HiLL PUBLISHING

Journal of Applied Mathematics and Computation, 2023, 7(1), 177-187
https://www.hillpublisher.com/journals/JAMC/
ISSN Online: 2576-0653
ISSN Print: 2576-0645

Effects of Zero Observations on Modelling Categorical Data

Tong Wang*, Wentao Li, Jing Ruan

Wenzhou Polytechnic, Wenzhou, Zhejiang, China.

How to cite this paper: Tong Wang, Wentao Li, Jing Ruan. (2023) Effects of Zero Observations on Modelling Categorical Data. *Journal of Applied Mathematics and Computation*, 7(1), 177-187.
DOI: 10.26855/jamc.2023.03.020

Received: March 15, 2023
Accepted: April 12, 2023
Published: May 6, 2023

***Corresponding author:** Tong Wang, Wenzhou Polytechnic, Wenzhou, Zhejiang, China.

Abstract

We study the maximum likelihood estimate by using the numerical method Newton Raphson under the Poisson sampling conditions. Log-linear model and logistic model are introduced here, where we both analyze the Independent model and Saturated model for the following studies. Two kinds of zero observations are considered when a model is fitting using MLE, they are sampling zeros and structure zeros. Take two-way contingency table for example, we obtain the correspondence from log-linear model to logistic model when zero entries a contingency table. Each term in a logistic regression model can correspond one term in log-linear model. We study the relationships between those terms through the standard errors in the presence of sampling zeros. We conclude that it is the numbers and positions of zeros that matters when it comes to the effects of zero observations. As for these problems, we analyze the reasons and proposed some useful suggestion to this.

Keywords

Categorical variables, contingency table, maximum likelihood estimation, log-linear model, logistic regression model

1. Introduction

Most data are in the form of categorical data, which is very common in social science. Studying such kind of data is getting increasingly popular, critical, and significant as well. There are many applications in the medical field and scientific area as well as machine learning. Generally speaking, the log-linear model and logistic model are the most popular and influential models for the analysis of such data points. Much attention has been paid to analyzing the large and sparse table, and many researchers are investigating count data with zeros cells in recent years. In practical, it is common for a table with many very small counts or zero counts.

If zero occurs in a table, the MLE might not be estimated because of the lack of information, which will cause the case of non-existence of MLE. It is not about the values of other remaining cell values, it is the number and position of zeros that affect the results. The patterns will be discussed here and some methods also given to solve these cases. In this article, we focus on the statistical inference to study the effects of zeros cells on modeling with two different regression models and see the connections between them. When one of the categorical variables are binary, a log-linear model then can be translated into a certain logistic regression one [1]. The method is Newton Raphson, which is the default method in glm(). Another numerical method IPF can obtain the MLE as well.

Section 2 describes log-linear models, using examples to illustrate it. And then detailed the methods for Sparse Tables. In section3, we show the correspondence between the log-linear model and the logistic model and derive the connections of the math formula.

2. Log-linear Models

We learn that log link function can be used in generalised regression models to implement log-linear regression. Agresti (2011) notes that the most popular and common way to use log-liner models is for categorical data [1]. For the log-linear chapter, we will demonstrate the formula for Poisson regression and the multinomial explanation in section 2.3. Section 2.5 will mainly talk about Newton Rapthon method to obtain maximum likelihood estimates of model parameters for a log-linear model. At the end of this chapter, a two-way contingency table will be as an example to show the cases where zero observations are in different cells. Here we will focus on zero cell counts for contingency table.

2.1 Distribution

In order to make some inferences for log-linear models and logistic models, we need to make some assumption for the data in the contingency table. In this article, we are assuming the categorical data are from the Poisson distribution, which is the best-known distribution for modeling count data. Poisson log-linear models are popular for modeling count data using GLMs. The equal, positive mean and variance for Poisson distribution is one of the most important properties. It is more common to take logarithm to the mean rather than modeling the data using GLMs directly. Agresti (2011) notes the log link for Poisson GLMs is in the canonical link. Therefore, the log mean is the nature parameter. It can be expressed as: $\log\mu = \alpha + \beta x$.

Two principal distributions will be explained and compared: Poisson distribution and Multinomial distribution. The most significant difference between them is the number of trials. The Multinomial distribution requires a fixed number of trials while another has no restriction. Here is an example that shows some count data do not come from a fixed number of trials. If y = the number of birth of babies in a national hospital during the following week, we can not say a specific number, and there is no limit on birth. Therefore, we'd like to use a Poisson log-linear model in this case.

