MARGINALLY-SPECIFIED MEAN MODELS FOR COUNTS WITH MIXTURE DISTRIBUTIONS

Habtamu Kassa Benecha

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics in the Gillings School of Global Public Health.

Chapel Hill
2016

Approved by:

John Preisser

Amy Herring

Brian Neelon

Kimon Divaris

Donglin Zeng

ABSTRACT

Habtamu Kassa Benecha: Marginally-specified Mean Models for Counts with Mixture Distributions
(Under the direction of John Preisser)

Counts from heterogeneous populations are often modeled using mixture distributions. These models assume that observations are generated from multiple unobserved subpopulations and estimate parameters having latent class interpretations. When interest is to make inferences about marginal means and incidence density ratios for the effects of risk factors in the overall population, regression coefficients obtained from common mixture models do not provide direct interpretations for these population-level parameters. While indirect techniques such as the use of post-modeling transformations may be employed to estimate the marginal effects of explanatory variables of interest, there are many instances where latent class model formulations fail to fully explain relationships between covariates and population-wide parameters (Preisser et al., 2012; Long et al., 2014). First, we employ two-component mixtures of non-degenerate count data distributions to estimate the overall effects of exposure variables on marginal means of zero-inflated and other heterogeneous counts. The models are examined using simulations and further applied to a double-blind dental caries incidence trial. Next, we develop a marginalized model for bivariate zero-inflated counts that allows the estimation of parameters for the overall effects of exposure variables on the marginal means of the two correlated outcomes. The model employs four-component mixture distributions and estimates marginally interpretable regression coefficients. We demonstrate the application of the method by using simulations and dental caries indices of primary and permanent teeth among children from a school-based fluoride mouthrinse study. Finally, extending earlier approaches, we propose an estimation

method for marginalized zero-inflated count models when covariates are missing at random. The method, which can also be applied to other missing data problems, is illustrated and compared with complete case analysis by using simulations and dental data.

To my mother, Serkalem Bogale.

ACKNOWLEDGMENTS

The greatest thanks go to my advisor Dr. John Preisser for his guidance and constant support in my research endeavors. I am very grateful for his insight, encouragement, patience, and for helping me grow as a biostatistician. I truly feel fortunate to have had the chance to work closely with him. I would also like to thank Dr. Amy Herring for her support in my research and for her invaluable contributions to my progress from my first year at UNC all the way through the completion of this dissertation.

Furthermore, I want to thank Drs. Brian Neelon, Kimon Divaris, Donglin Zeng and Kalyan Das for their insightful comments and their time. Special thanks go to Dr. Chirayath Suchindran for his encouragement, advice and support. I would also like to thank Dr. Lloyd Edwards for his mentorship and support, particularly during my first two years at UNC. Furthermore, I would like to thank Melissa Hobgood and the students, staff and faculty of the Department of Biostatistics for all the help I have received. Thanks also to the National Institute of Environmental Health Sciences and the Gary G. Koch Scholars Program for their financial support.

Finally, I would like to thank my family and friends for their love and for always being there for me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: LITERATURE REVIEW

1.1 Introduction

The analysis of counts generated from heterogeneous populations present special challenges to researchers. When data arise from several unobserved subpopulations, models based on standard probability distributions are often inadequate to explain observed variabilities (Wedel and DeSarbo, 1995; Frühwirth-Schnatter, 2005). One example would be the case of zero-inflated counts, where proportions of zero observations are higher than expected under standard distributions. Employing traditional distributions (such as the Poisson) to model such data often results in biased estimates and poor predictions (Lambert,1992). Instead, zero-inflated counts are commonly modeled by using two-component mixture distributions, hypothesizing that observations arise from two latent classes within the source population: one class provides only zeros and the other produces both zero and non-zero values. Such an approach is under the framework of finite mixture modeling, which partitions a source population into a number of unobserved classes or subpopulations and estimates parameters specific to the latent classes. Common models for counts with excess zeros such as zero-inflated Poisson (ZIP) regression utilize two-component mixtures consisting of a degenerate zero and a standard count distribution.

As in the univariate case, bivariate count outcomes with many zeros are commonly modeled through the use of mixture distributions that account for zero-inflation as well as the dependence between the outcomes. For example, Wang et al.(2003) employ a mixture of a bivariate Poisson distribution with a point mass at (0,0) to model counts of occupational injuries, and Li et al.(1999) propose a four-component mixture distribution for modeling

bivariate zero-inflated counts.

