An Alternatives Method for Fitting Logistic Regression to Grouped Data with Zero Counts

Nurin Dureh

A Thesis Submitted in Fulfillment of the Requirements for the Degree of

Doctor of Philosophy in Research Methodology

Prince of Songkla University

2015

Thesis Title An Alternative Method for Fitting Logistic Regression to Grouped Data with Zero Counts

Author Mrs. Nurin Dureh

Major Program Research Methodology

Major Advisor

.....

(Asst. Prof. Dr. Chamnein Choonpradub)

Co-advisor

.....

(Emeritus Prof. Dr. Don McNeil)

Examining Committee:

.....Chairperson

(Dr. Attachai Ueranantasun)

.....

(Asst. Prof. Dr. Chamnein Choonpradub)

.....

(Emeritus Prof. Dr. Don McNeil)

.....

(Assoc. Prof. Dr. Rohana Binti Jani)

.....

(Assoc. Prof. Dr. Halimah Binti Awang)

The Graduate School, Prince of Songkla University, has approved this thesis as fulfillment of the requirements for the Doctor of Philosophy Degree in Research Methodology.

.....

(Assoc. Prof. Dr. Teerapol Srichana)

Dean of Graduate School

This is to certify that the work here submitted is the result of the candidate's own investigations. Due acknowledgement has been made of any assistance received.

.....Signature

(Asst. Prof. Dr. Chamnein Choonpradub)

Major Advisor

.....Signature

(Mrs. Nurin Dureh)

Candidate

I hereby certify that this work has not been accepted in substance for any degree, and is not being currently submitted in candidature for any degree.

.....Signature

(Mrs. Nurin Dureh)

Candidate

ชื่อวิทยานิพนธ์ วิธีการสร้างโมเดลการถดถอยลอจิสติก กรณีข้อมูลแบบกลุ่ม เมื่อมีความถี่เป็นศูนย์
ผู้เขียน นางนูริน ดือเร๊ะ
สาขาวิชา วิธีวิทยาการวิจัย
ปีการศึกษา 2558

บทคัดย่อ

ตัวแบบการถดถอยลอจิสติกเป็นตัวแบบที่นิยมใช้สำหรับการทดสอบความสัมพันธ์ระหว่างตัวแปรต้นและตัวแปรตามที่มีจำนวนสองกลุ่ม แต่ในกรณีข้อมูลแบบกลุ่มมีความถี่เป็นศูนย์ ตัวแบบการถดถอยลอจิสติกไม่สามารถประมาณค่าพารามิเตอร์ได้หรือมีความผิดพลาดในการประมาณการ (non-convergence) การศึกษานี้เสนอวิธีการทางเลือกสำหรับการแก้ปัญหาดังกล่าว โดยใช้หลักการของการปรับค่าข้อมูลแทนการคิดค้นโปรแกรมหรือคำสั่งเฉพาะสำหรับการวิเคราะห์ข้อมูลลักษณะนี้ ใช้วิธีการแทนที่ความถี่ที่มีค่าเป็นศูนย์ด้วยค่าหนึ่ง และความถี่ที่อยู่ในกลุ่มตัวแปรต้นกลุ่มเดียวกันกับค่าศูนย์จะถูกเพิ่มค่าเป็นสองเท่า เรียกวิธีการนี้ว่า Data Modification (DM) จากนั้นจึงนำข้อมูลที่ได้จากการปรับค่าไปวิเคราะห์ด้วยโปรแกรมสำเร็จรูปที่มีอยู่ทั่วไป และเปรียบเทียบผลลัพธ์กับวิธีการที่มีอยู่เดิม คือวิธีการของ Firth ซึ่งใช้หลักการ penalized likelihood estimation ในการประมาณค่าพารามิเตอร์ ผลการศึกษาพบว่าวิธีการของ DM ให้ค่าระดับนัยสำคัญ (p-value) ที่ใกล้เคียงกับวิธีการที่มีอยู่เดิม

Thesis Title	An Alternative Method for Fitting Logistic Regression to Grouped Data with Zero Counts
Author	Mrs. Nurin Dureh
Major Program	Research Methodology
Academic Year	2015

ABSTRACT

Logistic regression is the commonly used for testing the association between binary outcome and a set of explanatory variables. When the data tables contains at least one zero count, the logistic regression does not converge. This study introduce an alternative method for solving the non-convergence problem in logistic regression. The method does not require any special software to be develop. 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 the existing method including the penalized likelihood suggested by Firth. Results show that the data modification method provides statistical significance of associations similar to Firth's method while using the standard logistic regression.

Acknowledgment

I would like to express my gratitude and deep appreciation to my advisor Asst. Dr. Chamnein Choonpradub and to my co-advisor Emeritus Professor Dr. Don McNeil for their considerable tuition and kind help throughout the completion of the thesis. I would like to thank Professor. Dr. Malcolm Hudson and Dr. Hilary Green for their very kindness help and very useful suggestions to my thesis. I would like to thank all of students in Research Methodology program for the encouragement. I would like to thank for The Royal Golden Jubilee Ph.D. Program under Thailand Research Fund for financial support and thank you to the Graduate School, Prince of Songkla University for partially funding this study. Finally, I would like to express my thankfulness to my family and all friends for cheering me up throughout this study.

Nurin Dureh

Content

List of Tables

List of Figures

CHAPTER 1

Introduction

Logistic regression is a method that have been widely use for testing the association in two by two tables. However, when any counts in table equal to zero, this method does not converge. In practice, there are existing method which can be solve this problem but using the special software. Therefore, this thesis suggest a new way to solving the problem by using the data modification instead of any statistical packages.

1.1 Rational for study

For analysis of 2 by 2 tables, the simplest case for contingency tables, there are many methods for getting p-values, some "exact", and they give widely varying results. The most common test is Pearson's chi-squared test, which is appropriate for sufficiently large sample sizes. It is inaccurate if any expected count is less than five (Mehta and Patel, 1997; Mehta and Senchaudhuri, 2003; Seneta and Phipps, 2001). In case of small sample sizes, Fisher's exact test is the most used (Mehta and Patel, 1997; Mehta and Senchaudhuri, 2003; Seneta and Phipps, 2001) and it is based on an "exact conditional approach". This approach can eliminate the nuisance parameter in the model under the null hypothesis by conditioning on its marginal totals (Mehrotra *et al*., 2003).

Another way to reduce the conservatism of Fisher's exact test is to consider an unconditional approach, such as Barnard's test. Barnard's test eliminates the nuisance parameter by taking its supremum over all possible values in the space of the null model (Lin and Yang, 2009; Lydersen *et al.*, 2009). Moreover, there is a concern

regarding the default use of Fisher's exact test so several alternative tests have been proposed (Biddle, 2011). These include Lancaster's mid-p test (King and Zeng, 2001; Lancaster, 1961), an adjustment to the Fisher's exact test that tend to have increased power while maintaining a Type I error rate close to the nominal level (Biddle, 2011; Lydersen, 2009). Liebermeister's test is also can be used in place of Fisher's exact test, and is less conservative than Fisher's test and just as easy to calculate (Seneta and Phipps, 2001). In addition, the "Conditional Binomial Exact Test" (CBET) is proposed as an alternative test for comparing binomial proportions estimated from samples of larger populations (Rice, 1988).

Logistic regression has been used commonly in contingency tables. It provides a more general method because it provides a model that accommodates more complex determinants. However, when one of any cell in the contingency table equal to zero, standard errors of parameters estimated by maximum likelihood method are too large and biased. Thus, logistic regression fails to converge (Bester and Hansen, 2005; Biddle *et al.*, 2011; Eyduran, 2008; Len and Yang, 2009; Rice, 1988; Sean, 2004; Seneta and Phipps, 2001). A procedure to solve this problem was proposed by Firth (1993). This method gives finite parameter estimates via penalized maximum likelihood (Firth, 1993; Heinze and Schemper, 2002; Heinze, 2009b; Heinze and Ploner, 2003a; Heinze and Ploner, 2004b). This method is available in statistical software such as SAS, S-PLUS and R (Heinze, 2006a; Heinze, 2009b; Heinze and Ploner, 2004b). However, the estimates from this procedure are biased away from zero (Heinze and Ploner, 2003a). Thus, this solution has limited use in practice because the bias may be quite substantial.

Tables with small count, especially with zero cell count thus lead to numerical problems (Brown, 1983), so it is important to identify the methods which provide the accurate results for particular data structures. Thus the main objective of this study is finding the method that provides more exact results using logistic regression. The procedure is based on making a small modification on contingency tables. The bias is reduced from logistic regression without changing the method as in Firth's procedure, but modifying the data instead, and thus not requiring any new software to be developed.

1.2 Review of literature

1.2.1 Convergence problem in logistic regression

In logistic regression, the well-known method for parameter estimation is "maximum likelihood estimation". A frequent problem in estimating logistic regression models is a failure of the likelihood maximization algorithm to converge. In several cases, this failure is a consequence of data patterns known as complete or quasi-complete separation. This section review the overview of maximum likelihood estimation including the example of data with separation with their possible solutions. The information were collected from Allison (2004) and Allison (2008).

Logistic maximum likelihood estimation

For a sample of n case $(i=1,\ldots,n)$, there are data on a dummy dependent variable y_i (with values of 1 and 0) and a vector of explanatory variables x_i (including a 1 for the intercept term). The logistic regression model states that:

$$\Pr(y_i=1|x_i) = \frac{1}{1+\exp(-\beta x_i)} \qquad (1.1)$$

where β is a vector of coefficients. Equivalently, we may write the model in "logit" form:

$$\mathrm{In}\left[\frac{\Pr(y_i = 1 \mid x_i)}{\Pr(y_i = 0 \mid x_i)}\right] = \beta \mathrm{x}_\mathrm{i} \qquad (1.2)$$

Assuming that the *n* cases are independent, the log-likelihood function for this model is:

$$\ell(\beta) = \sum_\mathrm{i} \beta \mathrm{x}_\mathrm{i} y_i - \sum_\mathrm{i} \mathrm{In}[1 + \exp(\beta \mathrm{x}_\mathrm{i})] \qquad (1.3)$$

The goal of maximum likelihood estimation is to find a set of values for β that maximize this function. One well-known approach to maximizing a function like this is to differentiate it with respect to β, set the derivative equal to 0, and then solve the resulting set of equations. The first derivative of the log-likelihood is:

$$\frac{\partial \ell(\beta)}{\partial(\beta)} = \sum_\mathrm{i} x_i y_i - \sum_\mathrm{i} x_i \hat{y}_i \qquad (1.4)$$

where $\hat{y}_i$ is the predicted value of y_i:

$$\hat{y}_i = \frac{1}{1 + \exp(-\beta \mathrm{x}_\mathrm{i})} \qquad (1.5)$$

The next step is to set the derivative equal to 0 and solve for β:

$$\sum_\mathbf{i} \mathbf{x_i} y_i - \sum_\mathbf{i} \mathbf{x_i} \hat{y}_i = 0 \qquad (1.6)$$

Because β is a vector, (1.6) is actually a set of equations, one for each of the parameters to be estimated. These equations are identical to the "normal" equations

for least-squares linear regression, except that by equation 1.4, y is a non-linear function of the x_i's rather than a linear function.

For some models and data (e.g., "saturated" model), the equation 1.6 can be explicitly solved for the ML estimator b. For example, suppose there is a single dichotomous x variable, so that the data can be arrayed in a 2 by 2 table, with observed cell frequencies f_{11}, f_{12}, f_{21}, and f_{22}. Then the ML estimator of the coefficient of x is given by the logarithm of the "cross-product ratio":

$$\hat{\beta} = \log\left[\frac{f_{11} f_{22}}{f_{12} f_{21}}\right] \qquad (1.7)$$

For most data and models, however equations 1.6 have no explicit solution. In such cases, the equations must be solved by numerical methods, of which there are many. The most popular numerical method is the Newton-Raphson algorithm. Let

U (β) be the vector of first derivatives of the log-likelihood with respect to β and let

I (β) be the matrix of second derivatives. That is,

$$\mathbf{U}(\beta) = \frac{\partial \ell(\beta)}{\partial \beta} = \sum_i \mathrm{x}_i y_i - \sum_i \mathrm{x}_i \hat{y}_i$$

$$\mathbf{I}(\beta) = \frac{\partial^2 \ell(\beta\beta}{\partial \beta \partial \beta'} = \sum_i \mathrm{x}_i \mathrm{x}_i' \hat{y}_i (1 - \hat{y}_i) \qquad (1.8)$$

The vector of first derivatives $\mathbf{U}(\beta)$ is sometimes called the *gradient* or *score* while the matrix of second derivatives $\mathbf{I}(\beta)$ is called the *Hessian.* The Newton-Raphson algorithm is then

$$\beta_{j+1} = \beta_j - \mathbf{I}^{-1}(\beta_j)\mathbf{U}(\beta_j) \qquad (1.9)$$

where $\mathbf{I}^{-1}$ is the inverse of $\mathbf{I}$.

To operationalize this algorithm, a set of starting values β_0 is required. Choice of starting values is not critical; usually, setting $\beta_0 = 0$ works fine. The starting values are substituted into the right-hand side of equation 1.9, which yields the results for the first iteration, β_1. These values are then substituted back into the right hand side, the first and second derivatives are recomputed, and the results is β_2. The process is repeated until the maximum change in each parameter estimate from one iteration to the next is less than some criterion, at which point we say that the algorithm has converged. Once we have the results of the final iteration, $\hat{\beta}$, by product of the Newton-Raphson algorithm is an estimate of the covariance matrix of the coefficients, which is just $-\mathbf{I}^{-1}(\beta)$. Estimates of the standard errors of the coefficients are obtained by taking the square roots of the main diagonal elements of this matrix.

What can go wrong?

A common problem in maximizing a function is the presence of local maxima. Fortunately, such problems cannot occur with logistic regression because the log-likelihood is globally concave, meaning that the function can have at most one maximum. Unfortunately, there are many situations in which the likelihood function has no maximum, in which case we say that the maximum likelihood estimate does not exist. Consider the set of data on 10 observations in Table 1.1, Y is binary outcome and X is continuous covariates.