The Poisson log-linear model is more straightforward than the Multinomial log-linear model when fitting the model using maximum likelihood methods. Because there is no sampling constraint for mean parameters, and the components of the random vector are independent. For Multinomial data, the total size n is fixed, so it poses a restriction on the such kind of sampling [2].

2.2 Zero Observations

Some problems will appear if zero entries in the contingency table when fitting a log-linear model [3], because the logarithm of zero does not exist which goes towards minus infinity. When it comes to zero observations ($n_i = 0$), sampling zeros (random zeros or accidentally empty cells) and structure zeros (fixed zeros or necessarily empty cells) [4] are here. Structure zeros are not part of the data set while the random zeros are [5]. When a model is fitting using MLE, Sampling zeros is by chance, while structural zeros are inevitable.

If we observed $n_i = 0$ but $\mu_i > 0$, this indicates that it belongs to sampling zeros. It is called sampling zeros if a zero observation has a positive expected value, if the zero observation has an expected value of zero, this empty cell is called structural zeros. As for sampling zeros, it is possible for any value in the cell count, but it can be only zero observations in terms of structure zeros. Tables with fixed zeros are called incomplete tables [6].

2.3 Definition

Definition 2.1 *Log-linear model is a generalized linear model which is a linear model for the log of expected cell counts. A log-linear model is also called a Poisson regression model [7].*

Firstly, we take a two-way contingency table as an example. When we fit a log-linear model in such a table, we need to study how the expected cell counts are affected by the row and column effects for independent models. As for the saturated model, the interactions among these variables need to be considered. So the log-linear model can be express as below. It is related to the row effects and column effects when we model the data in the contingency table.

Because we are using Poisson sampling with $N = 4$ independent cell counts $\{Y_{ij}\}$. The expected value is the mean of all the observations $\{n_{ij}\}$, which can be formulated as $\mu_{ij} = E(Y_{ij})$

The shorthand $Y \sim$ Poisson (λ) is used to indicate that the random variable Y has the Poisson distribution with positive parameter μ. A Poisson random variable Y with scale parameter μ in the exponential family has a probability mass function in Eqn 2.1 and the corresponding log-likelihood function in Eqn 2.2

$$f(y) = \frac{\mu^y e^{-\mu}}{y!} = \exp(-\mu)\left(\frac{1}{y!}\right)\exp(y\log\mu) \quad y = 0,1,2,\ldots . \tag{2.1}$$

$$L(\mu \mid y_1, y_2, ..., y_n) = -n\mu + \left(\sum_{i=1}^{n} y_i\right)\ln(\mu) + \ln\left(\prod_{i=1}^{n} y_i!\right) \tag{2.2}$$

As for modeling categorical data, another expression from Christenen (2006) for the log-linear model (independent and saturated) is shown in Eqn 2.3 and Eqn 2.4 [8].

(1) Independent Model

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y \tag{2.3}$$

There is a row effect λ_i^X and column effect λ_j^Y and there are no interactions. Hence, this is an independent model. The parameter constraint we set in section 2.4, λ should satisfy $\lambda_0^X = \lambda_0^Y = 0$

$$\begin{bmatrix} \log(\mu_{00}) \\ \log(\mu_{01}) \\ \log(\mu_{10}) \\ \log(\mu_{11}) \end{bmatrix} = X \begin{bmatrix} \lambda \\ \lambda^X \\ \lambda^Y \end{bmatrix}, \text{Where } X = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

Each row in the design matrix represents a cell count. Here we have four observations y_1,y_2,y_3,y_4.

(2) Saturated Model

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY} \tag{2.4}$$

$\lambda^{XY}{}_{ij}$ is the interaction between variable X and Y. It is called first-order interaction because it is related to two variables. The second-order interaction is the triple interaction among the three variables. A special terminology of the hierarchical model, which is terms in (Bishop) needs to be noted in this paper. When the higher-order exists in a model, all the lower-order must be included in this model. For example, $\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_{ij}^{XY}$, this is not a hierarchical model.

