HHS Public Access

Author manuscript

Biom J. Author manuscript; available in PMC 2019 April 08.

Published in final edited form as:

Biom J. 2018 July ; 60(4): 845–858. doi:10.1002/bimj.201600249.

Marginalized zero-inflated Poisson models with missing covariates

Habtamu K. Benecha[*], **John S. Preisser**[†], **Kimon Divaris**[‡], **Amy H. Herring**[§], and **Kalyan Das**[¶]

[*]USDA National Agricultural Statistics Service (NASS), Washington, DC.

[†]Department of Biostatistics, University of North Carolina, Chapel Hill, USA.

[‡]Department of Pediatric Dentistry, University of North Carolina, Chapel Hill, USA.

[§]Department of Statistical Science, Duke University, Durham, USA.

[¶]Department of Statistics, University of Calcutta, Kolkata, India.

Abstract

Unlike zero-inflated Poisson regression, marginalized zero-inflated Poisson (MZIP) models for counts with excess zeros provide estimates with direct interpretations for the overall effects of covariates on the marginal mean. In the presence of missing covariates, MZIP and many other count data models are ordinarily fitted using complete case analysis methods due to lack of appropriate statistical methods and software. This article presents an estimation method for MZIP models with missing covariates. The method, which is applicable to other missing data problems, is illustrated and compared with complete case analysis by using simulations and dental data on the caries preventive effects of a school-based fluoride mouthrinse program.

Keywords

Marginalized models; Missing at random; Missing data; Monte Carlo EM; Zero-inflation

1. Introduction

Counts collected in many applications often contain higher frequencies of zeros than assumed by the Poisson distribution. For example, in studies of dental caries (also known as tooth decay or cavities) among schoolchildren, counts of decayed, missing and filled tooth surfaces (dmfs) are typically zero for disproportionately high numbers of children (Lewsey and Thompson, 2004; Mwalili *et al.*, 2008; Preisser *et al.*, 2012; Long *et al.*, 2014; Divaris *et al.*, 2012; Albert *et al.*, 2014). Because of the inadequacy of Poisson models in such situations, ‘zero-inflated’ or ‘excess zero’ counts are often modeled through latent variables defining membership into one of two unobserved populations. Zero-inflated Poisson (ZIP) regression is the most common of such methods and assumes that zero counts arise either

from a ‘non susceptible’ or ‘perfect’ population that gives only zeros or from a ‘susceptible’, ‘imperfect’ population that produces both zero and positive counts according to a Poisson distribution (Lambert, 1992; Mullahy, 1986). ZIP has become a popular model for zero-inflated data after Lambert (1992) described the data generating process and applied it to defects in manufacturing. ZIP models commonly specify regression parameters for the probability of being from the ‘non-susceptible’ population and for the mean of the assumed Poisson distribution.

Although zero-inflated Poisson regression provides flexible modeling of counts with excess zeros, the resulting parameter estimates do not have direct interpretations for the overall population mean count. The limitations of ZIP models have been noted for the lack of regression coefficients having population-wide interpretations and for relying on hypothetical populations that may not be of interest to investigators (Albert *et al.*, 2014). In the dental caries example, while one set of ZIP parameters describes the probability that a child is from a non-susceptible, caries-free latent population, the other set of parameters explains the mean caries counts of children from a caries susceptible latent population. When interest is in estimating the effects of covariates on the overall mean caries count, regression coefficients obtained from such models can only be used through indirect methods using post-modeling calculations. In addition, ZIP model parameters are often inconvenient to use to estimate other important population parameters such as incidence density ratios (IDRs; Preisser *et al.*, 2012).

In order to estimate exposure effects on the overall population mean and allow for population-wide inferences, Long *et al.*(2014) propose marginalized zero-inflated Poisson (MZIP) models for independent responses, where regression parameters are estimated for the marginal mean by using maximum likelihood methods. While both ZIP and MZIP models define regression parameters for the probability of being from the ‘non-susceptible’ population, unlike ZIP, the second set of regression parameters in MZIP are linked directly to the overall population mean. Long *et al.*(2014) discuss parameter estimation methods for MZIP as well as their application in modeling counts of unprotected intercourse acts, and Preisser *et al.*(2016) describe marginalized models for counts with zero-inflated negative binomial distributions. Todem *et al.*(2016) estimate the effects of covariates on the marginal mean by using latent model formulations as well as by specifying regression parameters for the marginal mean.

While much of the statistical literature on zero-inflated data modeling treats covariates and outcomes as fully observed, data in practice often contain variables with missing values. In the absence of appropriate statistical software and methods to deal with incomplete data, modeling is typically done by using only cases with complete covariate and outcome data (Ibrahim *et al.*, 2005). However, this approach, known as complete case (CC) analysis, is valid only when missingness is independent of any observed and un-observed data. In generalized linear models where covariates are missing with ignorable missingness and their conditional distribution is log-concave, Ibrahim *et al.*(1999) propose a Monte Carlo EM (Wei and Tanner, 1990) algorithm to perform estimation. Although the method can be adapted to ZIP regression with missing covariates, it is not directly applicable to marginalized zero-inflated models because the corresponding conditional densities may not

be written as products of log-concave distributions. This paper extends the work of Ibrahim *et al.*(1999) to MZIP models with missing covariates and fully observed outcomes.

A motivation for the paper comes from a study carried out to evaluate the caries preventive effects of a school-based fluoride mouthrinse program among North Carolina (NC) schoolchildren (Divaris *et al.*, 2012). Because of missing covariate values in the study, MZIP models with complete case analysis discard data from a high proportion of children. The proposed approach makes use of all the available data for the estimation of parameters by assuming that observations were missing at random. Sections 2 and 3 review zero-inflated Poisson and marginalized zero-inflated Poisson models respectively. Section 4 describes Monte Carlo EM (MCEM) methods for MZIP models with missing covariates. Section 5 presents simulation studies that compare results from the proposed method with those from complete case analysis. Section 6 applies the new method to the NC schoolchildren data, and compares the results with complete case analysis and multiple imputation. We conclude with a discussion in Section 7.

2. Zero-inflated Poisson Models

Zero-inflated Poisson models assume that counts emanate either from a ‘susceptible’ population that gives zero and positive counts according to a Poisson distribution, or from a ‘non-susceptible’ population, which produces additional zeros (Lambert, 1992). Thus, while a subject with a positive count is considered as belonging to the ‘susceptible’ population, individuals with zero counts may belong to either of the two latent populations. Accordingly, a random count variable from the i^{th} subject, Y_i, takes zero or positive values as

$$Pr(Y_i = k) = \begin{cases} \psi_i + (1-\psi_i)\exp(-\mu_i), & k = 0 \\ (1-\psi_i)\dfrac{\exp(-\mu_i)\mu_i^k}{k!}, & k = 1, 2, \ldots \end{cases} \quad (1)$$

where ψ_i is the probability of being from the ‘non-susceptible’ population and μ_i is the Poisson mean corresponding to the ‘susceptible’ population. The marginal mean of a ZIP random variable is the Poisson mean multiplied by one minus the excess-zeros probability, in symbols $\nu_i = (1-\psi_i)\mu_i$, where $\nu_i = \mathrm{E}(Y_i)$. It can be seen from equation (1) that ZIP reduces to the standard Poisson distribution when $\psi_i = 0$. The joint distribution of n independent ZIP random variables is

$$f(\mathbf{y}|\boldsymbol{\psi}, \boldsymbol{\mu}) = \prod_{y_i=0}\left[\left(\frac{\psi_i}{1-\psi_i} + e^{-\mu_i}\right)(1-\psi_i)\right]\prod_{y_i>0}\left[(1-\psi_i)e^{-\mu_i}\mu_i^{y_i}/(y_i!)\right], \quad (2)$$

where $\mathbf{y}$, ψ, and $\boldsymbol{\mu}$ are $n\times 1$ vectors of the count outcomes $\{Y_i\}$ and the model parameters $\{\psi_i\}$ and $\{\mu_i\}$, respectively.

