Finite Mixture Logistic-Gaussian Model for Zero-Inflated Clustered Binary Data

John Kwagyan, Phd & Victor Apprey, PhD
Howard University College of Medicine,Washington, DC

Send correspondenc to Dr. John Kwagyan, jkwagyan@howard.edu

Abstract

We establish a finite mixture model for clustered binary data in which members of clusters in one latent class have a zero response with probability one; and clusters in a second latent class yield correlated outcomes. Response probabilities in terms of random effects models are formulated, and maximum marginal likelihood estimation procedures based on Gaussian quadrature are developed. Application to esophageal cancer data in Chinese families is presented.

KEYWORDS: Clustered binary data, Gaussian quadratures, Logistic-Gaussian model, random effect models, zero-inflated clustered data, zero-inflated models

1. INTRODUCTION

Binary response, such as the presence or absence of a disease are common in clinical reseach. In addition, correlated data arise in many application areas including studies of disease occurrence among family members, studies involving repeated measures of outcome on units, and studies involving group randomization. To account for correlation within clusters, random-effects models have been proposed and have been used in various applications for correlated binary data (Anderson and Aitkin, 1985; Prentice, 1988; Rosner, 1989). The presence of binary or count data with excess zeros are also a common phenomena in a wide variety of disciplines. A literature review on this (Ridout *and others*, 1998) cited examples from a variety of research areas including agriculture, econometrics, epidemiology, public health, medicine, and social work. For example, in laboratory litter studies, it may frequently happen that some animals are unaffected by treatment - the so-called non-response phenomena. In correlated grouped-time survival data, some groups of individuals may be immune to the event of interest. Moreover, in genetic studies, it is often suspected that only a small subgroup of patients may have a disease gene that would be linked to a disease marker. Consequently, in studying rare, genetic or familial diseases, data that are randomly sampled, will lead to many families that are largely devoid of individuals with the attribute. Mixture models (Brillinger and Preisler, 1983) provide a natural framework for unobserved heterogeneity in population studies and the overall distribution of disease occurrence in such data or similar outcomes should appropriately be a mixture.

Approaches for dealing with excess zero phenomena, notably the zero-inflated count models have seen rapid interest and development in recent years. This includes the zero-inflated Poisson model (Lambert, 1992) and the zero-inflated Binomial model (Hall, 2000) for count data with excess zeros. Fox(2013) proposed a multivariate zero-inflated Poisson–Gamma model for counts and processing times in modelling feedback behavior. Wang (2010) proposed a zero-inflated Poisson model to handle multivariate count data and zero-inflated Poisson models with random effects have been considered (Min and Agresti , 2005; Rabe-Hesketh and Skrondal, 2007). Hall(2000) introduced zero-inflated binomial model for count data and incorporated random effect to accommodate correlation of outcomes in a repeated measures design. Hur and others (2002), proposed a model for clustered count data with excess zeros for health outcomes research. Recently, (Loeys *and others,* 2012) gave a more general introduction, where Bayesian alternatives have been proposed.

The current paper builds on works of zero-inflated models and introduce a zero-inflated variance component model for clustered binary data with excess zero clusters. We consider an application of disease occurrence among families for motivation and development of our methods. For most applications in family studies, the outcome is a disease status -affected or not affected, and excess zero cluster of families is a common phenomena. For example, diseases that are rare or familial are more susceptible in certain families than others. In our case study of esophageal cancer in 2951 Chinese families (Kwagyan, 2001), 1580(53%) had no affected family members (see Table 1). While models for handling excess zeros for count data have been studied, and whiles the case study presents excess zero clusters, models that accommodate such data structure for clustered binary data with covariate effects have not been well developed.

We (i) introduce a finite mixture likelihood model for binary data with zero-inflated clusters, (ii) model the response probabilities in terms of random effects based on Gaussian distributional assumption to allow for investigation of between cluster heterogeneity and (iii) develop approximate maximum (marginal) likelihood procedure using Gaussian quadrature for estimation of parameters.