The matrix form of the saturated model which is formulated in 2.4 is $\log\mu = X\beta$, it can be expanded:

$$\begin{bmatrix} \log(\mu_{00}) \\ \log(\mu_{01}) \\ \log(\mu_{10}) \\ \log(\mu_{11}) \end{bmatrix} = X \begin{bmatrix} \lambda \\ \lambda^X \\ \lambda^Y \\ \lambda^{XY} \end{bmatrix}, \text{Where } X = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

2.4 Model Constraint

When it comes to modeling log-linear models for contingency tables, two common practice usually come into mind in social science. They are corner point and sum to zeros. Both of them aim at simplifying the calculation for the contingency table. Based on some papers, we found that Bishop et al. (1975), Goodman (1970) and Everitt (1992) and many others choose the former constraint for a model.

$$\sum_{i=1}^{r} \lambda_i^X = \sum_{j=1}^{s} \lambda_j^Y = \sum_{i=1}^{s} \lambda_{ij}^{XY} = 0 \; \lambda_i = \lambda_j = \quad \lambda_{ij} = 0 \quad = 1... \; r, j = 1... \; s, \tag{2.5}$$

While Plackett (1974) uses corner point constraints in his paper, and this constraint can be written as:

$$\lambda_i^X = \lambda_j^Y = \lambda_{rj}^{XY} = \lambda_{xs}^{XY} = 0 \qquad i = 0, j = 0. \tag{2.6}$$

The corner point constraint is the default in glm(). As for the independent model, we set all $\lambda^{XY}{}_{xj} = 0$ for whichever i or j is 0. However, this is not restricted to the saturated models while in such 2×2 contingency table, corner point constraint would be much more simpler than sum to zeros, which set all the factors λ with 0 subscripts to be 0 [8].

2.5 Estimation of parameters

In general, maximum likelihood estimations of the expected cell counts and model parameters are the most essential and vital analysis on modeling categorical data by log-linear models. We know that what is MLEs, it aims to find a value $\hat{\beta}$ that maximize the log-likelihood function (likelihood function). In order to obtain $\hat{\beta}$, we need to do some derivatives to the likelihood function.

$$U_j = \frac{\partial L(\beta)}{\partial \beta}$$

if $U = \begin{pmatrix} U_1 \\ U_2 \\ \cdots \\ U_P \end{pmatrix} = 0$. We then find MLEs. If we can not obtain the direct results from this equation, numerical methods should be used in this case. There are three methods for implementing maximum likelihood estimation. They are the analytical methods, grid search, and numerical ways. In fact, the first two methods are not practical in many cases. Thus, most statistical research will use the package with the numerical maximization. In this paper, we will focus on the numerical method.

Brown and Fuchs (1983) introduced the most two common way to obtain maximum likelihood estimates of log-linear model parameters. They are iterative proportional fitting method (IPF) and iteratively reweighted Newton Raphson method (NR). Both of them are numerical method. The Newton-Raphson process will be detailed in the following. Both of these methods can give the unique maximum likelihood estimators under the condition of the positive cells (n_i>0). However if there are some configurations of zeros in the contingency table, the data can fall on the boundary of the convex support, which means that MLE does not exist in this case, and some parameters may not be estimable [18]. Fienberg (1970) and Haberman (1974) note that it is the numbers and locations of zero counts in a table that can lead to different results, not the values of cells. From the view of Haberman (1974), many sufficient and important conditions for the existence of MLE are given.

2.5.1 Existence of MLEs

People have paid much attention to the problems of existence and uniqueness of the maximum likelihood estimates [9]. They also care about the shape of the likelihood function so that they can find the location of the maximum likelihood estimate. Only when the MLE exists, can the applications of the model analysis, interpretation of parameters, and model selections be valid [10]. The necessary and sufficient conditions of the existence of MLEs are also given in his paper. For example, the chi-square approximations for a large sample can be used to measure the goodness of fit which is utilized for model assessment like hypothesis testing and model selection [11]. When the MLE does not exist, we can not identify a specific model. The standard errors will not be defined as well, so the degrees of freedom has no practical meaning. Hence, if MLE does not exist, some statistical inference may not have a strong theoretical justification. This paper is studying the effects of zero configurations, so it is of much importance to identify the configurations of sampling zeros under the non-existence MLE cases.

2.5.2 Newton Raphson Method

Definition 2.2 *Newton Raphson Method is an iterative method for the problem of nonlinear regression, which is also known as Newton's method in numerical analysis [12].*

Agresti (2011) has detailed this numerical method. This numerical method is to start with a guess which is much close to the fact in your mind and let it is noted by $\widehat{\beta_0}$. In order to find the second guess, we need to find the maximum likelihood estimate in the neighbor of the point you guessed at the initial step by a second-degree polynomial. In terms of simplicity, let us call this adjustment A. so the new estimation in this numerical method could be expressed as

$$\widehat{\beta} = \widehat{\beta_0} + A \tag{2.7}$$

or this can be generally expressed as follows:

$$\widehat{\beta}^{t+1} = \widehat{\beta}^{t} + A \tag{2.8}$$

So on so forth, many guesses will be guessed in this manner. After many iterations, it will converge to a specific level, that is the maximum value we are searching. Using the notation of Agresti (2011) for a generalized regression model. We demonstrate the numerical methods below. We denote that likelihood function is $L(\beta)$, $\widehat{\beta_0}$ can be obtained when $L(\beta)$ is maximized.

