© Statistical Research and Training Center

مجله‌ی پژوهش‌های آماری ایران

J. Statist. Res. Iran **16** (2019): 397–407

دوره‌ی ۱۶، شماره‌ی ۲، پاییز و زمستان ۱۳۹۸، صص ۳۹۷–۴۰۷

پژوهشکده‌ی آمار

DOI: 10.52547/jsri.16.2.397

DOR: 20.1001.1.25385771.2020.16.2.7.7

Joint Modeling for Zero-Inflated Beta-Binomial and Normal Responses

Seyedeh Sedigheh Azimi*, Ehsan Bahrami Samani and Mojtaba Ganjali

Shahid Beheshti University

Received: 2021/09/01 Approved: 2021/12/01

Abstract. We present a new joint model with random effects for the correlated count with extra zero and continuous responses. In this model, we assume a Zero-Inflated Beta-Binomial distribution for the analysis of overdispersed binomial variable and a normal distribution for the analysis of continuous response. Furthermore, a full model likelihood function approach is used to obtain maximum likelihood estimates of the model parameters. We also evaluate the proposed model using the Monte Carlo simulation method. Finally, we fit the model to real data to find effective factors on mixed responses.

Keywords. Random effects; mixed response; the EM algorithm; population survey data.

MSC 2010: 62J05.

1 Introduction

In some statistical studies in different sciences, the purpose is to analyze mixed responses. For example, in economics, factors that simultaneously affect job position and wage are investigated. To this end, a population survey (PS) was conducted in 1985. In this study, we want to investigate whether

* Corresponding author

factors such as gender, work experience, and marital status simultaneously affect job position and wage. In this study, 385 individuals were considered. The job position (Y) is a sum of two Bernoulli variables: (1) Is the person a manager? (Yes = 1, No = 0) and (2) Is the person a Manufacturer? (Yes = 1, No = 0). Binomial data is a sum of identically distributed and independent Bernoulli experiments. In many sciences, Bernoulli experiments are not independent. Such data is called overdispersed binomial data. Therefore, these data no longer follows the Binomial distribution. In the PS study, the two Bernoulli variables are correlated at a 0.08 level (The p-value for testing $H_\circ$:$\rho = 0$ is 0.076), so we consider the job position as an overdispersed binomial variable. Furthermore, the left panel of Figure 1 shows the bar plot of the job position. The bar plot shows that many individuals are not in manager and Manufacturer occupations. A zero can occur from two sources: (1) One has the ability to have such a job position, but one does not and (2) One cannot have such a job position. Therefore, it seems that the Y response also has Zero-Inflated property. On the other hand, the right panel of Figure 1 shows the Histogram of the logarithm of the wage per hour. It shows that it follows the Normal distribution. The main purpose of this paper is to present a random effect model for the analysis of zero-inflated overdispersed binomial and normal mixed responses.

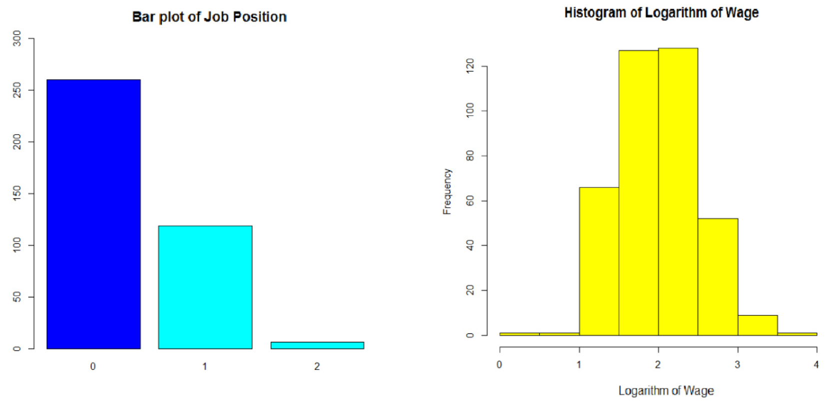

Figure 1. Left Panel: Bar Plots of Job Position. Right Panel: Histogram of Logarithm of Wage per Hour.