2. MODEL FOR ZERO-INFLATED CLUSTERED BINARY DATA

Suppose data is composed of N clusters each of size n_i, $i = 1, ..., N$ and a vector of binary responses $\mathbf{Y}_i = (Y_{i1}, ..., Y_{in_i})^T$ measured on it. Let $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2, .., \mathbf{Y}_N)^T$, then the $\mathbf{Y}_i's$ are independently distributed vectors. Let Π_0 and Π_1 represent two latent classes where Π_0 is the class (of families) whose members do not manifest the attribute under study and Π_1 the class (of families) whose members are susceptible to the attribute under study. In other words, we consider data situation in which excess number of zero vector of responses occur in some clusters. Further, suppose given a cluster (or family) from the class, Π_1, the probability of an outcome follows a Bernoulli distribution with probability of success, δ_{ij}. We shall call Π_0 the "*zero-vector state*" and Π_1 the "*Bernoulli state*". We define the unobserved random variable, $Z_i (Z_i = 0, 1)$, $i = 1, ...N$, such that

$$Z_i = \begin{cases} 0, \text{ with probability } 1 - \alpha_i \\ 1, \quad \text{ with probability } \alpha_i \end{cases}$$

Suppose further that $Z_i = 0$ when $\mathbf{Y}_i$ is generated from the "*zero-vector state*" and $Z_i = 1$ when $\mathbf{Y}_i$ comes from the "*Bernoulli state*". Thus $1 - \alpha_i$ is the probability that a randomly choosing cluster (family) comes from the "*zero-vector state*", Π_0, and α_i, the probability that it comes from the "*Bernoulli state*", Π_1.

Then for each outcome, Y_{ij},

$$Y_{ij} = \begin{cases} 0, \text{ with probability } 1 - \alpha_i \\ Ber(\delta_{ij}), \text{ with probability } \alpha_i \end{cases}$$

With this, joint distribution of the i-th cluster is derived via a mixture formulation as

$$\begin{aligned} P(\mathbf{Y}_i) &= P(Y_{i1} = y_{i1}, ..., Y_{in_i} = y_{in_i}) = E[P(y_{i1}, y_{i2}...., y_{in_i}|Z_i)] \\ &= P(Z_i = 0)P(\mathbf{Y}_i|Z_i = 0) + P(Z_i = 1)P(\mathbf{Y}_i|Z_i = 1) \\ &= (1 - \alpha_i)1_{[\mathbf{y}=0]} + \alpha_i P_1(\mathbf{y}_i) \qquad (1) \end{aligned}$$

By assumption,

$$1_{[\mathbf{y}=0]} = \begin{cases} 1, & \text{if } \sum_{j=1}^{n_i} y_j = 0 \\ 0, & \text{if } \sum_{j=1}^{n_i} y_i > 0 \end{cases} \Leftrightarrow \prod_{j=1}^{n_i}(1 - y_{ij}) \qquad (2)$$

Thus $1_{[y=0]}$ can be thought of as a degenerate (one point) distribution whose values are localized at 0.

In addition, we have by assumption

$$P_1(\mathbf{y}_i) = P_1(y_{i1}, ..., y_{in_i}) = \prod_{j=1}^{n_i} \delta_{ij}^{y_{ij}}(1 - \delta_{ij})^{1-y_{ij}} \qquad (3)$$

Substituting equation (2) and (3) into (1) we establish the joint distribution for the i-th cluster as

$$P(Y_{i1} = y_{i1}, ..., Y_{in_i} = y_{in_i}) = (1 - \alpha_i)\prod_{j=1}^{n_i}(1 - y_{ij}) + \alpha_i \prod_{j=1}^{n_i} \delta_{ij}^{y_{ij}}(1 - \delta_{ij})^{1-y_{ij}} \qquad (4)$$

Thus, the model we obtain is a mixture of a form of a degenerate distribution representing the class (of families) that do not manifest the attribute and an independent distribution representing the class (of families) whose members yield correlated outcomes. In effect, we have established a finite mixture model for clustered binary data in which members of clusters in one latent class have a zero response vector with probability one; and clusters in the other latent class yield correlated outcomes.