Numerical research needs two factors, which are directions and length. The meaning of length is how far you go in that direction. We specify the path according to the slope of the log likelihood function, which is the simplest way to update the guess.

$$\widehat{\beta}^{t+1} = \widehat{\beta}^{t} + \frac{\partial \ln(\beta)}{\partial(\beta^{t})} \tag{2.9}$$

Thus we need to make some adjustment to Eq.2.8. The new guess will not only depend on the direction towards the maximum but also the amount of step size (length is also noted before). After including how far for the current guess to move, changes should be made as:

$$\widehat{\beta}^{t+1} = \widehat{\beta}^{t} + \lambda^{t}\Delta^{t} \tag{2.10}$$

In Eq.2.10, the adjustment will consist of two elements:

1. Δ tells us the direction we should move towards the maximum likelihood estimation.
2. Λ tells us how far we should go each time along that direction we choose.

In general, the second derivative of log-likelihood function is often used to determine how far the current guess should move to reach the maximum. Newton Raphson method applies the Hession matrix, which can be expressed as

$$H = \frac{\partial^{2} \ln(\beta)}{\partial\beta\partial\beta^{t}} \tag{2.11}$$

Newton Rapthon method use the inverse of Hessian matrix. The guess is as follows:

$$\widehat{\beta}^{t+1} = \widehat{\beta}^{t} - \left(H^{t}\right)^{-1} \frac{\partial \ln(\beta)}{\partial\beta^{t}} \tag{2.12}$$

Where $(H^{t})^{-1}$ is the Hessian matrix evaluated at β^{t}.

Here I would like to introduce the information matrix in this paper. The definition of it is the negative of the expected value of the Hessian matrix, which is $[I(\theta)] = -E[H(\theta)]$. It tells us how much information could the data provided for unknown parameters. This paper aims to investigate how would the presence of zeros affect the results when modeling data. It is lack of some information when zero occurs in a contingency table. Hence, the data may provide less information than no zero cases.

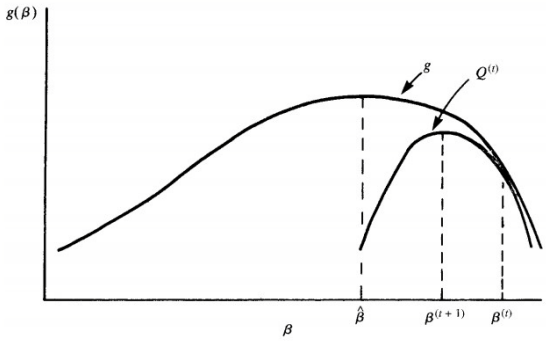

Figure 1. The convergence of Newton Raphson methods.

The shape of the likelihood function affects the maximum likelihood estimates to some extent. If $L(\beta)$ has more than 1 local maxima, this might cause a problem to the root-finding algorithm (Newton Raphson). We might get the wrong ML without a good guess for the initial step. The initial estimate would be extremely crucial in this case. Figure 1 is from (Agresti, 2011), which express the cycle of Newton Raphson in terms of only one single element. It is easy to understand how the arbitrary guess converged to a maximum likelihood estimate. Supposing the first guess is $\widehat{\beta}^{t}$, the MLE is when t reaches the limit. The convergence from $\widehat{\beta}^{t}$ to $\widehat{\beta}$ is described in Figure 1, which shows the second-order approximation for the initial guess.

2.6 Example for a two-way table

There are a lot of articles talking about the relationship between smoking and lung cancer. The number of deaths attributed to lung cancer is gradually increasing in the world. As for the United Kingdom, the mortality in England and Wales caused by the cancer of lung increased significantly, described by Doll & Hill (1950). The main reasons that result in

lung cancer come from two aspects. One is air pollution, and the other (more important) is smoking. This topic is a social concern, and it is hugely related to our interests [13]. So here I use this as an example to show how zero observations will affect log-linear models for contingency tables.

Table 1. Cross-Classification of Smoking by Lung Cancer

	Lung Cancer	
Smoker	Cases	Control
Yes	688	650
No	21	59

[*] Source: Based on data reported in Table IV, Doll R, Hill AB, British Med. J., 1950 Sep 30; 2 (4682):739.