Despite the flexibility that mixture distributions provide in modeling highly dispersed count data, interpretations of the regression parameters from such models are limited to the latent classes making up the study population. These parameters are not directly applicable to making inferences about the overall effects of covariates on the marginal mean. Even with the application of indirect methods of parameter estimation such as the use of post-modeling transformations, there are many instances where latent class model formulations fail to fully explain relationships between covariates and population-wide parameters (Preisser et al.,2012; Long et al., 2014).

The importance of models with marginally interpretable parameters for zero-inflated counts has long been recognized (Lambert, 1992; Long et al.,2014; Preisser et al.,2012, 2016; Albert et al., 2014). While the literature is scarce for bivariate zero-inflated counts, the development of marginalized models for univariate zero-inflated counts has been given attention in recent years. Based on the framework of the zero-inflated Poisson model likelihood function, Long et al.(2014) propose a maximum likelihood method to estimate regression parameters for marginal means of counts with excess zeros. Marginalized zero-inflated negative binomial models (Preisser et al., 2016) further estimate overall effects of covariates on marginal means of counts with zero-inflated negative binomial distributions. Todem et al.(2016) provide a general representation of two-part marginalized mean count models including distributions for bounded counts, e.g., the zero-inflated beta binomial distribution.

All these marginalized models assume that the count outcomes follow two-component mixtures consisting of a standard count distribution with a point-mass at zero. However, models employing degenerate distributions are sometimes inadequate to describe marginal means of counts from multiple source populations; data-generating mechanisms based on mixtures of non-degenerate count distributions could provide better fits for count data. In the first part of the dissertation, we expand the class of marginalized mixture models for zero-inflated and other heterogeneous count data to allow for greater model choice with

maximum likelihood estimation. In the second part, we propose a marginalized model for bivariate zero-inflated counts that provides directly interpretable regression parameters for the marginal means of the two correlated outcomes in the overall population.

While much of the statistical literature on zero-inflated data modeling treats covariates and outcomes as fully observed, missing data are a common occurrence in practice. 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, often referred to as complete case analysis, is valid only when the probability of missingness is independent of any observed and unobserved information. Even when complete case analysis is valid, estimates can be inefficient if too many observations are missing (Ibrahim et al., 1999, 2005). For problems where covariates are missing at random and their conditional distribution is log-concave, Ibrahim et al.(1999) propose a Monte Carlo EM (Wei and Tanner, 1990) algorithm to allow for maximum likelihood 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. In the third part of the dissertation, we extend the Monte Carlo EM approach to marginalized zero-inflated Poisson models with missing covariates

We conduct literature review in the remainder of Chapter 1. In Chapter 2, new marginalized models are developed for univariate zero-inflated and other heterogeneous count data. In Chapter 3, a marginalized model is proposed for two correlated count outcomes with excess zeros and Chapter 4 presents a Monte Carlo EM method for handling missing covariates in marginalized zero-inflated Poisson models. We present a conclusion in Chapter 5.

1.2 Mixture Models

Mixture distributions have been used to model observations with variabilities that are insufficiently explained by standard statistical models. An underlying assumption of such models is that variability in observations is due mainly to heterogeneity within the sampled population, which may contain a number of unobserved subpopulations of unknown proportions (Wedel and DeSarbo, 1995). In discrete modeling, a simple but popular mixture is that of the Poisson and gamma distributions (i.e, the negative binomial), which is commonly used to model counts with extra-Poisson dispersion. The Poisson and negative binomial distributions are also often mixed with a distribution degenerate at zero to model counts with much higher proportions of zeros than expected under either of these two standard distributions. These models presume that observations arise from a population containing two unobserved subpopulations; while one subpopulation produces only zero counts, observations from the other subpopulation can have zero or positive values. Because such assumptions lead to data generating mechanisms that conveniently explain heterogeneities in counts in various research problems, the two component mixture model has been given a lot of attention over the past few decades (Lambert, 1992; Mullahy, 1986; Heilbron, 1994; Böhning et al., 1999). Mixtures involving more than two component distributions have also been applied in the health sciences, medicine, genetics, economics, ecology and other areas (Wang et al., 1996, Morgan et al., 2014).