Table 1.1: Data exhibiting complete separation.

x	y	x	y
-5	0	1	1
-4	0	2	1
-3	0	3	1
-2	0	4	1
-1	0	5	1

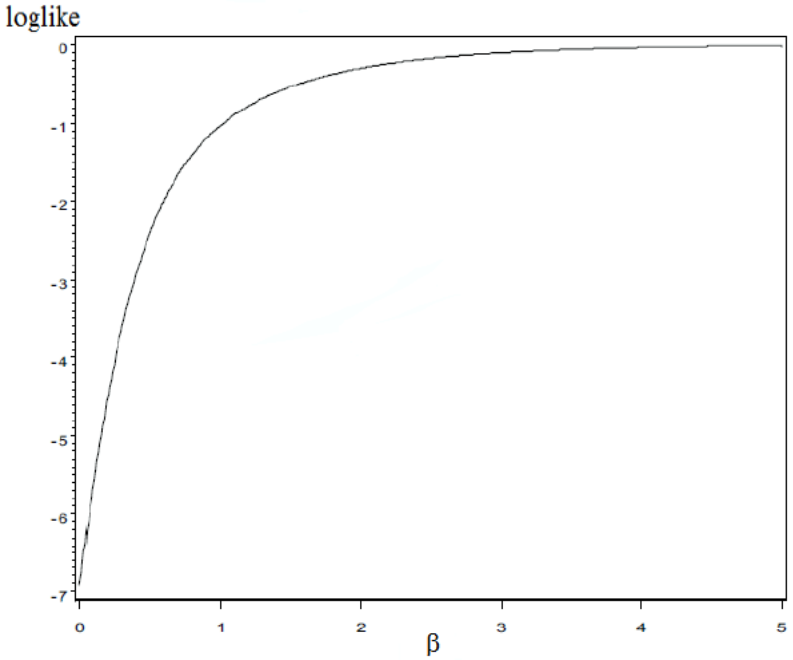


Figure 1.1: Log-likelihood as a function of the slope under complete separation

It is apparent that, although the log-likelihood is bounded above by 0, it does not reach a maximum as beta increases. We can make the log-likelihood as close to 0 as we choose by making beta sufficiently large. Hence, there is no maximum likelihood estimate.

This is an example of a problem known as complete separation (Albert and Anderson, 1984), which occurs whenever there exists some vector of coefficients b such that $y_i = 1$ whenever $bx_i > 0$ and $y_i = 0$ whenever $bx_i < 0$. In other words, complete separation occurs whenever a linear function of x can generate perfect predictions

of y. For our hypothetical data set, a simple linear function that satisfies this property is 0+1(x). That is, when x is greater than 0, y=1, and when x is less than 0, y=0.

A related problem is known as *quasi-complete separation*. This occurs when (a) there exists some coefficient vector b such that $bx_i \geq 0$ whenever $y_i = 1$, and $bx_i \leq 0$ whenever $y_i = 0$, and equality holds for at least one case in each category of the dependent variable. Table 1.2 displays a data set that satisfies this condition.

Table 1.2: Data exhibiting quasi-complete separation.

x	y	x	y
-5	0	1	1
-4	0	2	1
-3	0	3	1
-2	0	4	1
-1	0	5	1
0	0	0	1

The difference between this data set and the previous one is that there are two more observations, each with *x* values of 0 but having different values of y.

The log-likelihood function for these data, shown in Figure 1.2, is similar in shape to that in Figure 1.1. However, the asymptote for the curve is not 0, but a number that is approximately -1.39. In general, the log-likelihood function for quasi-complete separation will not approach 0, but some number lower than that. In any case, the curve has no maximum so, again, the maximum likelihood estimate does not exist.

Of the two conditions, complete and quasi-complete separation, the latter is far more common. It most often occurs when as explanatory variable x is a dummy variable and, for one value of x, either every case has the event y=1 or every case has the event y=0. Consider the following 2 by 2 table:

		y	
		1	0
x	1	5	0
	0	15	10

If we form the linear combination $c = 0 + (1)\,x$, we have $c \geq 0$ when $y = 1$ and $c \leq 0$ when $y = 0$. So the conditions of quasi-complete separation are satisfied.

To get some intuitive sense of why this leads to non-existence of maximum likelihood estimator, consider equation 1.7 which gives the maximum likelihood estimator of the slope coefficient for a 2 by 2 table. For our quasi-complete table, this would be undefined because there is a zero in the denominator. The same problem would occur if there were a zero in the numerator because the logarithm of zero is also undefined. If the table is altered to read:

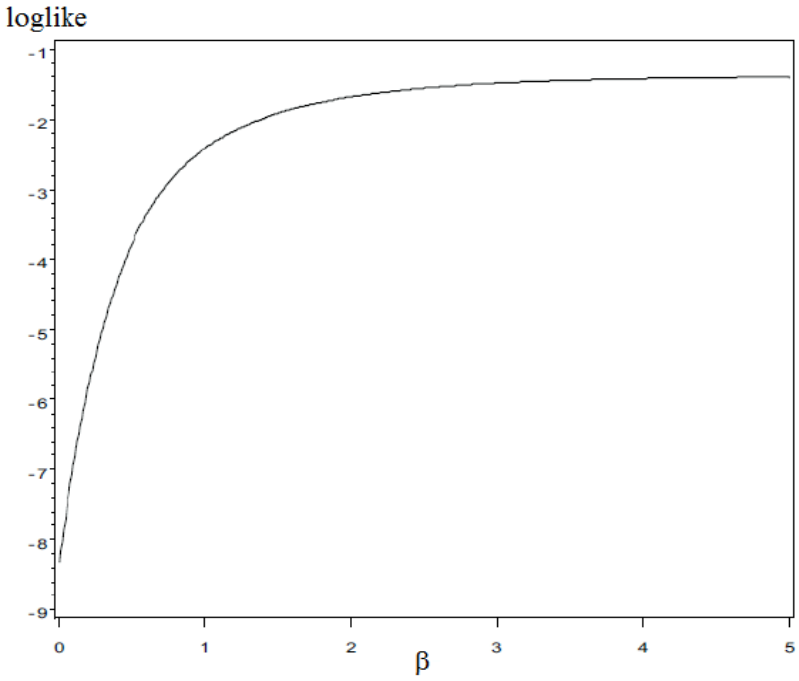


Figure 1.2: Log-likelihood as a function of the slope under quasi-complete separation

		y	
		1	0
x	1	5	0
	0	0	10

then there is complete separation with zeros in both the numerator and the denominator. So the general principle is evident. Whenever there is a zero in any cell of a 2 by 2 table, the maximum likelihood estimate of the logistic slope coefficient does not exist. This principle also extends to multiple independent variables:

For any dichotomous independent variable in a logistic regression, if there is a zero in the 2 by 2 table formed by that variable and the dependent variable, the ML estimate for the regression coefficient will not exist.

This is by far the most common cause of convergence failure in logistic regression. Obviously, it is more likely to occur when the sample size is small. Even in large samples, it will frequently occur when there are extreme splits on the frequency distribution of either the dependent or independent variables. Consider, for example, a logistic regression predicting the occurrence of a rare disease. Suppose further, that the explanatory variables include a set of seven dummy variables representing different age levels. It would not be terribly surprising if no one had the disease for at least one of the age levels, but this would produce quasi-complete separation.

1.2.2 Solutions for non-convergence problem in logistic regression

As mentioned by Santos and Barrios (2012), there are several solutions to solve the non-convergence problem. The most popular is dropping the separating variable(s) in the model. Clogg et al (1991) suggest to adding “artificial” data across the different patterns of categorical covariates and analysis is done on modified data. Another

approach is to use exact logistic regression that allows estimation of the coefficients even in the presence of empty cells and complete separation.

The exact logistic regression uses the method of conditional maximum likelihood in performing exact inference for a parameter yielding exact p-values rather than approximations. This method prevents infinite estimated odds ratios or confidence intervals with one side equal to infinity, see in Agresti (1996). However, Zorn (2005) revealed that the exact logistic regression may result to degenerate estimates when relatively sparse data or small number of observations in particular patterns of categorical predictors is present.

Another approach in solving the problem of separation of likelihood is the modified score procedure conducted by Firth (1993). The procedure modifies the maximum likelihood estimation by penalizing the score equation. The method reduces the bias on the maximum likelihood estimates for the coefficients of the logistic regression model. This approach are recommended by many studies as follows;

Heinze and Shamper (2002), suggested that a procedure by Firth originally developed to reduce the bias of maximum likelihood estimates and it is shown to provide an ideal solution to separation. It produced finite parameter estimates by means of penalized maximum likelihood estimation. Corresponding Wald test and confidence interval are available but it is shown that penalized likelihood ratio tests and profile penalized likelihood confidence intervals often preferable.

According to the suggestion by Heinze and Shamper (2002), Heinze and Ploner (2003a, 2004b) then developed a SAS macro and SPLUS library to make Firth method available from within of these widely used statistical software packages.

Heinze (2006), provide some examples of separation and near-separation in clinical data sets and give the options to analyses such data, including exact logistic regression and penalized likelihood approach. Both methods supply finite point estimates in case of separation. Profile penalized likelihood confidence intervals for parameters show excellent behavior in terms of coverage probability and provide higher power than exact confidence intervals.

Heinze (2009) study the data when the phenomenon of separation occurred. Example of two studies are given: the first one investigated whether primary graft dysfunction of lung transplants is associated with endothelin-1 mRNA expression measured in lung donors and in graft recipients. Second one is using the conditional logistic regression to analyses a randomized animal experiment in which animals were clustered into sets defined by equal follow-up time. The estimates obtained by a penalized likelihood approach have reduced bias compared to their maximum likelihood counterparts, and inference using penalized profile likelihood is straightforward.

Toshinari (2011) proposed new method based on the bootstrap resampling techniques and compare the true p-values for the likelihood ratio test, Wald test, Firth method and other testing methods. The Firth method has a good property that it gives the bias reduction of maximum likelihood estimation and the stable estimates are obtained even for the nearly quasi-separations case.

For more options, the user can applied the data modification into such cases of problem (separation/quasi-complete separation). Some studies mentioned about this approach as following;

Clogg *et al* (1991) considered the possibility of resolving the separation problem by adding "artificial" data across the different patterns of the categorical predictors and then conducting the analysis in the resulting data in the usual manner. This study consider sapling strategy as a possible solution to the separation problem. Since the separation problem usually arises from the existence of "patterns" among the data on the predictors, then it is possible that the problem is avoided if the likelihood of such "pattern" is minimized.

Similar approach, Gart and Zweifel (1967) and Haldane (1955) also suggest the data modification by adding .5 to every cell in table and the estimated odds ratio by this approach are behave well, mentioned by Agresti (2002).

Related the data modification approach, Parzen *et al* (2002) propose an estimate of the odds ratio in 2 by 2 table obtained from studies in which the row totals are fixed by design, such as phase II clinical trial. The estimation of the odds ratio is based on the median unbiased estimated of the probabilities of success in the 2 by 2 table compare with the estimated odds ratio by adding .5 to each cell of the table. Found that, the median unbiased estimate had smaller finite sample bias and larger mean square error.

Apart from those previous solution, there are another option to solve the non-convergence problem. Santos and Barrios (2012) propose a study by drawing the sample using ranked set sampling (RSS). An extensive simulation study was

conducted to assess the performance of logistic regression model fitted from ranked set samples and compared to those estimate using simple random samples (SRS). RSS performs best in small populations regardless of the distribution of the binary response variable in the population. As the sample and population sizes increase, the predictive ability under RSS also improves but it stabilizes to become comparable to SRS.

Since the solution of using penalized likelihood estimation (Firth method) is seem to be recommended. While the data modification approach are less used. Therefore, this study consider on the method involve data modification and provides the similar results to Firth method.

1.3 Objectives for studies

1. To measure the bias in the penalized likelihood method compared to other recommended tests of independence in 2 by 2 tables containing a zero and other small counts, when the true p-value is approximately 0.05.

2. To provide an alternative test based on making a small modification to the data similar to that used by Liebermeister’s and Lancaster’s mid-P tests.

3. To compare this method with p-values given by other recommended tests (CBET, Lancaster's mid-P, Pearson's, Fisher's exact test, Liebermiester's modification, Barnard’s test, Penalized likelihood ratio test, and logistic regression) when the true p-value is close to 0.05.

4. To provide an alternative to penalized likelihood that does not require special-purpose software, being able to run using ordinary logistic regression.

CHAPTER 2

Methodology

This chapter contains all the methods which involved in this study. Section 2.1 contains a brief description of the existing methods for testing the association in contingency tables. Section 2.2 contains the basic knowledge of the logistic regression with maximum likelihood estimation and logistic regression with penalize maximum likelihood estimation (Firth's procedure). Section 2.3 contains methods of data modification which including the existing approach (data augmentation by Clogg's and adding 0.5 to each cell by Agresti's) and an alternative idea with the similar approach, namely Data Modification, DM method. Section 2.4 contains more explanation for DM method.

2.1 Methods for testing the association in contingency tables

There are several methods for significance test for the association in 2 by 2 tables. A brief summary of these tests may describe as follows.

Table 2.1: The general counts of a 2 by 2 table.

		j		
		1	2	Total
i	1	*a*	*b*	m
	2	*c*	*d*	n
	Total	z	v	m+n

Suppose we have the 2 by 2 table as shown in Table 2.1. The most commonly used test statistics for association are Pearson's chi-squared test and Fisher's exact test.

1) Pearson's chi-squared test

This test is an asymptotic test that approximates the p-value by the upper tail probability of the chi-squared distributed with one degree of freedom (Lyderson *et al.*, 2009).