The probability of membership in the 'non-susceptible' population, ψ_i, and the mean μ_i of the Poisson part, are modeled as functions of covariates by using the logit and the log links as

$$logit(\psi_i) = \mathbf{z}_i'\gamma \quad \text{and} \quad log(\mu_i) = \mathbf{x}_i'\beta, \quad (3)$$

where $\mathbf{z}_i$ and $\mathbf{x}_i$ are $q \times 1$ and $p \times 1$ vectors of covariates for the i^{th} subject, and $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_q)'$ and $\beta = (\beta_1, \beta_2, \ldots, \beta_p)'$ are regression parameters. The ZIP model likelihood function is

$$L(\boldsymbol{\beta}, \boldsymbol{\gamma}/\mathbf{y}) = \prod_{i=1}^{n} \left\{1 + e^{\mathbf{z}_i'\gamma}\right\}^{-1} \left\{e^{\mathbf{z}_i'\gamma} + e^{-\exp(\mathbf{x}_i'\beta)}\right\}^{I(y_i = 0)} \left\{\frac{e^{-\exp(\mathbf{x}_i'\beta)} e^{y_i \mathbf{x}_i'\beta}}{y_i!}\right\}^{I(y_i > 0)} \quad (4)$$

In equation (4), $I(T)$ takes the value 1 if T is true and takes zero, otherwise. While interpretations of parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ in (3) pertain to the two latent populations, the overall marginal mean response, $E(y_i|\mathbf{z}_i, \mathbf{x}_i) = [(1 - \psi_i(\mathbf{z}_i)]\mu_i(\mathbf{x}_i)$, for the i^{th} subject may be estimated from the ZIP model by

$$E(y_i|\mathbf{z}_i, \mathbf{x}_i) = \frac{e^{\mathbf{x}_i'\beta}}{1 + e^{\mathbf{z}_i'\gamma}}, \quad (5)$$

by substituting $\psi = \text{logit}^{-1}(\mathbf{z}_i'\gamma)$ and $\mu_i = \exp(\mathbf{x}_i'\beta)$ from equation (3) into $\nu_i = (1 - \psi_i)\mu_i$. Given equation (5), the quantification of the effect of covariates on the marginal mean with suitable variance estimates may be difficult for many analysts in practice, and indeed many authors avoid making inferences on the marginal mean response or do so in error (Preisser *et al.*, 2012). In addition, when interest is in determining whether the effects of an exposure on ν_i are homogeneous across the levels of covariates, ZIP models usually do not provide the desired estimates (Long *et al.*, 2014).

3. Marginalized ZIP Models

In dental caries research, the marginal mean (ν_i) caries count is often of greater interest than the mean caries count μ_i of a susceptible latent class of individuals (Preisser *et al.*, 2017). Given that $\mu_i = \nu_i / (1 - \psi_i)$ by inversion of the equation $\nu_i = (1 - \psi_i)\mu_i$, an equivalent representation of the ZIP distribution in equation (1) is

$$Pr(Y_i = k) = \begin{cases} \psi_i + (1-\psi_i)\exp(-\nu_i/(1-\psi_i)), & k = 0 \\ (1-\psi_i)\dfrac{\exp(-\nu_i/(1-\psi_i))[\nu_i/(1-\psi_i)]^k}{k!}, & k = 1, 2, \ldots \end{cases} \quad (6)$$

and $Var(Y_i) = \nu_i\left[1 + \frac{\nu_i\psi_i}{1-\psi_i}\right]$.

It follows that the joint ZIP distribution in equation (2) can be expressed as a joint MZIP distribution as

$$f(\mathbf{y}|\boldsymbol{\psi}, \boldsymbol{\nu}) = \prod_{y_i=0}\left[\left(\frac{\psi_i}{1-\psi_i} + e^{-\nu_i/(1-\psi_i)}\right)(1-\psi_i)\right] \prod_{y_i>0}[(1-\psi_i)e^{-\nu_i/(1-\psi_i)}[\nu_i/(1-\psi_i)]^{y_i}/(y_i!)] \quad (7)$$

In order to allow direct inferences about the overall population from which zero-inflated counts are drawn, the MZIP model (Long *et al.*, 2014) links regression parameters directly to the marginal mean ν_i, while employing another set of parameters to model the probability of being an excess zero (i.e., ψ_i). For the i^{th} observation, MZIP relates ν_i and ψ_i with the independent variables as

$$logit(\psi_i) = \mathrm{z}_i'\gamma \quad \text{and} \quad log(\nu_i) = \mathrm{x}_i'\alpha. \quad (8)$$

In equation (8), ψ_i and $\boldsymbol{\gamma}$ have the same interpretation as in ZIP, and $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_p)'$ is a vector of regression parameters for ν_i having interpretations as log IDRs for the entire sampled population. For the i^{th} subject, the incidence density ratio (or incidence rate ratio) corresponding to a one unit increase in the j^{th} exposure variable, x_{ij}, from $x_{ij} = c$ to $x_{ij} = c + 1$ is,

$$IDR_i = \frac{E(Y_i|\mathbf{z}_i, x_{ij} = c+1, \tilde{\mathbf{x}}_i' = \tilde{\mathbf{x}}_i')}{E(Y_i|\mathbf{z}_i, x_{ij} = c, \tilde{\mathbf{x}}_i' = \tilde{\mathbf{x}}_i')} = e^{\alpha j}, \quad (9)$$

where $\tilde{\mathrm{x}}_i$ is the vector of covariates without x_{ij} (Long *et al.*, 2014). Substituting equation (8) into equation (7) gives the MZIP model log-likelihood function

$$\ell(\boldsymbol{\gamma},\boldsymbol{\alpha}|\mathbf{y}) = -\sum_{i=1}^{n}\log\left(1+e^{\mathbf{z}_i'\gamma}\right)+\sum_{i=1}^{n}I(y_i=0)\log\left\{e^{\mathbf{z}_i'\gamma}+e^{-\left(1+\exp\left(\mathbf{z}_i'\gamma\right)\right)\exp\left(\mathbf{x}_i'\alpha\right)}\right\}$$
$$+\sum_{i=1}^{n}I(y_i>0)\left\{-\left(1+e^{\mathbf{z}_i'\gamma}\right)e^{\mathbf{x}_i'\alpha}+y_i\log\left(1+e^{\mathbf{z}_i'\gamma}\right)+y_i\mathbf{x}_i'\alpha-\log y_i!\right\}.$$

Long *et al.*(2014) employ quasi-Newton optimization to obtain parameter estimates for complete data. The variance covariance matrix of the parameters is obtained by inverting the expected information matrix. For the case in which the counts are over-dispersed relative to ZIP, these authors estimate empirical sandwich standard errors.

4. Monte Carlo EM for Missing Covariates

The EM algorithm (Dempster, Laird and Rubin, 1977) has been an important method of estimation for models with incomplete data. Estimation involves iterations between the expectation and maximization steps; while the expectation or E-step of an iteration computes the expected value of the complete data log-likelihood conditional on the observed data and current parameter values, the maximization or M-step of EM maximizes the expected log-likelihood. Because the E-step is difficult to compute in many applications, the Monte Carlo EM algorithm (MCEM) of Wei and Tanner (1990) is often used to estimate the expected log-likelihood. MCEM computes the expected log-likelihood numerically by using Monte Carlo samples from the conditional distributions of the unobserved variables. Ibrahim *et al.*(1999) apply MCEM for missing covariates in parametric models by generating samples using the Gibbs sampler with adaptive rejection sampling (ARS) (Gilks and Wild, 1992). The ARS algorithm requires the conditional distribution of the missing covariates to be log-concave, and the method of Ibrahim *et al.*(1999) can be applied to any settings where the log-concavity criterion is met. In the case of MZIP models, which have complex likelihood functions involving two linear predictors, conditional distributions of missing covariates are generally not log-concave. We extend the Monte Carlo EM approach to MZIP models with missing covariates, where missingness is ignorable and the count outcome is fully observed.

Suppose that $y' = (y_1, y_2, \ldots y_n)$ is a vector of independent zero-inflated count outcomes from n subjects, and let $z_i' = (z_{i1}, z_{i2}, \ldots z_{iq})$ and $\mathrm{x}_i' = (x_{i1}, x_{i2}, \ldots x_{ip})$ be the covariate vectors in the MZIP model in equation (8). Because the linear predictors for the logit of ψ_i and the logarithm of ν_i typically contain one or more common covariates, $\mathbf{z}_i$ and $\mathbf{x}_i$ can be expressed as $\mathrm{z}_i = (\tilde{\mathrm{z}}_i', \mathrm{w}_i')'$ and $\mathrm{x}_i = (\tilde{\mathrm{x}}_i', \mathrm{w}_i')'$, where $\mathbf{w}_i$ represents covariates common to $\mathbf{z}_i$ and $\mathbf{x}_i$, while $\tilde{\mathrm{z}}_i$ and $\tilde{\mathrm{x}}_i$ denote covariates exclusive to $\mathbf{z}_i$ and $\mathbf{x}_i$ respectively. In the sense that covariates are partially missing for some subjects, the vector $\mathbf{u}_i' = (\tilde{\mathbf{z}}_i', \mathbf{w}_i', \tilde{\mathbf{x}}_i')$ of k distinct covariates from the i^{th} subject can also be written as in Ibrahim *et al.*(1999) as: $\mathbf{u}_i = (\mathbf{u}_i^{obs}, \mathbf{u}_i^{mis})$ with $\mathbf{u}_i^{obs}$ and $\mathbf{u}_i^{mis}$ representing the observed and the missing parts of $\mathbf{u}_i$ respectively. Using these notations, the observed data vector for the i^{th} subject is $(y_i, \mathbf{u}_i^{'obs}, \mathbf{r}_i')'$ where $\mathrm{r}_i' = (r_{i1}, r_{i2}, \ldots r_{ik})$ is a vector of missingness indicators for the k covariates and

$$r_{ij} = \begin{cases} 1, & \text{if the } j^{th} \text{component of } \mathbf{u}_i \text{is observed,} \\ 0, & \text{otherwise.} \end{cases} \quad (10)$$

