American Journal of Mathematics and Statistics 2023, 13(1): 44-59
DOI: 10.5923/j.ajms.20231301.02

Handling Missing Values in Covariate for Modeling Count Data with over Dispersion and Zero Inflation

Rajibul Mian[1,*], Sudhir Paul[2]

[1]Department of Medicine, McMaster University, Hamilton, ON L8N 3Z5 Canada
[2]Department of Mathematics and Statistics, University of Windsor, Windsor, ON N9B 3P4 Canada

Abstract The problem of regression analysis of count response data having information on some covariates missing may arise in some practical applications. Further complications, such as, over-dispersion and zero-inflation in the count responses, may also arise. In this paper we develop estimation procedure for the parameters of a zero inflated over/under dispersed count response model in the presence of missing covariates. A zero-inflated negative binomial model with missing covariate information is used. Obtaining maximum likelihood estimates by direct use of the log-likelihood involves multiple numerical integration. To avoid this we develop a weighted expectation maximization algorithm. A simulation study is conducted to investigate the properties of the estimates, in terms of bias, variance, mean squared errors (MSE) and coverage probability (CP). Further simulations are also conducted to study Robustness of the procedure for count data following other over-dispersed models, such as the log-normal mixture of the Poisson distribution. An example and a discussion are given.

Keywords Count Data, EM Algorithm, Missing Covariate Information, Over dispersion, Regression model, Zero inflation

1. Introduction

Discrete data in the form of counts arise in many health science disciplines such as biology and epidemiology. For examples of discrete count data see Deng and Paul (2000, 2005), Bohning, D., Dietz, E., Schlattmann, P., Mendonca, L., and Kirchner, U. (1999), Anscombe (1949); Bliss and Fisher (1953); Bliss and Owen (1958); McCaughran and Arnold (1976); Margolin, Kaplan and Zeiger (1981); Ross and Preece (1985)), Manton, Woodbury and Stallard (1981). The Poisson distribution is the most commonly used distribution for analysing count data. The Poisson distribution has a property that mean and the variance of the distribution are equal to each other. However, in many count data cases this property of the Poisson distribution does not hold, as extra dispersion (variation) is observed in the data, and thus Poisson distribution is not an ideal choice for analysing count data in many applications. The presence of extra dispersion in count data is common in many real life situations. To accommodate this extra dispersion situation in count data a well known model is the negative binomial distribution, which is very convenient and common in practice. For the applications of the negative binomial distribution see for example Engel (1984); Breslow (1984); Margolin et al. (1989); Lawless (1987); Manton et al. (1981). The negative binomial distribution has flexibility in its parameterization and has been used differently by different authors. For example, see Paul and Plackett (1978); Barnwal and Paul (1988); Paul and Banerjee (1998); Piegorsch (1990), Deng and Paul (2000, 2005). Often times a particular count (for example zero) may arise in the data more than the expected number. Count data with many zeros may not be explained properly by a model such as a Poisson distribution and a negative binomial distribution, so a zero inflated Poisson distribution and a zero inflated negative binomial distribution can be the ideal choice. For example see Deng and Paul (2000, 2005), Ridout, Demetrio and Hinde (1998), Williamson, Lin, Lyles and Hightower (2007). Count data in the presence of both extra dispersion as well as zero inflation can be analysed by a zero inflated negative binomial model. Extensive work has been done to fit zero-inflated and over-dispersed count data model to real life data. For example, see Ridout, Demetrio and Hinde (1998), Hinde and Demetrio (1998), Li, Lu, Park, Kim, Brinkley and Peterson (1999), Hall (2000), Lee, Wang and Yau (2001), Wang, Lee, Yau and Carrivick (2003), Lord, Washington and Ivan (2005), Jiang and Paul (2009), Cameron and Trivedi (2013). Also a lot of work has been done to test the presence of zero-inflation and/or over-dispersion. For example, see Mullahy (1997), Dean (1992), Greene (1994), Broek (1995), Deng and Paul (2000), Xie, He, and Goh (2001), Paul, Jiang, Rai and Balasooriya (2004), Williamson, Lin, Lyles and Hightower (2007).

* Corresponding author:
mianr@mcmaster.ca (Rajibul Mian)
Received: Oct. 6, 2022; Accepted: Oct. 30, 2022; Published: Apr. 15, 2023
Published online at http://journal.sapub.org/ajms

An example of count data in the presence of both extra dispersion as well as zero inflation can be found in Bohning, Dietz, Schlattmann, Mendonca, and Kirchner (1999). Bohning et al. (1999) present a set of data on a prospective study of dental status represented by decayed, missing and filled teeth (DMFT) index of school children from an urban area of Belo Horizonte (Brazil). DMFT index scores can range from 0 to 28 or 32 per individual. The tooth is considered as decayed, when a carious lesion or both carious lesion and a restoration are present. The tooth is considered as missing if the tooth has been extracted due to caries. If a temporary or permanent filling is present in the tooth, or the filling of the tooth is defective but not decayed, then the tooth is considered as a filled tooth. The total number of tooth of a person having these properties would be the DMFT index for the person. More details of DMFT index can be found in Cappelli and Mobley (2007). The DMFT index was observed for 797 children at the beginning and at the end of the study. For the purpose of illustration here we consider DMFT index observed at the beginning of the study which, when summarized in terms of index and its frequency, are (index, frequency): (0,172), (1,73), (2,96), (3,80), (4,95), (5,83), (6,85), (7,65), (8,48). The mean and the variance of these counts are 3.3237 and 6.6387, which show over-dispersion in the data. Further, the observed frequency of zeros is 172 as opposed to the expected frequency of $797 \times P(x = 0) = 797 \times (0.036010) = 28.71$ showing that the data are also zero-inflated under a Poisson model.

Regression analysis of count data may be further complicated by the existence of missing values either in the response variable and/or in the explanatory variables (covariates). Extensive work has been done on regression analysis of continuous response data with some missing covariates under normality assumption. See, for example, Rubin (1977), Little and Rubin (1987, 2002, 2014), Lipsitz and Ibrahim (1996), Ibrahim, Chen and Lipsitz (1999), Ibrahim, Chen, Lipsitz and Herring (2005), Sinha and Maiti (2007), Maiti and Pradhan (2009).

Some work on missing values has also been done on logistic regression analysis of binary data. See, for example, Ibrahim (1990), Lipsitz and Ibrahim (1996), Ibrahim and Lipsitz (1996), Ibrahim, Chen and Lipsitz (1999), Ibrahim, Chen and Lipsitz (2001), Sinha and Maiti (2007), Maiti and Pradhan (2009).

Rubin (1977) and Little and Rubin (1987, 2002, 2014) discuss various missingness mechanisms. If the missingness does not depend on observed data, then the missing data are called missing completely at random (MCAR). If the missing data mechanism depends only on observed data, then the data are missing at random (MAR). The MAR is also known as ignorable missing That is, in this case, the missing data mechanism is ignored. If the missing data mechanism depends on both observed and unobserved data, that is, failure to observe a value depends on the value that would have been observed, then the data are said to be missing not at random (MNAR) in which case the missingness is nonignorable. For more detailed discussion on missing data mechanism see Ibrahim et al. (2005).

The purpose of this paper is to develop estimation procedure for the parameters of a zero-inflated negative binomial (ZINB) model for count data when information on some covariates on some individuals are missing.

The problem of missing responses in ZINB model was dealt earlier by the same researchers in Mian and Paul (2016); and guided to this research.

Obtaining maximum likelihood estimates by direct use of the log-likelihood involves multiple numerical integration. To avoid this we develop a weighted expectation maximization algorithm following Ibrahim (1990). A simulation study is conducted to investigate the properties of the estimates, in terms of bias, variance, mean squared errors (MSE) and coverage probability (CP). Further simulations are also conducted to study Robustness of the procedure for count data following other over-dispersed models, such as the log-normal mixture of the Poisson distribution. The method is illustrated using the dental epidemiology data of Bohning et al. (1999) discussed above.

The procedure for the estimation of the parameters are developed in Section 2. Results of a simulation study is reported in Section 3. The illustrative example is given in Section 4 and a discussion leading to some conclusions is given in Section 5.

2. Estimation in Zero-Inflated and Over-Dispersed Count Data Regression Model with Missing Values in the Explanatory Variables

2.1. The ZINB Model

The zero-inflated negative binomial regression model (Deng and Paul, 2005) can be written as

$$f(y_i|x_i;\mu,c,\omega) = \begin{cases} \omega + (1-\omega)\left(\frac{1}{1+c\mu}\right)^{c^{-1}} & if\ y = 0, \\ (1-\omega)\frac{\Gamma(y+c^{-1})}{y!\Gamma(c^{-1})}\left(\frac{c\mu}{1+c\mu}\right)^{y}\left(\frac{1}{1+c\mu}\right)^{c^{-1}} & if\ y > 0 \end{cases} \quad (1)$$

with $E(Y) = (1-\omega)\mu$, and $Var(Y) = (1-\omega)\mu[1+(c+\omega)\mu]$, where ω is the zero-inflation parameter. We denote this distribution by $ZINB(\mu, c, \omega)$ distribution.

Suppose that data for the i^{th} of n subjects are (y_i, x_i), $i = 1, \dots, n$, which are realizations from $ZINB(\mu, c, \omega)$, where y_i represents the response variable and x_i represents a $p \times 1$ vector of covariates with the regression parameter $\beta = (\beta_1, \beta_2, \dots, \beta_p)$, such that $\mu_i = \exp(\sum_{j=1}^{p} X_{ij}\beta_j)$. Here β_1 is the intercept parameter in which case $X_{i1} = 1$ for all i.

2.2. Estimation of the Parameters with no Missing Values in Covariate

For complete data the likelihood function is

$$L(\beta, c, \omega|y_i) = \prod_{i=1}^{n} [\omega + (1-\omega)f(0;\mu_i, c, \omega)I_{\{y_i=0\}} + (1-\omega)f(y_i;\mu_i, c, \omega)I_{\{y_i>0\}}]. \quad (2)$$

Writing $\gamma = \omega/(1-\omega)$ the log likelihood, apart from a constant, can be written as

$$\begin{aligned} l(\beta, c, \gamma|y_i) &= \sum_{i=1}^{n} [-\log(1+\gamma) + \log[\gamma + f(0;\mu_i, c, \omega)]I_{\{y_i=0\}} + \log f(y_i;\mu_i, c, \omega)I_{\{y_i>0\}}] \\ &= \sum_{i=1}^{n} [-\log(1+\gamma) + \log[\gamma + \exp[-c^{-1}\log(1+\mu_i c)]]I_{\{y_i=0\}} \\ &\quad +[(y_i \log\mu_i - (y_i + c^{-1})\log(1+\mu_i c) + \sum_{l=1}^{y_i} [1+(l-1)c])]I_{\{y_i>0\}}]. \end{aligned} \quad (3)$$

The parameters β_j, c and γ can be estimated by directly maximizing the loglikelihood function (2.3) or by simultaneously solving the estimating equations Given in Appendix A.

2.3. Estimation of the Parameters with Missing Values in Covariate

2.3.1. Estimation under MCAR

In case of MCAR, missingness of the data do not depend on observed data and the subjects having the missing observations are deleted before the analysis. For estimation procedure the likelihood function remains same an given in equation (2.3) with reduced sample size having only complete observations.

2.3.2. Estimation under MAR

As some of the observations in some covariates (on some individuals) may be missing we write x_{ij} (the j^{th} covariate value of the i^{th} individual)

$$x_{ij} = \begin{cases} x_{o,ij} & if\ x_{ij}\ is\ observed, \\ x_{m,ij} & if\ x_{ij}\ is\ missing. \end{cases} \quad (4)$$

Putting this in $f(y_i|x_i;\mu, c, \omega)$ given in equation (1), the log-likelihood of y_i $i = 1, \ldots, n$ is

$$\begin{aligned} l(\psi|Y, X_{o,j}, X_{m,j}) &= \sum_{i=1}^{n} \log(f(y_i|x_{i,j}, \psi)) \\ &= \sum_{i=1}^{n} [-\log(1+\gamma) + \log[\gamma + f(0;\mu_i, c, \omega)]I_{\{y_i=0\}} \\ &\quad +\log f(y_i;\mu_i, c, \omega)I_{\{y_i>0\}}]. \end{aligned} \quad (5)$$

Note that $X_{o,j}$ and $X_{m,j}$, the observed and the missing values for the j^{th} covariate are involved in μ_i.