From Eqn (4), we find the first moment, the mean of Y_{ij}, as

$$\mu_{ij} = E(Y_{ij}) = P(Y_{ij} = 1) = \alpha_i \delta_{ij}$$

and the second moment, the variance of Y_{ij}, as

$$Var(Y_{ij}) = P(Y_{ij} = 1)\{1- \ P(Y_{ij} = 1)\} = {}_i\delta_{ij}(1- \ {}_i\delta_{ij}) = \mu_{ij}(1- \ \mu_{ij})$$

We notice that for $j \neq j'$,

$$\begin{aligned} P(Y_{ij} &= 1, Y_{ij'} = 1) = P(Y_{ij} = 1)P(Y_{ij'} = 1|Y_{ij} = 1) \\ &\Rightarrow \delta_{ij} = P(Y_{ij} = 1|Y_{ij'} = 1) \end{aligned}$$

And so δ_{ij} is simply the conditional probability of the outcome of one member given another member from the cluster has the attribute.

If $\alpha_i \rightarrow 1$, $\delta_{ij} \rightarrow \mu_{ij}$, the joint distribution, (Equation 4), of the ith cluster reduces to

$$P(Y_{i1} = y_{i1}, ..., Y_{in_i} = y_{n_i}) = \prod_{j=1}^{n_i} \delta_{ij}^{y_{ij}}(1-\delta_{ij})^{1-y_{ij}} \rightarrow \prod_{j=1}^{n_i} \mu_{ij}^{y_{ij}}(1-\mu_{ij})^{1-y_{ij}}$$

This is the standard logistic distribution, which in most applications is the null hypothesis of independence of outcomes within cluster. Thus, α may be interpreted as a measure of cluster dependence. *We shall term the parameter, α, the relative cluster dependence parameter.*

Now suppose the j-th subject has a vector of p individual-specific covariates, $\mathbf{X}_{ij} = (x_{ij1}, ..., x_{ijp})$ and let the i-th cluster has q cluster-specific covariates $\mathbf{W}_i = (w_{i1}, ..., w_{iq})$. The scientific objective is to characterize the dependence of $\mathbf{Y}_{ij}$ on $\mathbf{X}_{ij}$ and $\mathbf{W}_i$. The logit model with Gaussian random effects has been studied extensively (Anderson and Aitkin, 1985; Gilmour *and others*, 1985; Zeger & Karim, 1991; Pinheiro & Bates, 1995 & 2000) and will be adopted. We model the logit of the parameter, δ_{ij}, as

$$\delta_{ij}(\mathbf{X}) = P(Y_{ij} = 1|Y_{ij'} = 1, \mathbf{X}_{ij}) = \frac{1}{1 + \exp[\beta_0 + \boldsymbol{\beta}\mathbf{X}_{ij} + \gamma_i\ ]} \tag{5}$$

where $\lambda_i \backsim N(0, \sigma^2)$, is an unobservable random effect assumed to have a Gaussian distribution with mean zero and variance σ^2 to account for excess heterogeneity and within cluster correlation. In a similar manner, we can model the parameter α_i via a logit link as a function of cluster-specific covariates,$\mathbf{W}_i$, but without random effects. That is

$$\alpha_i(\mathbf{W}) = [P(Z_i = 1|\mathbf{W}_i)] = \frac{1}{1 + \exp[\lambda_0 + \boldsymbol{\lambda}\mathbf{W}_i\ ]}$$

This is necessary, because α_i has an embedded property to measure within cluster dependence, and heterogeneity is accounted for in the modeling for δ_{ij}. Hall (2000) suggested similar parametrization in a random effect zero-inated count model for a repeated measures design.

3. PARAMETER ESTIMATION

Different methods for estimating parameters in random effects models have been proposed by several authors (Zeger and Karim, 1991; McCulloch, 1997; Kuhn and Lavielle, 2005). When the dimension of the random effects is one or two, numerical integration techniques can be implemented reasonably easily and will be used (Im and Gianola**,** 1988; Crouch and Spiegelman, 1990; Hur *and others*, 2002).

In this application, we shall model $\alpha_i = \alpha$, a constant, without loss of generality. The constant relative dependence model can be regarded as the analogue of the desirable homoscedastic model in general linear models.