Finite mixture models partition a source population into $m \geq 2$ latent subpopulations and assume that the random variable of interest takes a value from the j^{th} subpopulation with a probability π_j. If Y_i is count random variable with observed value y_i, an m component mixture distribution can be defined for Y_i as (Frühwirth-Schnatter, 2005)

$$Pr(Y_i = y_i|\boldsymbol{\pi}, \boldsymbol{\theta}_i) = \sum_{j=1}^{m} \pi_j f_j(y_i|\boldsymbol{\theta}_{ij}), \tag{1.1}$$

where the components f_1, f_2, ..., f_m are probability mass functions of known distributions,

$\boldsymbol{\theta}_{ij}$ is the vector of parameters in f_j, $\boldsymbol{\theta}_i = (\boldsymbol{\theta}_{i1}, \boldsymbol{\theta}_{i2}, ..., \boldsymbol{\theta}_{im})$, and $\boldsymbol{\pi} = (\pi_1, \pi_2 ..., \pi_m)'$ is a vector of mixing probabilities with $0 \leq \pi_j \leq 1$ and $\sum_{j=1}^{m} \pi_j = 1$. The latent parameters π_j and $\boldsymbol{\theta}_{ij}$ corresponding to the j^{th} component are also estimated either as constants or as functions of covariates through convenient link functions. For example, if $\boldsymbol{\theta}_{ij}$ is a scalar and $\mathbf{x}_i$ is a vector of covariates from the i^{th} subject, then $\boldsymbol{\theta}_{ij}$ can be related to the covariates as

$$\boldsymbol{\theta}_{ij}(\boldsymbol{\beta}_j) = g^{-1}(\mathbf{x}_i' \boldsymbol{\beta}_j), \tag{1.2}$$

where $\boldsymbol{\beta}_j$ is a vector of regression parameters corresponding to the j^{th} component and g is a link function. While the mixture model in equation (1.1) imposes heterogeneity only through $f_j(y_i|\boldsymbol{\theta}_{ij})$, the mixing probabilities (i.e., π_j) may also be allowed to vary across individuals.

1.2.1 Poisson and Negative Binomial Mixtures

Finite Poisson mixtures are one of the popular mixture models for count data. In these models, $f_j, j = 1, 2, ..., m$ in equation (1.1) has the form

$$f_j(y_i|\mu_{ij}) = \frac{e^{-\mu_{ij}} \mu_{ij}^{y_i}}{y_i!}, \tag{1.3}$$

where μ_{ij} is a mean parameter. While earlier applications of Poisson mixtures estimate model parameters π_j and μ_{ij} as constants, Wang et al.(1996) introduce covariates to model the latent class mean parameters as

$$log(\mu_{ij}) = \mathbf{x}_i' \boldsymbol{\beta}_j, \; j = 1, 2, ..., m, \tag{1.4}$$

where $\mathbf{x}_i$ and $\boldsymbol{\beta}_j$ are as defined in equation (1.2). The model, which estimates the mixing parameters as constants, is identifiable when the design matrix is full rank. Wang et al.(1996)

implement the expectation maximization (EM) algorithm together with quasi-Newton maximization to perform estimation.

To account for extra-Poisson dispersion within each latent subpopulation, Ramaswamy, Anderson and DeSarbo (1994) propose negative binomial mixture models, for which the component distributions in equation (1.1) are negative binomial. That is,

$$f_j(y_i|\boldsymbol{\theta}_{ij}) = \frac{\Gamma(y_i+\alpha_j)}{y_i!\Gamma(\alpha_j)}\Big(\frac{\alpha_j}{\alpha_j+\mu_{ij}}\Big)^{\alpha_j}\Big(\frac{\mu_{ij}}{\alpha_j+\mu_{ij}}\Big)^{y_i}, \tag{1.5}$$

where μ_{ij} is the mean parameter, α_j is the dispersion parameter and $\boldsymbol{\theta}_{ij} = (\mu_{ij}, \alpha_j)$. Ramaswamy, Anderson and DeSarbo (1994) model the mean parameters as functions of covariates and estimate the mixing probabilities and dispersion parameters as constants using the EM algorithm.