In MAR, conditional probability of missing covariate values depends on observed values (y and x). Parameters of the missingness mechanism are completely separate and distinct from the parameters of the model (1). In likelihood based estimation considering MAR, missingness mechanism can be ignored from the likelihood and missing data (covariate) that are missing at random are often known as ignorable missing, but the subjects having these missing covariates can not be deleted before the analysis. (see Little and Rubin, 1987, 2002, 2014 and Ibrahim, Chen, Lipsitz and Herring, 2005 for detailed discussion).

Following Ibrahim (1990), and Lipsitz and Ibrahim (1996), we consider covariates $x_j, (j = 1,2, \ldots, p)$ are random variables with finite parameters $\alpha_j, (j = 1,2, \ldots, p)$. These covariates x_j have distribution that can be expressed by one dimensional conditional distribution as $p(x_j|\alpha_j) = p(x_j|x_{j-1}, \alpha_j)$ where $j = 1,2, \ldots, p$, more specifically

$$\begin{aligned} p(x_1, x_2, \ldots, x_p|\alpha_1, \alpha_2, \ldots, \alpha_p) &= p(x_p|x_1, \ldots, x_{p-1}, \alpha_p) \\ &\quad \times p(x_{p-1}|x_1, \ldots, x_{p-2}, \alpha_{p-1}) \\ &\quad \times \ldots p(x_2|x_1, \times \alpha_2)p(x_1, \alpha_1). \end{aligned} \quad (6)$$

Following Ibrahim (1990), suppose an arbitrary covariate $x_{i,j}$ have some missing observations and these missing observations are missing at random. For this covariate $x_{i,j}$ the probability of observing the missing observations (conditional on the response, y and the other completely observed covariates) does not depend on the missing covariate itself and any other covariate with unobserved observations, but may depend on the response as well as completely observed covariate. This flexible characteristics of MAR comes to an aid during estimation. Only if this probability of observing the missing observations depend on the response as well as completely observed covariate, then $p(x_1, x_2, \ldots, x_p|\alpha_1, \alpha_2, \ldots, \alpha_p)$ needs to incorporate in the main likelihood $l(\psi|Y, X_{o,j}, X_{m,j})$. Note that in practical regression problem, the covariates are usually very poorly dependent among each other, otherwise multicollinearity problems can be solved by using other statistical tools.

To incorporate this covariate distribution with the complete data loglikelihood of ψ, following Ibrahim (1990), we specify the joint distribution of (y_i, x_i) by using the conditional distribution of $(y_i|x_i)$ and the marginal distribution of x_i, that is

$p(y_i, x_i|\psi) = p(y_i|x_i, \psi) \times p(x_i|\alpha)$. Following Ibrahim (1990) we consider, $(y_i|x_i)$ are independent and $x_{i,j'}s$ are independently (for $j = 1,2, \dots, p$) and identically distributed for all n observations. Considering this, our likelihood becomes

$$l(\psi|Y, X_{o,j}, X_{m,j}) = \sum_{i=1}^{n} [-\log(1+\gamma) + \log[\gamma + f(0; \mu_i, c, \omega)] I_{\{y_i=0\}} + \log f(y_i; \mu_i, c, \omega) I_{\{y_i>0\}}] + \sum_{i=1}^{n} \log[p(x_1, x_2, \dots, x_p|\alpha_1, \alpha_2, \dots, \alpha_p)]. \quad (7)$$

In this loglikelihood, parameters of the count data model (ψ) and the covariates model $(\alpha's)$ are separate as well as distinct. This idea facilitates the separate maximization of the both parts of the likelihood. Moreover, covariates $x_{j'}s$ can be discrete or continuous or mixture of discrete and continuous. All the covariates in $p(x_1, x_2, \dots, x_p|\alpha_1, \alpha_2, \dots, \alpha_p)$ may not have missing observations, in that case, distributions of the completely observed covariates can be ignored (detailed discussion on this are available in Lipsitz and Ibrahim (1996), and Ibrahim et al. (2005)).

In this scenario, our main goal is to estimate the parameters of the count data model (ψ) by maximizing the following loglikelihood (Little and Rubin, 1987, 2002, 2014 p.89) with respect to the parameters ψ

$$l(\psi|Y, X_{o,j}) = \sum_{X_{m,j}} l(\psi|Y, X_{o,j}, X_{m,j}). \quad (8)$$

For continuous covariates or mixed covariates scenario (Ibrahim, Chen and Lipsitz, 1999) the loglikelihood become

$$l(\psi|Y, X_{o,j}) = \int_{X_{m,j}} l(\psi|Y, X_{o,j}, X_{m,j}) dX_{m,j}. \quad (9)$$

Direct maximization of $l(\psi|Y, X_{o,j}, X_{m,j})$ is not, in general, straight forward. However, the EM algorithm (Dempster, Larid and Rubin, 1977) is a very useful tool for obtaining maximum likelihood estimates with missing observations.

The EM algorithm uses two iterative steps known as the expectation-step (E-step) and the maximization-step (M-step). Following Little and Rubin (1987, 2002, 2014), the E-step provides the conditional expectation of the log-likelihood $l(\psi|y_i, x_{o,i,j}, x_{m,i,j})$ given the observed data $(y_i, x_{o,i,j})$ and current estimate of the parameters ψ.

Suppose we have a covariate with missing observations and A of the n observations of the covariate are observed and $B = n - A$ observations are missing and s be an arbitrary number of iterations during maximization of the log-likelihood, then the E-step of the EM algorithm for the i^{th} observation of the missing covariate for $(s+1)^{th}$ iteration can be written as

$$\begin{aligned} Q_i(\psi|\psi^{(s)}) &= E[l(\psi^{(s)}|y_i, x_{o,i,j}, x_{m,i,j})|y_i, x_{o,i,j}, \psi^{(s)}] \\ &= \sum_{x_{m,i,j}} l(\psi^{(s)}|y_i, x_{o,i,j}, x_{m,i,j}) P(x_{m,i,j}|y_i, x_{o,i,j}, \psi^{(s)}). \end{aligned} \quad (10)$$

For continuous covariates or mixed covariates scenario (Ibrahim, Chen, Lipsitz and Herring, 2005) $Q_i(\psi|\psi^{(s)})$ become

$$Q_i(\psi|\psi^{(s)}) = \int_{x_{m,i,j}} l(\psi^{(s)}|y_i, x_{o,i,j}, x_{m,i,j}) P(x_{m,i,j}|y_i, x_{o,i,j}, \psi^{(s)}) dx_{m,i,j}. \quad (11)$$

For all the observations, the E-step of EM algorithm for $(s+1)^{th}$ iteration is

$$Q(\psi|\psi^{(s)}) = \sum_{i=1}^{A} l(\psi^{(s)}|y_i, x_{i,j}) + \sum_{i=1}^{B} \sum_{x_{m,i,j}} l(\psi^{(s)}|y_i, x_{o,i,j}, x_{m,i,j}) P(x_{m,i,j}|y_i, x_{o,i,j}, \psi^{(s)}). \quad (12)$$

For all the observations in case continuous covariates or mixed covariates cases (Ibrahim, Chen, Lipsitz and Herring, 2005) $Q(\psi|\psi^{(s)})$ become

$$Q(\psi|\psi^{(s)}) = \sum_{i=1}^{A} l(\psi^{(s)}|y_i, x_{i,j}) + \sum_{i=1}^{B} \int_{x_{m,i,j}} l(\psi^{(s)}|y_i, x_{o,i,j}, x_{m,i,j}) P(x_{m,i,j}|y_i, x_{o,i,j}, \psi^{(s)}) dx_{m,i,j}. \quad (13)$$

Note for the situation in which there is no missing observations in covariates the EM algorithm requires only maximization of the first term on the right hand side.

Here $P(x_{m,i,j}|y_i, x_{o,i,j}, \psi^{(s)})$ is the conditional distribution of the missing covariate given the observed data and the current (s^{th} iteration) estimate of ψ. However, in many situations, $P(x_{m,i,j}|y_i, x_{o,i,j}, \psi^{(s)})$ may not always be available. Following Ibrahim, Chen, Lipsitz and Herring, 2005 and Sahu and Roberts, 1999, we can write $P(x_{m,i,j}|y_i, x_{o,i,j}, \psi^{(s)}) \propto P(y_i|x_{i,j}, \psi^{(s)}) P(x_{i,j}|\alpha^{(s)})$, where $P(y_i|x_{i,j}, \psi^{(s)})$ is the complete data distribution given in (1), $P(x_{i,j}|\alpha^{(s)})$ is the distribution for the covariates where the missing values exist and both have very elegant forms. For the i^{th} of the B missing observations of the covariate we take a sample $a_{i1}, a_{i2}, \dots, a_{im_i}$ from $P(x_{m,i,j}|y_i, x_{o,i,j}, \psi^{(s)})$ using Gibbs sampler (see Casella and George, 1992 for details). Then, following Ibrahim, Chen and Lipsitz (1999) and Ibrahim, Chen, Lipsitz and Herring (2005) $Q(\psi|\psi^{(s)})$ can be written as

$$Q(\psi|\psi^{(s)}) = \sum_{i=1}^{A} l(\psi^{(s)}|y_i, x_{i,j}) + \sum_{i=1}^{B} \frac{1}{m_i} \sum_{k=1}^{m_i} l(\psi^{(s)}|y_i, x_{o,i,j}, a_{ik}). \quad (14)$$

In the M-step of the EM algorithm, the $Q(\psi|\psi^{(s)})$ is maximized. Here maximizing $Q(\psi|\psi^{(s)})$ is analogous to maximization of complete data log likelihood where each incomplete covariate being replaced by m_i weighted observations. More details of EM algorithm by method of weights can be found in Ibrahim, 1990; Lipsitz and Ibrahim, 1996(a,b), Ibrahim, Chen and Lipsitz, 1999, 2001; Ibrahim, Chen, Lipsitz and Herring, 2005; Sinha and Maiti, 2007; Maiti and Pradhan, 2009.

Variance covariance matrix of the estimates of the parameters is calculated by inverting the observed information matrix at

convergence (Efron and Hinkley, 1978) which is

$$H_{\psi\psi'} = Q''(\psi|\psi^{(s)}) = \sum_{i=1}^{A} \frac{\partial^2}{\partial\psi\partial\psi'} l(\psi^{(s)}|y_i, x_{i,j}) + \sum_{i=1}^{B} \frac{1}{m_i} \sum_{k=1}^{m_i} \frac{\partial^2}{\partial\psi\partial\psi'} l(\psi^{(s)}|y_i, x_{o,i,j}, a_{ik}). \quad (15)$$

Expressions for the elements of H above are given in the Appendix.

2.3.3. Estimation under MNAR

If missing observation in the covariate are considered to be missing not at random (MNAR), the missingness depends on the values that would have been observed. In this case, probability of missing observations in covariate depends on the observed values of the other covariate, response and the values of the covariate that would have been observed. This missing data mechanism cannot be ignored and needs to be incorporated in the likelihood. The missing observations that follow this missing data mechanism are known as nonignorable missing. To incorporate this missing data mechanism in the data likelihood, it is natural to specify a parametric model for this mechanism. Let $r_{ij}, (i = 1,2, \dots, q)$ be a random vector of missingness indicator for the j^{th} covariate.

$$r_{ij} = \begin{cases} 0 & if\ x_{ij}\ is\ observed, \\ 1 & if\ x_{ij}\ is\ missing. \end{cases} \quad (16)$$

Following Ibrahim, Lipsitz and Chen (1999), we specify a one dimentional conditional distributions for r_{ij}

$$\begin{aligned} p(r_{1,j}, r_{2,j}, \dots, r_{q,j}|y_i, x_{i,j}, \nu_1, \nu_2, \dots, \nu_q) = \quad & p(r_{q,j}|r_{1,j}, \dots, r_{q-1,j}, y_i, x_{i,j}, \nu_q) \\ & \times p(r_{q-1,j}|r_{1,j}, \dots, r_{q-2,j}, y_i, x_{i,j}, \nu_{q-1}) \\ & \times \dots p(r_{2,j}|r_{1,j}, y_i, x_{i,j}, \nu_2) \times p(r_{1,j}, y_i, x_{i,j}, \nu_1). \end{aligned} \quad (17)$$

Here $\nu_1, \nu_2, \dots, \nu_q$ are the indexing parameters for the conditional distribution of r_{ij}. Highest value for q can be n. Logistic regression is a popular choice for the one dimensional distribution for r_{ij}

$$l(\nu|r_{ij}, y_i, x_{ij}) = \sum_{i=1}^{n} [r_{ij} * f(y_i, x_{ij}) - \log(1 + \exp(f(y_i, x_{ij})))]. \quad (18)$$

where $f(y_i, x_{ij}) = logit[p(r_{ij}|y_i, x_{ij}, \nu)] = \nu_0 + \nu_1 * y_i + \nu_2 * x_{i1} + \nu_3 x_{i2} + \dots + \nu_q x_{ip}$. Note that choice of variables for the model of r_{ij} is important. Often many variables in this model are not necessarily significant, and more importantly parameters in the model for r_{ij} are not the primary interest for estimation. Detailed discussion on this can be found in Ibrahim, Lipsitz and Chen (1999) and Ibrahim, Chen and Lipsitz (2001).