Let $\boldsymbol{\theta} = \{\alpha, \beta_0, \boldsymbol{\beta}, \sigma^2\}$ be the parameters to be estimated, then the conditional distribution of $\mathbf{Y}$ =$(Y_1, ..., Y_N)^T$, given γ_i is

$$\begin{aligned} P(\mathbf{Y}|\gamma_i, \boldsymbol{\theta}) &= \prod_{i=1}^{N} \left\{ (1-\alpha)\prod_{j=1}^{n_i}(1-y_{ij}) + \alpha \prod_{j=1}^{n_i} \delta_{ij}^{y_{ij}}[1-\delta_{ij}]^{1-y_{ij}} \right\} \\ \gamma_i &\sim N(0,\sigma^2) \end{aligned} \tag{6}$$

The (marginal) distribution of $\mathbf{Y}$ is found by integrating out of the conditional distribution equation (6) with respect to the unobserved random variable γ_i and is given by

$$P(\mathbf{Y}|\boldsymbol{\theta}) = \prod_{i=1}^{N} \left\{ (1-\alpha)\prod_{j=1}^{n_i}(1-y_{ij}) + \alpha \int_{-\infty}^{\infty} \left[\prod_{j=1}^{n_i} \delta_{ij}^{y_{ij}}[1-\delta_{ij}]^{1-y_{ij}} f(\gamma_i;\sigma^2) \right] d\gamma_i \right\}$$

For convenience, we consider $V_i = \gamma_i/\sigma$, then $V_i \sim N(0,1)$, and so

$$P(\mathbf{Y}|\sigma, \boldsymbol{\theta}) = \prod_{i=1}^{N} \left\{ (1-\alpha)\prod_{j=1}^{n_i}(1-y_{ij}) + \alpha \int_{-\infty}^{\infty} \left[\prod_{j=1}^{n_i} \delta_{ij}^{y_{ij}}(1-\delta_{ij})^{1-y_{ij}} \phi(V_i) \right] dV_i \right\} \tag{7}$$

where $\phi(.)$ is the standard normal density and

$$\delta_{ij}(\beta_0, \boldsymbol{\beta}, \sigma) = \frac{1}{1+\exp[-(\beta_0 + \boldsymbol{\beta}\mathbf{X} + \sigma V_i)]}$$

It is not possible to carry out the integration analytically, therefore numerical approximation is necessary. Since the integrals are over Gaussian densities, Gaussian quadrature would be appropriate and is used.

Using an M-point Hermite-Gaussian quadrature; an integral of the form $\int f(v)\phi(v)dv$, where $\phi(v)$ is the standard normal density is approximated by the weighted sum;

$$\int f(v)\phi(v)dv \approx \frac{1}{\sqrt{\pi}} \sum_{m=1}^{M} w_m f(\sqrt{2}v_m)$$

where v_m are the Gaussian quadrature points and w_m the associated weights. Applying this to the log-likelihood of equation (7), we have

$$l(\boldsymbol{\beta}, U; \mathbf{Y}) \approx \sum_{i=1}^{N} \log \left\{ (1-\alpha) \prod_{j=1}^{n_i} (1-y_{ij}) + \frac{\alpha}{\sqrt{\pi}} \sum_{m=1}^{M} w_m \left[\prod_{j=1}^{n_i} (\delta_{ijm})^{y_{ij}} (1-\delta_{jim})^{1-y_{ij}} \right] \right\}$$

where

$$\delta_{ijm}(\beta_0, \boldsymbol{\beta}, \sigma) = \frac{1}{1+\exp\{\beta_0 + \boldsymbol{\beta} X_{ij} + \sqrt{2}\sigma v_m\}}$$

Maximum likelihood using Newton-Raphson algorithm can be used to estimate the parameters. Im and Gianola (1988) recommends using a small number of quadrature points as possible. In principle, one could also use an EM algorithm combined with Gaussian quadrature (Bock and Aitkin, 1981; Im and Gianola, 1988), but notably, the EM has the disadvantage of not readily providing standard errors of the parameter estimates, and convergence is usually slow in the absence of close form solutions and therefore will not be discussed further in this application.