1.3 Analysis of Zero-inflated Counts

Oftentimes, counts collected in various research areas contain high proportions of zeros. One such area is dental caries research, where counts of decayed, missing and filled teeth (dmfs) are increasingly characterized by disproportionately high numbers of zeros (Lewsey and Thompson, 2004; Mwalili et al., 2008; Preisser et al., 2012; Albert et al., 2014). Because of the excess number of zero observations relative to what is expected under standard probability distributions, traditional generalized linear models do not sufficiently explain variability in such counts. For instance, while the Poisson distribution assumes equality of means and variances, the variances of zero-inflated counts are generally larger than the corresponding means. As a result, Poisson regression models tend to underestimate proportions of zeros and those of large positives when fitted to counts with excess zeros (Lambert, 1992).

Over the past few decades several methods have been proposed for the analysis of zero-inflated data (Lambert, 1992; Mullahy, 1986; Heilbron, 1994; Böhning et al., 1999). Most

of these models assume that counts originate from two latent subpopulations, and can in general be divided into two categories depending on how they treat the generation of zero and positive counts from the two latent groups. The first category of models, often called zero-inflated models (Long et al., 2014), presume that both zero and positive counts arise from one latent subpopulation according to a standard probability distribution, but extra zeros come from a second latent subpopulation based a distribution degenerate at zero. Zero-inflated Poisson (ZIP) regression is one of such models, and has been increasingly popular after Lambert (1992) described the data generating processes and applied it to defects in manufacturing processes. When zero-inflated counts show variabilities that are not attributed to excess zeros, the Poisson distribution in ZIP is often replaced by a negative binomial probability function, resulting in the zero-inflated negative binomial (ZINB) model. Hurdle or zero-altered models (Mullahy, 1986) comprise of the second category of estimation methods for zero-inflated data, where zero and positive counts are considered to come from two separate latent subpopulations. In hurdle models, regression parameters are often specified for the logit of the probability of a count being positive and the mean of the untruncated version of the distribution assumed for positive counts.

1.3.1 ZIP and ZINB Regression Models

Zero-inflated Poisson models assume that a count random variable follows a mixture of a Poisson distribution with a point mass at zero. Observations are thought of as arising from two different sources: while an 'imperfect' or 'susceptible' subpopulation gives rise to zero and positive counts based on a Poisson distribution, a 'perfect' or 'non-susceptible' subpopulation produces excess zero counts (Lambert, 1992; Long et al.,2014; Preisser et al., 2012). In dental caries studies among children, the 'non-susceptible' group can be considered to be the population of children not at risk of caries, from which only zero dmfs counts can be recorded. On the other hand, children in a 'susceptible' or at 'caries-risk' population can have zero or positive dmfs counts (Preisser et al., 2012). Given a sample of size n, ZIP

assumes that the random variable Y_i, $i = 1, 2, ..., n$, takes zero or positive values as follows (Long et al., 2014).

$$Y_i \sim \begin{cases} 0, & \text{with probability } \psi_i \\ \text{Poisson}(\mu_i), & \text{with probability } 1 - \psi_i \end{cases} \tag{1.6}$$

In (1.6), ψ_i is the probability of being from the 'perfect' or 'non-susciptible' subpopulation, and μ_i is the mean of the Poisson distribution corresponding to the 'imperfect' or 'susceptible' group. Considering ψ_i as a mixing probability, the distribution of Y_i can be written in the form of equation (1.1) as

$$Pr(Y_i = k) = \psi_i I(k = 0) + (1 - \psi_i) g(k|\mu_i), k = 0, 1, 2, ..., \tag{1.7}$$

where g is the Poisson mass function and $I(T)$ is an indicator variable taking the value 1 when T is true and the value 0 when T is false. Clearly, when the mixing parameter ψ_i is zero, ZIP reduces to the standard Poisson model. By using the logit and the log links, Lambert (1992) allows the probability of membership in the 'perfect' state, ψ_i, and the Poisson mean, μ_i, to depend on covariates as

$$logit(\psi_i) = \mathbf{z}_i'\boldsymbol{\gamma} \quad \text{and} \quad log(\mu_i) = \mathbf{x}_i'\boldsymbol{\beta} \tag{1.8}$$

In (1.8), $\mathbf{z}_i$ and $\mathbf{x}_i$ are $q \times 1$ and $p \times 1$ vectors of covariates for the i^{th} subject, and $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_q)$ and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$ are regression parameters. Usually, the set of covariates in $\mathbf{z}_i$ is a subset of those in $\mathbf{x}_i$.