Following Ibrahim, Lipsitz and Chen (1999), after incorporating the model for missingness mechanism $(l(\nu|r_{ij}, y_i, x_{ij}))$, the data loglikelihood become

$$\begin{aligned} l(\psi|Y, X_{o,j}, X_{m,j}) = \quad & \sum_{i=1}^{n} [-\log(1+\gamma) + \log[\gamma + f(0; \mu_i, c, \omega)] I_{\{y_i=0\}} \\ & + \log f(y_i; \mu_i, c, \omega) I_{\{y_i>0\}}] \\ & + \sum_{i=1}^{n} \log[p(x_1, x_2, \dots, x_p|\alpha_1, \alpha_2, \dots, \alpha_p)] \\ & + \sum_{i=1}^{n} [r_{ij} * f(y_i, x_{ij}) - \log(1 + \exp(f(y_i, x_{ij})))]. \end{aligned} \quad (19)$$

It is to be noted that three parts of this likelihood are separate and their parameters are distinct. This characteristics of the loglikelihood facilitates the separate maximization. Rest of the estimation procedure under MNAR remains exactly same as the estimation procedure under MAR.

3. Simulation Study

A simulation study was conducted to investigate the properties of the estimates, in terms of bias, variance, mean squared errors (MSE) and coverage probability (CP) of estimates. We use data under four scenarios: (i) data are observed completely, (ii) some observations in covariates are missing completely at random (MCAR), (iii) some observations in covariates are missing at random (MAR), and (iv) some observations in covariates are missing not at random (MNAR). Simulations are conducted for continuous as well as discrete covariate.

Responses are generated from the zero-inflated negative binomial model (1) with $\mu_i = \exp(\sum_{j=1}^{2} X_{ij}\beta_j)$ where $\beta_1 = 1$, $\beta_2 = -1$, and $c = 0.2$, $\omega = 0.2$. Note that β_1 is the intercept parameter, hence $x_{i1} = 1$. The explanatory variable x_{i2} was generated from $N(1.5,0.001)$ when covariate is considered to be continuous, and from $Binomial(0.5)$ in case of discrete covariate. We consider 5%, 10% and 25% missing observations in the explanatory variable. For empirical coverage probability we take nominal level $\alpha = 0.05$.

Simulation results for continuous covariate are given in Table 1 whereas results for discrete covariate are in Table 2.

Table 1. Properties (estimate, bias, variance, mse, coverage probability (cp)) of the estimates of the parameters, data simulated from NB (β_1, β_2, c, ω), based on 5000 simulation runs (continuous covariate)

2*n	2*	2*% missing	Missingness Mechanism															
			CC	MCAR	MAR	MNAR	CC	MCAR	MAR	MNAR	CC	MCAR	MAR	MNAR	CC	MCAR	MAR	MNAR
			$\beta_1 = 1$				$\beta_2 = -1$				$c = 0.2$				$\omega = 0.2$			
15*30	3*Estimate	5	1.0139	0.9822	0.9486	0.9601	-0.6852	-0.9748	-0.9902	-0.9997	0.0143	0.2261	0.2753	0.2576	0.1464	0.1175	0.1579	0.1641
		10	1.0139	0.9922	0.9486	0.9629	-0.6852	-1.0281	-0.9919	-1.0017	0.0143	0.2220	0.2670	0.2576	0.1464	0.1122	0.1604	0.1614
		25	1.0139	0.9811	0.9564	0.9566	-0.6852	-0.9601	-0.9902	-1.0072	0.0143	0.2122	0.2625	0.2616	0.1464	0.1125	0.1613	0.1596
	3*Bias	5	0.0139	-0.0178	-0.0514	-0.0399	0.3148	0.0252	0.0098	0.0003	-0.1857	0.0261	0.0753	0.0576	-0.0536	-0.0825	-0.0421	-0.0359
		10	0.0139	-0.0078	-0.0514	-0.0371	0.3148	-0.0281	0.0081	-0.0017	-0.1857	0.0220	0.0670	0.0576	-0.0536	-0.0878	-0.0396	-0.0386
		25	0.0139	-0.0189	-0.0436	-0.0434	0.3148	0.0399	0.0098	-0.0072	-0.1857	0.0122	0.0625	0.0616	-0.0536	-0.0875	-0.0387	-0.0404
	3*Variance	5	0.4006	1.3148	0.0364	0.0357	4.1113	14.2098	0.4204	0.4061	0.0930	0.5169	0.0217	0.0207	0.0318	0.4247	0.0075	0.0075
		10	0.4006	1.3516	0.0365	0.0354	4.1113	13.8158	0.3923	0.4227	0.0930	0.5578	0.0219	0.0206	0.0318	0.4587	0.0076	0.0075
		25	0.4006	1.3152	0.0361	0.0357	4.1113	13.5427	0.3855	0.4063	0.0930	0.6625	0.0214	0.0213	0.0318	0.4545	0.0076	0.0078
	3*MSE	5	0.3713	0.5268	0.0184	0.0126	4.0045	5.6536	0.0413	0.0278	0.0363	0.1059	0.0156	0.0102	0.0138	0.0430	0.0061	0.0044
		10	0.3713	0.5699	0.0180	0.0123	4.0045	5.7097	0.0348	0.0285	0.0363	0.0973	0.0135	0.0106	0.0138	0.0441	0.0055	0.0046
		25	0.3713	0.5681	0.0150	0.0139	4.0045	5.5389	0.0319	0.0320	0.0363	0.1040	0.0121	0.0116	0.0138	0.0422	0.0051	0.0050
	3*CP	5	0.9391	0.9918	0.9540	0.9728	0.9174	0.9928	1.0000	1.0000	0.9957	1.0000	0.9777	0.9916	0.9957	0.9813	0.9646	0.9665
		10	0.9391	0.9941	0.9559	0.9695	0.9174	0.9984	1.0000	1.0000	0.9957	0.9993	0.9830	0.9853	0.9957	0.9853	0.9716	0.9716
		25	0.9391	0.9901	0.9618	0.9681	0.9174	0.9965	1.0000	1.0000	0.9957	0.9986	0.9865	0.9907	0.9957	0.9830	0.9761	0.9670
15*50	3*Estimate	5	1.0191	0.9647	0.9601	0.9626	-0.7379	-0.9827	-0.9987	-0.9979	0.0330	0.2251	0.2618	0.2495	0.1756	0.1266	0.1573	0.1620
		10	1.0191	0.9743	0.9587	0.9691	-0.7379	-0.9960	-0.9903	-1.0010	0.0330	0.2268	0.2559	0.2478	0.1756	0.1291	0.1613	0.1593
		25	1.0191	0.9815	0.9616	0.9698	-0.7379	-1.0201	-0.9962	-1.0013	0.0330	0.2228	0.2528	0.2454	0.1756	0.1237	0.1598	0.1633
	3*Bias	5	0.0191	-0.0353	-0.0399	-0.0374	0.2621	0.0173	0.0013	0.0021	-0.1670	0.0251	0.0618	0.0495	-0.0244	-0.0734	-0.0427	-0.0380
		10	0.0191	-0.0257	-0.0413	-0.0309	0.2621	0.0040	0.0097	-0.0010	-0.1670	0.0268	0.0559	0.0478	-0.0244	-0.0709	-0.0387	-0.0407
		25	0.0191	-0.0185	-0.0384	-0.0302	0.2621	-0.0201	0.0038	-0.0013	-0.1670	0.0228	0.0528	0.0454	-0.0244	-0.0763	-0.0402	-0.0367
	3*Variance	5	0.3891	0.5854	0.0214	0.0211	3.9837	6.1372	0.2271	0.2278	0.0623	0.2699	0.0124	0.0120	0.0142	0.2157	0.0044	0.0044
		10	0.3891	0.6921	0.0216	0.0211	3.9837	7.0446	0.2227	0.2232	0.0623	0.2977	0.0125	0.0120	0.0142	0.2300	0.0044	0.0043
		25	0.3891	0.7689	0.0212	0.0212	3.9837	8.5434	0.2377	0.2183	0.0623	0.3287	0.0121	0.0122	0.0142	0.2558	0.0043	0.0043
	3*MSE	5	0.3769	0.3983	0.0133	0.0113	4.0116	4.1973	0.0316	0.0238	0.0324	0.0896	0.0120	0.0089	0.0070	0.0297	0.0054	0.0040
		10	0.3769	0.4473	0.0135	0.0099	4.0116	4.5101	0.0272	0.0234	0.0324	0.0917	0.0104	0.0083	0.0070	0.0300	0.0045	0.0042
		25	0.3769	0.4399	0.0124	0.0088	4.0116	4.8078	0.0254	0.0190	0.0324	0.0870	0.0096	0.0073	0.0070	0.0325	0.0044	0.0034
	3*CP	5	0.9360	0.9715	0.9563	0.9662	0.9260	0.9736	1.0000	1.0000	0.9915	0.9984	0.9399	0.9556	0.9744	0.9884	0.9140	0.9440
		10	0.9360	0.9769	0.9532	0.9676	0.9260	0.9793	1.0000	1.0000	0.9915	0.9998	0.9575	0.9582	0.9744	0.9842	0.9335	0.9425
		25	0.9360	0.9808	0.9594	0.9735	0.9260	0.9846	1.0000	1.0000	0.9915	0.9995	0.9526	0.9735	0.9744	0.9856	0.9377	0.9608
15*100	3*Estimate	5	0.9939	0.9895	0.9701	0.9722	-1.0023	-1.0173	-0.9957	-0.9991	0.1893	0.2066	0.2457	0.2405	0.1542	0.1501	0.1618	0.1615
		10	0.9939	0.9807	0.9677	0.9753	-1.0023	-0.9898	-0.9939	-0.9968	0.1893	0.2093	0.2436	0.2393	0.1542	0.1476	0.1608	0.1613
		25	0.9939	0.9766	0.9715	0.9718	-1.0023	-1.0063	-0.9958	-0.9945	0.1893	0.2148	0.2410	0.2410	0.1542	0.1444	0.1616	0.1625
	3*Bias	5	-0.0061	-0.0105	-0.0299	-0.0278	-0.0023	-0.0173	0.0043	0.0009	-0.0107	0.0066	0.0457	0.0405	-0.0458	-0.0499	-0.0382	-0.0385
		10	-0.0061	-0.0193	-0.0323	-0.0247	-0.0023	0.0102	0.0061	0.0032	-0.0107	0.0093	0.0436	0.0393	-0.0458	-0.0524	-0.0392	-0.0387
		25	-0.0061	-0.0234	-0.0285	-0.0282	-0.0023	-0.0063	0.0042	0.0055	-0.0107	0.0148	0.0410	0.0410	-0.0458	-0.0556	-0.0384	-0.0375
	3*Variance	5	0.3012	0.3308	0.0104	0.0104	3.2633	3.5050	0.1228	0.1103	0.0530	0.0718	0.0057	0.0058	0.0143	0.0333	0.0021	0.0021
		10	0.3012	0.3814	0.0106	0.0104	3.2633	4.0678	0.1050	0.1102	0.0530	0.0850	0.0060	0.0057	0.0143	0.0625	0.0021	0.0021