When covariate values are missing at random (Little and Rubin, 2002), the conditional distribution of $\mathbf{r}_i$ given the data is a function only of the observed data, i.e.,

$$Pr\left(\mathbf{r}_i \middle| y_i, \mathbf{u}_i^{obs}, \mathbf{u}_i^{mis}, \phi\right) \propto Pr\left(\mathbf{r}_i \middle| y_i, \mathbf{u}_i^{obs}, \phi\right),$$

where $\boldsymbol{\phi}$ is a vector of parameters. In addition, when $\boldsymbol{\phi}$ is distinct from the parameters in the joint distribution of $(y_i, \mathbf{u}_i)$, missingness is ignorable (Ibrahim *et al.*, 1999) and estimation can be done using the likelihood

$$\begin{aligned} L\left(\xi, \alpha, \gamma \middle| y, \mathbf{u}^{obs}, \mathbf{u}^{mis}\right) &= \prod_{i=1}^{n} Pr\left(y_i \middle| \mathbf{u}_i^{obs}, \mathbf{u}_i^{mis}, \alpha, \gamma\right) Pr\left(\mathbf{u}_i^{mis} \middle| \mathbf{u}_i^{obs}, \boldsymbol{\xi}\right) \quad (11) \\ &= \prod_{i=1}^{n} L_i\left(\boldsymbol{\xi}, \alpha, \gamma \middle| y_i, \mathbf{u}_i^{obs}, \mathbf{u}_i^{mis}\right), \end{aligned}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ are the regression parameters in equation (8), $\boldsymbol{\xi}$ is a vector of parameters in the joint distribution of the missing covariates, and $\mathbf{u}^{obs}$ and $\mathbf{u}^{mis}$ are the observed and the missing parts of covariates over all the n observations. From equation (11), the complete data log-likelihood $\ell\left(\boldsymbol{\theta} \middle| y, \mathbf{u}^{obs}, \mathbf{u}^{mis}\right)$ can be written as:

$$\ell\left(\boldsymbol{\theta} \middle| \mathbf{y}, \mathbf{u}^{obs}, \mathbf{u}^{mis}\right) = \sum_{i=1}^{n} \ell\left(\boldsymbol{\eta} \middle| y_i; \mathbf{u}_i^{obs}, \mathbf{u}_i^{mis}\right) + \sum_{i=1}^{n} \ell\left(\boldsymbol{\xi} \middle| \mathbf{u}_i^{mis}; \mathbf{u}_i^{obs}\right) \quad (12)$$

where, $\boldsymbol{\theta}' = (\alpha', \gamma', \xi')$, $\boldsymbol{\eta}' = (\alpha', \gamma')$, $\ell\left(\eta \middle| y_i; \mathbf{u}_i^{obs}, \mathbf{u}_i^{mis}\right) = log\left(Pr\left(y_i \middle| \mathbf{u}_i^{obs}, \mathbf{u}_i^{mis}, \boldsymbol{\eta}\right)\right)$, and $\ell\left(\xi \middle| \mathbf{u}_i^{mis}; \mathbf{u}_i^{obs}\right) = log\left(Pr\left(\mathbf{u}_i^{mis} \middle| \mathbf{u}_i^{obs}, \boldsymbol{\xi}\right)\right)$.

The observed data log-likelihood is obtained by integrating (summing) $\ell\left(\boldsymbol{\theta} \middle| y, \mathbf{u}^{\text{obs}}, \mathbf{u}^{mis}\right)$ over the domain of the missing covariates. However, direct estimation from the observed log-likelihood is difficult because the integral involves the conditional distribution of the MZIP outcome variable. An alternative method of estimation in such situations has been the EM algorithm, where, in the E-step, the expected value of the observed log-likelihood is estimated conditional on current parameter estimates and the observed data, and maximization is performed on the estimated log-likelihood. If the vector of parameter estimates at iteration t is $\boldsymbol{\theta}^{(t)}$, in the $(t+1)^{th}$ iteration, corresponding to the i^{th} subject, the E step of EM computes,

$$Q_i\left(\boldsymbol{\theta}\middle|\boldsymbol{\theta}^{(t)}\right) = E\left(\ell\left(\boldsymbol{\theta}\middle|y_i, \mathbf{u}_i^{obs}, \mathbf{u}_i^{mis}\right)\middle|y_i, \mathbf{u}_i^{obs}, \boldsymbol{\theta}^{(t)}\right) \quad (13)$$

Had the expectation in equation (13) been easily obtained, the M-step of EM would have maximized $Q_i\left(\boldsymbol{\theta}\middle|\boldsymbol{\theta}^{(t)}\right) = \sum_{i=1}^{n} Q_i\left(\boldsymbol{\theta}\middle|\boldsymbol{\theta}^{(t)}\right)$ to obtain the parameter estimates at iteration $(t + 1)$. However, because such expectations are difficult to compute for MZIP models, as in Ibrahim *et al.*(1999), we estimate the E-step using MCEM. At iteration $t + 1$, MCEM estimates $Q_i\left(\boldsymbol{\theta}\middle|\boldsymbol{\theta}^{(t)}\right)$ using Monte Carlo samples of size, say s, from the conditional distribution of the missing covariates given $y_i, \mathbf{u}_i^{obs}$ and the current parameter estimates, $\boldsymbol{\theta}^{(t)}$ by

$$Q_i\left(\boldsymbol{\theta}\middle|\boldsymbol{\theta}^{(t)}\right) = \frac{1}{s}\sum_{j=1}^{s} \ell\left(\boldsymbol{\theta}\middle|y_i, \mathrm{d}_{ij}, \mathbf{u}_i^{obs}\right)$$

where $\mathbf{d}_{i1}, \mathbf{d}_{i2}, \ldots, \mathbf{d}_{is}$ are vectors of samples from the conditional distribution of the missing covariates. After iteration t, the conditional distribution of the missing covariates, $Pr\left(\mathbf{u}_i^{mis}\middle|y_i, \mathbf{u}_i^{obs}, \boldsymbol{\theta}^{(t)}\right)$, can be written as,

$$Pr\left(\mathbf{u}_i^{mis}\middle|y_i, \mathbf{u}_i^{obs}, \boldsymbol{\theta}^{(t)}\right) = \frac{Pr\left(y_i\middle|\mathbf{u}_i^{obs}, \mathbf{u}_i^{mis}, \boldsymbol{\eta}^{(t)}\right) Pr\left(\mathbf{u}_i^{mis}\middle|\mathbf{u}_i^{obs}, \boldsymbol{\xi}^{(t)}\right)}{\int Pr\left(y_i\middle|\mathbf{u}_i^{obs}, \mathbf{u}_i^{mis}, \boldsymbol{\eta}^{(t)}\right) Pr\left(\mathbf{u}_i^{mis}\middle|\mathbf{u}_i^{obs}, \boldsymbol{\xi}^{(t)}\right) d\,\mathbf{u}_i^{mis}}. \quad (14)$$

For missing covariate problems in MZIP models, and in general for models where the log-concavity condition is not met, the adaptive rejection metropolis sampling (ARMS) algorithm of Gilks, Best and Tan (1995) allows sampling from the conditional distributions of the covariates in equation (14). ARMS is an extension of ARS for distributions that are not log-concave, and we employ the algorithm to generate Monte Carlo samples from conditional distributions of the missing covariates. After each E-step of EM, optimization can be performed on the estimated log-likelihood by using quasi-Newton methods.