Overdispersed binomial data analysis is one of the important topics in statistical and other sciences researches. One suitable distribution for analyzing these data is the Beta-Binomial distribution. This distribution compound of two Binomial and Beta distributions. Firstly, the idea formally proposed by Skellam (1948). Aitkin (1995) and Aitkin (1996) modelled overdispersed binomial data where they assume that random effects have a normal distribution or a discrete mixed distribution. Kim and Lee (2015) investigated the goodness of fit of the beta-binomial model for analyzing overdispersed binomial data.

On the other hand, in the random samples of discrete responses, excess zeros may be observed. Therefore, using standard distributions such as Poisson, Binomial, and Beta-Binomial distributions would not be appropriate for analyzing the response. This has led many researchers to develop suitable models for analyzing these data. For the first time, Lambert (1992) introduced a regression model for analyzing the zero-inflated Poisson response and examined the features of a zero-inflated discrete response. Hu et al. (2018) developed a zero-inflated beta-binomial model for modeling count data. Recently, some researchers have studied models for analyzing zero-inflated mixed responses. For example, Kassahun et al. (2012) introduced a joint model for the analysis of continuous and zero-inflated count mixed responses and analyzed real data via the model.

In the analysis of mixed responses, no model has been presented for the study of zero-inflated overdispersed binomial and normal mixed responses. In this paper, we present a model via random effects to analyze the mixed responses. It is also assumed that zero-inflated overdispersed binomial response follows Zero-Inflated Beta-Binomial distribution. Besides, the EM algorithm is used to estimate the Maximum Likelihood estimates (MLE) of parameters. By the Monte Carlo (MC) simulation study, we estimate the MLE of parameters and evaluate the estimation with precision criteria such as bias and standard error. Finally, we fit the proposed model to the PS data to find the factors which affect mixed responses.

The structure of this paper is as follows. In Section 2, the form of a Zero-Inflated Beta-Binomial distribution is reviewed. A joint model is proposed for analyzing zero-inflated overdispersed binomial and normal mixed responses. Also, we present the likelihood function of the model. In Section 3, the EM algorithm is intended for estimating MLE of parameters. A simulation study is conducted in Section 4 to evaluate estimates of parameters. Finally, fitting the model to real data is proposed in Section 5.

2 Joint Model and Likelihood

A Zero-Inflated Beta-Binomial distribution is a mixture of Beta-Binomial distribution and a degenerate distribution in zero state. Let ρ be the correlation parameter between Bernoulli experiments that random variable Y is a sum of them and π be mixture probability at zero state that is $P(Y=0)=\pi$. Then, the random variable Y that follows Zero-Inflated Beta-Binomial distribution with mean parameter μ will be denoted by $Y \sim \text{ZIBB}(m;\mu,\rho,\pi)$. The pmf of random variable Y is

$$P(Y=y|\mu,\rho,\pi)=\begin{cases}\pi+(1-\pi)P(Y=y|\mu,\rho) & ;y=0\\ (1-\pi)P(Y=y|\mu,\rho) & ;y=1,2,\ldots,m\end{cases}$$

where $0<\mu,\rho$ and $\pi<1$ and

$$P(Y=y|\mu,\rho)=\left(\begin{array}{c} m\\ y\end{array}\right)\frac{\beta\left(y+\mu\left(\rho^{-1}-1\right),m-y+(1-\mu)\left(\rho^{-1}-1\right)\right)}{\beta\left(\mu\left(\rho^{-1}-1\right),(1-\mu)\left(\rho^{-1}-1\right)\right)}.$$

In the following, we present a model for the analysis of zero-inflated overdispersed binomial and normal mixed responses. Let y_i $(i=1,\ldots,n)$ be independent random samples of $ZIBB(m;\mu_i,\rho,\pi_i)$. Also, let $z_1,\ldots,z_n$ be independent random samples of Normal distribution. A joint model with random effects for Zero-Inflated Beta-Binomial and normal mixed responses is defined as

$$\begin{cases} g(\mu_i)=X_i^{(1)'}\beta+W_i^{(1)'}b_i^{(1)},\\ h(\pi_i)=X_i^{(2)'}\lambda+W_i^{(2)'}\nu_i,\\ Z_i=X_i^{(3)'}\gamma+W_i^{(3)'}b_i^{(2)}+\epsilon_i\end{cases} \tag{1}$$