In Table 1, Lung cancer is the response variable Y and smoker is the exploratory variable X. Although we do not need to distinguish them when fitting log-linear models, it matters when fitting with the logistic model. There are 4 observations in this table, and y_1 is 688, y_2 is 21, y_3 is 650, y_4 is 59. This notation will show in the following. We first model the data in Table 1 with log-linear models, replacing non-zero counts with zero observations in turn, which may cause overparametrization.

When some of the cell counts are zeros in a table, it might cause a problem to the model parameters. Some of the parameters will not be estimable in that case. If a parameter has a substantial significant standard error, it is useless. We don't want a parameter with a large confidence interval, which makes no sense. If some estimators are with large standard error, overparametrization happens in this case [7]. According to the analysis above in section 2.5, the likelihood might change so that it might not have one point to make all the parameters to be maximized. This case will result in the singular of information matrix, which means some changes occur in the information matrix. The rank and determinant and eigenvalues are three main characteristic statistics for Fisher's information matrix.

Table 2 will show how zeros cells affect the fitting results and which parameter will be affected by a different pattern of zero counts.

Table 2. The effects of zero counts on information matrix for two-way table

	saturated model			independent model		
zero counts	Determinant	No.eigenvalue(0)	Parameters(big SE)	Determinant	No.eigenvalue(0)	Parameters(big SE)
none	5.54E+08	0	none	3.79E+07	0	none
y1	4.50E-04	1	all	1.06E+06	0	none
y2	1.48E-02	1	$\lambda X,\lambda XY$	2.76E+07	0	none
y3	4.78E-04	1	$\lambda Y,\lambda XY$	3.00E+06	0	none
y4	5.26E-03	1	λXY	9.52E+06	0	none
y1,y2	1.63E-15	2	all	2.86E-05	1	λ,λ^Y
y1,y3	5.26E-17	2	all	5.80E-07	1	λ,λ^X
y1.y4	5.79E-16	2	all	2.78E+05	0	none
y2.y3	1.72E-15	2	$\lambda X,\lambda Y,\lambda XY$	2.21E+06	0	none
y2,y4	1.90E-14	2	$\lambda X,\lambda XY$	1.84E-04	1	λX
y3,y4	6.14E-16	2	$\lambda Y,\lambda XY$	6.44E-06	1	λY
y1,y2,y3	5.17E-28	3	all	6.42E-18	2	all
y2,y3,y4	6.02E-27	3	all	5.13E-17	2	all
y1,y3,y4	1.84E-28	3	all	1.17E-18	2	all

From Table 2, the results can be seen for two log-linear models, and they are different according to the different pattern of zeros. As for full mode, when no zero entries, the determinant of Fishers information matrix is positive and big

enough, all the parameters are estimable. However, once one zero enters the table no matter where it is, it will cause problems to the estimation of parameters. Any two zero counts contain y_1 will make all the model parameters inestimable when we have two zero observations in a table, where two eigenvalues are close to 0. As for three zero observations, all the parameter are with significant standard errors because of the lack of some information. In this case, three eigenvalues are close to 0, and the determinant of Fishers information matrix is extremely small, up to 10^{-28}.

From an independent model, we can see the consequence zeros counts caused is not that severe as the saturated model. If there is only one zero observation in the table, it will not cause problems to model fitting. The determinant and eigenvalues as well as the nonestimable parameters will keep the similar result as no zeros cases. As for two zeros cases, the results are totally different from the saturated model. When the diagonal elements y_1, y_4 and y_2, y_3 are zero counts, the properties of information matrix will keep a similar result as no zero cases. The properties are the positive and big determinant, and no eigenvalues are near 0, so all the parameters can be estimated in this case. When the remaining two cell counts are zeros, some of the model parameters can be not estimated, and there is one eigenvalue which is close to 0. The results of three zero observations will make all the parameters inestimable as well with two eigenvalues near 0.

From this two-way contingency table, we conclude that it is the numbers and positions of zeros that matters when it comes to the effects of zero observations.

2.7 Methods for Sparse Tables

As Friedlander (2016) notes in his article, when zero entries in a table, the data can fall on the boundary of the convex support where some numerical methods like IPF and NR will have problems in converge and become unstable. Log-likelihood function is a concave function, and we can estimate the MLE at the top of this curve. If some data fall on the convex support, the likelihood function may change. If the log-likelihood function is relatively flat, it may cause some problems to locate the maximum likelihood estimate [28]. The information matrix will change as well because the information matrix is defined as the amount of information that we can know from the data about the unknown parameters [14].