The formula used to calculate the Pearson's chi-squared test is

$$\chi^2 = \frac{m+n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} \qquad (2.1)$$

In general, the p-value is defined as the probability of the test statistic *T* being equal to or more extreme than its value for the observed table *(t_{obs})*, therefore, the approximate p-value for Pearson's chi-squared test is (Lyderson *et al.*, 2009).

$$\text{p-value} = P(\chi^2 \geq t_{obs})$$

2) Fisher's exact test

This procedure is a conservative test (i.e. the p-value tends to be too large), but commonly applied test when the sample sizes are small. The formula used to calculate p-value of Fisher's exact test is (Seneta and Phipps, 2001)

$$P_F = \sum_{r \geq a} \binom{m}{r}\binom{n}{z-r} \Big/ \binom{m+n}{z} \qquad (2.2)$$

Therefore, alternative test statistics which can be used in place of Fisher's exact test with small counts are needed. The following procedures are recommended test for reducing the conservativeness of the p-value from Fisher's exact test.

3) Liebermeister's test

To calculate the p-value, use formula as follows.

$$P_L = \sum_{r \geq a+1} \binom{m+1}{r} \binom{n+1}{z+1-r} \Big/ \binom{m+n+2}{z+1} \qquad (2.3)$$

4) Lancaster's mid-P test

From (2.2), we may write P_F or P_F (a), as Lancaster's mid-P test is Fisher's exact test adjusted so the formula for the p-value is (Eyduran, 2008)

$$P_M = \left[P_F(a) + P_F(a+1)\right] / 2 \qquad (2.4)$$

Another procedure which based on the exact approach for comparing the equality of two binomial probabilities is Barnard's exact test. For small sample size, statistically significant exact p-value produced by Barnard's method is no accident and more powerful than Fisher's (Lyderson *et al.*, 2009). A brief description of this procedure is as follows.

5) Barnard's exact test

Suppose τ = { X : X is a 2 x 2 table as in Table 2.1}

Barnard's test is an unconditional test. It generates the exact distribution of T(X) by considering all the tables $X \in \tau$. Barnard suggested that we calculate p (π) for all possible values of $\pi \in (0,1)$. Barnard's exact p-value is defined as (Mehta and Senchaudhuri, 2003)

$$P_B = \sup\{p(\pi): \pi \in (0,1)\} \qquad (2.5)$$

Barnard's test and a simplified version of Barnard's test have higher power, but are considered too computationally intensive for practical use (Lyderson *et al.*, 2009).

6) Conditional Binomial Exact Test (CBET)

This test is derived from the joint distribution of two binomial samples and conditioned by the estimate of the probability of success *p* based on the combined samples (Phipps, 2003)

For more advanced analysis in testing the association in contingency table, logistic regression is commonly used. However, when any cells in table equal to zero, using the maximum likelihood estimation is not preferable. Thus, there is a procedure to solve this problem by using the penalized maximum likelihood estimation. The summary of these two procedures is as follows.

2.2 Logistic regression

Logistic regression is the commonly use method for analysis the association between dichotomous responses variables and explanatory variables, which can be either continuous or categorical. These models very useful since covariates can be included in the model to account for variability and to determine the effect of covariates on the response variable.

2.2.1 Logistic Regression with Maximum Likelihood Estimation

Consider the logistic regression model

$$\text{Prob}(y_i = 1 | x_i, \beta) = \pi_i = \{1 + \exp(-\Sigma_{r=1}^{k} x_{ir}\beta_r)\}^{-1} \qquad (2.6)$$

With i =1, . . . , n, yi ∈ {0, 1} denoting the binary outcome variable, x_{i1} =1 denoting the constant, and x_{ir} (i =1, . . . , n; r =2, . . . , k) referring to *n* observations on k − 1 independent covariates. Maximum likelihood estimates $\hat{\beta}_r$ of regression parameters $\beta_r (r = 1,...,k)$, which can be interpreted as log odds ratio estimates, are obtained as solutions to the score equations

$\partial \log L / \partial \beta_r = U(\beta_r) = \sum_{i=1}^{n} (y_i - \pi_i) x_{ir} = 0$ where log L is the log-likelihood function (Heinze, 2006).

2.2.2 Logistic regression with Penalized Maximum Likelihood Estimation (PMLE)

In order to reduce the small bias of those estimates by maximum likelihood method. Firth (1993) suggested the procedure which based on the estimation on the modified score equations.

$$U(\beta_r)^* \equiv U(\beta_r) + \frac{1}{2} \text{trace}[I(\beta)^{-1} \{\partial I(\beta) \mid \partial(\beta_r)\}] = 0 \qquad (r = 1,\ldots,k) \qquad (2.7)$$

Where $I(\beta^{-1})$ is the inverse of the information matrix evaluated at β. The modified score function $U(\beta_r)^*$ is related to the penalized log-likelihood and likelihood functions.

$$\log L\ (\beta_r)^* = \log L(\beta) + \frac{1}{2} \log |I(\beta)| \quad \text{and} \quad L\ (\beta_r)^* = L(\beta) |I(\beta)|^{1/2}$$

If Firth's general idea is applied to a logistic regression model

$$\text{Prob}\ (y_i = 1 \mid x_i, \beta) = \pi_i = \{1 + \exp(-\textstyle\sum_{r=1}^{k} x_{ir} \beta_r)\}^{-1}$$

Then the score equation $U(\beta_r) = \sum_{i=1}^{n}(y_i - \pi_i)x_{ir} = 0$ is replaced by the modified score equation

$$U(\beta_r)^* = \sum_{i=1}^{n}\left\{ y_i - \pi_i + h_i\left(\frac{1}{2} - \pi_i\right)\right\} x_{ir} = 0 \qquad (r=1,\ldots,k)$$

where the h_i's are the *i*th diagonal elements of the 'hat' matrix

$H=W^{1/2}X(X^TWX)^{-1}X^TW^{1/2}$ with $W = \text{diag}\{\pi_i(1- \pi_i)\}$

2.3 Data modification for non-convergence problem in logistic regression

This section contains a data manipulation approach for solving the non-convergence problem in logistic regression, which occur when the data set contain a zero counts, which known as 'separation'. This situation occur if the responses and non-responses can be perfectly separated by a single risk factor or by a non-trivial linear combination of risk factors (Albert and Anderson, 1984). There are many options for solving this problem as following:

1. Omission of that risk factor from the model.
2. Changing to a different type of model.
3. Use of an ad hoc adjustment (data manipulation)
4. Exact logistic regression
5. Standard analysis with $\hat{\beta}_{factor}$ set to a 'high' value. (Heinze and Schamper, 2002)

The method suggested in this study is focused on the data modification/data manipulation. However, there are the previous studies related to the data manipulation which proposed by Clogg *et al* (1991) and Agresti (2002). The briefly ideas for these two procedure are as follows.

2.3.1 Data Augmentation by Clogg et al (1991)

Let $\bar{p} = \sum_{i=1}^{n} y_i / n$ with $y_i \in \{0,1\}$, k is the number of parameters to be estimated. The basic idea is to add $\bar{p}\, k/g$ artificial responses and (1- $\bar{p}$)k/g artificial non-responses to each of the *g* groups of distinct risk factor patterns, and then to do a standard analysis on the augmented data set.

Suppose the contingency 2 by 2 table contain counts as in Table 2.2 (A), then the augmented data are shown as in Table 2.2 (B). (Heinze and Schamper, 2002)

Table 2.2: The example of contingency tables with data augmentation by Clogg *et al* (1991) (with $\bar{p}$ = 14/18, k=2, g=2)

A

y	x		Total
	1	2	
1	8	6	14
2	0	4	18
Total	8	10	18

⇒

B

y	x	
	1	2
1	8+(14/18)	6+(14/18)
2	0 + (4/18)	4+(4/18)

2.3.2 Adding 0.5 to cell frequencies of contingency tables

Suppose we have a table as shown in Table 2.1. The sample odds ratio $\hat{\theta} = ad/bc$ for a 2 by 2 table equals to 0 or ∞ if any cell equal to zero and it is undefined if both entries in a row or column are zero. Gart (1966) suggested the simple adjustment by replace a, b, c, d by {a+ 0.5, b+05, c+0.5 and d+0.5} in the estimator and standard error. In terms of bias and mean-squared error, Gart and Zweiful (1967) and Haldane (1956) showed that the amended estimator is

$$\tilde{\theta} = \frac{(a+0.5)(d+0.5)}{(b+0.5)(c+0.5)} \qquad (2.8)$$

And log $\tilde{\theta}$ behave well (Agresti, 2002)

2.3.3 Data modification method (DM)

The DM method is another choice of data adjustment. It is similar to the standard approach in equation 2.1 and equation 2.2. This study 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).

The DM method is the data modification applying a duplication method similar used by Lynn and McNeil (1995). Suppose that there is no responses in the data table (zero frequency count), we assumed to observed one case by doubling the sizes of the same group of risk factor (explanatory variable).

Suppose we have a data of 2 by 2 tables containing counts *a*, *b*, *c*, and *d* as shown in Table 2.2, which suppose *c=0*. Using DM method, we simply replace zero by 1 and doubling the corresponding cell *a*. For more example, let's see the table below.

Table 2.3: The study of testing the association between urinary tract infection (y) and diaphragm used (x)

A

y	x	
	yes	no
yes	7(a)	123 (b)
no	0 (c)	109 (d)

⟹

B

y	x	
	yes	no
yes	14 (a*)	123 (b)
no	1 (c*)	109 (d)

The example of the data in Table 2.3 show that, there are no cases of women with the uninfected urinary tract and use of diaphragm. With applied the DM method, the modified data will be shown as in Table 2.3 (B). We assumed to get one women without urinary tract infection by collected another seven women who using the diaphragm.

2.4. Hypothesis testing and 95% confidence intervals for DM method

In a 2 by 2 tables with counts *a, b, c* and *d* as in Table 2.3, the sample odds ratio (OR) $\hat{\theta} = ad/bc$ equals 0 or ∞ if any count is 0, then DM's estimator of the OR for counts in Table 2.3 (B) is

$$\hat{\theta} = \left(\frac{a^* \times d}{b \times c^*} \right) \tag{2.9}$$

Logistic regression with applied the DM method provide the p-value determined by Wald's test. The null hypothesis of the logistic regression model is H_0: β=0, testing no association between outcome and explanatory variables, where β= log (θ) is the

log (OR). Then $\hat{\beta} = \log(a^* d / bc^*)$, using the Mantel Haenszel test (McNeil, 1996), the standard error

$$\mathrm{SE}(\hat{\beta}) = \sqrt{\frac{1}{a^*} + \frac{1}{b} + \frac{1}{c^*} + \frac{1}{d}} \tag{2.10}$$

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

However, the standard errors of the log(OR) from the DM method give a too narrow confidence intervals as a consequence of the increased sample size. 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 $n_i * \pi_I$ where n_i is the total number for each group of independent variables ($n_1 = a + c$, $n_2 = b + d$) and π_i is the fitted probability of the successful outcome Y=1 for a modified data table. The new standard error is then calculated as

$$SE(\hat{\beta}) = \sqrt{\frac{1}{\hat{a}} + \frac{1}{\hat{b}} + \frac{1}{\hat{c}} + \frac{1}{\hat{d}}} \qquad (2.11)$$

CHAPTER 3

Applications

This chapter present the application of the DM method with the existing methods based on various data structures, including the simulated data and the real data set. The following section were including in this chapter. First section, provide the appropriate methods for testing the association in 2 by 2 tables containing a zero count by comparing the results for the existing methods. Section 3.2, comparison the results for DM method to the existing method using the simulation data set. Section 3.3, illustrate the use of DM method by comparing the results (including p-values, standard errors of log odds ratios and also 95% confidence intervals) of logistic regression with maximum likelihood estimation (using DM method), logistic regression with penalized maximum likelihood estimation (Firth's method) using the real data set. These information were appeared in Dureh et al (2015) and Dureh et al (2016).

3.1 Comparing methods for testing the association in 2 by 2 tables with zero counts

Since there are several methods available for testing the association in 2 by 2 tables. The main objective of this thesis is related to handle table with zero counts. Following example shows to handle zero counts using simulated data.

The simulated data in Table 3.1 comprising 72 two by two tables were created based on the condition that one cell is always equal to zero and the rest are small counts that make the averaged p-value from Pearson's chi-square test and Fisher's exact test close to 0.05.

The p-value equal to 0.05 was used as a reference p-value for comparison. We selected these 72 tables because they cover all such tables that fail to satisfy the sample size requirement in Pearson's chi-squared test that all expected counts are at least 5. These two methods were selected because they are most commonly preferred for testing an association in categorical data. Pearson's chi-square test is the conventional method for testing independence and Fisher's exact test is the preferred method when the sample sizes are too small. Therefore, using the averaged p-value from these two methods as a reference value is acceptable.

Table 3.1: Cell counts in 72 two by two tables where one cell contains zero and which gives averaged p-value close to 0.05

Table	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
a	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2
b	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
c	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d	18	27	37	46	55	64	74	83	92	6	8	10	13	15	17	20	22	25
Table	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
a	3	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4	4	4
b	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
c	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d	4	5	6	8	9	10	11	12	13	3	4	5	6	8	8	10	11	12
Table	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54
a	5	5	5	5	5	5	5	5	5	6	6	6	6	6	6	6	6	6
b	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
c	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d	2	3	4	5	6	7	7	8	9	2	3	4	4	5	6	6	7	8
Table	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72
a	7	7	7	7	7	7	7	7	7	8	8	8	8	8	8	8	8	8
b	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
c	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d	1	2	3	4	4	5	5	6	6	2	2	3	3	4	4	5	6	6

Figure 3.1 show that the p-values from each test not entirely consistent but almost all are between 0.01 and 0.2. A group including Pearson's chi-squared test with the continuity correction, Pearson's uncorrected chi-squared with Monte Carlo simulation test, Fisher's exact test as well as Barnard's test give p-values higher than 0.05. Pearson's uncorrected chi-squared test and Liebermeister's test tend to give p-values lower than 0.05 and tend to increase when sample size is increased. CBET, Lancaster mid-P test and Penalized maximum likelihood estimates gave p-values close to 0.05. There were no outliers distant from 0.05.