where $g(.)$ and $h(.)$ are link functions and vectors $X_i^{(1)}$, $X_i^{(2)}$ and $X_i^{(3)}$ are the ith rows of design matrices and vectors $W_i^{(1)}$, $W_i^{(2)}$ and $W_i^{(3)}$ are three subsets of design matrices. β, λ and γ are coefficient vectors. Also, let $b_i^{(1)}$, $b_i^{(2)}$ and ν_i be vectors of random effects for $i=1,\ldots,n$ where $\nu_i \overset{iid}{\sim} MVN(\underline{0},\Sigma_\nu)$ and $(b_i^{(1)'},\ b_i^{(2)'})' \overset{iid}{\sim} MVN(\underline{0},\Sigma_b)$. The correlation between individual responses be made by $b_i^{(1)}$s and $b_i^{(2)}$s, which lead to independence of random variables Y_i and Z_i given random effects $b_i^{(1)}$ and $b_i^{(2)}$. Furthermore, let $\epsilon_i \overset{iid}{\sim} N(0,\sigma^2)$.

Also, ϵ_i, ν_i and $(b_i^{(1)'},\ b_i^{(2)'})'$ are mutually independent. For identifiability of model we assume that σ^2 is fixed (Wang (2013)).

The likelihood function for the model is

$$L = \prod_{i=1}^{n} \int P\left(Y_i = y_i | b_i^{(1)}, \nu_i, \eta\right) f_{Z_i}\left(z_i | b_i^{(2)}, \eta\right) \\ \times f\left(b_i^{(1)}, b_i^{(2)} | \eta\right) f\left(\nu_i|\eta\right) db_i^{(1)} db_i^{(2)} d\nu_i$$

where $\eta = (\beta', \lambda', \gamma', \Sigma_b, \Sigma_\nu, \rho)'$ is the vector of parameters. Also, $f\left(b_i^{(1)}, b_i^{(2)}|\eta\right)$ is the joint pdf of random effects $b_i^{(1)}$ and $b_i^{(2)}$ and $f\left(\nu_i|\eta\right)$ is the pdf of random effect ν_i. Furthermore, $f_{Z_i}\left(z_i|b_i^{(2)}, \eta\right)$ is the pdf of random variable Z_i where $Z_i|b_i^{(2)}, \eta \sim N\left(x_i^{(3)'}\gamma + w_i^{(3)'} b_i^{(2)}, \sigma^2\right)$ and $P\left(Y_i = y_i|b_i^{(1)}, \nu_i, \eta\right)$ is the conditional pmf of ZIBB $(m; \mu_i, \rho, \pi_i)$ distribution.

3 Estimation: The EM Algorithm

For estimating parameters via the EM algorithm, we defined latent variable U_i as

$$U_i = \begin{cases} 1; & y_i = 0 \\ 0; & y_i = 1, 2, \ldots, m \end{cases}$$

where $P(U_i = 1) = \pi_i$. Therefore, in using the EM algorithm vector $(y_i, z_i)'$ is incomplete data and vector $(y_i, z_i, u_i, b_i)'$ is complete data where $b_i = (b_i^{(1)'}, b_i^{(2)'}, \nu_i')'$ for $i = 1, \ldots, n$. If we assume conditional distribution of $Y_i|U_i = 1$ is degenerated in zero state with the indicator pmf $I_0(y_i)$, and $Y_i|U_i = 0$ follows BB$(m; \mu_i, \rho)$ distribution, joint pmf of vector $(y_i, u_i)'$ is given as

$$P\left(Y_i = y_i, U_i = u_i | b_i^{(1)}, \nu_i\right) = [\pi_i]^{u_i} \left[(1 - \pi_i) P(Y_i = y_i)\right]^{1 - u_i}.$$

Therefore, the complete data likelihood function is given by

$$f(y, z, u, b) = f(b) f\left(y, z, u | b\right)$$

where $y = (y_1, \ldots, y_n)^{'}$, $z = (z_1, \ldots, z_n)^{'}$, $u = (u_1, \ldots, u_n)^{'}$, $b = \left(b_1^{'}, \ldots, b_n^{'}\right)^{'}$ and

$$f\left(y, z, u|b\right) \quad = \quad \prod_{i=1}^{n} P\left(Y_i = y_i, U_i = u_i|b_i\right) f_{Z_i}\left(z_i|b_i\right).$$