We know that log-linear regression can not directly work with zero entries. The premise of its application is positive cell counts [15]. When zero entries into the table, here are some other approaches we can take into consideration if we still prefer to use log-linear regression to model the data.

2.7.1 Add a Small Value to Tables

Durch (2016) and Agresti (2011) and Everitt (1992) all note this method to deal with the cases where zero entries in a table. There are two types of adding a small value to a contingency table with zero observations [22, 28]. One is only adding a small value like 0.5 or 1 to zero cell counts, and the other is for all the cells before doing some sensitivity analysis or modeling. Therefore, we can guarantee that all the observations are positive so that the log-linear regression can be applied.

Everitt (1992) notes that this method does not work when there are many zero observations in a table. If we add some constant to zero cells in that case, this will automatically increase the sample size. Some disadvantages are summarised in the following. Firstly, the standard error will give an incorrect confidence interval, resulting from the increased sample size [16]. Secondly, this method, which is called data modification, may have some bias as an estimator of standard (underestimated). Finally, if it is not the case of sampling zeros but a structure zero, this may have s sever impact on the results, because this is an incomplete table now [17].

2.7.2 A Function of Model Parameters

Let's first introduce parameter redundant. If not all the parameters can be estimated when fitting a model, it is called the parameter redundant. To deal with the case, the popular and accurate method is a symbolic method which can detect the true number of estimable parameters as well as the exact combinations of them like the summations [18].

2.7.3 A Smaller Subset of Model Parameters

There is a lack of information in a sparse table, so we can not estimate everything. However, we can estimate those can be estimable and ignore non-estimable parameters, which is also called extend maximum likelihood estimator. There are two steps to perform extended maximum likelihood estimation, which is what we called a smaller subset in this paragraph. The first thing is to identify which sampling zeros will cause problems to estimate the parameters using MLE. Secondly, we should ignore those zeros we identified in step1 and fit a log-linear model for the remaining ones [19].

3. Logistic Regression

Logistic regression is the most useful and essential for modeling categorical variable [1]. It has been widely used in many fields in recent years. The most common use is for social science and research, it is getting increasingly popular in

the financial market as well. Logistic regression also plays a vital role in genetics.

3.1 Definition

Definition 3.1 *Agresti (2002) pointed out that Logistic regression is a linear model for the logit transformation of a binary parameter. It is usually used when the dependent variable is binary as regression analysis.*

For the binary response Y and exploratory X, $p_i = P(Y = 1|X = i)$. The general logistic regression model is $P_i = \frac{\exp(\beta_0 + \beta_1 x)}{1 - \exp(\beta_0 + \beta_1 x)}$. The general definition of log odds which is also called the logit is shown as:

$$\log it(p_i) = \log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 x_i$$

We are focusing on categorical data in this paper, not the covariates, so we need to distinguish them to some extent. If it is independent logistic regression, then Y is independent of X:

$$logit(P(Y = 1)) = \beta \tag{3.1}$$

If it is a saturated logistic regression model, we can express the log odds with an alternative way which is shown in Eqn 3.2.

$$\log it\left(P\left(Y = 1\right)\right) = \log\left(\frac{p_i}{1 - p_i}\right) = \beta + \beta^X \tag{3.2}$$

3.2 Examples for Two Way Contingency Tables

The examples we choose is the same as log-linear, and we still study the relationship between the presence of lung cancer and the smoker people. The results can be used to compare. As we analyzed in section 2, we again analyze the information matrix, including the rank, the determinant ,and the number of eigenvalues which is close to zero and the parameters to be estimated with big stand errors.

Table 3. The effects of zeros counts on information matrix for two-way table for logistic regression

Saturated model for two-way table				
zero counts	rank	determinant	eigenvalues	parameters
no zeros	2	5176.885	0	none
y1	2	6.43E-06	1	all
y2	2	2.73E-03	1	βX
y3	2	6.43E-06	1	all
y4	2	2.63E-03	1	βX
y1,y2	2	1.40E-19	2	all
y1,y3	1	1.55E+01	0	none
y1,y4	2	1.35E-19	2	all
y2,y3	2	1.40E-19	2	all
y2,y4	1	3.34E+02	0	none
y3,y4	2	1.35E-10	2	all

Here we choose lung cancer Y as response variable, which only has two outcomes 0 and 1, let smoker be other variables. We can see from the table that when there are no zeros in the table, all the parameters will be estimable in this case. Table 3 shows the results that β_X is with big standard error under y_2 is 0. When y_3 is 0. The two components as correspondence in the log-linear model, are not estimable, and all the parameters are with significant standard error. The results are in correspondence with our expectation. When Y_4 is 0, only the interaction term can not be estimated in log-linear model. The corresponding parameter β^X is definitely with big stand error.