		25	0.3012	0.4468	0.0104	0.0104	3.2633	4.8191	0.1147	0.1178	0.0530	0.1147	0.0057	0.0057	0.0143	0.0712	0.0021	0.0021
	3*MSE	5	0.2892	0.2837	0.0087	0.0075	3.1011	3.0263	0.0199	0.0167	0.0331	0.0441	0.0072	0.0062	0.0094	0.0116	0.0037	0.0033
		10	0.2892	0.3300	0.0093	0.0057	3.1011	3.4871	0.0175	0.0148	0.0331	0.0497	0.0069	0.0057	0.0094	0.0145	0.0035	0.0031
		25	0.2892	0.3557	0.0079	0.0077	3.1011	3.7945	0.0171	0.0158	0.0331	0.0612	0.0065	0.0062	0.0094	0.0174	0.0033	0.0032
	3*CP	5	0.9479	0.9536	0.9557	0.9596	0.9461	0.9485	1.0000	1.0000	0.9692	0.9774	0.9281	0.9415	0.9913	0.9900	0.9030	0.9213
		10	0.9479	0.9535	0.9483	0.9660	0.9461	0.9519	1.0000	1.0000	0.9692	0.9792	0.9349	0.9495	0.9913	0.9907	0.9181	0.9331
		25	0.9479	0.9594	0.9602	0.9600	0.9461	0.9604	1.0000	1.0000	0.9692	0.9909	0.9437	0.9432	0.9913	0.9880	0.9217	0.9264

Table 2. Properties (estimate, bias, variance, mse, coverage probability (cp)) of the estimates of the parameters, data simulated from NB (β_1, β_2, c, ω), based on 5000 simulation runs (discrete covariate)

2*n	2*	2*% missing	Missingness Mechanism															
			CC	MCAR	MAR	MNAR	CC	MCAR	MAR	MNAR	CC	MCAR	MAR	MNAR	CC	MCAR	MAR	MNAR
			$\beta_1 = 1$				$\beta_2 = -1$				$c = 0.2$				$\omega = 0.2$			
15*30	3*Estimate	5	0.9858	0.9792	0.9430	0.9513	-1.1393	-1.0842	-1.0014	-0.9924	0.2041	0.2172	0.2324	0.2638	0.1373	0.1212	0.1671	0.1634
		10	0.9858	0.9737	0.9342	0.9583	-1.1393	-1.0628	-0.9974	-0.9987	0.2041	0.2174	0.2403	0.2626	0.1373	0.1170	0.1681	0.1598
		25	0.9858	0.9850	0.9251	0.9574	-1.1393	-1.1237	-0.9952	-0.9869	0.2041	0.2205	0.2371	0.2631	0.1373	0.1168	0.1702	0.1655
	3*Bias	5	-0.0142	-0.0208	-0.0570	-0.0487	-0.1393	-0.0842	-0.0014	0.0076	0.0041	0.0172	0.0324	0.0638	-0.0627	-0.0788	-0.0329	-0.0366
		10	-0.0142	-0.0263	-0.0658	-0.0417	-0.1393	-0.0628	0.0026	0.0013	0.0041	0.0174	0.0403	0.0626	-0.0627	-0.0830	-0.0319	-0.0402
		25	-0.0142	-0.0150	-0.0749	-0.0426	-0.1393	-0.1237	0.0048	0.0131	0.0041	0.0205	0.0371	0.0631	-0.0627	-0.0832	-0.0298	-0.0345
	3*Variance	5	0.0685	0.1355	0.0376	0.0373	0.3523	0.2327	0.1251	0.1478	0.2025	0.5133	0.0207	0.0196	0.0539	0.2915	0.0082	0.0078
		10	0.0685	0.1481	0.0396	0.0406	0.3523	0.2235	0.1095	0.1086	0.2025	0.5187	0.0233	0.0238	0.0539	0.3622	0.0087	0.0089
		25	0.0685	0.1528	0.0434	0.0403	0.3523	0.3097	0.0874	0.1084	0.2025	0.6163	0.0301	0.0239	0.0539	0.3507	0.0104	0.0089
	3*MSE	5	0.0476	0.0766	0.0232	0.0160	0.3661	0.2459	0.0455	0.0306	0.0613	0.0959	0.0129	0.0117	0.0186	0.0335	0.0067	0.0048
		10	0.0476	0.0827	0.0271	0.0126	0.3661	0.2413	0.0519	0.0292	0.0613	0.0912	0.0158	0.0109	0.0186	0.0402	0.0070	0.0050
		25	0.0476	0.0865	0.0290	0.0138	0.3661	0.3333	0.0574	0.0286	0.0613	0.1004	0.0165	0.0111	0.0186	0.0359	0.0073	0.0046
	3*CP	5	0.9524	0.9406	0.9299	0.9642	0.9746	0.9533	0.9962	1.0000	0.9966	0.9995	0.9737	0.9816	0.9757	0.9768	0.9260	0.9694
		10	0.9524	0.9465	0.9174	0.9753	0.9746	0.9490	0.9946	1.0000	0.9966	0.9997	0.9788	0.9866	0.9757	0.9802	0.9364	0.9711
		25	0.9524	0.9428	0.9317	0.9700	0.9746	0.9621	0.9796	1.0000	0.9966	0.9997	0.9857	0.9889	0.9757	0.9771	0.9449	0.9834
15*50	3*Estimate	5	0.9813	0.9690	0.9413	0.9647	-1.0293	-1.0301	-1.0046	-0.9914	0.2032	0.2245	0.2404	0.2472	0.1428	0.1287	0.1672	0.1646
		10	0.9813	0.9709	0.9321	0.9656	-1.0293	-1.0390	-0.9944	-0.9966	0.2032	0.2198	0.2368	0.2497	0.1428	0.1313	0.1707	0.1638
		25	0.9813	0.9762	0.9167	0.9569	-1.0293	-1.0407	-0.9964	-0.9790	0.2032	0.2142	0.2432	0.2615	0.1428	0.1286	0.1663	0.1637
	3*Bias	5	-0.0187	-0.0310	-0.0587	-0.0353	-0.0293	-0.0301	-0.0046	0.0086	0.0032	0.0245	0.0404	0.0472	-0.0572	-0.0713	-0.0328	-0.0354
		10	-0.0187	-0.0291	-0.0679	-0.0344	-0.0293	-0.0390	0.0056	0.0034	0.0032	0.0198	0.0368	0.0497	-0.0572	-0.0687	-0.0293	-0.0362
		25	-0.0187	-0.0238	-0.0833	-0.0431	-0.0293	-0.0407	0.0036	0.0210	0.0032	0.0142	0.0432	0.0615	-0.0572	-0.0714	-0.0337	-0.0363
	3*Variance	5	0.0552	0.0795	0.0236	0.0233	0.1126	0.1172	0.0642	0.0681	0.1534	0.2770	0.0134	0.0124	0.0406	0.1665	0.0051	0.0048
		10	0.0552	0.0726	0.0217	0.0235	0.1126	0.1280	0.0834	0.0656	0.1534	0.2786	0.0111	0.0129	0.0406	0.1857	0.0046	0.0049
		25	0.0552	0.0922	0.0223	0.0216	0.1126	0.1480	0.0805	0.0952	0.1534	0.3392	0.0116	0.0105	0.0406	0.2263	0.0048	0.0044
	3*MSE	5	0.0424	0.0554	0.0248	0.0096	0.1228	0.1234	0.0541	0.0177	0.0600	0.0820	0.0162	0.0069	0.0160	0.0272	0.0073	0.0032
		10	0.0424	0.0540	0.0285	0.0098	0.1228	0.1349	0.0614	0.0206	0.0600	0.0868	0.0170	0.0080	0.0160	0.0261	0.0077	0.0035
		25	0.0424	0.0601	0.0339	0.0135	0.1228	0.1579	0.0627	0.0273	0.0600	0.0849	0.0185	0.0104	0.0160	0.0282	0.0082	0.0042
	3*CP	5	0.9513	0.9566	0.9098	0.9755	0.9428	0.9459	0.9470	0.9918	0.9949	0.9979	0.8956	0.9816	0.9782	0.9750	0.8721	0.9663
		10	0.9513	0.9501	0.8900	0.9713	0.9428	0.9463	0.9760	0.9836	0.9949	0.9957	0.8492	0.9661	0.9782	0.9766	0.8514	0.9630
		25	0.9513	0.9516	0.8659	0.9615	0.9428	0.9456	0.9738	1.0000	0.9949	0.9987	0.8550	0.9412	0.9782	0.9805	0.8511	0.9466

15*100	3*Estimate	5	0.9864	0.9840	0.9311	0.9708	-1.0088	-1.0190	-0.9974	-0.9940	0.2052	0.2106	0.2423	0.2408	0.1538	0.1521	0.1692	0.1641
		10	0.9864	0.9814	0.9292	0.9694	-1.0088	-1.0100	-1.0020	-0.9928	0.2052	0.2097	0.2385	0.2441	0.1538	0.1455	0.1704	0.1635
		25	0.9864	0.9766	0.9238	0.9590	-1.0088	-1.0123	-1.0010	-0.9748	0.2052	0.2145	0.2450	0.2521	0.1538	0.1445	0.1698	0.1620
	3*Bias	5	-0.0136	-0.0160	-0.0689	-0.0292	-0.0088	-0.0190	0.0026	0.0060	0.0052	0.0106	0.0423	0.0408	-0.0462	-0.0479	-0.0308	-0.0359
		10	-0.0136	-0.0186	-0.0708	-0.0306	-0.0088	-0.0100	-0.0020	0.0072	0.0052	0.0097	0.0385	0.0441	-0.0462	-0.0545	-0.0296	-0.0365
		25	-0.0136	-0.0234	-0.0762	-0.0410	-0.0088	-0.0123	-0.0010	0.0252	0.0052	0.0145	0.0450	0.0521	-0.0462	-0.0555	-0.0302	-0.0380
	3*Variance	5	0.0259	0.0250	0.0120	0.0116	0.0535	0.0602	0.0305	0.0315	0.0577	0.0592	0.0069	0.0062	0.0119	0.0153	0.0026	0.0024
		10	0.0259	0.0385	0.0119	0.0117	0.0535	0.0572	0.0313	0.0310	0.0577	0.0994	0.0067	0.0063	0.0119	0.0537	0.0025	0.0024
		25	0.0259	0.0383	0.0115	0.0120	0.0535	0.0687	0.0351	0.0294	0.0577	0.1009	0.0062	0.0068	0.0119	0.0329	0.0024	0.0025
	3*MSE	5	0.0241	0.0231	0.0282	0.0063	0.0576	0.0620	0.0592	0.0129	0.0374	0.0394	0.0176	0.0051	0.0080	0.0085	0.0076	0.0028
		10	0.0241	0.0344	0.0293	0.0079	0.0576	0.0578	0.0633	0.0162	0.0374	0.0547	0.0182	0.0063	0.0080	0.0147	0.0079	0.0031
		25	0.0241	0.0325	0.0320	0.0129	0.0576	0.0677	0.0643	0.0218	0.0374	0.0535	0.0193	0.0081	0.0080	0.0134	0.0081	0.0038
	3*CP	5	0.9534	0.9578	0.8540	0.9689	0.9424	0.9506	0.7926	0.9796	0.9578	0.9570	0.7839	0.9624	0.9861	0.9825	0.7579	0.9495
		10	0.9534	0.9525	0.8526	0.9571	0.9424	0.9448	0.7813	0.9636	0.9578	0.9801	0.7655	0.9507	0.9861	0.9817	0.7476	0.9368
		25	0.9534	0.9553	0.8374	0.9416	0.9424	0.9537	0.8000	0.9416	0.9578	0.9831	0.7445	0.9299	0.9861	0.9809	0.7346	0.9118

Simulation results for continuous covariate shows that estimation for the regression parameters performs better under all 3 missingness mechanism compared to complete case analysis. Estimation under MAR and MNAR shows better performance compared to MCAR by showing lower variance and MSE. As sample size increases these trend of performance remain similar though estimation under complete case perform better which is not surprising. We observe some degree of overestimation for the over dispersion parameter and we assume over dispersion might be influenced by covariates and a mathematical model for the over dispersion might be necessary but ignored here due to complexity and the scope of the research. We also noticed that estimation for zero inflation parameter shows more bias under CC, MCAR, and MAR but as sample size increases the bias also reduces. In terms of coverage probability estimates of all parameters often seem some what liberal (empirical coverage is larger than the nominal coverage of 95%). Estimates and its properties remain meaningfully stable under all three percentage of missing, makes the estimation process robust. For discrete covariate, estimation performance remain more or less similar except we noticed that estimation under MCAR, MAR and MNAR performs much better for the parameter associated with the variable that have missing values compared to the parameters for the variables that do not have non missing values.

Simulation results for the zero-inflated over-dispersed count data regression model under continuous covariate and discrete covariate show the similar nature of the parameters, except we noticed that estimation under MCAR, MAR and MNAR performs much better for the parameter associated with the variable that have missing values compared to the parameters for the variables that do not have non missing values. We also noticed that discrete covariate has effect on over-dispersion and shows relatively stable results at least under complete case compared to the results under continuous covariate.