Given the maximum likelihood estimates $\hat{\boldsymbol{\theta}}$ obtained from MCEM, the observed information matrix $I\left(\hat{\boldsymbol{\theta}}\right)$ is obtained (Wei and Tanner, 1990; Ibrahim *et al.*, 1999; Louis, 1982) by using Monte Carlo samples $\mathbf{d}_{i1}, \mathbf{d}_{i2}, \ldots, \mathbf{d}_{is}$ as

$$I\left(\hat{\boldsymbol{\theta}}\right) = -\sum_{i=1}^{n}\frac{1}{s}\sum_{j=1}^{s}\left.\frac{\partial^2 \ell\left(\boldsymbol{\theta}\middle| y_i, \mathbf{d}_{ij}, \mathbf{u}_i^{obs}\right)}{\partial \boldsymbol{\theta}^2}\right|_{\left(\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}\right)} \quad (15)$$
$$-\sum_{i=1}^{n}\frac{1}{s}\sum_{j=1}^{s}\left.\frac{\partial \ell\left(\boldsymbol{\theta}\middle| y_i, \mathbf{d}_{ij}, \mathbf{u}_i^{obs}\right)}{\partial \boldsymbol{\theta}}\left\{\frac{\partial \ell\left(\boldsymbol{\theta}\middle| y_i, \mathbf{d}_{ij}, \mathbf{u}_i^{obs}\right)}{\partial \boldsymbol{\theta}}\right\}',\right|_{\left(\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}\right)}$$
$$+\sum_{i=1}^{n}\left\{\frac{1}{s}\sum_{j=1}^{s}\frac{\partial \ell\left(\boldsymbol{\theta}\middle| y_i, \mathbf{d}_{ij}, \mathbf{u}_i^{obs}\right)}{\partial \boldsymbol{\theta}}\right\}\left\{\frac{1}{s}\sum_{j=1}^{s}\left.\frac{\partial \ell\left(\boldsymbol{\theta}\middle| y_i, \mathbf{d}_{ij}, \mathbf{u}_i^{obs}\right)}{\partial \boldsymbol{\theta}}\right\}',\right|_{\left(\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}\right)}$$

Standard errors of parameter estimates are calculated by

$$\mathrm{se}\left(\hat{\boldsymbol{\theta}}\right) = \sqrt{\mathrm{diagonal}\left\{\left[I\left(\hat{\boldsymbol{\theta}}\right)\right]^{-1}\right\}} \quad (16)$$

which means that the square roots of the diagonal elements of $\left[I\left(\hat{\boldsymbol{\theta}}\right)\right]^{-1}$ were taken to be the standard errors for the elements of $\hat{\boldsymbol{\theta}}$.

5. Simulation Studies

Simulations were carried out to assess the performance of the MCEM method relative to CC analysis for MZIP models involving one and two missing covariates. Complete case analysis provides a practical reference given that it is the standard method in practice. In the first set of simulations, samples of sizes $n = 250$, $n = 500$ and $n = 1000$ zero-inflated counts were generated from equation (1), with $\mu_i = \nu_i/(1-\psi_i)$ and (ψ_i, ν_i) defined by

$$\begin{aligned} logit(\psi_i) &= \gamma_0 + \gamma_1 x_{i1} + \gamma_2 x_{i2} \quad (17)\\ log(\nu_i) &= \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2}, \end{aligned}$$

where $(\gamma_0, \gamma_1, \gamma_2) = (1, -1, 1)$, $(\alpha_0, \alpha_1, \alpha_2) = (1, -1, 1)$, $x_{i2} \sim N(\tau, \sigma^2)$ with $\tau = 0.25$ and $\sigma^2 = 1, x_{i1} \sim N(\omega_0 + \omega_1 x_{i2}, k^2)$ with $\omega_0 = 1, \omega_1 = 1$ and $\kappa^2 = 1$. Covariate x_{i2} was fully observed, and missing data were generated for x_{i1} with the missingness mechanism depending only on the fully observed variables y_i and x_{i2}. Denote the missingness indicator for x_{i1} by r_i such that $r_i = 1$ when x_{i1} is observed and $r_i = 0$ when x_{i1} is missing. The missingness indicators were generated from the logistic model

$$logit(Pr(r_i = 0)) = \phi_0 + \phi_1 y_i + \phi_2 x_{i2}, \quad (18)$$

with $(\phi_0, \phi_1, \phi_2) = (0.5, 1, -1)$ and x_{i1} was set to missing whenever $r_i = 0$. The mean percentages of missing values for the simulations with sample sizes 250, 500 and 1000 were respectively 34.4%, 34.5% and 34.5%. The true MZIP model in equation (17) was fitted using the MCEM method where linear regression was used to model the missing covariate as a function of the observed covariate. For each sample size, simulations were performed using 500 replications. In all simulation scenarios, including those described immediately below, the number of Monte Carlo iterations used within each iteration of EM was 1000.

The second set of simulations involve MZIP models with three covariates, two of which were missing at random. The sample size in each scenario was 1000 and 500 replications were used. Specifically, the count y_i was generated from the model

$$\begin{aligned} \text{logit}(\psi_i) &= \gamma_0 + \gamma_1 x_{i1} + \gamma_2 x_{i2} + \gamma_3 x_{i3} \\ \log(\nu_i) &= \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \alpha_3 x_{i3}, \end{aligned} \tag{19}$$

where $(\gamma_0, \gamma_1, \gamma_2, \gamma_3) = (0.5, -0.5, -0.5, 0.5)$, $(\alpha_0, \alpha_1, \alpha_2, \alpha_3) = (0.5, -0.5, -0.5, 0.5)$, $x_{i3} \sim \text{Exponential}(\lambda)$ with $\lambda = 1, x_{i2} \sim N(\mu_2, \sigma_2^2)$ with $\mu_2 = 0, \sigma_2^2 = 1$, and $x_{i1} \sim N(\omega_0 + \omega_1 x_{i2}, k^2)$ with $\omega_0 = 0.5$, $\omega_1 = -0.5$, and $\kappa^2 = 1$. Variable x_{i3} was fully observed and missing values were generated for x_{i1} and x_{i2} with missingness probabilities that are dependent on the fully observed variables y_i and x_{i3}. Missing data were generated based on the following models, assuming missingness is at random.

$$\begin{aligned} logit(Pr(r_{i1} = 0)) &= \phi_{01} + \phi_{11} y_i + \phi_{21} x_{i3} \\ logit(Pr(r_{i2} = 0)) &= \phi_{02} + \phi_{12} y_i + \phi_{22} x_{i3}, \end{aligned} \tag{20}$$

where $r_{ij} = 1$ when $x_{ij} (j = 1, 2)$ is observed and $r_{ij} = 0$ when x_{ij} is missing. The true MZIP model in equation (19) was fitted using the MCEM method where the missing covariates were modeled by using their true distributions. Simulations were performed under two different scenarios for the missing data probabilities in equation (20). In Senario 1, the parameters were specified as $(\phi_{01}, \phi_{11}, \phi_{21}) = (-0.25, 0.25, -2)$, $(\phi_{02}, \phi_{12}, \phi_{22}) = (0.25, -0.25, -2)$, and under Scenario 2, $(\phi_{01}, \phi_{11}, \phi_{21}) = (-2, -1, 1)$ and $(\phi_{02}, \phi_{12}, \phi_{22}) = (-1, -1, -1)$. The minimum and the maximum percentages of observations with at least one missing covariate in Scenario 1 were respectively 36.2 and 45.6 with a mean of 41.0. For Scenario 1, percentages of observations missing $\mathbf{x}_1$ and $\mathbf{x}_2$ range from 22.9 to 30.7 and from 17.0 to 24.7, respectively. The minimum and the maximum percentages of observations with at least one missing covariate in Scenario 2 were 26.6 and 34.2, respectively, with a mean of 30.1.

It can be seen from the two tables that percent relative biases and MSEs of estimates from MCEM are uniformly smaller than those from the CC analysis. In Table 1, MCEM provided estimated standard errors with small bias when the simulation standard deviation is used as

the true standard deviation, whereas CC analysis underestimated the standard errors for γ_1 and γ_2. However, both methods gave estimated standard errors with little biases for the parameters in the marginal mean model, which are the parameters of primary interest.

As a sensitivity analysis, the simulations under Scenario 2 were repeated by increasing the number of Monte Carlo iterations within each step of EM from 1000 to 2000 while keeping all other conditions the same. These results (Table S1) support the use of 1000 Monte Carlo iterations.

Additional sets of simulations were performed by using four covariates in an MZIP model with two of them having potentially missing values to assess sensitivities of the MCEM approach when the model for a covariate is misspecified, when the proportion of zero counts varies, and when the MZIP model is fitted to data generated from marginalized zero-inflated negative binomial (MZINB) models (Preisser *et al.*, 2017). Tables S2 and S3 in the supplementary material show that the MCEM method generally performed better than CC analysis with regard to bias and mean squares error under a misspecified model for one of the missing covariates, for different levels of zero-inflation, or for a misspecified MZIP model.