4. APPLICATION TO ESOPHAGEAL CANCER DATA IN CHINESE FAMILIES

Esophageal cancer, a gastointestinal cancer, is one of the deadliest cancers worldwide because of its extremely aggressive nature and poor survival rate and are commonly seen in China than in the US (Li, 1982; Khuroo a*nd others*, 1992; Rasool, 2012), with incidence rates of 20 to 30 times higher. This application involves the study of esophageal cancer in 2951 randomly sampled nuclear families collected in Yangcheng County, Shanxi Province, Peoples Republic of China (Kwagyan, 2001). The main objective of the study was to assess the presence of familial aggregation of esophageal cancer. In this analysis, we consider as a cluster, the nuclear family unit, and assess correlation of the disease adjusting for measured risk factors. The outcome variable is whether an individual is affected with esophageal cancer or not. **Table 1** summarizes the distribution of number of affected individuals by family size. Of the 2951 families, 1371(47%) had at least one affected member and 1580(53%) had no affected members- presenting a substantial number of "excess zero" clusters. The respondent within a family has correlated outcomes which are influenced in part or wholly by the cluster as well as the variables on the individual respondent. The following covariates are available: SEX is coded as 0 for $female$ and 1 for *male*, AGE is years centered at 50, $SMOKE$ is coded as 0 for *nonsmoker* and 1 for *smoker*, and $ALCOHOL$ is coded 0 for *nondrinker* and 1 for *drinker*. The outcome variable, Y, is coded 1 if an individual is affected with esophageal cancer and 0 otherwise.

Table 1 : Distribution of number of affecteds by family size

	Number with esophageal cancer								
Family Size	0	1	2	3	4	5	6	7	Total
3	435	151	34	3					623
4	536	215	52	16	6				819
5	335	203	81	34	8	0			659
6	159	155	59	27	1	4	0		412
7	74	95	46	12	5	3	0	1	232
8	30	54	25	11	2	3	1	0	129
9	6	21	10	3	1	1	0	0	43
10	4	8	6	2	1	1	1	0	23
11	0	3	1	1	0	1	1	0	8
12	0	1	0	1	0	0	0	0	2
13	1	0	0	0	0	0	0	0	1
Total	1580	906	314	110	24	13	3	1	2951

Assuming a single random effect, the general (full) model for predicting an individual's response to esophageal cancer, accounting for potential family to family heterogeneity whiles adjusting for the excess zero clusters is given as:

$$\begin{aligned} logit[\delta_{ij}|\mathbf{X},\lambda_i] &= \beta_0+\beta_1 Sex+\beta_2 Age+\beta_3 Smoke+\beta_4 Alcohol+\lambda_i,\ \ \lambda_i \backsim N(0,\sigma^2) \\ \alpha &= \text{constant} \end{aligned}$$

The following describes specific fitted models and GEE for comparison.

1. *Model I: Standard Logistic Regression Model:* This is the complete independence model, which is our basic null hypothesis.

$$\begin{aligned} logit[\mu_{ij}|\mathbf{X}] &= \beta_0+\beta_1 Sex+\beta_2 Age+\beta_3 Smoke+\beta_4 Alcohol \\ \alpha &= 1 \end{aligned}$$

2. *Model II: Logistic-Gaussian Model:* This is a random effects model that assumes correlation of outcomes within families arises from unobserved random variation accross families.

$$\begin{aligned} logit[\delta_{ij}|\mathbf{X}, \lambda_i] &= \beta_0 + \beta_1 Sex + \beta_2 Age + \beta_3 Smoke + \beta_4 Alcohol + \lambda_i, \quad \lambda_i \backsim N(0, \sigma^2) \\ \alpha &= 1 \end{aligned}$$

3. Model III: Zero-inflated Logistic Model: This is a fixed effect model that assumes correlation of outcomes within families whiles adjusting for excess zero clusters.

$$\begin{aligned} logit[\delta_{ij}|\mathbf{X}, \lambda_i] &= \beta_0 + \beta_1 Sex + \beta_2 Age + \beta_3 Smoke + \beta_4 Alcohol \\ \alpha &= \text{constant} \end{aligned}$$

4. *Model IV: Zero-inflated Logistic-Gaussian Model.* This is a random effect model, that assumes correlation of outcomes within families and further tests for unobserved variation across families whiles adjusting for excess zero clusters.

$$\begin{aligned} logit[\delta_{ij}|\mathbf{X}, \lambda_i] &= \beta_0 + \beta_1 Sex + \beta_2 Age + \beta_3 Smoke + \beta_4 Alcohol + \lambda_i, \lambda_i \backsim N(0, \sigma^2) \\ \alpha &= \text{constant} \end{aligned}$$

Computations of the proposed models were performed using computer programs we developed which was linked with the likelihood optimization software MULTIMAX (Bonney *and others*, 1997). Brillinger and Preisler (1983) reported that results do not change much for quadrature points, M > 8 and so for computations, M $= 9$ was employed to complete the analysis.