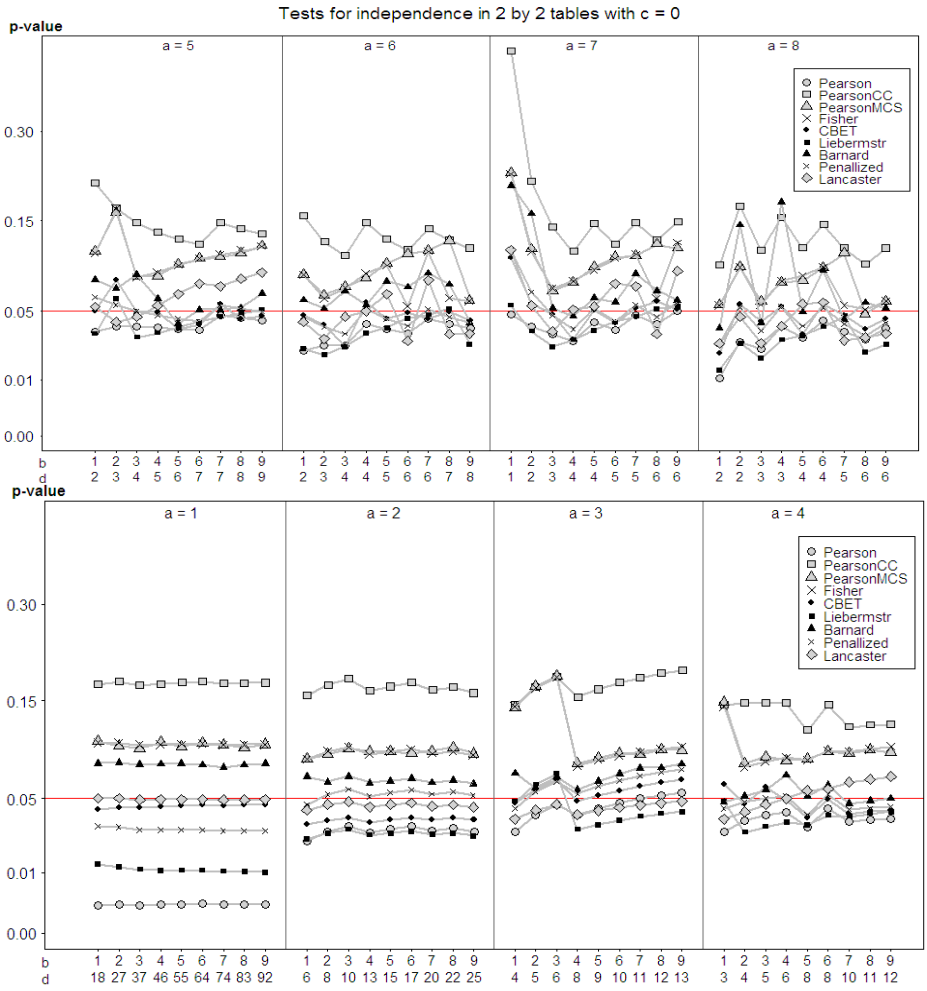


Figure 3.1: p-values from the recommended tests using data in two by two tables with c= 0

3.2 DM method for simulation 2 by 2 tables

3.2.1 Comparison the p-values from DM method to the existing method

To illustrate the used of the DM method, we simulate the 2 by 2 contingency tables. Each table contains at least one zero count. The 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 $\lambda = N\pi = 3$. However, *c* was forced to be a zero count since our purpose is to study the use of the DM method. 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 3.2 shows the level of agreement of the p-values from DM method, Fisher's exact test Lancaster's mid-p test, Agresti's method and Clogg's method are compared to p-values for Firth's method. 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. Agresti's method and data augmentation by Clogg's tends to have a larger p-values compare to Firth's. Similar to Fisher's exact test and Lancaster's mid-P test also giving a larger p-values when compare to Firth's. This is consistent with Fisher's test being more conservative than Firth's test.

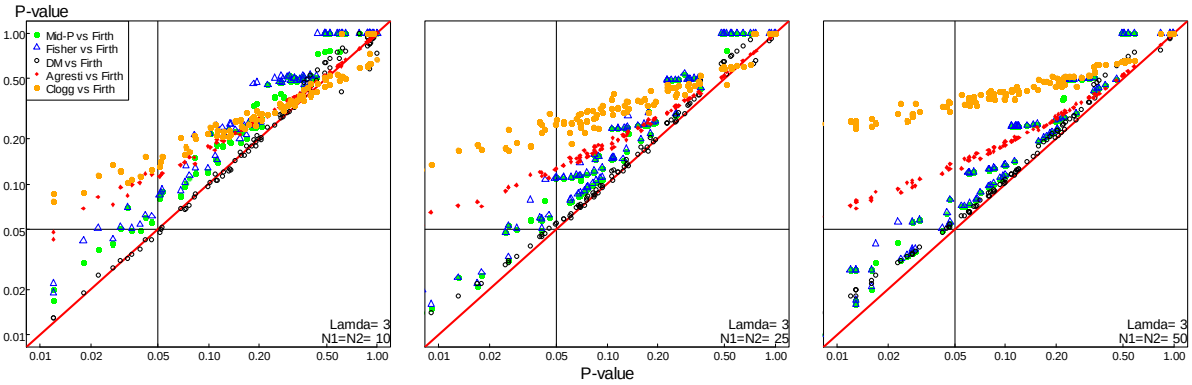


Figure 3.2: P-values from Fisher's exact test, Lancaster's mid-p test, Agresti's method, DM method and Clogg's data augmentation compared with Firth's method.

Figure 3.3 shows the differences of the p-values for those previous methods with Firth's. This finding confirm that the DM method give the closest p-values with Firth's method. While the p-values for Clogg's method have a big difference with Firth's. Even though, the DM and Clogg's are using the similar approach.

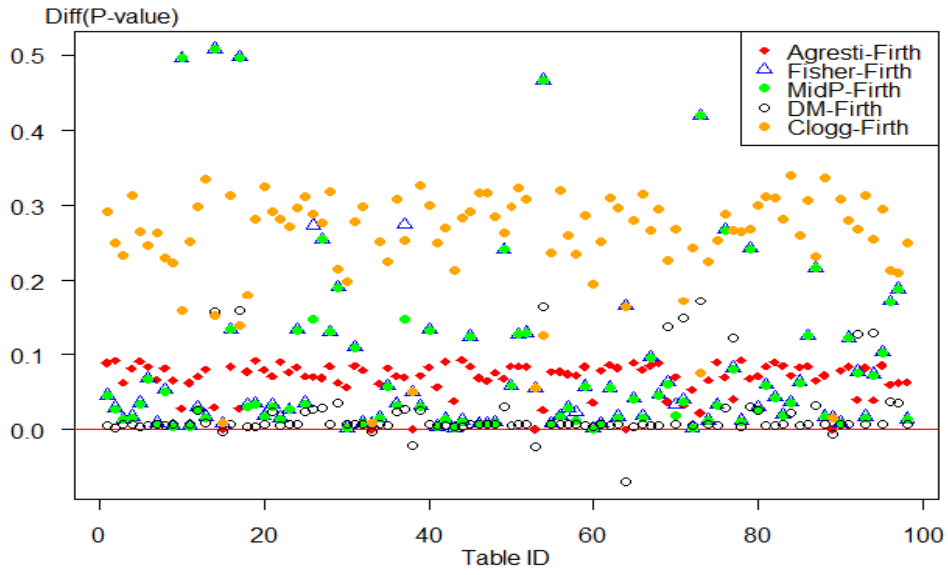


Figure 3.3: The differences of the p-values from Fisher's exact test, Lancaster's mid-p test, Agresti's method, DM method and Clogg method compared with Firth's method.

3.2.2 Comparison the percentages of correctly identified p-values

The 2 by 2 tables were obtained by generating binomial random numbers using R software. Suppose we have a binary response variable where Y=1 denotes a success and Y=0 otherwise. We also have a binary covariate X, also with values 0 or 1. If p_{ij} is the probability of a successful outcome, $P(Y|X)$, the logistic regression model is given by:

$$P(Y \mid X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \tag{3.1}$$

And $$\text{Logit}\,(p_{ij}) = \beta_0 + \beta_1 X \tag{3.2}$$

The logistic regression model for a 2 by 2 table can be shown as in Table 3.2:

Table 3.2: The general probabilities given by logistic regression model

Y	X	
	1	0
1	P(Y=1|X=1)	P(Y=1|X=0)
0	1-P(Y=1|X=1)	1- P(Y=1|X=0)

If β_1=0 in equation (3.2), then the rows and columns of the 2 by 2 table are independent. In the following we simulate data based on condition β_1=0 and β_0=-3, for example. Using equation (3.2), the p_{ij} for this independent model may represented by Table 3.3.

Table 3.3: The probabilities given by the logistic regression model, using β_1=0, β_0=-3

Y	X	
	1	0
1	0.0474	0.9526
0	0.9526	0.0474

Moreover, we also simulate data tables where the row and columns are dependent/nearly independent. According to the logistic regression model, we calculate the p_{ij} using

β_1=0.5, β_0=-3 to generate another set of tables. Table 3.4 below is in accordance with this model.

Table 3.4: The probabilities given by logistic regression model using β_1=0.5, β_0=-3

Y	X	
	1	0
1	0.0759	0.9526
0	0.9241	0.0474

The sample sizes of the simulated data sets vary from 10 to 100 with equal sizes for groups X=0 and X=1. For the purposes of this study, only the data from tables which include zero counts are selected. An example of data tables is shown in Table 3.5 (A) with different sample sizes for X=1 and X=0 and the corresponding data tables after data modification (DM) are shown in Table 3.5 (B).

Table 3.5: The counts in simulated 2 by 2 tables which include zero counts (A) and corresponding tables (B) modified according to the DM method

A

Table ID	a	b	c	d
1	10	10	0	0
2	9	10	1	0
3	10	9	0	1
4	8	10	2	0
:				
:				
103	100	96	0	4
104	96	100	4	0
105	100	97	0	3
106	94	100	6	0
107	100	95	0	5
108	97	100	3	0

B

Table ID	a	b	c	d
1	20	20	1	1
2	9	20	1	1
3	20	9	1	1
4	8	20	2	1
:				
:				
103	200	96	1	4
104	96	200	4	1
105	200	97	1	3
106	94	200	6	1
107	200	95	1	5
108	97	200	3	1

Figure 3.4 shows the percentages of correctly identified the p-value for DM method, Lancaster's mid-P test and Firth's method. Figure 3.4 (A) shows the percentage of three methods correctly identified that the explanatory variable (X) and outcome (Y) variable are dependent, that is the test produced a p-value less than 0.05. The DM method yielded the highest percentage (approximately 46%) of correct identification of dependence, Lancaster's mid-P test and Firth's method correctly identified dependence in 38 percent and 41 percent of cases, respectively.

Figure 3.4 (B) shows the results for the data where the explanatory variable (X) and outcome (Y) variable are assumed to be independent. We found that Firth's method correctly identified the highest percentage, approximately 81 percent of independent cases the p-values, while the DM method correctly identified only 63 percent.

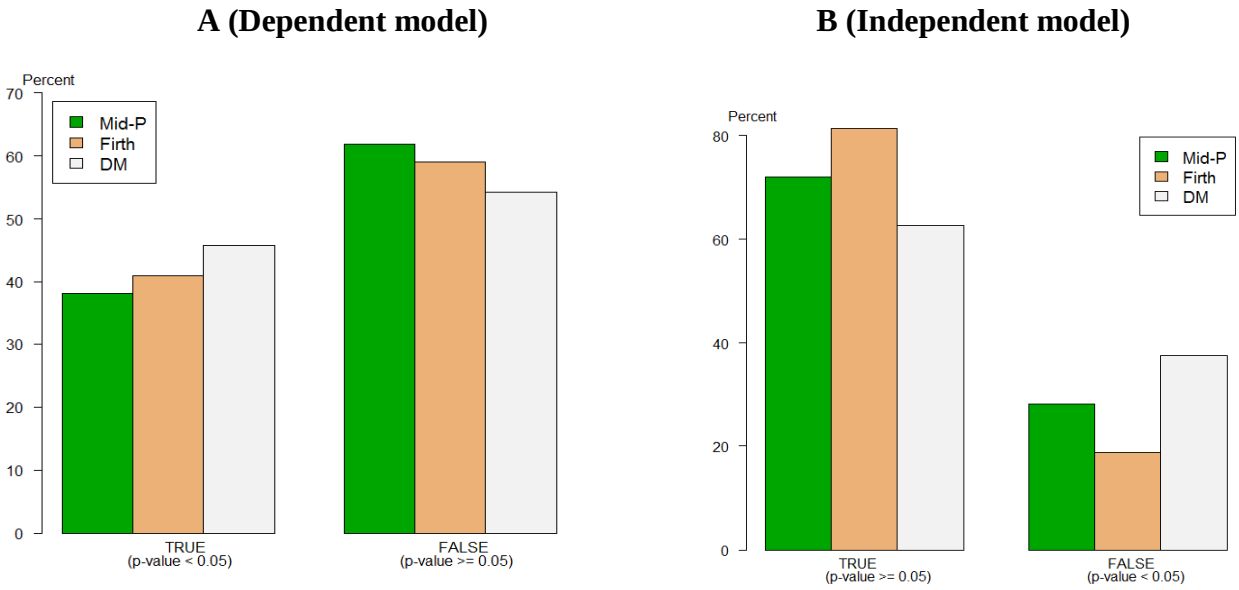


Figure 3.4: Percentages times the methods correctly identified p-values

3.3 DM method for real data set (2 by n table)

For more illustration, we applied the DM method with the data set 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 congenital (chapter Q) causes versus other causes. These data are listed in the left panels of Table 3.6 with modified data for using the DM method asterisked in the right panel.