To address the substantial underestimation of standard errors in the MCEM-MZIP approach when counts are overdispersed (Table S3), an additional simulation study was performed to evaluate use of the standard sandwich estimator of the variances of marginal mean MZIP regression parameter estimates (Long et al., 2014) in the MCEM approach. For the MZIP model with two covariates where one of them has potentially missing values, the sandwich estimator is less biased than the variance estimator in equation (16) relative to the gold standard Monte Carlo standard deviation; however significant bias remains when the true model is MZINB (Table S4).

6. Application to a School-based Fluoride Mouthrinse Program

The methods developed in this article are illustrated using data collected to assess the caries preventive effects of a school-based fluoride mouthrinse program (FMR) in North Carolina (NC) schools. The data were obtained from the 2003–04 NC Oral Health Survey and involve 1363 children in grades from 1 to 5. The main exposure variable was the parent-reported number of years of participation in the FMR program (years) and the number of decayed and filled primary teeth (dfs) was an outcome variable of interest. Previously, Divaris *et al.*(2012) fitted zero-inflated negative binomial (ZINB) models to the data based on the 677 children who had complete covariate and outcome values. In this paper, we consider 1094 children with possibly missing information on two covariates (i.e., number of years of participation and family income) but with complete data on the outcome variable and the remaining covariates in the data. Of the 1094 children, 191 (17.5%) had only years missing, 180 (16.5%) had only income missing and 46 (4.20%) children had both years and income missing. Approximately, 50% of the dfs counts are zero (Figure 1). Based on prior work by Divaris *et al.*(2012), we used linear predictors of the following form:

$$\begin{aligned} logit(\psi_i) &= \gamma_0 + \gamma_1 x_{i1} + \gamma_2 x_{i2} + \gamma_3 x_{i3} + \gamma_4 x_{i4} + \gamma_5 x_{i5} + \gamma_6 x_{i6} \\ &\quad + \gamma_7 x_{i7} + \gamma_8 x_{i8} + \gamma_9 x_{i9} + \gamma_{10} x_{i10} + \gamma_{11} x_{i11} \\ log(\nu_i) &= \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \alpha_3 x_{i3} + \alpha_4 x_{i4} + \alpha_5 x_{i5} + \alpha_6 x_{i6} \\ &\quad + \alpha_7 x_{i7} + \alpha_8 x_{i8} + \alpha_9 x_{i9} + \alpha_{10} x_{i10} + \alpha_{11} x_{i11}, \end{aligned} \quad (21)$$

where ψ_i is the probability that the i^{th} child came from a caries free population, ν_i is the marginal mean caries count, x_{i1} is years divided by 3, x_{i2} is a binary indicator of whether the child is African American (1=yes, 0 =no), x_{i3} is a binary indicator of whether the child is of other non-Caucasian race (1=yes, 0 =no), x_{i4} is the child's brushing frequency (1= less than once a day, 2= once a day & 3= more than once a day), x_{i5} is family income in $ 10,000, x_{i6} is an indicator for availability of established dental home (1=yes, 0=no), x_{i7} is an indicator for availability of dental care when needed (1=yes, 0=no), x_{i8}, x_{i9} and x_{i10} are respectively age centered at the mean, its square and cubic values, and x_{i11} is an indicator for whether the child had sealants (1=yes, 0=no).

To apply the MCEM method to the data, the joint probability function of the two missing covariates was written as a product of two univariate exponential densities. As the values of years and income are non-negative and the corresponding observed data are skewed, exponential distributions seem to be appropriate to model the two missing covariates. Conditional on income and five of the observed covariates, the value of years from the i^{th} subject was assumed to have an exponential distribution with rate λ_{i1}, where

$$\lambda_{i1} = \exp(\xi_{10} + \xi_{12}x_{i2} + \xi_{13}x_{i3} + \xi_{15}x_{i5} + \xi_{16}x_{i6} + \xi_{17}x_{i7} + \xi_{18}x_{i8}). \quad (22)$$

Similarly, income was modeled using the exponential distribution with the rate parameter λ_{5i} depending on observed covariates as

$$\lambda_{i5} = \exp(\xi_{50} + \xi_{52}x_{i2} + \xi_{53}x_{i3} + \xi_{56}x_{i6} + \xi_{57}x_{i7} + \xi_{58}x_{i8}). \quad (23)$$

Based on the two exponential models and following Lipsitz and Ibrahim (1996), the joint distribution of the missing covariates years (x_{i1}) and income (x_{i5}) was obtained using equation (24).

$$\begin{aligned} Pr(x_{i1}, x_{i5}|\lambda_{i1}, \lambda_{i5}) &= Pr(x_{i1}|x_{i5}, \lambda_{i1}) Pr(x_{i5}|\lambda_{i5}) \\ &= \lambda_{i1} e^{-\lambda_{i1} x_{i1}} \lambda_{i5} e^{-\lambda_{i5} x_{i5}} \end{aligned} \quad (24)$$

Estimates from complete case analysis were used as starting values of the EM algorithm and $s = 500$ Monte Carlo iterations were used within each EM iteration. For comparison,

multiple imputation was performed by using SAS software (SAS Institute, 2015) and employing fully conditional specifications for the missing covariates. The conditional specifications involve a linear regression of years on the observed covariates in equation (22), and a linear regression of income on the covariates used in equation (23). The procedure PROC MI was used in SAS with the FCS REGPMM statement. The number of imputations was 20 and the predictive mean matching method was used to impute values. The criteria for convergence in the MCEM analyses was that the squared distance between the k^{th} and $(k+5)^{th}$ iterations should be less than 0.0001. The total computation time was close to 22 hours. The sample size, the proportion of observations with missing covariates and the total number of covariates in the MZIP model possibly contributed to the slow convergence in the MCEM estimation. Considering the number of parameters in the MZIP model for the FMR data, the threshold we set for convergence (i.e., 0.0001) may also be too small and increasing it to 0.01 or 0.001 can possibly reduce the computation time by several hours. MCMC samples based on ARMS can also be generated in the R software using packages such as ‘dlm’ and ‘HI’.

Table 3 reports results for the covariate models in equations (22) and (23). Table 4 shows parameter estimates and the corresponding standard errors from MCEM, multiple imputation and CC analysis. There is little difference between the MCEM and CC estimates of years in the marginal mean model, and most of the other covariates in the model also have similar estimates under the two approaches. A notable difference between the MCEM and CC analysis is that the two methods provide estimates of opposite signs for years and age in the zero-inflation model. For these covariates, MCEM and multiple imputation provide estimates of the same signs. Based on the MCEM analysis, the incidence rate ratio for the overall effect of three years participation in the FMR program is estimated as exp(−0.099)=0.906 with 95% CI (0.753, 1.089). Thus, conditional on covariates, the mean caries count ν_i for a child in the overall population with three years participation in the FMR program is approximately 90.6% of the mean caries count of a child with zero years of participation. In contrast, based on the CC analysis, the incidence density ratio for the overall effect of three years participation in the FMR program is estimated as exp(−0.084)= 0.919 with 95% CI (0.789, 1.071). However, the results from both MCEM and CC methods show that there was no statistically significant treatment effect as evidenced by the inclusion of 1.0 in the confidence intervals of IDR.

To further evaluate the performances of the MCEM, MI and CC methods, randomized quantile residuals (Dunn and Smyth, 1996) were computed from the three approaches based on the 677 subjects with complete data. It can be seen from plots of these residuals (Figure S1 in the supplementary material) that the MZIP model provided poor fits to the data under all the three approaches. The MZINB model may provide a better fit to the dfs counts that appear to be overdispersed.

While the observed lack-of-fit shows that the MZIP model is inadequate for predicting dfs counts in the FMR study data, Long et al. (2014) found that MZIP regression may still be valid for estimating associations between covariates and count outcomes even when the true model is MZINB provided that empirical sandwich standard errors are used for the MZIP estimates based on complete data. Empirical sandwich standard errors, constructed using the

observed information matrix and score vectors, were computed for the MCEM based estimates of the MZIP model for the FMR data (Table S5). The sandwich standard errors are larger than the model based standard errors for all parameters of the model, which is consistent with the previously reported simulation results (Table S4). However, while the sandwich estimators are likely less biased than the model-based ones based on equations (15) and (16), the simulation results suggest that they probably under-estimate the true standard errors in the FMR data analysis based on lack of fit of the MZIP model shown by the randomized quantile residuals plots. Further study or adaptation of the sandwich estimator to the MCEM setting is required.