Then, the complete data log likelihood function has kernel

$$\log\left[f\left(y, z, u|b\right)\right] = \sum_{i=1}^{n}\{u_i \log \pi_i + \left(1 - u_i\right)\left[\log\left(1 - \pi_i\right) + \log P\left(Y_i = y_i\right)\right] + \log f_{Z_i}(z_i|b_i^{(2)})\} \equiv \sum_{i=1}^{n} l_c\left(\eta; y_i, z_i, u_i|b_i\right).$$

We find MLE of η parameters vector using $log\left[f\left(y, z, u|b\right)\right]$ via the EM algorithm.

In the $(r+1)$th iteration of the EM algorithm, we compute

$$Q(\eta|\eta^r) = E\left[\log f\left(y, z, u|b, \eta\right)|y, z, \eta^r\right]$$

where the Expectation is with respect to the joint distribution of U and b given y, z and η^r, the parameter estimates based on the rth iteration. This expectation can be taken in two steps,

$$Q(\eta|\eta^r) = E_b\left[E_{U|b,\eta^r}\{\log f\left(y, z, u|b, \eta\right)|b, \eta^r\}|y, z, \eta^r\right]$$

where inner conditional Expectation is Expectation with respect to $U|b, \eta^r$, which is a linear function of U. Therefore, the elements of vector U is replaced by $u_i^r(b_i^{(1)}, \nu_i)$ which is defined as

$$u_i^r(b_i^{(1)}, \nu_i) \quad = \quad E\left(U_i|b_i^{(1)}, \nu_i, \eta^r\right) = \frac{\pi_i^r}{\pi_i^r + P\left(Y_i = y_i|\eta^r\right)\left(1 - \pi_i^r\right)}.$$

We now need to take the Expectation with respect to the distribution of $b_i|y_i, z_i, \eta^r$ to complete the E step. Dropping terms that do not involve η and are therefore irrelevant in the M step, it follows that $Q(\eta|\eta^r)$ equals

$$Q(\eta|\eta^r) \quad = \quad \sum_{i=1}^{n} \int l_c\left(\eta; y_i, z_i, u_i^r(b_i^{(1)}, \nu_i)|b_i\right) f\left(b_i|y_i, z_i, \eta^r\right) db_i. \qquad (2)$$

Integrals given in equation (2) are only with respect to the random effects and can not be evaluated directly, so they are approximated by the Monte Carlo method where $j = 1, 2, \ldots, M = 20$. Random variables are generated according to $\nu_i^{(j)} \overset{iid}{\sim} MVN(\underline{0}, \Sigma_\nu)$ and $(b_i^{(1)'}, \ b_i^{(2)'})'^{(j)} \overset{iid}{\sim} MVN(\underline{0}, \Sigma_b)$.

To complete the M step, we maximize $Q(\eta|\eta^r)$ with respect to η. If MLE of parameters are available, the EM algorithm will be certainly converged because the property of the EM algorithm give us assurance that the sequence $\hat{\eta}^{r+1}$ converges to the incomplete data MLE as $r \to \infty$ (Casella and Berger (2001)). When implementing the EM algorithm we need to check the convergence of the resulting chain. Several convergence diagnostics have been proposed. The E and M Steps are iterated until $|\frac{\log L(\hat{\eta}^{r+1}|y) - \log L(\hat{\eta}^r|y)}{\log L(\hat{\eta}^r|y)}|$ takes an arbitrary small value ε, then the EM algorithm will be converged.