Table 3.6: 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.*

In comparison, we found that DM, Clogg data augmentation and adding 0.5 to each cell by Agresti are return the similar results for the coefficients, standard errors and p-values to Firth's method. However, for DM method, we have to re-calculate the standard error for the coefficient (with the asterisk in Table 3.7) according to the increasing number of sample sizes.

Table 3.7: Logistic regression analysis of number of child deaths from congenital and other causes

Variable	Firth's method			DM				Clogg's			Agresti's	
	coef	se(coef)	p-value	coef	se(coef)	p-value	coef	se(coef)	p-value	coef	se(coef)	p-value
Intercept	-0.999	0.651	0.089	-1.099	0.667	0.099	-1.045	0.648	0.105	-0.998	0.625	0.110
Perinatal inside hospital	0	-	-	0	-	-	0	-	-	-	-	-
Congenital inside hospital	-0.947	1.863	0.531	-0.693	1.792*	0.585	-1.509	2.164	0.486	-0.947	1.636	0.563
External+ inside hospital	1.319	0.720	0.046	1.427	0.735	0.052	1.369	0.714	0.055	1.319	0.696	0.058
All causes outside hospital	1.129	0.716	0.087	1.232	0.731	0.092	1.177	0.710	0.097	1.129	0.692	0.103

Note: The standard error with asterisk (*) is the adjusted standard errors.

3.4 DM method for real data set (2 by 2^p table)

This illustration considered on the data structure with one binary outcome and many binary explanatory variables. 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).

As mentioned in the chapter 1 for the numerical problem in logistic regression. Table 3.8 shows the results given by logistic regression with the data set contain zero counts. Variable *dia* return a very large standard error which leading to an infinite 95% confidence intervals.

Table 3.8: Logistic regression analysis of condom use and first-time urinary infection study for original data

Variable	coef	SE(coef)	OR(95% CI)	p-value
Intercept	0.13	0.49	1.14 (0.44, 2.97)	0.794
age	-1.16	0.43	0.31 (0.13,0.73)	0.007
oc	-0.07	0.45	0.93 (0.38, 2.24)	0.870
vic	2.41	0.57	11.09 (2.63, 33.86)	<0.001
vicl	-2.25	0.56	0.11 (0.04, 0.32)	<0.001
vis	-0.82	0.42	0.44 (0.19, 1.01)	0.052
dia	16.73	799.43	18517467.13 (0,Inf)	0.983

When the data were applied with DM method, the logistic regression model provide a better results for those coefficients, standard error, 95% confidence interval and including p-values. Both DM method and Firth's method give a 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 as shown in Table 3.9.

Table 3.9: Logistic regression analysis of condom use and first-time urinary infection study for DM method and Firth's method

Variable	DM				Firth's method			
	coef	SE(coef)	OR (95% CI)	p-value	coef	SE(coef)	OR (95% CI)	p-value
Intercept	0.21	0.48	1.23 (0.48, 3.15)	0.659	0.12	0.49	1.13 (0.44, 2.92)	0.802
age	-1.07	0.39	0.34 (0.16,0.75)	0.007	-1.11	0.42	0.33 (0.14,0.76)	0.006
oc	-0.15	0.43	0.86 (0.37,2.02)	0.731	-0.07	0.44	0.93 (0.39,2.23)	0.875
vic	2.04	0.51	7.72 (2.85,20.94)	<0.001	2.27	0.55	9.67 (3.30,28.33)	<0.001
vicl	-1.92	0.50	0.15 (0.06,0.39)	<0.001	-2.11	0.54	0.12 (0.04,0.35)	<0.001
vis	-0.81	0.41	0.45 (0.20,1.00)	0.048	-0.79	0.42	0.45 (0.20,1.03)	0.054
dia	1.16	1.04	3.18 (0.41,24.54)	0.052	3.10	1.67	22.11(0.83,589.36)	0.005

CHAPTER 4

Conclusion and Discussion

This chapter comprising two sections, first section 4.1 includes discussion and conclusion of the study. Section 4.2, and 4.3, mention about the limitation and recommendation of the study, respectively.

4.1 Conclusion and discussion

This study introduced an alternative method for solving the problem of non-convergence in logistic regression when the two by two table contain a zero count. This problem is widely known but need to use special statistical package which conducted by Firth (1993). Therefore, this thesis suggest the simpler way which involved the data modification by replace the zero count by one and doubling the corresponding non-zero counts, called as “data modification (DM)” method. Using logistic regression applying the DM method yeilds a similar results (p-values) with Firth’s method. Firth’s method is the recommended method for solving the non-convergence problem in logistic regression (Heinze and Schemper, 2002; Eyduran, 2008).

This DM method uses logistic regression with the maximum likelihood estimates, the well-known method for parameter estimation and has the advantage of not requiring 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, 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. Thus, we may conclude that the DM method is useful and preferable for solving the non-convergence problem in logistic regression.

4.2 Recommendation of study

1. This study does not compare any actual bias of the estimated parameter between DM method and other existing methods. This point should be staged in the next study.

2. The DM method is appropriate handing zero counts, especially in Randomize Controlled Trial (RCT), rare diseases and other clinical data set.

4.3 Limitation of study

The application of DM method in this study is based the categorical covariates and has not been applied with continuous covariates. The DM method might also be applicable with continuous covariates, but this possibility needs to be considered in further study.

References

Agresti, A., and Hitchcock, D.B. 2005. Bayesian Inference for Categorical Data Analysis. Statistical Method and Application. 14.

Agresti, A. 2002. Categorical data analysis. 2nd edition. Hoboken, NJ: John Wiley & Son, New York, U.S.A.

Agresti, A. 1996. An Introduction to Categorical data analysis, NY: John Wiley & Son, New York, U.S.A.

Aitkin, M., and Chadwick, T. 2003. Bayesian Analysis of 2x2 Contingency Tables from Comparative Trials. School of Mathermatics and Statistics, University of Newcastle UK. 1-11.

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

Allison, P.D. 2004. Numerical Issues in Statistical Computing for the Social Scientist. John Wiley & Son, NewYork, U.S.A., Chapter 9, pp. 219-225.

Allison, P.D. 2008. Convergence Failure in Logistic Regression. Statistical and Data analysis, SAS Global Forum. 380.

Bester, C.L., and Hansen, C. 2005. Bias reduction for Bayesian and frequentist estimators. University of Chicago. 1-45.

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.

Brown, M.B. 1983. On Maximum likelihood estimation in sparse contingency tables. Computational Statistics Data analysis. 1: 3-15.

Chen, L.S., and Lin, C.Y. 2009. Bayesian P-values for Testing Independence in 2x2 Contingency Tables. Communications in Statistics-Theory and Methods. 38: 1635-1648.

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 Bayesian 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 Science. 42.

Dureh, N., Choonpradub, C., and Tongkumchum, P. 2016. An Alternative Method for Logistic Regression on Contingency Tables with Zero Cell Counts. Songklanakarin Journal of Science and Technology. 38. (in press).

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

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

Gart, J.J., and Zweifel, J.R. 1967. On the bias of various estimators of the logit and its variance with application to quantal bioassay. Biometrika. 54: 181-187.

Haldane, J.B.S. 1956. The estimation of viabilities. Journal of Genetics. 54: 294-296.

Heinze, G., and Schemper, M. 2002. A solution to the problem of separation in logistic regression. Statistic in Medicine. 21: 2409-2419.Heinze, G., and Ploner, M. 2003a. Fixing the non-convergence bug in logistic regression with SPLUS and SAS. Computer methods and Programs in Biomedicine. 71: 181-187.

Heinze, G., and Ploner, M. 2004b. Technical report 2: A SAS macro, S-PLUS library and R package to perform logistic regression without convergence problems. Department of Medical Computer Sciences, Medical University of Vienna. 5-39.

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

Heinze, G. 2009b. Avoiding infinite estimates in logistic regression- theory, solutions, examples. Available: http://www. hal.archives-ouvertes.fr/docs /00/38/66/ 39/ PDF /p80.pdf. [January 12, 2012].

Toshinari, K. 2011. Inference for the coefficient parameters of the logistic regression from a small sample. Proceeding in 58th World Statistical Congress, Dublin. 5340-5343.

King, G., and Zeng, L. 2001. Logistic Regression in Rare Events Data. Political Analysis. 9: 131-163.

Lancaster, H.O. 1961. Significance Tests in Discrete Distributions. Journal of the American Statistical Association. 56.

Lin, C.Y., and Yang, M.C. 2009. Improved p-Value Tests for Comparing Two Independent Binomial Proportions. Communications in Statistics-Simulation and Computation. 38: 78-91.

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

Lydersen, S., Fagerland, M.W., and Laake, P. 2009. Tutorial in Biostatistic Recommended tests for association in 2x2 tables. Statistics in Medicine. 28: 1159-1175.

McNeil, D. 1996. Epidemiological Research Method. John Wiley & Sons. New York.

Mehta, C.R., and Patel, N.R. 1997. Exact Inference for Categorical Data. Cambridge, MA: Harvard University and Cytel Software Corporation. 1-38.

Mehta, C.R., and Patel, N.R. 2011. IBM SPSS Exact Tests. IBM Corporation, Chicago, U.S.A. 11-29.

Mehta, C.R., and Senchaudhuri, P. 2003. Conditional versus Unconditional Exact Tests for Comparing Two Binomials. Available: http://www.cytel. com/papers/twobinomials.pdf [January, 9, 2011].

Mehrotra, D.V., Chan, I.S., and Berger, R.L. 2003. A cautionary note on exact unconditional inference for a difference between two independent binomial proportions. Biometrics. 59: 441-450.

Pattaraarchachai, J., Rao, C., Polprasert, W., Porapakkham, Y., Pao-in, W., Singwerathum, N., and Lopez, A.D. 2010. Cause-specific mortality patterns

among hospital deaths in Thailand: validating routine death certification. Population Health Metrics. **8**:12.

Parzen, M., Lipsitz, S., Ibrahhim, J., and Klar, N. 2002. An Estimate of the Odds Ratio That Always Exists. Journal of Computational and Graphical Statistics. 11: 420-436.

Phipps, M.C. 2003. Liebermeister's quasi-exact test for two binomials. Proceedings of the 2003 Hawaii International conference on statistics and related fields, Second Hawaii International Conference on Statistics and Related Fields, T. Gregson and D. Yang (eds.), University of Hawaii, West Oahu, Hawaii. 1-10.

Polprasert, W., Rao, C., Adair, T., Pattaraarchachai, J., Porapakkham, Y., and Lopez, A.D. 2010. Cause-of death ascertainment for deaths that occur outside hospitals in Thailand: application of verbal autopsy methods. Population Health Metrics. 8:13.

Porapakkham, Y., Rao, C., Pattaraarchachai, J., Polprasert, W., Vos, T., Adair, T., and Lopez, A.D. 2010. Estimated causes of death in Thailand, 2005: implications for health policy. Population Health Metrics. 8:14.

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.

Santos, K.C., and Barrios, E.B. 2012. Predictive Accuracy of Fitted Logistic Regression Model Using Ranked Set Sample. UPSS Working Paper No. 2012-02. School of Statistics, University of the Philippines Diliman.

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

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

Zorn, C. 2005. A Solution to Separation in Binary Response Models. Political Analysis. 13: 157-170.

Appendix 1

Comparing Tests for Association in Two by Two Tables with Zero Cell Counts

Nurin Dureh*[a], Chamnien Choonpradub [a] and Phattrawan Tongkumchum [a]

[a] Department of Mathematics and Computer Sciences, Faculty of Science and Technology, Prince of Songkla University. Pattani Campus, Thailand.

*Author for correspondence; e-mail: dnurin@gmail.com

ABSTRACT

This study compared the tests for association in two by two tables with zero cell counts. Pearson's uncorrected chi-squared test, Pearson's chi-squared test with the continuity correction, Pearson's uncorrected chi-squared Monte Carlo simulation test, Fisher's exact test, the Conditional Binomial Exact Test (CBET), Barnard's exact test, Liebermeister's test, Lancaster's mid P-test and logistic regression with penalized maximum likelihood were considered. Criteria used 72 two by two tables with smallest counts and average p-value for Fisher's exact test and Pearson's uncorrected chi-squared test close to 0.05. CBET, Lancaster's test, and the penalized maximum likelihood test give similar p-values closest to 0.05, suggesting that these three methods can be recommended for testing association in two by two tables with zero cell counts.

Keywords: zero cell, two by two table, association.

1. INTRODUCTION

Several methods have been recommended for analysis of association in two by two tables. The most common test is Pearson's chi-squared test, which is appropriate for sufficiently large sample sizes. It is inaccurate if any expected count is less than five [1, 2, 3]. In cases of small sample sizes, Fisher's exact test is recommended [1, 2, 3]. This method eliminates the nuisance parameter in the model under the null hypothesis by conditioning on its marginal totals [4] but is conservative. Another way to reduce the conservatism of Fisher's exact test is to consider an unconditional approach, such as Barnard's test, which eliminates the nuisance parameter by taking its supreme value over all possible values in the space of the null model [5, 6]. Several alternative tests also have been proposed [7]. These include Lancaster's mid-P test [8], an adjustment to the Fisher's exact test that tends to have increased power while maintaining a Type I error rate close to the nominal level [7, 9]. The Liebermeister's test also can be used in place of Fisher's exact test, and is less conservative than Fisher's test and just as easy to calculate [3]. In addition, the "conditional binomial exact test" (CBET) is proposed as an alternative test for comparing binomial proportions estimated from samples of larger populations [10].

Logistic regression provides a more general method because it provides a model that accommodates more complex determinants. However, when one of the four cells in the two by two table is equal to zero, maximum likelihood estimates fail to converge [3, 5, 7, 10, 11, 12, 13]. A solution to this problem was proposed by Firth [14], giving finite parameter estimates based on penalized maximum likelihood [14, 15, 16, 17, 18]. This method is available in statistical software such as SAS, S-PLUS and R [17,

19]. However, these estimates are biased away from zero [16], so it is important to know how substantial these biases are.