Finally, a sensitivity analysis was performed to evaluate any changes in the MCEM and MI model estimates when a different model is specified for one of the covariates. Specifically, the MCEM estimates of years with and without income in the model equation (22) for years were similar in the marginal mean model in equation (21), but its standard error was smaller with income removed from the model for years. On the other hand, α_1 and its standard error from MI decreased substantially when income was removed from the conditional specification of years (Table S6).

7. Discussion

Marginalized zero-inflated Poisson models provide direct inferences about exposure effects on the overall population average of a count outcome with excess zeros. Extending the method of Ibrahim *et al.*(1999), this article has presented a Monte Carlo EM based method to analyze MZIP data when one or more covariates are missing at random and the count outcome is fully observed. The method can be applied to MZIP and other models where the conditional distributions of covariates are not log-concave. The proposed method uses adaptive rejection metropolis algorithm with Gibbs sampling to generate Monte Carlo samples from conditional distributions of missing covariates. While previously proposed approaches generate samples using adaptive rejection sampling, such methods are limited to models where the conditional distributions of the missing covariates are log-concave.

Simulations performed using various sample sizes and models with one and two missing covariates showed that estimates from the MCEM method have smaller mean squared errors compared to those from complete case analysis. In addition, percent relative biases of parameter estimates from the MCEM method were generally smaller than those obtained from CC analysis. In a simulation analysis where the model for one of two missing covariates was misspecified, the MCEM approach performed better than CC analysis with regard to the biases and mean squared errors of the parameter estimates. Simulations also showed that estimates from MCEM have smaller mean squared errors than those from CC analysis when the MZIP model is fitted to data generated from the marginalized ZINB distribution.

Monte Carlo EM as well as multiple imputation and complete case analysis methods were applied to real data obtained from a school-based fluoride mouthrinse study among schoolchildren. These analyses showed that, except for a few cases, estimates from the MCEM method have smaller standard errors compared to the corresponding estimates from

the CC approach. On the other hand, the MCEM and MI methods provided similar estimates and standard errors. However, high values of randomized quantile residuals under all the three methods suggest that MZIP may not be an appropriate model for the data. The MZINB model with the same linear predictors may provide a better fit to the data. Until MCEM for missing covariates is extended to MZINB or a consistent and robust variance estimator for MCEM is developed for MZIP, the best marginalized analysis of overdispersed count data with missing covariates may be based on MZINB with multiple imputation (e.g., Table S5).

The MCEM method has some notable limitations. First, one has to specify a distribution for the missing covariates and the validity of estimates is dependent on the suitability of the assumed distribution. Since misspecification of the covariate distribution can introduce new biases in the estimates of MZIP models, special attention should be given to modeling the covariates (Ibrahim et al., 1999; Ibrahim et al., 2005). In the analysis of the FMR data, the MCEM method was slightly less sensitive to covariate model misspecification than MI. A second concern is computation time of MCEM, which took almost 22 hours for the NC FMR data.

Marginalized zero-inflated Poisson regression models have application to longitudinal data (Long *et al.*, 2015) and to zero-inflated counts with varying exposure times. The ZINB model with random effects (Yau *et al.*, 2003) has been extended to the MZINB model with random effects (Burgette *et al.*, 2017). For the problem of varying exposures, standard ZIP models typically include an offset in the Poisson part of the model only (Lee et al., 2001). Beatschmann *et al.* (2013) proposes a modified zero inflated count data model where the probability of an extra zero is derived from an underlying duration model with Weibull hazard rate. In equation (8), an offset can be included in the marginal mean part of the MZIP model. Unlike the standard ZIP, separate adjustment of the first component of equation (8) for varying exposure times is unnecessary because the marginal mean includes both random and excess zeros.

Other areas for future research in MZIP models are variable selection (Wang *et al.*, 2015), model assessment and influence diagnostics. The residual plots in Figure S1 suggested possible systematic and isolated departures from the fit of MZIP to the fluoride mouthrinse data. The development of local influence diagnostics (Conceição *et al.*, 2013; Rakhmawati *et al.*, 2016) could identify observations having exceptionally large influences on model estimates. As in the comparison of standard zero-inflated models, care must be taken in applying likelihood ratio tests as test statistics may not follow conventional chisquare distributions (Böhning, 1998). A score test for comparing MZIP to MZINB models having a common covariate structure has recently been developed (Inan *et al.*, 2017). Statistical tests for zero-inflation (Ugarte *et al.*, 2004) relative to MZIP and MZINB models could also be developed.

Several extensions of marginalized ZIP models have been proposed. For example, the zero-inflated generalized Poisson (ZIGP; Famoye and Singh, 2006) allows under-or over-dispersion whereas the MZINB model only allows over-dispersion. A marginalized ZIGP with the specification in equation (8) and an underlying generalized Poisson distribution has recently been proposed (Famoye and Preisser, 2017). A further possible extension could be a

marginalized version of the zero-modified Poisson (Conceição *et al.*, 2013) regression model that allows both zero-inflation and zero-deflation. Böhning (1998) has noted that ZIP models are special cases of general finite mixtures of non-degenerate distributions, such as the Poisson-Poisson mixture. Benecha *et al.*(2017) have proposed marginalized finite mixture models for count data from multiple source populations have non-degenerate component distributions.

Finally, the problem of zero-inflation is not confined to counts. Böhning and Alfò (2016) cite several examples and discuss issues in the analysis of zero-inflated continuous responses. Typically, in semi-continuous data problems, a two-part model is used to model the probability of a zero response in combination with a model for the mean response among positive outcomes. Smith *et al.*(2014) proposed a marginalized two-part model where the second of these components is replaced with a model part for the marginal mean.

In conclusion, missing data may arise among covariates, responses or both in all of the above regression situations, and appropriate procedures are needed to address them. In this sense, the relative merits of MCEM and alternative missing data approaches such as multiple imputation and weighting require further development in MZIP models and their extensions, most notably MZINB models. Multiple imputation procedures for zero-inflated Poisson counts (Kleinke and Reinecke, 2013; Pahel *et al.*, 2011) could possibly be adapted to MZIP models when responses are missing.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

Albert J, Wang W and Nelson S (2014). Estimating overall exposure effects for zero-inflated regression models with application to dental caries. Statistical Methods in Medical Research 23, 257–278. [PubMed: 21908419]

Baetschmann G and Winkelmann R (2013). Modeling zero-inflated count data when exposure varies: With an application to tumor counts. Biometrical Journal 55, 679–686. [PubMed: 24003010]

Benecha H, Neelon B, Divaris K and Preisser JS (2017). Marginalized mixture models for count data from multiple source populations. Journal of Statistical Distributions and Applications 4:3. [PubMed: 28446995]

Böhning D (1998). Zero-Inflated Poisson Models and C.A.MAN: A Tutorial Collection of Evidence. Biometrical Journal 40, 833–843.

Böhning 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. Series A 162, 195–209.

Böhning D and Alfò M (2016). Editorial: Special issue on models for continuous data with a spike at zero. Biometrical Journal 58, 255–258. [PubMed: 26927408]

Burgette JM, Preisser JS, Weinberger M, King RS, Lee JY and Rozier RG (2017). Impact of Early Head Start in North Carolina on dental use among children younger than 3 years. American Journal of Public Health 107, 614–620. [PubMed: 28207343]

Chen XD and Fu YZ (2010). Model selection for zero-inflated regression with missing covariates. Computational Statistics & Data Analysis 136,1360–1375.

Conceição KS, Andrade MG and Louzada F (2013). Zero-modified Poisson model: Bayesian approach, influence diagnostics, and an application to a Brazilian leptospirosis notification data. Biometrical Journal 55, 661–678. [PubMed: 23564691]

Dempster AP, Laird NM and Rubin DB(1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society. Series B 39, 138.

Divaris K, Rozier RG and King RS (2012). Effectiveness of a school-based fluoride mouthrinse program. Journal of Dental Research 91, 282–287. [PubMed: 22202124]

Dunn PK and Smyth K (1996). Randomized quantile residuals. Journal of Computational and Graphical Statistics 5, 236244.

Famoye F and Preisser J (2017). Marginalized zero-inflated generalized Poisson regression. Journal of Applied Statistics DOI: 10.1080/02664763.2017.1364717.

Famoye F and Singh KP (2006). Zero-inflated generalized Poisson regression model with applications to domestic violence data. Journal of Data Science 4, 117–130.

Ghosh SK, Mukhopadhyay P and Lu JC (2006). Bayesian analysis of zero-inflated regression models. Journal of Statistical Planning and Inference 136,1360–1375.

Gilks WR, Best NG and Tan KKC (1995). Adaptive rejection Metropolis sampling within Gibbs sampling. Journal of the Royal Statistical Society. Series C 44, 455–472.

Gilks WR and Wild P(1992). Adaptive rejection sampling for Gibbs sampling. Journal of the Royal Statistical Society. Series C 41, 337–348.