The above results are for data which come from a zero-inflated negative binomial $NB(\mu, c, \omega)$ distribution. In order to see whether similar properties of the estimates hold when over-dispersed data are generated from another distribution rather than the $NB(\mu, c)$ distribution. Such a distribution that has been used earlier by others (Lawless, 1987 and Paul and Banergee, 1998) is the log-normal (m, σ^2) mixture of the Poisson distribution with $m = \log(\mu) - \frac{1}{2}\log(c+1)$ and $\sigma^2 = \log(c+1)$, where μ and c are the parameters of the $NB(\mu, c)$. In the situation in which there are covariates we take $\mu_i = \exp(\sum_{j=1}^{p} X_{ij}\beta_j)$. For more details of generating data from the log-normal mixture of the Poisson distribution see Lawless (1987).

The parameter values used to simulate data from the zero-inflated log-normal mixture of the Poisson distribution were the same as those used to generate data from the zero-inflated negative binomial distribution. We also used the same percentages of missing data as those in the previous case.

Results of the simulation study of the zero-inflated log-normal mixture of the Poisson distributed data are given in Table 3 and Table 4. Fortunately, we arrived at very similar conclusions of the results given in Table 1 and Table 2. This shows, perhaps, that the results will remain similar irrespective of the mechanism in which over-dispersed count data are generated.

Table 3. Properties (estimate, bias, variance, mse, coverage probability (cp)) of the estimates of the parameters, data simulated from lognormal mixture of Poisson ($\beta_1,\ \beta_2,\ c,\ \omega$), based on 5000 simulation runs (continuous covariate)

2*n	2*	2*% missing	Missingness Mechanism															
			CC	MCAR	MAR	MNAR	CC	MCAR	MAR	MNAR	CC	MCAR	MAR	MNAR	CC	MCAR	MAR	MNAR
			$\beta_1 = 1$				$\beta_2 = -1$				$c = 0.2$				$\omega = 0.2$			
15*30	3*Estimate	5	1.1208	0.9607	0.9282	0.9625	-1.0893	-0.9557	-0.9927	-1.0041	0.0204	0.2307	0.2417	0.2579	0.1404	0.1055	0.1721	0.1571
		10	1.1208	0.9900	0.9265	0.9628	-1.0893	-1.0489	-1.0021	-1.0025	0.0204	0.2293	0.2408	0.2568	0.1404	0.1098	0.1742	0.1603
		25	1.1208	1.0029	0.9250	0.9590	-1.0893	-1.0481	-0.9989	-1.0008	0.0204	0.2159	0.2396	0.2567	0.1404	0.1070	0.1701	0.1569
	3*Bias	5	0.1208	-0.0393	-0.0718	-0.0375	-0.0893	0.0443	0.0073	-0.0041	-0.1796	0.0307	0.0417	0.0579	-0.0596	-0.0945	-0.0279	-0.0429
		10	0.1208	-0.0100	-0.0735	-0.0372	-0.0893	-0.0489	-0.0021	-0.0025	-0.1796	0.0293	0.0408	0.0568	-0.0596	-0.0902	-0.0258	-0.0397
		25	0.1208	0.0029	-0.0750	-0.0410	-0.0893	-0.0481	0.0011	-0.0008	-0.1796	0.0159	0.0396	0.0567	-0.0596	-0.0930	-0.0299	-0.0431
	3*Variance	5	0.4405	1.4921	0.0355	0.0354	4.2194	15.2630	0.3771	0.4019	0.0890	0.5477	0.0216	0.0211	0.0263	0.4045	0.0078	0.0077
		10	0.4405	1.0679	0.0349	0.0357	4.2194	11.4875	0.4031	0.3428	0.0890	0.5657	0.0214	0.0216	0.0263	0.4113	0.0080	0.0079
		25	0.4405	1.4024	0.0351	0.0352	4.2194	13.5096	0.3746	0.4116	0.0890	0.6872	0.0214	0.0206	0.0263	0.5278	0.0080	0.0076
	3*MSE	5	0.4627	0.5955	0.0297	0.0120	4.1845	5.8639	0.0652	0.0289	0.0365	0.1077	0.0186	0.0106	0.0165	0.0450	0.0082	0.0049
		10	0.4627	0.4733	0.0295	0.0119	4.1845	5.2109	0.0625	0.0251	0.0365	0.1076	0.0181	0.0097	0.0165	0.0420	0.0076	0.0044
		25	0.4627	0.5870	0.0299	0.0138	4.1845	5.5613	0.0606	0.0290	0.0365	0.1091	0.0176	0.0106	0.0165	0.0480	0.0077	0.0051
	3*CP	5	0.9231	0.9949	0.9038	0.9743	0.9231	0.9966	1.0000	1.0000	0.9960	0.9997	0.9729	0.9861	0.9838	0.9878	0.9154	0.9690
		10	0.9231	0.9899	0.9082	0.9710	0.9231	0.9936	1.0000	1.0000	0.9960	0.9997	0.9721	0.9936	0.9838	0.9850	0.9306	0.9817
		25	0.9231	0.9918	0.9079	0.9626	0.9231	0.9961	1.0000	1.0000	0.9960	1.0000	0.9722	0.9861	0.9838	0.9846	0.9297	0.9669
15*50	3*Estimate	5	1.0674	0.9699	0.9192	0.9686	-0.9804	-1.0316	-1.0004	-0.9975	0.0496	0.2346	0.2388	0.2501	0.1603	0.1179	0.1688	0.1595
		10	1.0674	0.9545	0.9208	0.9682	-0.9804	-0.9726	-1.0013	-1.0011	0.0496	0.2400	0.2428	0.2494	0.1603	0.1186	0.1719	0.1596
		25	1.0674	0.9568	0.9289	0.9600	-0.9804	-0.9954	-0.9993	-0.9933	0.0496	0.2355	0.2421	0.2505	0.1603	0.1131	0.1684	0.1589
	3*Bias	5	0.0674	-0.0301	-0.0808	-0.0314	0.0196	-0.0316	-0.0004	0.0025	-0.1504	0.0346	0.0388	0.0501	-0.0397	-0.0821	-0.0312	-0.0405
		10	0.0674	-0.0455	-0.0792	-0.0318	0.0196	0.0274	-0.0013	-0.0011	-0.1504	0.0400	0.0428	0.0494	-0.0397	-0.0814	-0.0281	-0.0404
		25	0.0674	-0.0432	-0.0711	-0.0400	0.0196	0.0046	0.0007	0.0067	-0.1504	0.0355	0.0421	0.0505	-0.0397	-0.0869	-0.0316	-0.0411
	3*Variance	5	0.3500	0.7714	0.0208	0.0209	3.9758	8.2864	0.2390	0.2397	0.0486	0.2834	0.0125	0.0119	0.0148	0.2386	0.0046	0.0043
		10	0.3500	0.8912	0.0212	0.0210	3.9758	9.6218	0.2128	0.2205	0.0486	0.3065	0.0130	0.0121	0.0148	0.2689	0.0048	0.0044
		25	0.3500	0.9124	0.0207	0.0211	3.9758	8.6431	0.2520	0.2294	0.0486	0.4236	0.0122	0.0124	0.0148	0.3714	0.0046	0.0044
	3*MSE	5	0.3384	0.4526	0.0330	0.0097	3.7474	4.8559	0.0618	0.0225	0.0299	0.0937	0.0180	0.0088	0.0102	0.0349	0.0079	0.0041
		10	0.3384	0.4853	0.0327	0.0102	3.7474	5.1041	0.0623	0.0229	0.0299	0.0921	0.0180	0.0084	0.0102	0.0341	0.0077	0.0041
		25	0.3384	0.5215	0.0299	0.0133	3.7474	4.9612	0.0626	0.0247	0.0299	0.1135	0.0182	0.0091	0.0102	0.0434	0.0080	0.0044
	3*CP	5	0.9244	0.9853	0.8712	0.9685	0.9276	0.9858	1.0000	1.0000	0.9947	0.9995	0.8826	0.9548	0.9808	0.9874	0.8534	0.9474
		10	0.9244	0.9894	0.8678	0.9676	0.9276	0.9909	1.0000	1.0000	0.9947	0.9995	0.8959	0.9643	0.9808	0.9904	0.8770	0.9427
		25	0.9244	0.9852	0.8840	0.9546	0.9276	0.9870	1.0000	1.0000	0.9947	0.9998	0.8779	0.9599	0.9808	0.9868	0.8604	0.9409
15*100	3*Estimate	5	0.9700	0.9552	0.9289	0.9717	-1.0101	-0.9784	-0.9963	-0.9972	0.2219	0.2305	0.2425	0.2388	0.1393	0.1348	0.1707	0.1614
		10	0.9700	0.9702	0.9347	0.9708	-1.0101	-1.0137	-1.0014	-0.9959	0.2219	0.2299	0.2329	0.2416	0.1393	0.1374	0.1698	0.1586
		25	0.9700	0.9794	0.9266	0.9693	-1.0101	-1.0547	-1.0000	-0.9931	0.2219	0.2273	0.2439	0.2420	0.1393	0.1331	0.1674	0.1587
	3*Bias	5	-0.0300	-0.0448	-0.0711	-0.0283	-0.0101	0.0216	0.0037	0.0028	0.0219	0.0305	0.0425	0.0388	-0.0607	-0.0652	-0.0293	-0.0386
		10	-0.0300	-0.0298	-0.0653	-0.0292	-0.0101	-0.0137	-0.0014	0.0041	0.0219	0.0299	0.0329	0.0416	-0.0607	-0.0626	-0.0302	-0.0414
		25	-0.0300	-0.0206	-0.0734	-0.0307	-0.0101	-0.0547	0.0000	0.0069	0.0219	0.0273	0.0439	0.0420	-0.0607	-0.0669	-0.0326	-0.0413
	3*Variance	5	0.2857	0.3575	0.0104	0.0104	3.0139	3.7771	0.1135	0.1100	0.0667	0.0916	0.0062	0.0058	0.0193	0.0768	0.0023	0.0021
		10	0.2857	0.4525	0.0103	0.0104	3.0139	4.7370	0.1079	0.1071	0.0667	0.0935	0.0061	0.0059	0.0193	0.0426	0.0022	0.0021

		25	0.2857	0.3862	0.0104	0.0103	3.0139	4.1403	0.1167	0.1182	0.0667	0.1148	0.0061	0.0057	0.0193	0.0828	0.0023	0.0021
	3*MSE	5	0.2650	0.3029	0.0299	0.0079	2.8223	3.1692	0.0639	0.0156	0.0438	0.0583	0.0186	0.0059	0.0123	0.0187	0.0080	0.0032
		10	0.2650	0.3610	0.0256	0.0081	2.8223	3.7645	0.0560	0.0171	0.0438	0.0571	0.0157	0.0065	0.0123	0.0161	0.0072	0.0036
		25	0.2650	0.3232	0.0297	0.0085	2.8223	3.4626	0.0632	0.0187	0.0438	0.0639	0.0183	0.0067	0.0123	0.0209	0.0078	0.0037
	3*CP	5	0.9498	0.9550	0.8415	0.9620	0.9513	0.9557	0.9913	1.0000	0.9774	0.9818	0.7584	0.9451	0.9899	0.9910	0.7345	0.9335
		10	0.9498	0.9652	0.8572	0.9546	0.9513	0.9633	0.9890	1.0000	0.9774	0.9880	0.7957	0.9329	0.9899	0.9910	0.7617	0.9154
		25	0.9498	0.9576	0.8352	0.9524	0.9513	0.9578	0.9940	1.0000	0.9774	0.9896	0.7709	0.9302	0.9899	0.9902	0.7310	0.9080

Table 4. Properties (estimate, bias, variance, mse, coverage probability (cp)) of the estimates of the parameters, data simulated from lognormal mixture of Poisson ($\beta_1,\ \beta_2,\ c,\ \omega$), based on 5000 simulation runs (discrete covariate)