The variance and the marginal mean of a ZIP random variable Y_i are, $Var(Y_i|\mathbf{z}_i, \mathbf{x}_i) = \mu_i(1 - \psi_i) + \mu_i^2\psi_i(1 - \psi_i)$ and $E(Y_i|\mathbf{z}_i, \mathbf{x}_i) = \mu_i(1 - \psi_i)$ (Böhning, 1999; Long et al., 2014). While the mean and the variance are equal when $\psi_i = 0$ (i.e., for standard Poisson models), the variance is always greater than the mean for zero inflated counts (i.e., when $\psi_i > 0$).

For problems where ψ_i and μ_i are believed to be related, Lambert (1992) specifies shared regression coefficients to model the two latent parameters.

$$log(\mu_i) = \mathbf{x}_i'\boldsymbol{\beta} \quad \text{and} \quad logit(\psi_i) = \tau\mathbf{x}_i'\boldsymbol{\beta}, \tag{1.9}$$

where τ is a parameter to be estimated. Note that the specification in (1.9) reduces the number of regression parameters by almost half.

To estimate the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ in equation (1.8), Lambert (1992) employs the EM algorithm on a complete data log-likelihood function involving a binary latent variable that defines membership in either of the two latent subpopulations. For the shared parameter ZIP model in equation (1.9), estimation is performed using Newton-Raphson algorithm.

Zero-inflated negative binomial models are similarly formulated as ZIP by using a negative binomial probability mass function g in (1.7). In addition to zero-inflation, ZINB models allow for the handling of overdispersion caused by unobserved heterogeneities.

1.4 Models for Bivariate Zero-inflated Counts

While much of the literature on zero-inflated counts is focused on univariate outcomes, studies sometimes involve two or more correlated and zero-inflated count variables (Divaris et al., 2012; Li et al., 1999; Wang et al., 2003). When two dependent random variables take higher proportions of zeros than expected under standard bivariate count distributions, modeling requires accounting for zero-inflation and the dependence between the outcomes. To model counts of occupational injuries, Wang et al.(2003) employ a two-component mixture of a bivariate Poisson distribution with a point mass at (0,0) and perform estimation using the EM algorithm. Under this model, excess zeros arise from a 'non-susceptible' subpopulation with a probability ψ, and with probability $1 - \psi$, components of the bivariate outcome take zero and positive values from a 'susceptible' subpopulation according to a bivariate Poisson distribution. For general multivariate zero-inflated counts, Li et al.(1999)

propose mixtures comprising a multivariate distribution degenerate at zero values, a multivariate Poisson distribution and a number of univariate Poisson distributions. For the bivariate case, they assume that a zero-inflated random variable (Y_1, Y_2) arises either from a distribution degenerate at $(0,0)$, from a bivariate Poisson distribution, or from a bivariate distribution with one component degenerate at 0 and the other having a standard Poisson mass function. That is,

$$(Y_1, Y_2) \sim \begin{cases} (0,0), \text{with probability } p_0 \\ (\text{Poisson}(\lambda_1), 0), \text{with probability } p_1 \\ (0, \text{Poisson}(\lambda_2)), \text{with probability } p_2 \\ \text{Bivariate Poisson}(\lambda_{10}, \lambda_{20}, \lambda_{00}), \text{ with probability } p_3, \end{cases} \tag{1.10}$$

where $p_k \geq 0$, $k = 0,1,2,3$, $\sum_{k=0}^{3} p_k = 1$, and λ_1, λ_2, λ_{10}, λ_{20}, $\lambda_{00} > 0$. The bivariate distribution in (1.10) reduces to the standard bivariate Poisson model for $p_0 = p_1 = p_2 = 0$. When $\lambda_1 = \lambda_{10} + \lambda_{00}$ and $\lambda_2 = \lambda_{20} + \lambda_{00}$ in equation (1.10), the marginal distributions of Y_1 and Y_2 become univariate ZIP. That is,

$$\Pr(Y_t = k) = \begin{cases} (1 - p_t - p_3) + (p_t + p_3)\exp(-\lambda_t), & k = 0 \\ (p_t + p_3)\frac{\exp(-\lambda_t)\lambda_t^k}{k!}, & k = 1, 2, ... \end{cases} \tag{1.11}$$

where $t = 1, 2$. Li et al.(1999) employ directional grid search approaches (Powell, 1964) and methods of moments to obtain maximum likelihood estimates of model parameters.