Heagerty PJ (1999). Marginally specified logistic-normal models for longitudinal binary data. Biometrics 55, 688–698. [PubMed: 11314994]

Heilbron D (1994). Zero-altered and other regression models for count data with added zeros. Biometrical Journal 36, 531–547.

Ibrahim JG (1990). Incomplete data in generalized linear models. Journal of the American Statistical Association 85,765–769.

Ibrahim JG, Chen M-H and Lipsitz SR (1999). Monte Carlo EM for missing covariates in parametric regression models. Biometrics 55,591–596. [PubMed: 11318219]

Ibrahim JG,Chen M-H, Lipsitz SR and Herring AH (2005). Missing data methods for generalized linear models: A comparative review. Journal of the American Statistical Association 100,332–346.

Ibrahim JG, Lipsitz SR and Chen M-H (1999b). Missing covariates in generalized linear models when the missing data mechanism is nonignorable. Journal of the Royal Statistical Society. Series B 61,173–190.

Inan G, Preisser J and Das K (2017). A Score Test for Testing a Marginalized Zero-Inflated Poisson Regression Model Against a Marginalized Zero-Inflated Negative Binomial Regression Model. Journal of Agricultural, Biological, and Environmental Statistics DOI: 10.1007/s13253-017-0314-5.

Kleinke K and Reinecke J (2013). Multiple imputation of incomplete zero-inflated count data.Statistica Neerlandica 67, 311–336.

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

Lee AH, Wang K and Yau KKW (2001). Analysis of zero-inflated Poisson incorporating extent of exposure. Biometrical Journal 43, 963975.

Lewsey JD and Thomson WM (2004). The utility of the zero-inflated Poisson and zero-inflated negative binomial models: a case study of cross-sectional and longitudinal DMF data examining the effect of socio-economic status. Community Dentistry and Oral Epidemiology 32, 183–9. [PubMed: 15151688]

Lipsitz SR and Ibrahim JG (1996). A conditional model for incomplete covariates in parametric regression models. Biometrika 83, 916–922.

Little RJA and Rubin DB (2002). Statistical analysis with missing data (2nd ed.) New York: Wiley.

Little RJA and Schluchter M (1985). Maximum likelihood estimation for mixed continuous and categorical data with missing values. Biometrika 72, 497–512.

Long DL, Preisser JS, Herring AH and Golin CE (2014). A marginalized zero-inflated Poisson regression model with overall exposure effects. Statistics in Medicine 33, 5151–5165. [PubMed: 25220537]

Long DL, Preisser JS, Herring AH and Golin C (2015). A Marginalized Zero-Inflated Poisson Regression Model with Random Effects. Journal of the Royal Statistical Society, Series C 64, 815–830.

Louis T (1982). Finding the observed information matrix when using the EM algorithm. Journal of the Royal Statistical Society. Series B 44, 226–233.

McCulloch C (1997). Maximum likelihood algorithms for generalized linear mixed models. Journal of the American Statistical Association 92, 162–170.

McCulloch C and Searle S (2001). Generalized, linear, and mixed models New York: Wiley.

Mullahy J (1986). Specification and testing of some modified count data models. Journal of Econometrics 33, 341–365

Mwalili SM, Lesaffre 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. [PubMed: 17698937]

Pahel BT, Preisser JS, Stearns SC, Rozier RG (2011). Multiple imputation of dental caries data using a zero inflated Poisson regression model. Journal of Public Health Dentistry 71, 71–78. [PubMed: 20880027]

Petris G(2014). Bayesian and Likelihood Analysis of Dynamic Linear Models. R package version 1.1–4 Retrieved from https://cran.r-project.org/web/packages/dlm/index.html

Preisser JS, Das K, Long DL and Divaris K (2016). A Marginalized zero-inflated negative binomial regression model with application to dental caries. Statistics in Medicine 35, 1722–1735. [PubMed: 26568034]

Preisser JS, Long DL and Stamm JW (2017). Matching the Statistical Model to the Research Question for Dental Caries Indices with Many Zero Counts". Caries Research 51:198–208. [PubMed: 28291962]

Preisser JS, Stamm JW, Long DL and Kincade M (2012). Review and recommendations for zero-inflated count regression modeling of dental caries indices in epidemiological studies. Caries Research 46, 413–423. [PubMed: 22710271]

Rakhmawati TW, Molenberghs G, Verbeke G and Faes C (2016). Local influence diagnostics for hierarchical count data models with overdispersion and excess zeros. Biometrical Journal 58, 1390–1408. [PubMed: 27356111]

Smith VA, Preisser JS, Neelon B and Maciejewski ML (2014). A marginalized two-part model for semicontinuous data. Statistics in Medicine 33 4891–4903. [PubMed: 25043491]

Todem D, Kim K, and Hsu WW (2016). Marginal mean models for zero-inflated count data. Biometrics DOI: 10.1111/biom.12492.

Ugarte MD, Iba´n´ez B and Militino AF (2004). Testing for Poisson Zero Inflation in Disease Mapping. Biometrical Journal 46, 526–539.

Wang Z, Ma S, and Wang C-Y (2015).Variable selection for zero-inflated and overdispersed data with application to health care demand in Germany. Biometrical Journal 57, 867–884. [PubMed: 26059498]

Wei GCG and Tanner MA (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. Journal of the American Statistical Association 85, 699704.

Yau KKW, Wang K and Lee AH (2003). Zero-Inflated Negative Binomial Mixed Regression Modeling of Over-Dispersed Count Data with Extra Zeros. Biometrical Journal 45, 437–452.

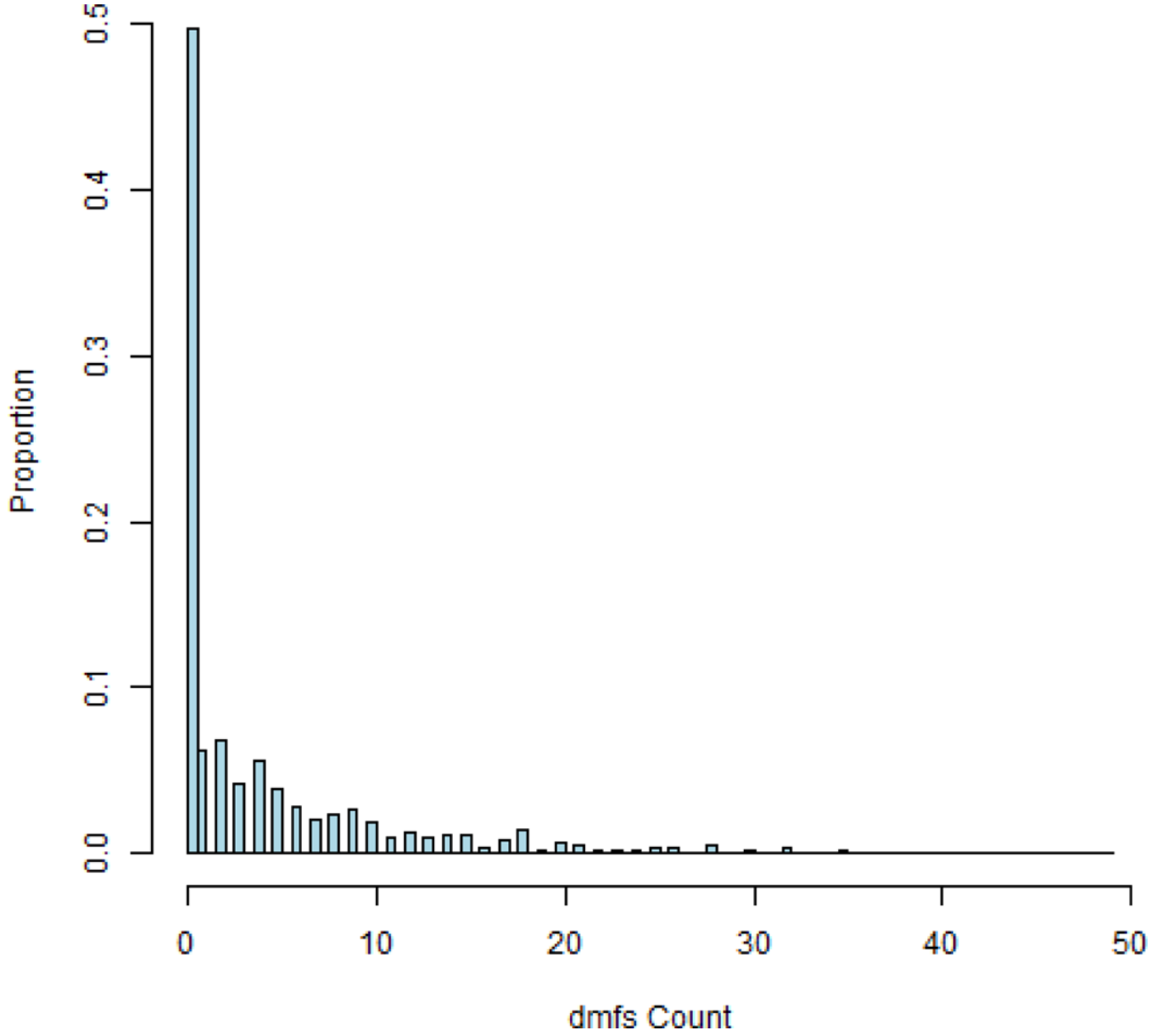

Figure 1:
Distribution of dmfs counts from 1094 children grades 1 to 5 participating in a school-based fluoride mouthrinse program.