2*n	2*	2*% missing	Missingness Mechanism															
			CC	MCAR	MAR	MNAR	CC	MCAR	MAR	MNAR	CC	MCAR	MAR	MNAR	CC	MCAR	MAR	MNAR
			$\beta_1 = 1$				$\beta_2 = -1$				$c = 0.2$				$\omega = 0.2$			
15*30	3*Estimate	5	0.9784	0.9625	0.938	0.9564	-1.0929	-1.0492	-1.0006	-0.9902	0.2067	0.2366	0.2337	0.2638	0.1310	0.1048	0.1655	0.1636
		10	0.9784	0.9641	0.941	0.9591	-1.0929	-1.0381	-1.0045	-1.0018	0.2067	0.2354	0.2361	0.2615	0.1310	0.1059	0.1683	0.1588
		25	0.9784	0.9667	0.9314	0.9543	-1.0929	-1.046	-1.0039	-0.9759	0.2067	0.2242	0.2415	0.2656	0.1310	0.1068	0.1713	0.1599
	3*Bias	5	-0.0216	-0.0375	-0.062	-0.0436	-0.0929	-0.0492	-0.0006	0.0098	0.0067	0.0366	0.0337	0.0638	-0.069	-0.0952	-0.0345	-0.0364
		10	-0.0216	-0.0359	-0.059	-0.0409	-0.0929	-0.0381	-0.0045	-0.0018	0.0067	0.0354	0.0361	0.0615	-0.069	-0.0941	-0.0317	-0.0412
		25	-0.0216	-0.0333	-0.0686	-0.0457	-0.0929	-0.0460	-0.0039	0.0241	0.0067	0.0242	0.0415	0.0656	-0.069	-0.0932	-0.0287	-0.0401
	3*Variance	5	0.0864	0.1936	0.0382	0.0374	0.2345	0.2050	0.1162	0.1455	0.2625	0.7212	0.0216	0.0200	0.0745	0.5529	0.0087	0.0079
		10	0.0864	0.2250	0.0445	0.0379	0.2345	0.2205	0.0835	0.1344	0.2625	0.8407	0.0330	0.0208	0.0745	0.5389	0.0104	0.0082
		25	0.0864	0.2753	0.0387	0.0398	0.2345	0.2704	0.1176	0.1076	0.2625	0.9482	0.0225	0.0242	0.0745	0.6088	0.0088	0.0090
	3*MSE	5	0.0550	0.1011	0.0265	0.0134	0.2640	0.2196	0.0504	0.0296	0.0706	0.1182	0.0150	0.0116	0.0213	0.0516	0.0066	0.0047
		10	0.0550	0.1205	0.0237	0.0122	0.2640	0.2306	0.0545	0.0289	0.0706	0.1232	0.0152	0.0110	0.0213	0.0494	0.0069	0.0047
		25	0.0550	0.1360	0.0273	0.0150	0.2640	0.2751	0.0607	0.0314	0.0706	0.1103	0.0176	0.0123	0.0213	0.0459	0.0077	0.0051
	3*CP	5	0.9606	0.9530	0.9155	0.9694	0.9546	0.9509	0.9918	1.0000	0.9988	1.0000	0.9715	0.9785	0.9760	0.9811	0.9270	0.9704
		10	0.9606	0.9484	0.9458	0.9758	0.9546	0.9373	0.9792	1.0000	0.9988	0.9994	0.9880	0.9800	0.9760	0.9815	0.9414	0.9758
		25	0.9606	0.9477	0.9248	0.9625	0.9546	0.9517	0.9902	1.0000	0.9988	1.0000	0.9716	0.9818	0.9760	0.9790	0.9237	0.9700
15*50	3*Estimate	5	0.9621	0.9639	0.9425	0.9603	-1.0349	-1.0396	-1.0031	-0.9949	0.2206	0.2229	0.2359	0.2550	0.1315	0.1283	0.1688	0.1616
		10	0.9621	0.9582	0.9283	0.9626	-1.0349	-1.0196	-1.0094	-0.9853	0.2206	0.2372	0.2390	0.2541	0.1315	0.1174	0.1692	0.1606
		25	0.9621	0.9523	0.9266	0.9610	-1.0349	-0.9994	-1.0198	-0.9834	0.2206	0.2394	0.2381	0.2564	0.1315	0.1099	0.1696	0.1620
	3*Bias	5	-0.0379	-0.0361	-0.0575	-0.0397	-0.0349	-0.0396	-0.0031	0.0051	0.0206	0.0229	0.0359	0.0550	-0.0685	-0.0717	-0.0312	-0.0384
		10	-0.0379	-0.0418	-0.0717	-0.0374	-0.0349	-0.0196	-0.0094	0.0147	0.0206	0.0372	0.0390	0.0541	-0.0685	-0.0826	-0.0308	-0.0394
		25	-0.0379	-0.0477	-0.0734	-0.0390	-0.0349	0.0006	-0.0198	0.0166	0.0206	0.0394	0.0381	0.0564	-0.0685	-0.0901	-0.0304	-0.0380
	3*Variance	5	0.0583	0.0573	0.0244	0.0238	0.1139	0.1343	0.0577	0.0630	0.1668	0.2119	0.0146	0.0135	0.0454	0.1481	0.0054	0.0051
		10	0.0583	0.0960	0.0226	0.0243	0.1139	0.1199	0.0731	0.0584	0.1668	0.3579	0.0123	0.0143	0.0454	0.2760	0.0049	0.0053
		25	0.0583	0.2098	0.0240	0.0229	0.1139	0.1737	0.0634	0.0713	0.1668	0.7453	0.0140	0.0122	0.0454	0.5505	0.0053	0.0048
	3*MSE	5	0.0446	0.0443	0.0230	0.0117	0.1253	0.1473	0.0540	0.0234	0.0657	0.0869	0.0152	0.0091	0.0182	0.0255	0.0069	0.0039
		10	0.0446	0.0627	0.0292	0.0110	0.1253	0.1247	0.0588	0.0237	0.0657	0.0998	0.0172	0.0091	0.0182	0.0343	0.0076	0.0042
		25	0.0446	0.1212	0.0300	0.0118	0.1253	0.1777	0.0635	0.0262	0.0657	0.1230	0.0180	0.0095	0.0182	0.0464	0.0079	0.0042
	3*CP	5	0.9588	0.9608	0.9216	0.9634	0.9423	0.9441	0.9254	0.9791	0.9958	0.9954	0.9205	0.9634	0.9782	0.9810	0.8832	0.9582
		10	0.9588	0.9545	0.8918	0.9711	0.9423	0.9488	0.961	0.9786	0.9958	0.9982	0.8797	0.9775	0.9782	0.9785	0.8720	0.9539
		25	0.9588	0.9480	0.8914	0.9639	0.9423	0.9416	0.9392	0.9873	0.9958	1.0000	0.8990	0.9470	0.9782	0.9788	0.8609	0.9470

15*100	3*Estimate	5	0.9724	0.9710	0.9317	0.9738	-1.0143	-1.0033	-1.0019	-0.9987	0.2140	0.2179	0.2385	0.2404	0.1459	0.1435	0.1666	0.1640
		10	0.9724	0.9697	0.9323	0.9711	-1.0143	-1.0119	-1.0018	-0.9898	0.2140	0.2186	0.2421	0.2403	0.1459	0.1398	0.1681	0.1648
		25	0.9724	0.9701	0.9204	0.9657	-1.0143	-1.0219	-1.0003	-0.9757	0.2140	0.2207	0.2436	0.2487	0.1459	0.1361	0.1691	0.1649
	3*Bias	5	-0.0276	-0.0290	-0.0683	-0.0262	-0.0143	-0.0033	-0.0019	0.0013	0.0140	0.0179	0.0385	0.0404	-0.0541	-0.0565	-0.0334	-0.0360
		10	-0.0276	-0.0303	-0.0677	-0.0289	-0.0143	-0.0119	-0.0018	0.0102	0.0140	0.0186	0.0421	0.0403	-0.0541	-0.0602	-0.0319	-0.0352
		25	-0.0276	-0.0299	-0.0796	-0.0343	-0.0143	-0.0219	-0.0003	0.0243	0.0140	0.0207	0.0436	0.0487	-0.0541	-0.0639	-0.0309	-0.0351
	3*Variance	5	0.0225	0.0290	0.0120	0.0114	0.0590	0.0562	0.0294	0.0334	0.0509	0.0722	0.0071	0.0059	0.0104	0.0198	0.0026	0.0023
		10	0.0225	0.0342	0.0108	0.0107	0.0590	0.0581	0.0427	0.0407	0.0509	0.0927	0.0054	0.0051	0.0104	0.0522	0.0022	0.0021
		25	0.0225	0.0392	0.0118	0.0111	0.0590	0.0684	0.0327	0.0374	0.0509	0.1083	0.0066	0.0057	0.0104	0.0578	0.0025	0.0022
	3*MSE	5	0.0219	0.0264	0.0281	0.0066	0.0630	0.0580	0.0601	0.0146	0.0371	0.0472	0.0173	0.0056	0.0088	0.0110	0.0079	0.0029
		10	0.0219	0.0314	0.0276	0.0073	0.0630	0.0614	0.0634	0.0138	0.0371	0.0534	0.0182	0.0052	0.0088	0.0157	0.0080	0.0027
		25	0.0219	0.0339	0.0333	0.0098	0.0630	0.0711	0.0633	0.0187	0.0371	0.059	0.0186	0.0070	0.0088	0.0165	0.0080	0.0032
	3*CP	5	0.9603	0.9604	0.8540	0.9692	0.9413	0.9479	0.7862	0.9703	0.9451	0.9688	0.7851	0.9480	0.9766	0.9823	0.7605	0.9374
		10	0.9603	0.9626	0.8530	0.9629	0.9413	0.9485	0.8464	0.9829	0.9451	0.9735	0.7489	0.9558	0.9766	0.9808	0.7280	0.9357
		25	0.9603	0.9590	0.8351	0.9569	0.9413	0.9470	0.7910	0.9621	0.9451	0.9852	0.7610	0.9348	0.9766	0.9823	0.7518	0.9222

Table 5. Estimates and Standard Errors of the parameters for DMFT data with covariates

Percentage missingness		$\hat{\beta}$	$SE(\hat{\beta})$	$\hat{\beta}_G$	$SE(\hat{\beta}_G)$	$\hat{\beta}_{E(1)}$	$SE(\hat{\beta}_{E(1)})$	$\hat{\beta}_{E(2)}$	$SE(\hat{\beta}_{E(2)})$
Complete data	0%	0.3863	0.1234	0.0487	0.0517	0.2884	0.0837	0.2407	0.0858
3*MCAR	5%	-0.2509	0.2168	0.2202	0.0685	0.8052	0.1382	0.6646	0.1244
	10%	-0.0608	0.1885	0.2257	0.0686	0.7136	0.1259	0.5549	0.1160
	25%	0.3753	0.1538	0.1736	0.0604	0.3478	0.0992	0.3131	0.1010
3*MAR	5%	0.4447	0.2354	0.0864	0.0654	0.4333	0.1696	0.3555	0.1315
	10%	0.1176	0.1658	0.1673	0.0613	0.5340	0.1071	0.4189	0.1030
	25%	0.3074	0.2274	0.1087	0.0804	0.5196	0.1335	0.4373	0.1213
3*MNAR	5%	0.4674	0.1897	0.1590	0.0607	0.4322	0.1279	0.3439	0.1147
	10%	0.3146	0.2106	0.1374	0.0594	0.5256	0.1288	0.5010	0.1226
	25%	0.2519	0.2636	0.1323	0.0682	0.4014	0.1494	0.3343	0.1360
Percentage missingness		$\hat{\beta}_{S(1)}$	$SE(\hat{\beta}_{S(1)})$	$\hat{\beta}_{S(2)}$	$SE(\hat{\beta}_{S(2)})$	$\hat{\beta}_{S(3)}$	$SE(\hat{\beta}_{S(3)})$	$\hat{\beta}_{S(4)}$	$SE(\hat{\beta}_{S(4)})$
Complete data	0%	0.8927	0.1148	0.7948	0.1040	0.9724	0.1126	0.9187	0.1056
3*MCAR	5%	0.9654	0.1505	0.9961	0.1540	0.8989	0.1396	1.0009	0.1405
	10%	0.7937	0.1387	0.9279	0.1421	0.7768	0.1249	0.8583	0.1326
	25%	0.7715	0.1429	0.6760	0.1207	0.7764	0.1291	0.7724	0.1279
3*MAR	5%	0.5412	0.1951	0.6655	0.1620	0.6927	0.2359	0.6524	0.1871
	10%	0.8640	0.1384	0.8504	0.1291	0.8382	0.1246	0.9136	0.1271
	25%	0.5348	0.1503	0.7502	0.1849	0.6413	0.1559	0.6872	0.1470
3*MNAR	5%	0.4702	0.1413	0.5773	0.1427	0.5816	0.1461	0.6319	0.1479
	10%	0.5449	0.1576	0.7189	0.1579	0.5736	0.1416	0.6782	0.1644
	25%	0.6513	0.1565	0.8826	0.2017	0.8134	0.2037	0.8215	0.1515