Results of the fitted models are summarized in **Table 2**. The table contains values of the parameter estimates with their standard errors, the likelihood ratio chi-square statistics to test the significance of differences between the null hypothesis of independence and the hypothesis of dependence. Within family dependence is described by the magnitude of the relative dependence parameter, α, and excess familial heterogeneity by the magnitude of σ, the variance component parameter. The significance of the individual estimates is judged by t-test based on the standard errors and 95% CI. The final (most parsimonious) model was selected based on the Akaike information criterion (AIC), defined as,

$$AIC =- 2 * log(\text{likelihood}) + 2 * (\text{number of parameters})$$

As expected, the likelihood ratio chi-square statistics show significance of the logistic-Gaussian model (likelihood ratio $\chi^2(1) = 6.76, p < 0.01$), zero-inflated (fixed effect) model (likelihood ratio $\chi^2(1) = 34.02, p < 0.0001$) and the zero-inflated logistic-Gaussian model (likelihood ratio $\chi^2(2) = 46.92, p < 0.0001$) compared with the standard logistic (independence) model. Both zero-inflated models (Model III and Model IV) fit the data better than the logistic-Gausian model (Model II) based on the AIC. The likelihood ratio test indicates that the zero-inflated (random effect) logistic-Gaussian model fits the data better than the zero-inflated (fixed effect) model (likelihood ratio $\chi^2(1) = 15.16, p < 0.001$), indicating significant variation of outcomes across families. It should be noted that the use of the likelihood-ratio test for testing variance components has been called into question, with some advocating halved p-values for such testing of variance component parameters (Snijders and Bosker ,1999). In this analysis, the difference in log-likelihood values between the zero-inflated fixed effect model and random-effect models is large, relative to the number of degrees of freedom, so that the preference of the random effect model is unequivocal. The zero-inflated logistic-Gaussian model was therefore the best fitted model and selected. For this model, the maximum likelihood estimate±SE of the relative dependence parameter, α, was 0.849 ± 0.025; the 95% confidence interval is estimated from

$$0.795 \pm (1.96)(0.027) \text{ as } (0.7421,\ 0.8479)$$

which excludes 1, and so we conclude there is significant dependence of outcomes in these families. This suggests that the data was sampled from a population where the aggregation of esophageal cancer is higher than that from the general population. The estimated ± SE of σ, the variance component parameter is 0.771 ± 0.086 with a 95% CI of $(0.605, 0.939)$. which excludes 0. Thus the data further suggests some degree of heterogeneity of outcomes across families.

At the individual covariate level, sex and age have positive significance while alcohol has negative significance. Smoking was not significant at the 5% level. We conclude that the males were at a higher risk of getting esophageal cancer than the females and also that it is more prevalent in older people. The negative effect of alcohol means that, it has a propensity to lower the risk of esophageal cancer. Although the amount (number of drinks) of alcohol drank is not available for this analysis, we can speculate that the significance could be due to moderation in drinking.

In comparison, the parameter estimate of the mean risk, β_0, is much larger in the logistic (independence) model and the logistic-Gaussian model than the zero-inflated models. The parameter estimates in both fixed and random-effect zero-inflated models are relatively close for sex, age, and smoke. However, the parameter estimate for alcohol in the random-effects model is appreciably smaller (in absolute terms) compared to the estimate in the fixed-effects model. In summary we conclude that esophageal cancer aggregates in the families sampled.

5. CONCLUSION

This paper has been concerned with development of a finite mixture likelihood formulation for clustered binary data with zero-inflated clusters, a data structure in which all members of clusters in one latent class have a zero response with probability one; and clusters in the other latent class yield correlated outcomes. Response probabilities in terms of random effects models and approximate maximum (marginal) likelihood estimation procedures based on Gaussian quadrature for regression analysis were developed. The development, albeit being straight forward and based on simple analytic formulation, is novel and well suited for areas of application including public health and biomedical research. For example, in correlated grouped-time survival data (Hedeker *and others*, 2000), some groups of individuals may be immune to the event of interest. In public health research, patients clustered within a physician office may not have the outcome of interest. The case analysis of the esophageal cancer data demonstrated that, the proposed zero-inflated (fixed effect) logistic model improved the fit over the standard logistic regression model and the logistic-Gaussian model, and it also illustrated that the zero-inflated (random-effect) logistic-Gaussian model fits the data significantly better than the fixed-effect model. Thus, the random effects model provides a useful tool for analyzing clustered binary data with excess-zero clusters. The proposed model provides accounts for within cluster (family) dependence, provides a good portrayal of cluster (family) differences while adjusting for excess zero clusters. The relative errors incurred by ignoring the adjustment of excess-zeros can be problematic even for a small number of groups with zeros, if a traditional modeling methods are used (Gupta *and others*, 1996).