Table 1:

Comparison of MCEM and CC estimates with one potentially missing covariate and one always observed covariate.

		MCEM				Complete Case			
S Size	[1]Parm	Percent Rel. Bias	Sim. Std	Mean SE	MSE	Percent Rel. Bias	Sim. Std	Mean SE	MSE
250	α_0	−0.7	0.185	0.184	0.034	34.6	0.171	0.165	0.149
	α_1	1.2	0.149	0.151	0.022	7.9	0.142	0.139	0.026
	α_2	0.7	0.174	0.184	0.030	25.8	0.171	0.169	0.096
	γ_0	0.2	0.232	0.232	0.054	−52.7	0.251	0.239	0.340
	γ_1	1.1	0.215	0.211	0.046	4.8	0.244	0.189	0.062
	γ_2	1.5	0.254	0.258	0.065	−26.5	0.297	0.254	0.158
500	α_0	−1.3	0.132	0.130	0.018	34.3	0.122	0.116	0.133
	α_1	−0.3	0.107	0.106	0.012	6.6	0.103	0.098	0.015
	α_2	−0.2	0.128	0.128	0.016	24.7	0.123	0.119	0.076
	γ_0	0.8	0.161	0.162	0.026	−52.3	0.173	0.167	0.304
	γ_1	1.3	0.144	0.139	0.021	4.1	0.162	0.129	0.028
	γ_2	1.4	0.173	0.172	0.030	−24.9	0.200	0.174	0.102
1000	α_0	−1.2	0.094	0.091	0.009	34.3	0.086	0.082	0.125
	α_1	−0.6	0.076	0.075	0.006	6.3	0.072	0.069	0.009
	α_2	−0.5	0.090	0.091	0.008	24.6	0.084	0.084	0.067
	γ_0	1.1	0.115	0.114	0.013	−52.0	0.119	0.118	0.284
	γ_1	1.3	0.098	0.094	0.010	4.1	0.109	0.089	0.014
	γ_2	1.4	0.118	0.118	0.014	−24.4	0.128	0.121	0.076

−22Parm = True values: $\{\gamma_0, \gamma_1, \gamma_2\}$ = (1,−1, 1), and $\{\alpha_0, \alpha_1, \alpha_2\}$ = (1,−1, 1) On average 34.4%, 34.5% and 34.5% of the observations with a missing covariate value corresponding to sample size 250, 500 and 1000.

Table 2:

Comparison of MCEM and CC estimates for scenario where two covariates are potentially missing and one covariate is always observed ($n = 1000$).

	MCEM				Complete Case			
Par	**Percent Rel. Bias**	**Sim. Std**	**Mean SE**	**MSE**	**Percent Rel. Bias**	**Sim. Std**	**Mean SE.**	**MSE**
Scenario 1: Mean= 41.0 % of observations missing at least one covariate value and true model is MZIP, 45.6% excess zeros								
α_0	0.1	0.092	0.088	0.008	−87.6	0.156	0.157	0.216
α_1	1.5	0.070	0.066	0.005	−65.3	0.084	0.090	0.113
α_2	2.2	0.073	0.072	0.005	−64.9	0.094	0.099	0.114
α_3	1.3	0.079	0.075	0.006	−38.3	0.108	0.110	0.048
γ_0	2.7	0.123	0.115	0.015	83.2	0.193	0.173	0.210
γ_1	9.0	0.100	0.088	0.012	74.4	0.108	0.101	0.150
γ_2	6.2	0.099	0.095	0.011	74.4	0.121	0.111	0.153
γ_3	2.7	0.094	0.088	0.009	47.9	0.130	0.115	0.074
Scenario 2: Mean= 30.1% of observations missing at least one covariate value and true model is MZIP, 40.2% excess zeros								
α_0	−1.4	0.087	0.084	0.008	35.3	0.084	0.080	0.038
α_1	−1.4	0.060	0.059	0.004	8.2	0.056	0.054	0.005
α_2	0.1	0.062	0.065	0.004	9.1	0.058	0.059	0.005
α_3	−1.1	0.071	0.070	0.005	35.5	0.074	0.068	0.037
γ_0	1.1	0.110	0.106	0.012	−66.4	0.122	0.115	0.125
γ_1	1.7	0.073	0.073	0.005	2.1	0.077	0.073	0.006
γ_2	0.2	0.075	0.079	0.006	1.1	0.080	0.081	0.007
γ_3	1.3	0.080	0.080	0.006	−36.7	0.097	0.089	0.043

Table 3:

Parameter estimates and standard errors from the models of missing covariates years and income in the NC FMR data analysis.

		MCEM		Complete Case	
Variable	**Parameter**	**Estimate**	**SE**	**Estimate**	**SE**
Model for years					
Intercept	ξ_{10}	1.437	0.143	1.516	0.131
African American	ξ_{12}	0.250	0.091	0.300	0.100
Other race	ξ_{13}	−0.127	0.191	−0.105	0.218
Fam. income	ξ_{15}	−0.001	0.031	−0.013	0.021
Dental home	ξ_{16}	0.035	0.099	−0.007	0.109
No access	ξ_{17}	−0.129	0.091	−0.222	0.102
Age	ξ_{18}	−0.677	0.029	−0.675	0.034
Model for income					
Intercept	ξ_{50}	−1.183	0.094	−1.184	0.119
African American	ξ_{52}	0.335	0.077	0.321	0.097
Other race	ξ_{53}	0.216	0.174	0.239	0.209
Dental home	ξ_{56}	−0.273	0.091	−0.322	0.115
No access	ξ_{57}	0.221	0.088	0.210	0.108
Age	ξ_{58}	0.019	0.020	0.023	0.024

Table 4:

MZIP estimates and standard errors for the NC FMR data from MCEM, multiple imputation and complete case analyses.

	MCEM		Multiple Imputation		Complete Case	
Variable	**Estimate**	**SE**	**Estimate**	**SE**	**Estimate**	**SE**
Marginal mean model						
Intercept	1.726	0.144	1.680	0.164	1.468	0.166
Years	−0.099	0.094	−0.015	0.106	−0.084	0.078
African American	−0.451	0.073	−0.380	0.082	−0.464	0.080
Other race	−0.598	0.178	−0.622	0.191	−0.974	0.268
Brushing freq.	−0.095	0.046	−0.106	0.051	−0.017	0.054
Fam. income	−0.196	0.016	−0.153	0.020	−0.213	0.018
Dental home	0.307	0.078	0.254	0.084	0.359	0.089
No access	0.316	0.070	0.273	0.078	0.426	0.069
Age	−0.035	0.045	−0.026	0.047	−0.120	0.046
Age-sq	−0.037	0.014	−0.024	0.014	−0.044	0.014
Age-cu	−0.033	0.009	−0.035	0.009	−0.019	0.009
Sealants	0.771	0.052	0.675	0.073	0.954	0.072
Zero-inflation model						
Intercept	−1.707	0.347	−1.128	0.365	−1.229	0.402
Years	−0.010	0.164	−0.056	0.147	0.138	0.154
African American	0.687	0.154	0.401	0.169	0.711	0.172
Other race	1.300	0.277	1.157	0.286	1.875	0.418
Brushing freq.	0.196	0.098	0.207	0.109	0.063	0.126
Fam. income	0.440	0.028	0.259	0.042	0.428	0.036
Dental home	−0.148	0.148	−0.009	0.165	−0.276	0.201
No access	−0.041	0.151	−0.056	0.165	−0.472	0.178
Age	−0.071	0.076	−0.120	0.082	0.105	0.088
Age-sq	0.026	0.026	−0.010	0.026	0.053	0.029
Age-cu	0.062	0.013	0.066	0.014	0.035	0.016
Sealants	−1.290	0.099	−1.030	0.142	−1.434	0.144