Tables with zero cell count thus lead to numerical problems [20], so it is important to identify the methods which provide the most accurate results for particular data structures. With this information, researchers can select the appropriate method for their studies. Thus the main objective of this study was to compare the results when using recommended tests for association in two by two tables with small zero cell counts.

2. MATERIALS AND METHODS

Tests for an association in two by two table

There are several methods for testing the association in 2 by 2 tables. Suppose that the 2 by 2 table contain counts as in Table 1, a brief summary of computing a p-value of these tests describes as follows.

Table 1: The general counts of a 2 by 2 table.

		j		
		1	2	Total
i	1	a	b	m
	2	c	d	n
	Total	z	v	m+n

Pearson's uncorrected chi-squared test

The functional form of this test is

$$\chi^2_p = \frac{m + n(ad - bc)^2}{(a+b)(c+b)(a+c)(b+d)} \quad (1)$$

In general, the *p*-value is defined as the probability of the test statistic *T* being equal to or more extreme than its value for the observed table *(t_{obs})*, therefore, the approximate p-value for Pearson's chi-squared test is [6]

$$\text{p-value} = P(\chi^2 \geq t_{obs})$$

Pearson's chi-squared test with the continuity correction (Pearson's CC)

A continuity correction for the Pearson's chi-squared test was proposed by Yate (1984). The corresponding formula for Pearson's CC test is

$$\chi^2_{PCC} = \frac{n(abs(ad - bc) - n/2)^2}{(a+b)(c+d)(a+c)(b+d)}. \quad (2)$$

Pearson's uncorrected chi-squared test with Monte Carlo Simulation

This test uses a reference set of 10,000 samples to compute the p-value for Pearson's uncorrected chi-squared test in (1)

Fisher's Exact Test

The Fisher's Exact P-value is obtained by conditioning on the observed total successes [3]. If r is the observed value in 2 by 2 table, which can be greater or equal to cell a, thus the formula is

$$P_F = \sum_{r \geq a} \binom{m}{r} \binom{n}{z-r} \Big/ \binom{m+n}{z} \qquad (3)$$

Alternative test statistics which can be used in place of Fisher's exact test with small counts are as follows.

Liebermeister's test

This test is the quasi-exact test for two binomials. It proposed by adjusting the observed table and can be obtained by hand calculator very simply as the formula

$$P_L = \sum_{r \geq a+1} \binom{m+1}{r} \binom{n+1}{z+1-r} \Big/ \binom{m+n+2}{z+1} \qquad (4)$$

Lancaster's mid-P test

From (2), we may write Fisher's Exact P-value as P_F or P_F (a). As Lancaster's mid-P test is Fisher's exact test adjusted so the formula for this p-value is [3].

$$P_M = \left[P_F(a) + P_F(a+1)\right]/2 \qquad (5)$$

Barnard's exact test

Suppose τ = {X: X is a 2 by 2 table as in Table 1}

Barnard's exact test is an unconditional test. Suppose T(X) is a "discrepancy measure" or test statistic that measures how discrepant any table X is relative to the type of table one would expect under the null hypothesis. It generates the exact distribution of T(X) by considering all the tables $X \in \tau$. If p (π) is the exact p-value for any given π, Barnard suggested that we calculate p(π) for all possible values of

$\pi \in (0,1)$ and choose the value π which maximized p(π), thus, Barnard's exact p-value is defined as [2]

$$P_B = \sup \{p(\pi): \pi \in (0,1)\} \quad (6)$$

Conditional Binomial Exact Test (CBET)

This test is derived from the joint distribution of two binomial samples and conditioned by the estimate of the probability of success p based on the combined samples [10]

Data Simulation

Data comprising 72 two by two tables were created based on the condition that one cell is always equal to zero and the rest are small counts that make the averaged p-value from Pearson's chi-square test and Fisher's exact test close to 0.05. We selected these 72 tables because they cover all such tables that fail to satisfy the sample size requirement in Pearson's chi-squared test that all expected counts are at least 5. We selected these two methods because they are most commonly preferred for testing an association in categorical data. Pearson's chi-square test is the conventioned method for testing independence and Fisher's exact test is the preferred method when the sample sizes are too small. Therefore, using the averaged p-value from these two methods as a reference value is acceptable.

Table 1: Cell counts in 72 two by two tables where one cell contains zero and which gives averaged p-value close to 0.05.

Table	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
a	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2
b	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
c	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d	18	27	37	46	55	64	74	83	92	6	8	10	13	15	17	20	22	25
Table	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
a	3	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4	4	4
b	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
c	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d	4	5	6	8	9	10	11	12	13	3	4	5	6	8	8	10	11	12
Table	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54
a	5	5	5	5	5	5	5	5	5	6	6	6	6	6	6	6	6	6
b	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
c	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d	2	3	4	5	6	7	7	8	9	2	3	4	4	5	6	6	7	8
Table	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72
a	7	7	7	7	7	7	7	7	7	8	8	8	8	8	8	8	8	8
b	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
c	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d	1	2	3	4	4	5	5	6	6	2	2	3	3	4	4	5	6	6

The nine methods used for testing association in the two by two tables were Pearson's uncorrected chi-squared test, Pearson's chi-squared test with the continuity correction (Pearson's CC), Pearson's uncorrected chi-squared test with Monte Carlo Simulation, Fisher's exact test, Liebermeister's test, Lancaster's mid-P test , Barnard's exact test, Conditional Binomial Exact Test (CBET) and Logistic regression with penalized maximum likelihood estimates. The reference p-value used was p=0.05 based on the average p-value of Pearson's uncorrected chi-squared test and Fisher's exact test.

3. RESULTS

The p-values from the nine methods applied to tables 1-36 are shown in Figure 1 and Figure 2 shows p-values from tables 37 - 72. The solid line represents p-values equal to 0.05 and each dotted line denotes p-values for each test.

The p-values from each test not entirely consistent but almost all are between 0.01 and 0.2. A group including Pearson's chi-squared test with the continuity correction, Pearson's uncorrected chi-squared with Monte Carlo simulation test, Fisher's exact test as well as Barnard's test give p-values higher than 0.05. Pearson's uncorrected chi-squared test and Liebermeister's test tend to give p-values lower than 0.05 and tend to increase when sample size is increased. CBET, Lancaster mid-P test and Penalized maximum likelihood estimates gave p-values close to 0.05. There were no outliers distant from 0.05 for any of these three methods.

Figure 1: P-values from the recommended tests using data in two by two tables with c= 0 and a is 1, 2, 3 and 4.

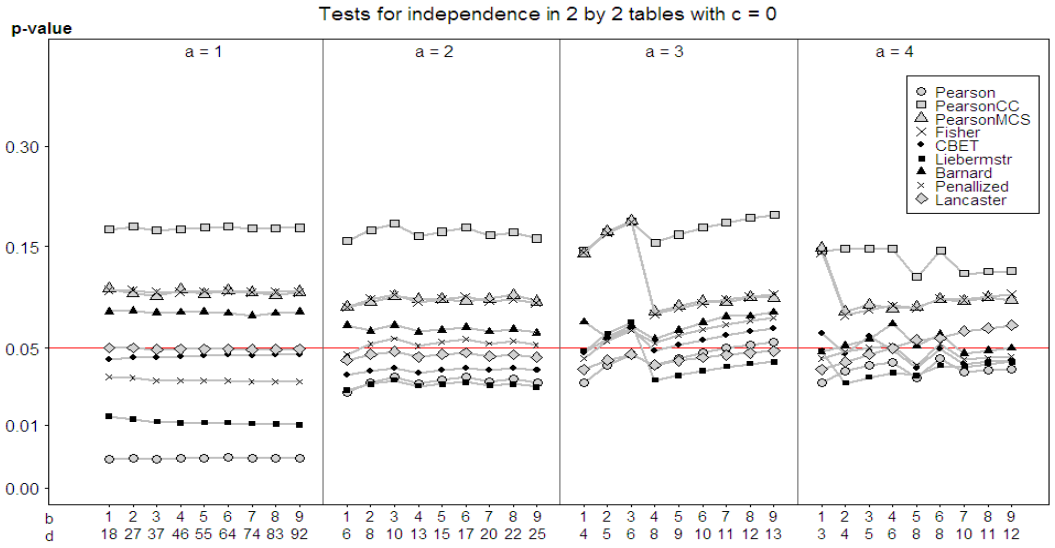
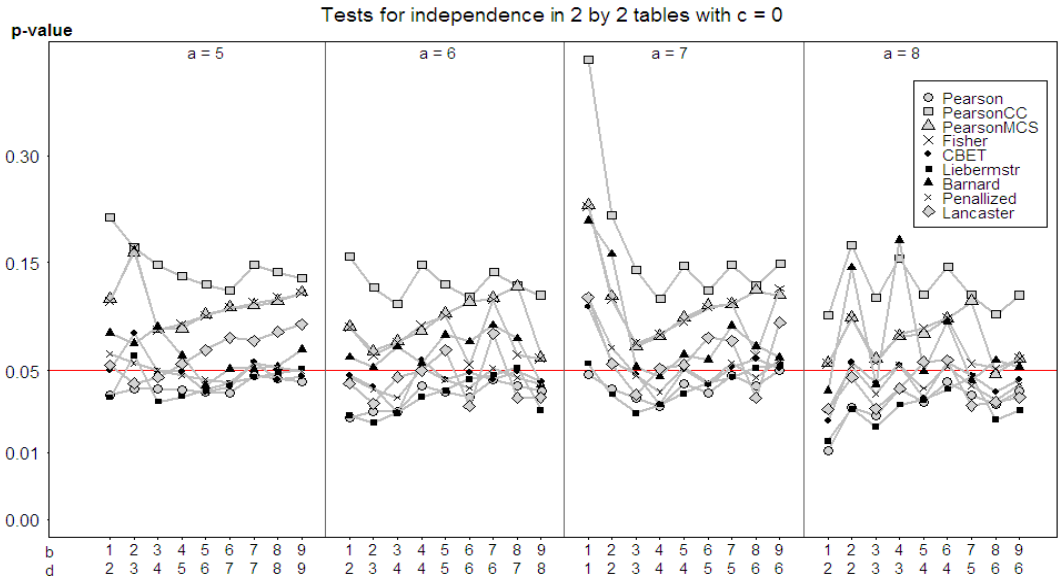


Figure 2: P-values from the recommended tests using data in two by two tables with c= 0 and a is 5, 6, 7 and 8.



For comparison, Table 2 displays the average p-values of 72 tables from those nine tests. It is clear that Lancaster's mid-P test, the Conditional Binomial Exact Test (CBET) and penalized maximum likelihood gave p-values closest to 0.05 thereby providing the most exact results, at least when averaged. Pearson's uncorrected chi-squared test and Liebermeister's test gave averaged p-values lower than 0.05. On the other hand, Fisher's exact test, Pearson's chi-squared test with the continuity correction, Pearson's uncorrected chi-squared with Monte Carlo simulation test and Barnard's exact test gave large biases. All tended to have a large p-value.

Table 2 Average p-values from nine recommended tests for testing the association in two by two tables with zero cell count.

Test	Averaged p-value
Pearson's uncorrected chi-square	0.0320
Pearson's corrected chi-square	0.1528
Pearson's uncorrected chi-square Monte Carlo Simulation	0.0970
Fisher's exact test	0.0939
Conditional Binomial Exact Test (CBET)	0.0469
Barnard's exact test	0.0694
Liebermeister's test	0.0332
Penalized Maximum Likelihood	0.0478
Lancaster's mid-P test	0.0501

Child Deaths from Perinatal Originating Conditions on Thai Provinces Example

The data shown in Table 3 are the numbers of child deaths from perinatal originating conditions separated groups of provinces, based on the Thai 2005 Verbal Autopsy (VA) study [21, 22, 23, 24]. With the question "Is the number of deaths from perinatal originating conditions in Chumporn province are different from other provinces in Thailand ? ". The results are shown in Table 4.

Table 3: Number of child death from perinatal originating conditions by province

Provinces	Cause of death		Total
	Other	Perinatal	
Other	84	59	143
Chumporn	6	0	6

A p-values from those nine recommended tests show that the tests including Pearson's uncorrected chi-square test, CBET, Liebermeister's test and Penalized Maximum Likelihood gave the significance results with p-values 0.0429, 0.0483, 0.0423 and 0.0441, respectively. Meaning that Chumporn province have the number of child deaths from perinatal originating conditions different from other provinces.

Table 4 Average p-values from nine recommended tests for testing the association of provinces and child death from perinatal originating conditions

Test	Averaged p-value
Pearson's uncorrected chi-square	0.0429
Pearson's corrected chi-square	0.1100
Pearson's uncorrected chi-square Monte Carlo Simulation	0.0838
Fisher's exact test	0.0815
Conditional Binomial Exact Test (CBET)	0.0483
Barnard's exact test	0.0541
Liebermeister's test	0.0423
Penalized Maximum Likelihood	0.0441
Lancaster's mid-P test	0.0588

4. CONCLUSION AND DISCUSSION

This study compared the accuracy of nine separate tests of the association in two by two tables, where one cell contained a zero count, using a reference p-value equal to 0.05. When comparing the individual p-value with the reference p-value, most of the tests gave p-values in the range from 0.01 and 0.2. This study showed that the methods of Pearson's chi-squared test and Fisher's exact test were not appropriate for this condition of a zero count in a two by two tables, because of the high p-values resulting from their application.

Lancaster's mid-P test, Conditional Binomial Exact Test and method of penalized maximum likelihood estimates were identified as acceptable and clearly preferable in testing the association in two by two tables with zero counts. These three methods consistently produced results close to the reference (p=0.05), in average as shown in Table 2 and in range as shown in Figure 1 and 2.

For CBET, this conforms to the finding by Rice (1988) that CBET can be used in place of Fisher's exact test when analyzing contingency tables that compare binomial proportions estimated from samples of larger populations. In addition, this study can also recommend the use of Lancaster's mid-P test and Penalized maximum likelihood estimates. The three methods, Conditional Binomial Exact Test, Lancaster's mid-P test and penalized maximum likelihood estimates can be recommended in cases of testing the association in two by two tables with zero cell counts. Since the main objective of this study was not identify the best method but to compare the results when using recommended tests for association in two by two tables, this study can not be concluded which method is the best. The researchers have to considered due to many conditions, for example, the important of data, software avaiability, simplicity of calculation, the situation of each study etc.