To our knowledge, the work we have established in this paper is the first likelihood formulation for modelling clustered binary data with zero-inflated clusters. It is possible, however, that closer scrutiny, practical considerations and numerical studies would suggest modifications and/or refinements to the methods discussed. In conclusion, we remark that the proposed finite mixture model for correlated binary data is suitable and computationally tractable for the regression analysis of binary data with zero-inflated clusters with covariate effects.

ACKNOWLEDGEMENT

This project has been funded in part with Federal funds (UL1TR000101 previously UL1RR031975) from the National Center for Advancing Translational Sciences (NCATS), NIH, through the Clinical and Translational Science Awards Program (CTSA), by the National Institute on Minority Health and Health Disparities/NIH Award Number G12MD007597 and by National Instituttte of Diabetes and Digestive and Kidney Diseases/NIH grant 1R21DK100875-01A1. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

REFERENCES

ANDERSON, D. A., and AITKIN, M (1985), Variance Component Models with Binary Response: Interviewer Variability. *Journal of the Royal Statistical Society, Series B,* **47**, *203-210*

BOCK, R. D., & AITKIN, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46(4), 443-459.*

BONNEY, G. E., APPREY, V., KWAGYAN, J., & AMFOH, K. (1997). MULTIMAX-A computer package for MULTI-objective MAXimization with applications in genetics and epidemiology. *In Am. J. Human genetics (Vol. 61, No. 4, pp. A193-A193)*

CROUCH AC, SPIELGELMAN E (1990). The Evaluation of Integrals of the form $\int f(t)\exp(-t^2)dt$: Application to Logistic-Normal Models. *J. Am Stat. Assoc,* **85**, *464-46*

DIGGLE P. J, LIANG K.Y, ZEGER S. L. (1994). Analysis of longitudinal data. *Oxford University Press.*

FITZMAURICE G. M., LAIRD N. M, ROTNITZKY.(1993). Regression models for Discrete Longitudinal responses. *Statistical Sciences: 8, 284-309.*

FOX, J. P. (2013). Multivariate zero-inated modeling with latent predictors: Modeling feedback behavior. *Computational Statistics & Data Analysis, 68, 361-374.*

GILMOUR A. R, ANDERSON R. D, RAE A. L (1985). The Analysis of Binomial Data by a Generalized Linear Mixed Model. *Biometrika,* **72**, *593-599.*

GUPTA, P., GUPTA, R., TRIPATHI, R., (1996). Analysis of zero-adjusted count data *Computational Statistics and Data Analysis 23, 207–218, 1996.*

HALL, D. B. (2000). Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics, 56(4), 1030-1039.*

HEDEKER, D., SIDDIQUI, O., AND HU, F. B. (2000). Random-effects regression analysis of correlated grouped-time survival data. *Statistical Methods in Medical Research, 9(2), 161-179.*

HINDE, J. (1982). Compound Poisson regression models. *In GLIM 82: Proceedings of the International Conference on Generalised Linear Models (pp. 109-121). Springer New York.*

HUR, K., HEDEKER, D., HENDERSON, W., KHURI, S., & DALEY, J. (2002). Modeling clustered count data with excess zeros in health care outcomes research. Health Services and Outcomes Research Methodology, 3(1), 5-20.

IM, S., & GIANOLA, D. (1988). Mixed models for binomial data with an application to lamb mortality. *Applied Statistics, 196-204.*

KHUROO MS, ZARGAR SA, MAHAJAN R, BANDAY MA. (1992) High incidence of oesophageal and gastric cancer in Kashmir in a population with special personal and dietary habits. *Gut 33: 11-15*

KUHN, E., & LAVIELLE, M. (2005). Maximum likelihood estimation in nonlinear mixed effects models. *Computational Statistics & Data Analysis, 49(4), 1020-1038.*

KWAGYAN J. (2001). Further Investigation of the disposition Model for correlated binary outcomes. *Ph.D Thesis, Department of Statistics, Temple University, Philadelphia, PA*

LAMBERT, D., (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* 34, 1–14.