4 Simulation Study

In this section, we conduct a simulation study using the MC method to evaluate the MLE of parameters. Also, we use the EM algorithm to estimate parameters. The model given in (1) is general. We consider a sub model of it to evaluate the proposed model. Let

$$\begin{cases} \text{logit}(\mu_i) = \beta_0 + \beta_1 X_{1i} + \sigma_b b_i^*, \\ \Phi^{-1}(\pi_i) = \lambda_0 + \lambda_1 X_{2i}, \\ Z_i = \gamma_0 + \gamma_1 X_{3i} + \sigma_b b_i^* + \epsilon_i \end{cases} \tag{3}$$

In the model given in (3), we assume random effects in μ_i and Z_i equal to b_i^* to make correlation between Zero-Inflated Beta-Binomial response and normal response. Also, we assume the error variance to be fixed and equal one ($\sigma^2 = 1$). The vector of parameters of model is

$$\eta = (\beta_0, \ \beta_1, \ \lambda_0, \ \lambda_1, \ \gamma_0, \ \gamma_1, \sigma_b, \rho)'.$$

For simulation study, we generate covariates from Normal distribution. Also, ϵ_is and b_i^*s are generated from Normal standard distribution for $i = 1, \ldots, n$. We also generate responses Y_i and Z_i using $ZIBB(m; \mu_i, \rho, \pi_i)$ and Normal distributions which satisfying equation (3), respectively.

The results of the simulation study have been presented in Table 1 with 200 iterations for MC simulation. Also, the EM algorithm is converged with stopping boundary $\varepsilon = 0.0001$. The results show that estimates are near to real values. Also, for most parameters, estimates are closer to their real values with increasing n. Furthermore, standard errors of parameters are small, and most of them decrease with increasing n.

5 Application

In the introduction, PS data were introduced. The purpose of this study is to determine factors such as gender, the number of years of work experience (Experience) and marital status (MS) on the mixed responses of job position, and the logarithm of wage per hour (Z). In these data set, the number of men is slightly higher than the number of women. Also, there are more married people. On the other hand, the number of people with a low number of years of work experience is higher.

In the following, we investigate the assumptions for fitting the proposed model to PS data. Response Y is a sum of two Bernoulli variables. The success probabilities of Bernoulli experiments are equal at a 0.06 level (p-value = 0.055). Furthermore, the Generalized Likelihood Ratio test shows that response Y has overdispersion and Zero-Inflated properties at a 5% level. Also, the Kolmogorov-Smirnov test was performed at a 5% level, with a p-value of 0.165, indicating that the response Z follows the Normal distribution.

We fit the model given in (3) where $X_i^{(1)} = X_i^{(2)} = X_i^{(3)} = (1,\ Gender_i,\ MS_i,\ Experience_i)'$ are the ith rows of design matrices. The result of fitting the model presented in TABLE 2. The result shows that covariates gender, Marital status, and Experience are not significant for the job position at a 5% level. Also, intercept and the covariate Marital Status are significant for the mixed probability (π_i) at a 5% level; that is, if an individual is married, the mixed probability increases. Furthermore, intercept and three covariates are significant for the wage per hour at a 5% level. Fixing other covariates and random effect, if an individual is married, the wage per hour increases, if an individual is a female, the wage per hour decreases, and if an individual has a year of work experience more than another individual, the wage per hour increases.

Table 1. Results using ZIBB - Normal joint model

		m=5					
		n=50		n=100		n=500	
Parameter	True value	Est.	S.E.	Est.	S.E.	Est.	S.E.
β_0	0.500	0.496	0.055	0.498	0.057	0.500	0.055
β_1	1.000	1.000	0.064	1.000	0.054	0.993	0.055
λ_0	3.000	3.475	0.120	3.470	0.120	3.402	0.097
λ_1	1.000	1.540	0.093	1.492	0.078	1.396	0.058
γ_0	0.500	0.546	1.222	0.536	0.818	0.522	0.370
γ_1	1.000	1.018	0.402	1.011	0.268	1.008	0.121
σ_b	1.000	0.931	0.225	0.949	0.157	0.962	0.056
ρ	0.500	0.503	0.057	0.498	0.056	0.492	0.055
		m=10					
		n=50		n=100		n=500	
Parameter	True value	Est.	S.E.	Est.	S.E.	Est.	S.E.
β_0	0.500	0.498	0.061	0.492	0.058	0.504	0.060
β_1	1.000	0.997	0.059	0.995	0.057	1.004	0.056
λ_0	3.000	3.500	0.170	3.459	0.117	3.418	0.132
λ_1	1.000	1.540	0.093	1.497	0.075	1.391	0.071
γ_0	0.500	0.575	1.251	0.619	0.904	0.473	0.371
γ_1	1.000	1.015	0.417	1.040	0.294	0.991	0.123
σ_b	1.000	0.898	0.233	0.953	0.163	0.976	0.060
ρ	0.500	0.502	0.058	0.502	0.056	0.499	0.056
		m=20					
		n=50		n=100		n=500	
Parameter	True value	Est.	S.E.	Est.	S.E.	Est.	S.E.
β_0	0.500	0.507	0.058	0.496	0.057	0.491	0.054
β_1	1.000	1.002	0.056	1.005	0.056	0.999	0.056
λ_0	3.000	3.498	0.143	3.471	0.149	3.403	0.143
λ_1	1.000	1.532	0.094	1.488	0.087	1.386	0.063
γ_0	0.500	0.349	1.153	0.609	0.974	0.501	0.322
γ_1	1.000	0.947	0.382	1.033	0.317	0.999	0.106
σ_b	1.000	0.872	0.244	0.961	0.166	0.977	0.067
ρ	0.500	0.504	0.058	0.504	0.058	0.499	0.056

