, as judged by Firth's method. In examples other than small 2 by 2 tables this bias was less evident, as regression coefficients as well as SE's more closely agreed.

Acknowledgements

This research received financial support from the Thailand Research Fund through the Royal Golden Jubilee Ph.D. Program. We are grateful to Emeritus Professor Don McNeil for his guidance and to referees for their helpful comments.

References

Aitkin, M. and Chadwick, T. 2004. Bayesian analysis of 2x2 contingency tables from comparative trials. Proceeding of 24th Conference on Applied Statistics in Ireland, Gulway, Ireland, May 12-14, 2004.

Agresti, A. 2002. Categorical data analysis. John Wiley and Son, New York, U.S.A., pp. 70-71.

Albert, A. and Anderson, J.A. 1984. On the Maximum Likelihood Estimates in Logistic Regression Models. Biometrika. 71, 1-10.

Bester, C.L. and Hansen, C. 2005. Bias reduction for Bayesian and frequentist estimators. Working paper, Graduate School of Business, University of Chicago, U.S.A.

Biddle, D.A. and Morris, S.B. 2011. Using Lancaster's mid-P correction to the Fisher's exact test for adverse impact analysis. Journal of Applied Psychology. 96, 956-965.

Clogg, C.C., Rubin, D.B., Schenker, N., Schultz, B. and Weidman, L. 1991. Multiple Imputation of Industry and Occupation Codes in Census Public-use Samples Using Baysian Logistic Regression. Journal of the American Statistical Association. 86, 68-78.

Dureh, N., Choonpradub, C. and Tongkumchum, P. 2015. Comparing Tests for Association in Two-by-Two Tables with Zero Cell Counts. Chiang Mai Journal of Sciences. 42, 1031-1037.

Eyduran, E. 2008. Usage of Penalized Maximum Likelihood Estimation Method in Medical Research: An Alternative to Maximum Likelihood Estimation Mehod. Journal of Research in Medical Sciences. 13, 325-330.

Firth, D. 1993. Bias reduction of maximum likelihood estimates. Biometrika. 80, 27-38.

Foxman, B., Marsh, J., Gillespie, B., Rubin, N., Kopman, J.S. and Spear, S. 1997. Condom Use and First-Time Urinary Tract Infection. Epidemiology. 8, 637-641.

Heinze, G. and Schemper, M. 2002. A solution to the problem of separation in logistic regression. Statistic in Medicine. 21, 2409-2419.

Heinze G. 2006. A comparative investigation of methods for logistic regression with separated or nearly separated data. Statistics in Medicine. 25, 4216-4226.

Heinze, G. 2009. Avoiding infinite estimates in logistic regression- theory, solutions, examples. Medical University of Vienna. Available from: https://www.researchgate.net/publication/255594296_Avoiding_infinite_ estimates _in_logistic_regression__theory_solutions_ examples. [October 13, 2014].

Heinze, G. and Ploner, M. 2004. A SAS macro, S-PLUS library and R package to perform logistic regression without convergence problems. Technical report, Medical University of Vienna, Department of Medical Computer Sciences, Section of Clinical Biometrics.

Lunn, M. and McNeil, D. 1995. Applying Cox Regression to Competing Risks. Biometrics. 51, 524-532.

McNeil, D. 1996. Epidemiological Research Method. John Wiley and Sons. New York, U.S.A.

Rao, C., Porapakkham, Y., Pattaraarchachai, J., Polprasert, W., Swampunyalert, W. and Lopez A.D. 2010. Verifying causes of death in Thailand: rationale and methods for empirical investigation. Population Health Metrics. 8, 11.

Rice, W.R. 1988. A New Probability Model for Determining Exact P-values for 2x2 Contingency Tables When Comparing Binomial Proportions. Biometrics. 44, 1-22.

Sean, R.E. 2004. What is Bayesian statistics?, Nature Biotechnology. 22, 1177-1178.

Seneta, E. and Phipps, M.C. 2001. On the Comparison of Two Observed Frequencies. Biometrical Journal. 43, 23-43.