LI JY. (1982) Epidemiology of esophageal cancer in China. *Natl Cancer Inst Monog*r 62: 113-120

LOEYS, T., MOERKERKE, B., DE SMET, O., BUYSSE, A., (2012). The analysis of zero-inflated count data: beyond zero-inflated Poisson regression. *British J of Math and Stat Psy 65, 163–180.*

MCCULLAGH P, NELDER J. A. (1990). Generalized linear models. *Chapman and Hall.*

MCCULLOCH, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *JASA, 92(437), 162-170.*

MIN, Y., AGRESTI, A., (2005). Random effect models for repeated measures of zeroinflated count data. *Statistical Modeling* 5, 1–19.

PRENTICE R. L(1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics 44, 1033-1048*

QAQISH B. F, LIANG K. Y(1992). Marginal models for correlated binary responses with multiple classes and multiple levels of nesting. *Biometrics 48: 939-950*
RABE-HESKETH, S., SKRONDAL, A., (2007). Multilevel and latent variable modeling with composite links and exploded likelihoods. Psychometrika 72, 123–140.
RIDEOUT, M., DEMETRIO, C. G., AND HINDE, J. (1998). Models for count data with many zeros. In Proceedings of the XIXth International Biometric Conference (Vol. 19, pp. 179-192).
ROSNER B. (1989). Multivariate methods for clustered binary data with more than one level of nesting. *Journal American Statistical Association 84: 373-380.*
RASOOL S, A GANAI B, SYED SAMEER A, MASSOD A**.** (2012) Esophageal cancer: associated factors with special reference to the Kashmir Valley. *Tumori , 98: 191-203*
SKRONDAL, A., RABE-HESKETH, S**.,** (2004). Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models. *Chapman & Hall, Boca Raton, Florida.*
SNIJDERS, T. and BOSKER, R**.**, (1999) . Multilevel analysis: An introduction to basic and advanced multilevel modeling, pg: *90-91. Sage, Thousand Oaks, CA.*
ZEGER S. L, KARIM R (1991). Generalized Linear Models With Random Effects; A Gibbs Sampling Approach. *Journal American Statistical Association,* **86***, 79-86*
ZHAO L. P, PRENTIC R. L (1990). Correlated binary regression using a quadratic exponential model. *Biometrika 77: 642-648*
WANG, L., 2010. IRT-ZIP modeling for multivariate zero-inated count data. Journal of Educational and Behavioral Statistics 35, 671–692

Table 2: Estimates and Standard Errors of the Regression Analysis of Esophageal Cancer in Chinese Families

	Model I: Standard Logistic	Model II: Standard Logistic-Gaussian	Model III: Zero-inflated Logistic	Model IV: Zero-inflated Logistic-Gaussian
Parameter	Estimate ± SE	Estimate ± SE	Estimate ± SE	Estimate ±SE
Mean Risk: β_0	-5.235 ±0.116	-4.494 ± 0.109	-4.152 ±0.109	-4.272 ± 0.141
Individual Covariates				
Sex: β_1	0.982 ± 0.052	0.908 ± 0.059	0.834 ± 0.058	0.811 ± 0.059
Age: β_2	0.037 ± 0.003	0.038 ± 0.002	0.037 ± 0.002	0.036 ± 0.004
Alcohol: β_3	-1.175±0.171	-1.145 ± 0.173	-1.046 ±0.169	-0.968 ± 0.146
Smoke: β_4	0.056 ± 0.066	0.057± 0.067	0.058 ± 0.059	0.064 ± 0.053
Dependence parameter: α	-	-	0.816 ±0.023	0.749± 0.024
Variance Component : σ		0.732 ± 0.074		0.841 ±0.085
-2*Log(Likelihood)	10749.08	10736.06	10717.32	10702.16
AIC	10759.08	10738.06	10729.32	10714.16
*Likelihood Ratio χ^2(df)		13.02(1)	31.76(1)	46.92(2)
*p-value		< 0.001	< 0.0001	< 0.0001