Table 2. Results of fitting the model to PS data

Parameter	Est.	S.E.
job position		
β_0	-1.680	0.398
Gender	-0.340	0.354
MS	0.004	0.013
Experience	0.229	0.387
mixed probability		
λ_0	6.250	0.206
Gender	0.332	0.183
MS	-0.019	0.007
Experience	0.147	0.203
logarithm of wage per hour		
γ_0	1.940	0.053
Gender	-0.289	0.049
MS	0.004	0.002
Experience	0.137	0.054
σ_b	0.131	0.057
ρ	0.145	0.066

Discussion and Conclusions

In this paper, we proposed a random effects joint model for analyzing Zero-Inflated Beta-Binomial and normal mixed responses. In a simulation study, we estimate parameters via the EM algorithm. It shows that the MLE of parameters are near to their real values. Also, we fit the proposed model to the PS data and find factors that simultaneously affect mixed responses. We could extend the proposed model for modeling mixed responses with missing values.

References

Aitkin, M. (1995). *NPML Estimation of the Mixing Distribution in General Statistical Models with Unobserved Random Variation.* (eds.G.U.H. Seeber, B.J. Francis, R. Hatzinger, G. Steckel-Berger), Springer-Verlag, Berlin, 1-9.

Aitkin, M. (1996). A General Maximum Likelihood Analysis of Overdispersion in Generalized Linear Models. *Statistics and Computing*, **6**, 251-262.

Casella, G., and Berger, R.L. (2001). *Statistical Inference*, 2nd Edition. Duxbury Press, Pacific Grove.

Hu, T., Gallins, P., and Zhou, Y-H. (2018). A Zero-inflated Beta-binomial Model for Microbiome Data Analysis. *Stat.*, **7**, e185. DOI:10.1002/sta4.185.

Kassahun, W., Neyens, T., Molenberghs, G., Faes, C., and Verbeke G. (2012). Modeling Overdispersed Longitudinal Binary Data Using a Combined Beta and Normal Random-Effects Model. *Archives of Public Health*, **70**. DOI: 10.1186/0778-7367-70-7.

Kim, J., and Lee, J.H. (2015). The Validation of a Beta-binomial Model for Overdispersed Binomial Data. *Communications in Statistics - Simulation and Computation*, **46**, 807-814. DOI: 10.1080/03610918.2014.96009.

Lambert, D. (1992). Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. *Technometrics*, **34**, 1-14.

Skellam, J.G. (1948). A Probability Distribution Derived from the Binomial Distribution by Regarding the Probability of a Success as Variable Between the Sets of Trials. *Journal of the Royal Statistical Society, Series B*, **10**, 25-261.

Wang, W. (2013). Identifiability of Linear Mixed Effects Models. *Electron. J. Stat.*, **7**, 244-263.

Seyyedeh Sedigheh Azimi
Department of Statistics,
Shahid Beheshti University,
Tehran, Iran.
email: *sazimi04@gmail.com*

Ehsan Bahrami Samani
Department of Statistics,
Shahid Beheshti University,
Tehran, Iran.
email: *ehsan_bahrami_samani@yahoo.com*

Mojtaba Ganjali
Department of Statistics,
Shahid Beheshti University,
Tehran, Iran.
email: *m-ganjali@sbu.ac.ir*