ACKNOWLEDGEMENTS

This research received full 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.

REFERENCES

[1] Mehta C.R. and Patel NR., Exact Inference for Categorical Data, Harvard University and Cytel Software Corporation., 1997.

[2] Mehta C.R. and Senchaudhuri P, Conditional versus Unconditional Exact Tests for Comparing Two Binomials, Cytel Software Corporation, Cambridge., 2003.

[3] Seneta E. and Phipps M.C., On the Comparison of Two Observed Frequencies. *Biometrical. J.,* 2001; **43**: 23-43.

[4] Mehrotra D.V., Chan I.S. and Berger R.L., A cautionary note on exact unconditional inference for a difference between two independent binomial proportions, *Biometrics*., 2003; **59**: 441-450.

[5] Lin C.Y. and Yang M.C., Improved p-Value Tests for Comparing Two Independent Binomial Proportions, *Commun. Stat. Simulat.,* 2009; **38**: 78-91.

[6] Lydersen S., Fagerland M.W. and Laake P., Tutorial in Biostatistic Recommended tests for association in 2x2 tables, *Stat. Med*., 2009; **28:** 1159-1175.

[7] Biddle D.A., Morris and Scott B, Using Lancaster's mid-P correction to the Fisher's exact test for adverse impact analysis, *J. Appl. Phychol.*, 2011; **96**: 956-965.

[8] Lancaster H.O., Significance Tests in Discrete Distributions. *J. Amer. Statist. Assoc.,* 1961; **56**.

[9] Chen L.S., and Lin C.Y., Bayesian P-values for Testing Independence in 2 by 2 Contingency Tables, *Commun. Stat. Theory.,* 2009; **38**: 1635-1648.

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

[11] Aitkin M. and Chadwick T., Bayesian analysis of 2x2 contingency tables from comparative trials, School of Mathematics and Statistics, University of Newcastle UK., 2003.

[12] Bester C.L. and Hansen C., Bias reduction for Bayesian and frequentist estimators, University of Chicago., 2005.

[13] Sean R.E., What is Bayesian statistics?, *Nat. Biotechnol*., 2004; **22**.

[14] Firth D., Bias reduction of maximum likelihood estimates, *Biometrika*., 1993; **80**: 27-38.

[15] Eyduran E., Usage of Penalized Maximum Likelihood Estimation Method in Medical Research: An Alternative to Maximum Likelihood Estimation Mehod, Jrms., 2008; **13**: 325- 330.

[16] Heinze G. and Schemper M., A solution to the problem of separation in logistic regression, *Stat. Med*., 2002; **21**: 2409-2419.

[17] Heinze G. and Ploner M., Fixing the nonconvergence bug in logistic regression with SPLUS and SAS, *Comput. Meth.Prog. Bio*., 2003; **71**: 181-187.

[18] Heinze G., Avoiding infinite estimates in logistic regression- theory, solutions, examples, Medical University of Vienna., 2009.

[19] Heinze G. and Ploner M., A SAS macro, S-PLUS library and R package to perform logistic regression without convergence problems, Medical University of Vienna., 2004.

[20] Brown M.B., On Maximum likelihood estimation in sparse contingency tables, *Comput. Stat. Data. An*., 1983; **1**: 3-15.

[21] Pattaraarchachai J., Rao C., Polprasert W., Porapakkham Y., Pao-in W., Singwerathum N. and Lopez A.D., Cause-specific mortality patterns among hospital deaths in Thailand: validating routine death certification, *Population Health Metrics.,* 2010; **8**:12.

[22] Polprasert W., Rao C., Adair T., Pattaraarchachai J., Porapakkham Y. and Lopez A.D., Cause-of death ascertainment for deaths that occur outside hospitals in Thailand: application of verbal autopsy methods, *Population Health Metrics.,* 2010; **8**:13.

[23] Porapakkham Y., Rao C., Pattaraarchachai J., Polprasert W., Vos T., Adair T. and Lopez A.D., Estimated causes of death in Thailand, 2005: implications for health policy, *Population Health Metrics.,* 2010; **8**:14.

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

Appendix 2

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.

*Corresponding Author; e-mail: dnurin@gmail.com Tel: +66 84 1974442

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 (2014), where several methods for testing association in two-by-two tables containing at least one small count

(possibly zero) were compared. The result showed that the Conditional Binomial Exact Test (Rice, 1988), Lancaster's mid-P test (Biddle and Moris, 2011) and the penalized maximum likelihood (Firth, 1993) had 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 (Aitkin and Chadwick, 2003; Albert and Anderson, 1984; 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.

2. Methodology

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,…x_p), which can be discrete, continuous or a combination. If π is the probability of a successful outcome, Pr(Y=1|X), the logistic regression model is given by:

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

or $$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$$

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.

Data modification (DM)

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 odds 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.

Table 1: The 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

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 (** $\frac{a^* \times d}{b \times c^*}$ **).** 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 $z = \log(\frac{a^* \times d}{b \times c^*}) / SE = \frac{\hat{\beta}}{SE(\hat{\beta})}$

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 $n_i * \pi_i$ where n_i is the total number for each group of independent variables ($n_1 = a + c$, $n_2 = b + d$) and π_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 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.

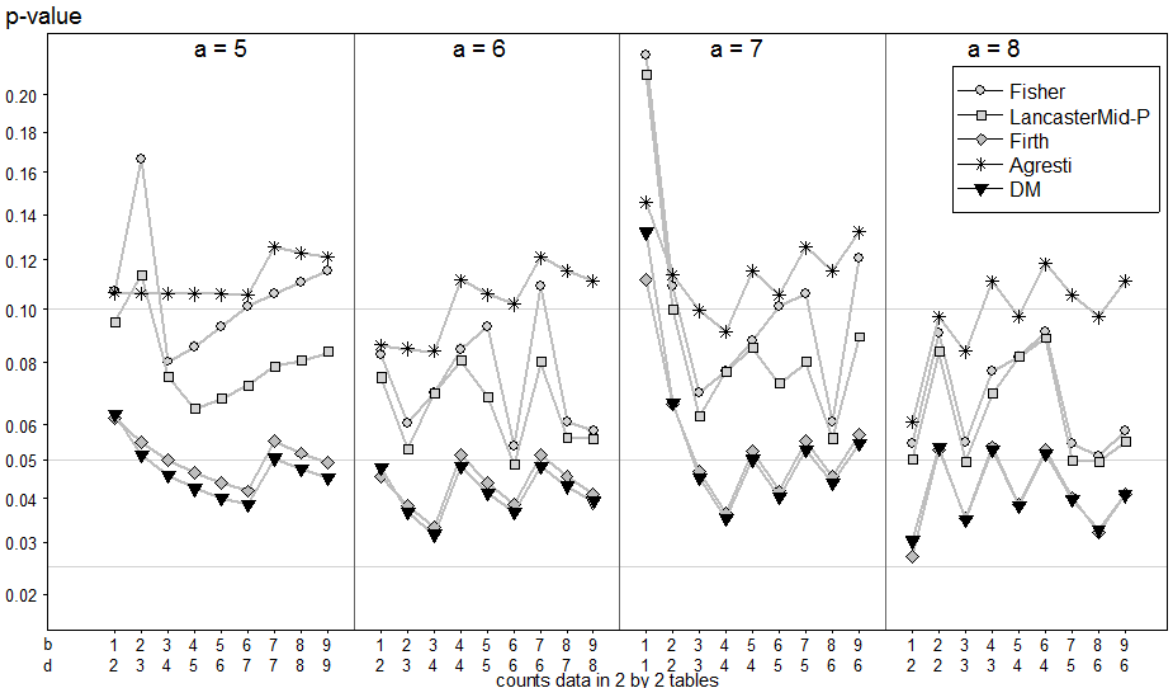


Figure 1: P-values of test for independence in two-by-two tables with a zero count for 36 tables with specified values of the counts (a, b, c, d).

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.

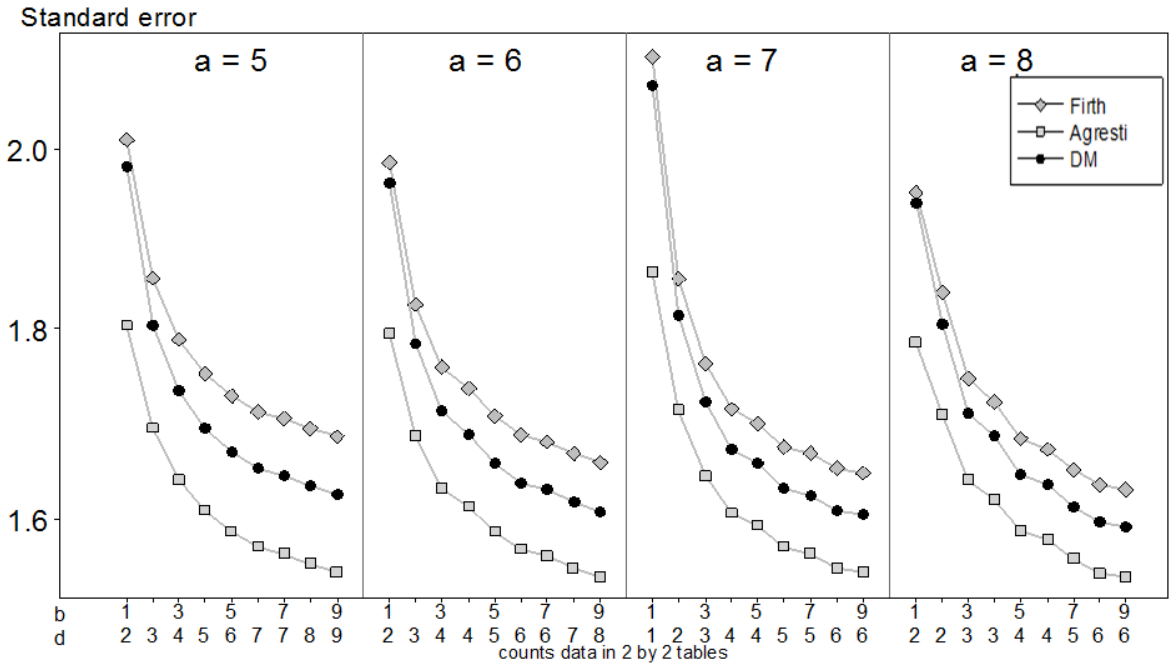


Figure 2: Standard error of log odds ratio of test for independence in two-by-two tables with a zero count for 36 tables with specified values of the counts (a, b, c, d).

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 $\lambda = N\pi = 3$. However, *c* was forced to be a zero count since our purpose is to study the use of the DM 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.

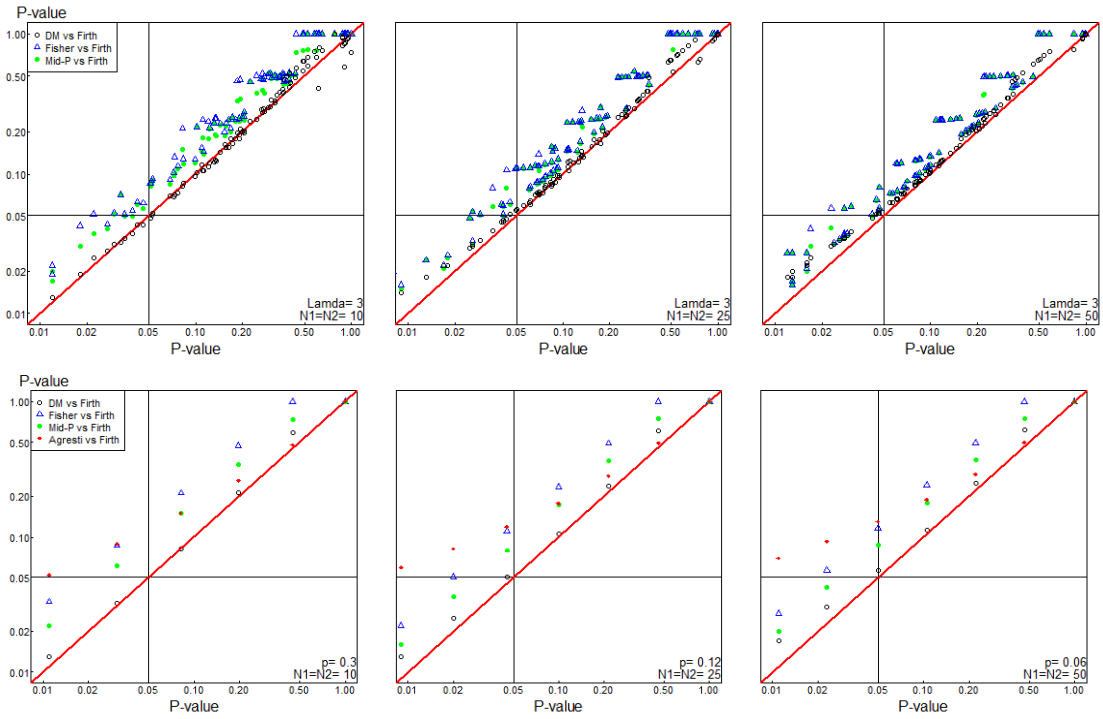


Figure 3: P-values from Fisher's exact test, Lancaster's mid-p test and DM method compared with Firth's method

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.

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

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

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.

Table3: 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.*

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

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 Conclusion

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, 12-14 May 2004.

Agresti, A. 2002. Categorical data analysis. John Wiley & 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.

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, CC. Rubin, DB. 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. 2014. Comparing Tests for Association in Two-by-Two Tables with Zero Cell Counts. Chiang Mai Journal of Sciences. 42. (in press).

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: 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 & Sons. New York.

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.