Percentage missingness		$\hat{\beta}_{S(5)}$	$SE(\hat{\beta}_{S(5)})$	$\hat{c}$	$SE(\hat{c})$	$\hat{\omega}$	$SE(\hat{\omega})$	$\hat{E}(y)$	$\hat{V}ar(y)$
Complete data	0%	0.8889	0.1059	0.1327	0.0351	0.1760	0.0159	3.4882	8.0463
3*MCAR	5%	1.0450	0.1483	0.2551	0.0860	0.1543	0.0227	3.5015	9.4302
	10%	0.9470	0.1359	0.2146	0.0660	0.1567	0.0203	3.4889	8.8842
	25%	0.8257	0.1297	0.1484	0.0492	0.1897	0.0195	3.4618	8.5149
3*MAR	5%	0.6490	0.1301	0.2130	0.1512	0.1519	0.0315	3.5511	8.9869
	10%	0.9565	0.1328	0.2041	0.0648	0.1430	0.0183	3.6047	8.9114
	25%	0.7537	0.1810	0.2148	0.1287	0.1470	0.0309	3.5762	8.5042
3*MNAR	5%	0.5989	0.1389	0.2132	0.0912	0.1479	0.0228	3.5487	9.0101
	10%	0.6635	0.1591	0.2202	0.0986	0.1644	0.0244	3.4658	9.1306
	25%	0.9845	0.2294	0.2471	0.1457	0.1515	0.0328	3.6101	9.0244

4. An Illustrative Example

We now analyze a set of data from a prospective study of dental status of school children from Bohning et al. (1999). The children were all 7 years of age at the beginning of the study. Dental status were measured by the decayed, missing and filled teeth (DMFT) index. Only the eight deciduous molars were considered so the smallest possible value of the DMFT index is 0 and the largest is 8. The prospective study was for a period of two years. The DMFT index was obtained at the beginning of the study and also at the end of the study.

The data also involved 3 categorical covariates: gender having two categories (0 - female, 1 - male), ethnic group having three categories (1 - dark, 2 - white, 3 - black) and school having six categories (1 - oral health education, 2 - all four methods together, 3 - control school (no prevention measure), 4 - enrichment of the school diet with ricebran, 5 - mouthrinse with 0.2% NaF-solution, 6 - oral hygiene).

For the purpose illustration of our method we deal with the DMFT index data obtained at the beginning of the study (as in Deng and Paul, 2005). The DMFT index data at the beginning of the study are: (index, frequency): (0,172), (1,73), (2,96), (3,80), (4,95), (5,83), (6,85), (7,65), (8,48). We then fitted a zero-inflated negative binomial model to the complete data and data with missing observations in covariate. To obtain data with missing observations in covariate we randomly deleted a certain percentage (5%, 10%, 25%) of the observed covariate gender. The model fitted was $\mu = \exp(\beta + \beta_G I(Gender = 1) + \beta_{E(1)} I(Ethnic = 1) + \beta_{E(2)} I(Ethnic = 2) + \beta_{S(1)} I(School = 1) + \beta_{S(2)} I(School = 2) + \beta_{S(3)} I(School = 3) + \beta_{S(4)} I(School = 4) + \beta_{S(5)} I(School = 5))$, where β represents the intercept parameter and β_G represents the regression parameter for gender, $\beta_{E(1)}$ and $\beta_{E(2)}$ represent the regression parameters for the ethnic groups 1 and 2, and $\beta_{S(1)}$, $\beta_{S(2)}$, $\beta_{S(3)}$, $\beta_{S(4)}$, and $\beta_{S(5)}$ represent the regression parameters for school 1, school 2, school 3, school 4, and school 5 respectively.

The estimates of the mean parameter μ, where $\mu_i = \exp(\sum_{j=1}^{p} X_{ij}\beta_j)$ the over dispersion parameter c and the zero inflation parameter ω based on the zero-inflated negative binomial model, under different percentages of missingness, and their corresponding standard errors are presented in Table 5. It is to note that the estimates of the parameters μ, c and ω and the corresponding standard errors changes with the amount of missingness in the covariate (this is expected as it depends on which observations have remained in the final data set). In general, the standard errors of the estimates are larger than those under complete data. However, estimates of $\hat{E}(Y)$ do not vary much irrespective of the percentage missing and the missing data mechanism. The same comment applies to $\hat{V}ar(Y)$, although for $\hat{V}ar(Y)$ is a bit higher under MNAR.

5. Discussion

We have developed estimation procedure for the parameters of a zero inflated negative binomial model in presence of missing observations in covariate. We applied a weighted expectation- maximization algorithm (Ibrahim, 1990) for the maximum likelihood estimation of the parameters. Although missing data methodologies have been developed extensively in literature, the current development for the estimation of the parameters of a zero inflated negative binomial model in presence of missing covariate is new.

The overall message of the simulation study is that estimation procedure performs in a stabilized fashion irrespective of the missingness mechanism and percentage of missing observation though the performance is relatively noticeable for the smaller sample size.

6. Appendix

6.1. Appendix A: Estimating Equations for the Parameters $\boldsymbol{\beta_j}$, c and $\boldsymbol{\gamma}$ in the Model (2.3)

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \left[\left[\frac{-(1+\mu c)^{-1}\exp\left[(-c^{-1}\log(1+\mu c))\right]}{\gamma+\exp\left[(-c^{-1}\log(1+\mu c))\right]}\right] I_{\{y_i=0\}} + \left[\frac{y_1}{\mu} - \frac{c(y_1+c^{-1})}{1+\mu c}\right] I_{\{y_i>0\}}\right] \frac{\partial \mu_i}{\partial \beta_j} = 0, \tag{20}$$

where $\frac{\partial \mu_i}{\partial \beta_j} = X_{ij}\exp(\sum_{j=1}^{p} X_{ij}\beta_j)$,

$$\begin{aligned}\frac{\partial l}{\partial c} = \sum_{i=1}^{n} \Bigg[& \frac{\left[-\mu c^{-1}(1+\mu c)^{-1} + c^{-2}\log(1+\mu c)\right]\exp\left[(-c^{-1}\log(1+\mu c))\right]}{\gamma+\exp\left[(-c^{-1}\log(1+\mu c))\right]} I_{\{y_i=0\}} \\ & + \left[\mu(y_i + c^{-1})(1+\mu c)^{-1} - c^{-2}\log(1+\mu c) + \sum_{l=1}^{y_i}(l-1)\right] I_{\{y_i>0\}}\Bigg] = 0,\end{aligned} \tag{21}$$

and

$$\frac{\partial l}{\partial \gamma} = \sum_{i=1}^{n} \left[-(1+\gamma)^{-1} + \left[\gamma + \exp[(-c^{-1}\log(1+\mu c))]\right]^{-1} I_{\{y_i=0\}} + 0 I_{\{y_i>0\}}\right] = 0. \tag{22}$$

6.2. Appendix B: The Gibb Sampler

To use the Gibbs sampler we need to generate each sample point of $a_{i1}, a_{i2}, \ldots, a_{im_i}$ by using Gibbs sequence. For example Gibbs sequence for a_{i1} is

$$\begin{aligned}
&a_{i1}^{(1)} \sim P(y_i^{(0)} | x_i^{(0)}, \psi^{(0)}) \\
&a_{i1}^{(2)} \sim P(y_i^{(0)} | x_i^{(0)}, \psi^{(0)}, a_{i1}^{(1)}) \\
&a_{i1}^{(3)} \sim P(y_i^{(0)} | x_i^{(0)}, \psi^{(0)}, a_{i1}^{(1)}, a_{i1}^{(2)}) \\
&a_{i1}^{(4)} \sim P(y_i^{(0)} | x_i^{(0)}, \psi^{(0)}, a_{i1}^{(1)}, a_{i1}^{(2)}, a_{i1}^{(3)}) \\
&\cdots \\
&a_{i1}^{(k)} \sim P(y_i^{(0)} | x_i^{(0)}, \psi^{(0)}, a_{i1}^{(1)}, a_{i1}^{(2)}, a_{i1}^{(3)}, \cdots, a_{i1}^{(k-1)}).
\end{aligned}$$

For large K, $a_{i1}^{(k)} = a_{i1}$. According to Sahu and Roberts (1999) $a_{i1}, a_{i2}, \ldots, a_{im_i}$ can be considered as a block and can obtained from $P(y_i^{(0)} | x_i^{(0)}, \psi^{(0)})$. In this scenario, for each missing response, samples are considered as a block. For example if there are 5 missing response, then there are 5 blocks. Sahu and Roberts (1999) also mentioned that most practical cases, missing observations are independent of parameters and considers as a single block. In this case, 5 missing observations can be treated as a single block. In our model, missing responses are independent of parameters and hence we follow Sahu and Roberts (1999) for Gibbs sampling. We stop the sequence and obtain the required sample for which the absolute deviation of parameters between two consecutive steps become minimal. Extensive explanation of Gibbs sampler are available in Casella and George (1992) and Sahu and Roberts (1999).

Appendix C: Elements of the Observed Information Matrix

From equation (14) we have

$$Q(\psi|\psi^{(s)}) = \sum_{i=1}^{A} l(\psi^{(s)}|y_i, x_i) + \sum_{i=1}^{B} \frac{1}{m_i}\sum_{k=1}^{m_i} l(\psi^{(s)}|y_i, x_{o,i}, a_{ik}). \tag{23}$$

Maximizing $Q(\psi|\psi^{(s)})$ is analogous to maximization of complete data log likelihood, $l(\beta, c, \gamma|y_i)$ in (3) where each incomplete response being replaced by m_i weighted observations. The elements of the observed information matrix are as given below.

$$\begin{aligned}\frac{\partial^2 l}{\partial \beta_j^2} = \sum_{i=1}^{n} \Bigg[& \frac{\frac{1}{(1+\mu c)^2}\exp(-c^{-1}\log(1+\mu c))(\gamma + c(\gamma+\exp(-c^{-1}\log(1+\mu c))))}{[\gamma+\exp(-c^{-1}\log(1+\mu c))]^2}\left[\frac{\partial \mu_i}{\partial \beta_j}\right]^2 I_{\{y_i=0\}} \\ & + \left[\frac{-y_i}{\mu^2} + \frac{c^2(y_i+c^{-1})}{(1+\mu c)^2}\right]\left[\frac{\partial \mu_i}{\partial \beta_j}\right]^2 I_{\{y_i>0\}} \\ & + \left[\frac{-(1+\mu c)^{-1}\exp\left[(-c^{-1}\log(1+\mu c))\right]}{\gamma+\exp\left[(-c^{-1}\log(1+\mu c))\right]} I_{\{y_i=0\}}\right. \\ & \left. + \left[\frac{y_1}{\mu} - \frac{c(y_1+c^{-1})}{1+\mu c}\right] I_{\{y_i>0\}}\right] \frac{\partial^2 \mu_i}{\partial \beta_j^2}\Bigg]\end{aligned} \tag{24}$$

$$\frac{\partial^2 l}{\partial \beta_j \beta_j'} = \sum_{i=1}^{n} \left[\frac{\frac{1}{(1+\mu c)^2}\exp(-c^{-1}\log(1+\mu c))(\gamma + c(\gamma + \exp(-c^{-1}\log(1+\mu c))))}{[\gamma + \exp(-c^{-1}\log(1+\mu c))]^2}\right]\left[\frac{\partial \mu_i}{\partial \beta_j}\frac{\partial \mu_i}{\partial \beta_j'}\right] I_{\{y_i=0\}}$$
$$+\left[\frac{-y_i}{\mu^2} + \frac{c^2(y_i + c^{-1})}{(1+\mu c)^2}\right]\left[\frac{\partial \mu_i}{\partial \beta_j}\frac{\partial \mu_i}{\partial \beta_j'}\right] I_{\{y_i>0\}}$$
$$+\left[\frac{-(1+\mu c)^{-1}\exp[(-c^{-1}\log(1+\mu c))]}{\gamma + \exp[(-c^{-1}\log(1+\mu c))]} I_{\{y_i=0\}}\right.$$
$$\left.+\left[\frac{y_1}{\mu} - \frac{c(y_1 + c^{-1})}{1+\mu c}\right] I_{\{y_i>0\}}\right]\frac{\partial^2 \mu_i}{\partial \beta_j \beta_j'} \tag{25}$$