As for two zero observations, There are two special cases (y_1, y_3 *and* y_2, y_4 are zeros) where β^X is NA value. This means that it is lack of information to obtain the value of β^X. Only intercept can be estimated in this case. As for one zero case, when y_1 or y_3 is zero, all the two parameters are not estimable. The intercept can be estimated in the remaining cases where y_2 is zero or y_4 is zero. If two Zero observations come into the contingency table. All parameters will not be estimable except for the two special cases which have been mentioned at the beginning of this paragraph if we change the response variable and explanatory variable. The results are somewhat different from what we obtained above.

3.3 The Correspondence with Log-linear Model

Many researches like Birch (1963), Bishop et al. (1975) and Agresti (2011) pointed out these two models have some equivalences. It makes no sense if we analyze the independent model for a two-dimensional table, as the Eqn 3.1 shows there is only the intercept β in logistic regression model. Hence, we focus on the saturated model based on table 1. From Eqn 2.4 and Eqn 3.2, we can obtain the log odds for logistic regression:

$$\begin{aligned} logit(P(Y=1)) &= log\left(\frac{P(Y=1|X=x)}{P(Y=0|X=x)}\right) \\ &= log\left(\frac{P(Y=1,X=x)}{P(Y=0,X=x)}\right) \\ &= log(\mu_{i1}) - log(\mu_{i0}) \\ &= (\lambda + \lambda_i^X + \lambda_1^Y + \lambda_{i1}^{XY}) - (\lambda + \lambda_i^X + \lambda_0^Y + \lambda_{i0}^{XY}) \\ &= (\lambda_1^Y - \lambda_0^Y) + (\lambda_{i1}^{XY} - \lambda_{i0}^{XY}) \\ &= \beta + \beta^X \end{aligned}$$

The above equation has connected the two regression models for categorical data. In addition, the corner point model constraint is setting all the elements which are with a subscript 0 will be 0, like $\lambda_0^X = 0, \lambda_{0j}^{XY} = 0$. Then the connection between two models are $\beta = \lambda_1^Y, \beta^X = \lambda_{i1}^{XY}$. From these two equations, we expect that if the parameter which is the main column effect λ_1^Y is not estimable in log-linear model, the corresponding parameter β would not be estimable as well. If the parameter which is the interaction term λ_{i1}^{XY} can not be estimated in the log-linear model, the corresponding parameter β^X would also be with a big standard error.

In a two-way table, we use λ^Y to denote the column effect instead of λ_1^Y because of the corner point $\lambda_0^Y = 0$. The similar notation for other parameters. According to the Tables 2 and Table 3, we can concluded that: If there is only one zero observation in a table, it does have the specific correspondence as we expected. However, when two zeros entry in the table, this correspondence depends on the positions of zeros. If two zeros are in the same row, model parameters will be affected by this in the log-linear model, while this is not the case when we model the same data using logistic regression. In this special case, the only intercept can be estimated, and another parameter is an NA value which may contribute to the lack of information. However, all parameters will not be estimable except for the two special cases. The correspondence exists for all the remaining cases.

4. Conclusions

In this paper, we use log-linear model and the logistic model for the contingency table. The independent model and saturated model are considered for each type of regression model as well. The results differ according to a different model, not a different regression model. In terms of a two-way contingency table with a log-linear model, all the parameters will not be estimable for the saturated model if y_1 or two zeros which contain y_1. However, this is not the case for the independent model. It does not have any effect on the parameters if there is only one zero appears in the table. There are also two special cases, the diagonal zeros, for a two-way contingency table, where two zero observations do not influence the parameter estimation. It is the numbers and positions of zero observations that matters to the estimation of model parameters. In addition to listing the non-estimable parameters and the corresponding standard errors, some properties of information matrix including the rank, the determinant and the number of eigenvalues which is near 0 are considered as well. Because this will change as well if zero counts affect the likelihood function and MLE will not exist in this case. Some practical methods are also proposed to deal with these cases, for example, adding a small value to guarantee the

positive frequencies or making a function of the parameters so that the combination is estimable. After comparing the parameters for two different regression models, we found that the correspondence between the log-linear model and the logistic model are the same as the math formula shows. In order to draw an ordinary conclusion which is for all cases, we still model the multidimensional table as a comparison. We found the results are consistent with what we obtain from the simplest two-way contingency table.

Funding

This paper is supported by Wenzhou Polytechnic (XJ2022001801).