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

Appendix 3

Comparing Methods for Testing Association in Tables with Zero Cell Counts Using Logistic Regression

Nurin Dureh[1], Chamnein Choonpradub[2] and Hilary Green[3]

[1]Ph.D Candidate, Department of Mathematics and Computer sciences,
Prince of Songkla University, Pattani, 94000, Thailand.
Email: dnurin@gmail.com

[2]Lecturer, Department of Mathematics and Computer sciences,
Prince of Songkla University, Pattani, 94000, Thailand
Email: cchamnein.c@gmail.com

[3]Lecturer, Department of Statistis, Macquarie University, NSW, Australia.
Email: Hilary.green@mq.edu.ac

Abstract

Logistic regression is one of the most useful methods used to describe the relationship between a binary dependent variable and a set of independent variables. However, when any of the counts are zero, a non-convergence problem will occur. A procedure for solving such a problem has been proposed by Firth (1993), and provides finite parameter estimates based on penalized maximum likelihood. This study suggests a simpler method which involves modifying the data by replacing the zero count by one and doubling the corresponding non-zero count. Results show that this simple data modification method gives similar results to those from the Firth's procedure.

Keywords: Logistic regression, zero counts, Firth’s procedure, non-convergence

1 Introduction

Logistic regression is one of the most used methods for modeling and testing the association between a binary outcome and one or more determinants. However, when one of the four cells in a two by two table of counts is equal to zero, maximum likelihood estimates of the model parameters fail to converge [1,3,5]. A solution to this problem was proposed by Firth (1993), giving finite parameter estimates based on penalized maximum likelihood. This method is available in statistical software packages such as SAS, S-PLUS and R [7,8]. Since this method requires special software we considered the possibility of simply modifying the data rather than using Firth's method. Lunn and McNeil (1995) used a similar simple approach for modeling competing risks in survival analysis. Agresti (2002) and Rubin (2002, Chapter 2), also reccommended modification in preference to new methodology when cell counts are small or the data are incomplete.

2 Method

The data modification method (DM) proposed in this paper is an improvement from the standard approach suggested by Agresti (2002) where 0.5 is added to each cell in a 2 by 2 table. To deal with zero count problem in logistic regression, we introduce a new simple method for which the statistical significance determined by conventional procedures aligns closely with Firth's method.

Suppose the 2 by 2 table contains counts *a*, *b*, *c*, *d* with all possible cases of zero count as in Table 1A. The DM method simply replaces zero by 1 and double the count in the corresponding cell. The modified tables contain count with asterisk as in Table 1B.

Table 1. The general counts of a 2 by 2 table with a zero count (1A) and modified table (1B)

1A

y	x	
	1	0
1	a=0	b
0	c	d

→

1B

y	x	
	1	0
1	a*=1	b
0	c*=c+c	d

y	x	
	1	0
1	a	b=0
0	c	d

→

y	x	
	1	0
1	a	b*=1
0	c	d*=d+d

y	x	
	1	0
1	a	b
0	c=0	d

→

y	x	
	1	0
1	a*=a+a	b
0	c*=1	d

y	x	
	1	0
1	a	b
0	c	d=0

→

y	x	
	1	0
1	a	b*=b+b
0	c	d*=1

2.1 Simulation data for 2 by 2 tables with zero count

The 2 by 2 tables were obtained by generating binomial random numbers using R software. Suppose we have a binary response variable where Y=1 denotes a success and Y=0 otherwise. We also have a binary covariate X, also with values 0 or 1. If p_{ij} is the probability of a successful outcome, P(Y|X), the logistic regression model is given by:

$$P(Y \mid X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \quad (1)$$

And $$\text{Logit}\,(p_{ij}) = \beta_0 + \beta_1 X \quad (2)$$

The logistic regression model for a 2 by 2 table can be shown as in Table 2:

Table 2. The general probabilities given by logistic regression model

Y	X	
	1	0
1	P(Y=1|X=1)	P(Y=1|X=0)
0	1-P(Y=1|X=1)	1- P(Y=1|X=0)

If β_1=0 in Eqn 2, then the rows and columns of the 2 by 2 table are independent. In the following we simulate data based on condition β_1=0 and β_0=-3, for example. Using Eqn 2, the p_{ij} for this independent model may represented by Table 3.

Table 3. The probabilities given by the logistic regression model, using β_1=0, β_0=-3

Y	X	
	1	0
1	0.0474	0.9526
0	0.9526	0.0474

Moreover, we also simulate data tables where the row and colums are dependent/nearly independent. According to the logistic regression model, we calculate the p_{ij} using β_1=0.5, β_0=-3 to generate another set of tables. Table 4 below is in accordance with this model.

Table 4. The probabilities given by logistic regression model using β_1=0.5, β_0=-3

Y	X	
	1	0
1	0.0759	0.9526
0	0.9241	0.0474

The sample sizes of the simulated data sets vary from 10 to 100 with equal sizes for groups X=0 and X=1. For the purposes of this study, only the data from tables which include zero counts are selected. An example of data tables is shown in Table 5A with different sample sizes for X=1 and X=0 and the corresponding data tables after data modification (DM) are shown in Table 5B.

Table 5. The counts in simulated 2 by 2 tables which include zero counts (5A) and corresponding tables (5B) modified according to the DM method

5A

Table ID	a	b	c	d
1	10	10	0	0
2	9	10	1	0
3	10	9	0	1
4	8	10	2	0
:				
:				
103	100	96	0	4
104	96	100	4	0
105	100	97	0	3
106	94	100	6	0
107	100	95	0	5
108	97	100	3	0

5B

Table ID	a	b	c	d
1	20	20	1	1
2	9	20	1	1
3	20	9	1	1
4	8	20	2	1
:				
:				
103	200	96	1	4
104	96	200	4	1
105	200	97	1	3
106	94	200	6	1
107	200	95	1	5
108	97	200	3	1

3 Results

3.1. Comparison the percentages of correctly identified p-values

Figure 1 shows the percentage of times each of three methods correctly identified that the explanatory variable (X) and outcome (Y) variable are dependant, that is,that the test produced a p-value less than 0.05. While the results from our simulation study yielded the highest percentage (approximately 46%) of correct identification of dependence, Lancaster's mid-P test and Firth's method correctly identified dependence in 38% and 41% of cases, respectively.

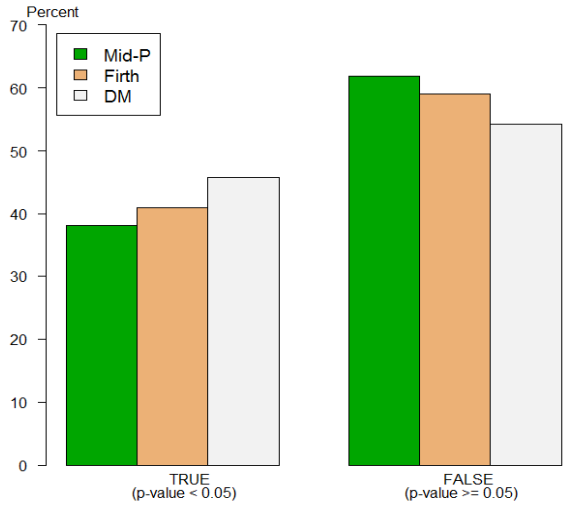


Figure 1. Percentages of times the methods correctly identified a dependent model

When these methods were applied to data where the explanatory variable (X) and outcome (Y) variable are independent, that is, that the test produced a p-value greater than 0.05, we found that Firth's method correctly identified the highest percentage, approximately 81% of independent cases the p-values, while the DM method correctly identified only 63% as shown in Figure 2.

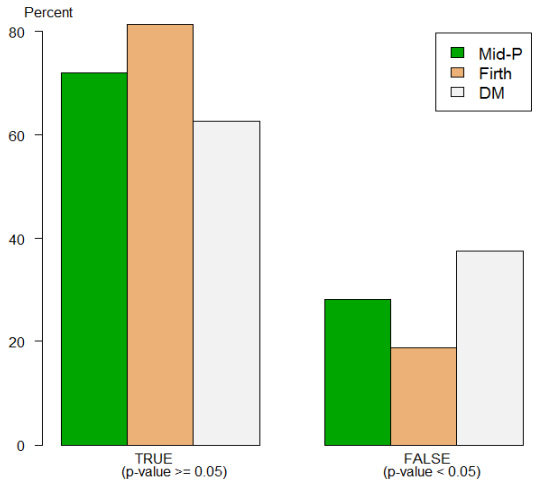


Figure 2. Percentage times the methods correctly identified an independent model

3.2 Comparison of the p-values from DM, Fisher and Lancaster's mid-P test to Firth's

Figure 3 shows a comparison of the p-values from DM, Fisher's exact test and Lancaster's mid-P test to Firth's method. The red diagonal line indicates that p-values from the first three methods are in agreement with those obtained using Firth's method. The finding shows that the p-values from DM method appear closest to Firth's and some of them are in complete agreement both on the data for generated from a dependent model (Table 3) and from an independent model (Table 4).

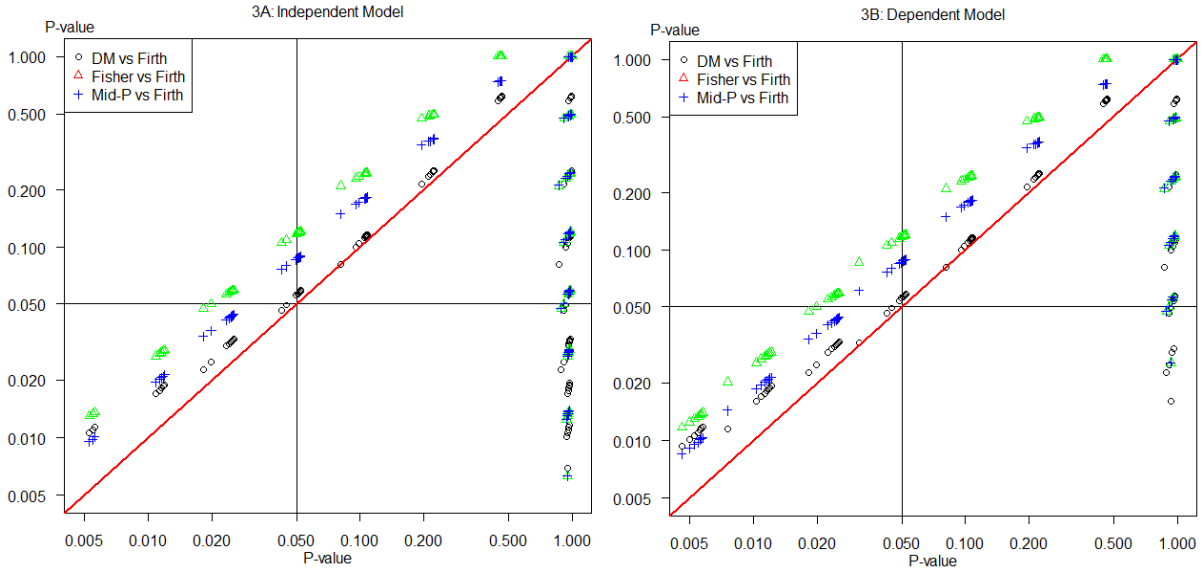


Figure 3. Comparison of the p-values from DM method, Fisher's exact test and Lancaster's mid-P test with Firth's method for 2 by 2 tables simulated from the independent model and dependent model.

4 Conclusion

In this study, we have introduced a simple method for solving the non-convergence problem in logistic regression, resulting from zero counts. We have made a simple modification to the data and compared the results to those obtained from Firth's procedure and others. Firth's method (FL) has previously been recommended for the analysis data with such a problem [6]. From this study, we found that the data modification (DM) method gave similar p-values as Firth's method. In addition, when compared the percentages of correctly indentified p-values for logistic regression, we also found that the tests of independence carried out on data modified by the DM method, produced p- values which were more likely to be in agreement with those produced by the same tests applied to data modified using Firth's method, compared to p-values obtained by Lancaster's mid-p test carried out on the data. Furthermore, the p- values from the tests on the DM data were closer in value to those from the tests on the data modified using Firth's method, than those obtained from both Fisher's exact text and Lancaster's mid-P test.

5 Reference

1. Aitkin M. and Chadwick T., Bayesian analysis of 2x2 contingency tables from comparative trials, School of Mathematics and Statistics, University of Newcastle UK., 2003.

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

3. Bester C.L. and Hansen C., Bias reduction for Bayesian and frequentist estimators, University of Chicago., 2005.

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

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

6. Heinze G. and Schemper M., A solution to the problem of separation in logistic regression, *Stat. Med*., 2002; **21**: 2409-2419.

7. Heinze G. and Ploner M., Fixing the nonconvergence bug in logistic regression with SPLUS and SAS, *Comput. Meth.Prog. Bio*., 2003; **71**: 181-187.

8. Heinze G. and Ploner M., A SAS macro, S-PLUS library and R package to perform logistic regression without convergence problems, Medical University of Vienna., 2004.

9. Little, R.J. and Rubin, D.B. 2002. Statistical Analysis with Missing Data. John Wilay& Sons. New York.

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

Vitae

Name: Mrs. Nurin Dureh

Student ID: 5520330003

Education Attainment:

Degree	Name of institution	Year
B.Sc. (Applied Mathematics)	Prince of Songkla University	2006
M.S. (Research Methodology)	Prince of Songkla University	2010

Scholarship Awards during Enrolment

The Royal Golden Jubilee Ph.D. Program. The Thailand Research Fund.

Work-Position and address:

Lecture

Department of Mathematics and Computer Sciences, Prince of Songkla University.

List of Publications:

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 (4).

Dureh, N., Choonpradub, C., and Tongkumchum, P. 2016. An Alternative Method for Logistic Regression on Contingency Tables with Zero Cell Counts. Songklanakarin Journal of Science and Technology. 38 (2).

Proceeding:

Dureh, N., Choonpradub, C., and Green, H. 2015. Comparing Methods for Testing Association in Tables with Zero Cell Counts Using Logistic Regression. The 2nd International Conference on Computing, Mathematics and Statistics 2015. Malaysia.