$$\frac{\partial^2 l}{\partial \beta_j \partial c} = \sum_{i=1}^{n} \left[\left(\frac{1}{1+\mu c}\right)\exp(-c^{-1}\log(1+\mu c))\right][\gamma + \exp(-c^{-1}\log(1+\mu c))]$$
$$\left[c^{-1}\left(\frac{\mu}{1+\mu c}\right) - c^{-2}\log(1+\mu c) + \frac{\mu}{1+\mu c}\right]$$
$$+[\exp(-c^{-1}\log(1+\mu c))]\left[-c^{-1}\left(\frac{\mu}{1+\mu c}\right) + c^{-2}\log(1+\mu c)\right]]$$
$$[[\gamma + \exp(-c^{-1}\log(1+\mu c))]^2]^{-1}\left[\frac{\partial \mu_i}{\partial \beta_j}\right] I_{\{y_i=0\}}$$
$$+\frac{-(1+\mu c)[c(y_i - c^{-2}) + (y_i + c^{-1})] - \mu c(y_i + c^{-1})}{(1+\mu c)^2}\left[\frac{\partial \mu_i}{\partial \beta_j}\right] I_{\{y_i>0\}}] \tag{26}$$

$$\frac{\partial^2 l}{\partial \beta_j \partial \gamma} = \sum_{i=1}^{n} \left[\frac{\left(\frac{1}{1+\mu c}\right)\exp(-c^{-1}\log(1+\mu c))}{[\gamma + \exp(-c^{-1}\log(1+\mu c))]^2}\left[\frac{\partial \mu_i}{\partial \beta_j}\right] I_{\{y_i=0\}} + 0 I_{\{y_i>0\}}\right] \tag{27}$$

$$\frac{\partial^2 l}{\partial c^2} = \sum_{i=1}^{n} [\exp(-c^{-1}\log(1+\mu c))][[\gamma + \exp(-c^{-1}\log(1+\mu c))]$$
$$\left[-c^{-1}\frac{-\mu^2}{(1+\mu c)^2} + 2c^{-2}\frac{\mu}{1+\mu c} + (-2)c^{-3}\log(1+\mu c)\right.$$
$$+\left[c^{-1}\left(\frac{\mu}{1+\mu c}\right) - c^{-2}\log(1+\mu c)\right]^2]$$
$$-\exp(-c^{-1}\log(1+\mu c))\left[c^{-1}\left(\frac{\mu}{1+\mu c}\right) - c^{-2}\log(1+\mu c)\right]^2]$$
$$[[\gamma + \exp(-c^{-1}\log(1+\mu c))]^2]^{-1} I_{\{y_i=0\}}$$
$$+\left[(y_i + c^{-1})\frac{-\mu^2}{(1+\mu c)^2} + (-2)c^{-2}\frac{\mu}{1+\mu c} + 2c^{-3}\log(1+\mu c)\right] I_{\{y_i>0\}}] \tag{28}$$

$$\frac{\partial^2 l}{\partial c \partial \gamma} = \sum_{i=1}^{n} [\exp(-c^{-1}\log(1+\mu c))]\left[c^{-1}\left(\frac{\mu}{1+\mu c}\right) - c^{-2}\log(1+\mu c)\right]$$
$$[[\gamma + \exp(-c^{-1}\log(1+\mu c))]^2]^{-1} I_{\{y_i=0\}} + 0 I_{\{y_i>0\}}] \tag{29}$$

$$\frac{\partial^2 l}{\partial \gamma^2} = \sum_{i=1}^{n} [(1+\gamma)^{-2} - [[\gamma + \exp(-c^{-1}\log(1+\mu c))]^2]^{-1} I_{\{y_i=0\}} + 0 I_{\{y_i>0\}}] \tag{30}$$

ACKNOWLEDGEMENTS

This research was partially supported by the Natural Science and Engineering Research Council of Canada.

REFERENCES

[1] Mian, R. and Paul, S. (2016). Estimation for zero-inflated over-dispersed count data model with missing response. *Statistics in Medicine* 35, 5603-5624.

[2] Minami, M., Cody, C.E. L-. and Verdesoto, M. R-. (2007). Modeling shark bycatch: The zero-inflated negative binomial regression model with smoothing. *Fisheries Research* 84, 210-221.

[3] Mwalili, S. M., Lasaffre, E. and Declerck, D. (2008). The zero-inflated negative binomial regression model with correction for misclassification: an example in caries research. *Statistical Methods in Medical Research* 17, 123-139.

[4] Ibrahim, J. G. (1990). Incomplete Data in Generalized Linear Model. *J. Amer. Statist. Assoc.* 85, 765–769.

[5] Bohning, D., Dietz, E., Schlattmann, P., Mendonca, L., and Kirchner, U. (1999). The Zero-Inflated Poisson Model and the Decayed, Missing and Filled Teeth Index in Dental Epidemiology. *Journal of the Royal Statistical Society* A 162, 195-209.

[6] Paul, S. R. and Plackett, R. L. (1978). Inference sensitivity for Poisson mixtures. *Biometrika* 65, 591-602.

[7] Barnwal, R. K. and Paul, S. R. (1988). Analysis of one-way layout of count data with negative binomial variation. *Biometrika* 75, 215-22.

[8] Paul, S. R. and Banergee, T. (1998). Analysis of Two-Way Layout of Count Data Involving Multiple Counts in Each Cell. *J. Amer. Statist. Assoc.* 93, 1419-1429.

[9] Piegorsch, W. W. (1990). Maximum likelihood estimation for the negative binomial dispersion parameter. *Biometrics* 46, 863-867.

[10] Prentice, R. L. (1986). Binary Regression Using an Extended Beta-Binomial Distribution, With Discussion of Correlation

Induced by Covariate Measurement Errors. *J. Amer. Statist. Assoc.* 81, 321-327.

[11] Deng, D., and Paul, S. R. (2005). Score Tests for Zero Inflation and Over Dispersion in Generalized Linear Models. *Statistica Sinica* 15, 257-276.

[12] Ridout, M, Demetrio, C. G. B. and Hinde, J. (1998). Models for count data with many zeros. *International Biometric Conference*, Cape Town.

[13] Hinde, J., and Demetrio, C. G. B. (1998). Overdispersion: Models and estimation. *Computational Statistics and Data Analysis* 27, 151-170.

[14] Li, C-S, Lu, J-C, Park, J., Kim, K., Brinkley, P. A. and Peterson, J. P. (1999). Multivariate Zero-Inflated Poisson Models and Their Applications. *Technometrics* 41, 29-38.

[15] Hall, B. H. (2000). A Note on the Bias in Herfindahl-type Measures Based on Count Data. University of California at Berkeley and NBER.

[16] Lee, A. H., Wang, K., and Yau, K. K. W. (2001). Analysis of Zero-Inflated Poisson Data Incorporating Extent of Exposure. *Biometrical Journal* 43, 963-975.

[17] Wang, K., Lee, A. H., Yau, K. K. W. and Carrivick, P. J. W. (2003). A bivariate zero-inflated Poisson regression model to analyze occupational injuries. *Accident Analysis & Prevention* 35, 625-629.

[18] Lord, D., Washington, S. P., and Ivan, J. N. (2005). Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. *Accident Analysis & Prevention* 37, 35-46.

[19] Jiang, X. and Paul, S. R. (2009). Analysis of covariance of zero-inflated paired count data using a zero-inflated bivariate Poisson regression model. *Calcutta Statistical Bulletin (Special Volume)* 61, 113-124.

[20] Cameron, A. C., and Trivedi, P. K. (2013). Regression analysis of count data. Cambridge University Press.

[21] Mullahy, J. (1997). Instrumental-Variable Estimation of Count Data Models: Applications to Models of Cigarette Smoking Behavior. *The Review of Economics and Statistics* 79, 586-593.

[22] Dean, c. b. (1992). Testing for Overdispersion in Poisson and Binomial Regression Models. *J. Amer. Statist. Assoc.* 87, 451-457.

[23] Greene, W. H. (1994). Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regression Models. New York University, Unpublished research paper.

[24] Broek, J. V. D. (1995). A Score Test for Zero Inflation in a Poisson Distribution. *Biometrics* 51, 738-743.

[25] Deng, D., and Paul, S. R. (2000). Score Tests for Zero Inflation in Generalized Linear Models. *The Canadian Journal of Statistics* 87, 451-457.

[26] Xie, M., He, B., and Goh, T. N. (2001). Zero-inflated Poisson model in statistical process control. *Computational Statistics and Data Analysis* 38, 191-201.

[27] Paul, S. R., Jiang, X., Rai, S. N. and Balasooriya, U. (2004). Test of treatment effect in pre-drug and post-drug count data with zero-inflation. *Statistics in medicine* 23, 1541-1554.

[28] Williamson, J. M., Lin, H-M, Lyles, R. H., and Hightower, A. W. (2007). Power Calculations for ZIP and ZINB Models. *Journal of Data Science* 5, 519-534.

[29] Rubin, D. B. (1977). Formalizing subjective notions about the effect of nonrespondents in sample surveys. *J. Amer. Statist. Assoc.* 72, 538-543.

[30] Little, R. J. A., and Rubin, D. B. (1987, 2002, 2014). Statistical Analysis With Missing Data. New York: Wiley, 2nd ed.

[31] Anderson, T. W. and Taylor, J. B. (1976). Strong Consistency of Least Squares Estimates in Normal Linear Regression. *The Annals of Statistics* 4, 788-790.

[32] Geweke, J. (1986). Inference in the Inequality Constrained Normal Linear Regression Model. *Journal of Applied Econometrics* 1, 117-141.

[33] Raftery, A. E., Madigan, D. and Hoeting, J. A. (1997). Bayesian Model Averaging for Linear Regression Models. *J. Amer. Statist. Assoc.* 92, 179-191.

[34] Chen, J., Hubbard, S. and Rubin, Y. (2001). Estimating the hydraulic conductivity at the south oyster site from geophysical tomographic data using Bayesian Techniques based on the normal linear regression model. *Water Resources Research* 37, 1603-1613.

[35] Kelly, B. C. (2007). Some aspects of measurement error in linear regression of astronomical data. *The Astrophysical Journal* 665, 1489-1506.

[36] Zhang, C-H and Huang, J. (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics* 36, 1567-1594.

[37] Lipsitz, S. R., and Ibrahim, J. G. (1996). A conditional model for incomplete covariates in parametric regression models. *Biometrika* 83, 916-922.

[38] Ibrahim, J. G., and Lipsitz, S. R. (1996). Parameter Estimation From Incomplete Data in Binomial Regression When the Missing Data Mechanism Is Nonignorable. *Biometrics* 52, 1071-1078.

[39] Ibrahim, J. G., Chen, M.-H., and Lipsitz, S. R. (1999). Monte Carlo EM for Missing Covariates in Parametric Regression Models. *Biometrics* 55, 591-596.

[40] Ibrahim, J. G., Chen, M.-H., and Lipsitz, S. R. (2001). Missing Responses in Generalized Linear Mixed Models When the Missing Data Mechanism Is Nonignorable. *Biometrika* 88, 551-556.

[41] Ibrahim, J. G., Chen, M.-H., and Lipsitz, S. R., and Herring, A. H. (2005). Missing-Data Methods for Generalized Linear Models. *J. Amer. Statist. Assoc.* 100, 332-346.

[42] Sinha, S and Maiti, T (2007). Analysis of matched case-control data in presence of nonignorable missing exposure. *Biometrics* 64, 106-114.

[43] Maiti, T., and Pradhan, V. (2009). Bias Reduction and a Solution for Separation of Logistic Regression with Missing Covariates. *Biometrics* 65, 1262-1269.

[44] Dempster, A. P., Larid, N. M. and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM

Algorithm. *Journal of the Royal Statistical Society* B 39, 1-38.

[45] Sahu, S. K. and Roberts, G. O. (1999). On convergence of the EM algorithm and the Gibbs sampler. *Statistics and Computing* 9, 55-64.

[46] Casella, G. and George, E. L. (1992). Explaining the Gibbs Sampler. *The American Statistician* 46, 167-174.

[47] Efron, B. and Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika* 65, 457-87.

[48] Lawless, J. F. (1987). Negative Binomial and Mixed Poisson Regression. *The Canadian Journal of Statistics* 15, 209-225.