References

[1] A. Agresti and M. Kateri. *Categorical data analysis*. Springer, 2011.

[2] R. Baker, M. Clarke, and P. Lane. Zero entries in contingency tables. *Computational Statistics & Data Analysis*, 3:33–45, 1985.

[3] M. Birch. Maximum likelihood in three-way contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 25(1):220–233, 1963.

[4] Y. Bishop, S. Fienberg, and P. Holland. Discrete multivariate analysis mit press. *Cambridge, Mass*, 1975.

[5] Y. M. Bishop. Full contingency tables, logits, and split contingency tables. *Biometrics*, pages 383–399, 1969.

[6] Y. M. Bishop, S. E. Fienberg, and P. W. Holland. *Discrete multivariate analysis: theory and practice*. Springer Science & Business Media, 2007.

[7] M. B. Brown and C. Fuchs. On maximum likelihood estimation in sparse contingency tables. *Computational Statistics & Data Analysis*, 1:3–15, 1983.

[8] R. Christensen. *Log-linear models and logistic regression*. Springer Science & Business Media, 2006.

[9] D. J. Cole. Determining parameter redundancy of multi-state mark–recapture models for sea birds. *Journal of Ornithology*, 152(2):305–315, 2012.

[10] R. Doll and A. B. Hill. Smoking and carcinoma of the lung. *British medical journal*, 2(4682):739, 1950.

[11] N. Dureh, C. Choonpradub, and P. Tongkumchum. An alternative method for logistic regression on contingency tables with zero cell counts. *Songklanakarin Journal of Science & Technology*, 38(2), 2016.

[12] B. S. Everitt. *The analysis of contingency tables*. Chapman and Hall/CRC, 1992.

[13] S. S. Far, M. Papathomas, and R. King. Parameter redundancy and the existence of maximum likelihood estimates in log-linear models. *arXiv preprint arXiv:1902.10009*, 2019.

[14] S. E. Fienberg et al. An iterative procedure for estimation in contingency tables. *The Annals of Mathematical Statistics*, 41(3):907–917, 1970.

[15] S. E. Fienberg and A. Rinaldo. Three centuries of categorical data analysis: Log-linear models and maximum likelihood estimation. *Journal of Statistical Planning and Inference*, 137(11):3430– 3445, 2007.

[16] S. E. Fienberg and A. Rinaldo. Maximum likelihood estimation in log-linear mod-els—supplementary material. Technical report, Citeseer, 2012.

[17] S. E. Fienberg, A. Rinaldo, et al. Maximum likelihood estimation in log-linear models. *The Annals of Statistics*, 40(2):996–1023, 2012.

[18] M. Friedlander. Fitting log-linear models in sparse contingency tables using the emleloglin r package. *arXiv preprint arXiv:1611.07505*, 2016.

[19] S. J. Haberman. The analysis of frequency data. 1974.

[20] B. Halpin. Modelling categorical data: Loglinear models and logistic regression. *G&S, Limerick University, August*, pages 26–30, 2002.

[21] D. Holt. Log-linear models for contingency table analysis: On the interpretation of parameters. *Sociological Methods & Research*, 7(3):330–336, 1979.

[22] T.-w. Hu. The fitting of log-regression equation when some observations in the regressand are zero or negative. *Metroeconomica*, 24(1):86–90, 1972.

[23] W. Jing and M. Papathomas. On the correspondence of deviances and maximum likelihood and interval estimates from log-linear to logistic regression modelling. *arXiv preprint arXiv:1711.10440*, 2017.

[24] J. B. Lang. On the comparison of multinomial and poisson log-linear models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):253–266, 1996.

[25] J. I. Myung and D. J. Navarro. Information matrix. *Encyclopedia of Statistics in Behavioral Science*, 2005.

[26] M. Papathomas. On the correspondence from bayesian log-linear modelling to logistic regression modelling with g-priors. *Test*, 27(1):197–220, 2018.

[27] R. L. Plackett. The analysis of categorical data. 1974.

[28] W.-Y. Poon, M.-L. Tang, and S.-J. Wang. Influence measures in contingency tables with application in sampling zeros. *Sociological methods & research*, 31(4):439–452, 2003.

[29] M. Ridout, C. G. Dem´etrio, and J. Hinde. Models for count data with many zeros. In *Proceedings of the XIXth international biometric conference*, volume 19, pages 179–192. International Biometric Society Invited Papers. Cape Town, South Africa, 1998.

[30] A. Rinaldo. Computing maximum likelihood estimates in log-linear models. 2005.