When covariates are used to model bivariate zero-inflated Poisson counts, linear predictors are specified for the mean parameters and the mixing probabilities, for example, as $\log(\lambda_{10i}) = \mathbf{x}'_{1i}\boldsymbol{\alpha}_1$, $\log(\lambda_{20i}) = \mathbf{x}'_{2i}\boldsymbol{\alpha}_2$, $\log(\lambda_{00i}) = \mathbf{x}'_{3i}\boldsymbol{\alpha}_3$, $\log(p_{0i}/p_{3i}) = \mathbf{x}'_{4i}\boldsymbol{\gamma}_0$, $\log(p_{1i}/p_{3i}) = \mathbf{x}'_{5i}\boldsymbol{\gamma}_1$ and $\log(p_{2i}/p_{3i}) = \mathbf{x}'_{6i}\boldsymbol{\gamma}_2$, where $\mathbf{x}_{1i}$, ..., $\mathbf{x}_{6i}$ are vectors of covariates from the i^{th} individual, and $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$, $\boldsymbol{\alpha}_3$, $\boldsymbol{\gamma}_0$, $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ are vectors of parameters (Li et al., 1999; Majundar and Gries 2010). Because the model parameters have latent class interpretations, one has

to employ post-modeling transformations to estimate the effects of covariates on the overall population means $\nu_{1i} = E(Y_{1i})$ and $\nu_{2i} = E(Y_{2i})$. The marginal means and the model parameters can be related by

$$\nu_{1i} = (p_{1i} + p_{3i})(\lambda_{00i} + \lambda_{10i}) = \frac{(e^{\mathbf{x}'_{1i}\boldsymbol{\alpha}_1} + e^{\mathbf{x}'_{3i}\boldsymbol{\alpha}_3})(1 + e^{\mathbf{x}'_{5i}\boldsymbol{\gamma}_1})}{1 + e^{\mathbf{x}'_{4i}\boldsymbol{\gamma}_0} + e^{\mathbf{x}'_{5i}\boldsymbol{\gamma}_1} + e^{\mathbf{x}'_{6i}\boldsymbol{\gamma}_2}} \tag{1.12}$$

$$\nu_{2i} = (p_{2i} + p_{3i})(\lambda_{00i} + \lambda_{20i}) = \frac{(e^{\mathbf{x}'_{2i}\boldsymbol{\alpha}_2} + e^{\mathbf{x}'_{3i}\boldsymbol{\alpha}_3})(1 + e^{\mathbf{x}'_{6i}\boldsymbol{\gamma}_2})}{1 + e^{\mathbf{x}'_{4i}\boldsymbol{\gamma}_0} + e^{\mathbf{x}'_{5i}\boldsymbol{\gamma}_1} + e^{\mathbf{x}'_{6i}\boldsymbol{\gamma}_2}}$$

Although ν_{1i} and ν_{2i} could be estimated at fixed covariate values by using equations (1.12), the quantification of the relationship between covariates and the marginal means with suitable variance estimates may be difficult in practice. In addition, when interest is in determining whether the effects of an exposure on ν_{1i} or ν_{2i} are homogeneous across the levels of covariates, existing bivariate zero-inflated models usually do not provide the desired estimates as in the case of traditional zero-inflated models for univariate counts (Long et al., 2014).

1.5 Inference About the Overall Population

While traditional models for zero-inflated counts provide flexible frameworks of estimation, regression coefficients from these methods do not have straightforward interpretations in explaining the effects of covariates on the overall marginal mean count in the sampled population. The limitations of such modeling approaches in quantifying important population-level parameters has long been acknowledged (Preisser et al., 2012, Long et al., 2014, Albert et al., 2014). Lambert (1992) discusses the difficulty of predicting changes in the marginal mean, $E(Y) = (1 - \psi)\mu$, when an exposure variable increases both ψ and μ in ZIP models. Shortcomings of the latent coefficients in explaining exposure effects on population-wide parameters are not limited to marginal mean counts. When interest is in determining effects of an exposure variable on population level parameters such as incidence

density ratios, it has been indicated that ZIP and ZINB models may not always provide the desired estimates (Long et al., 2014). For example, consider a clinical trial where the ZIP regression in equation (1.8) is used to model a zero-inflated outcome variable with $\mathbf{z}_i = \mathbf{x}_i$. From the relation $\nu_i = \mu_i(1-\psi_i)$, where $\nu_i = E(Y_i|\mathbf{x}_i)$, the overall mean for the i^{th} subject is

$$\nu_i = \frac{e^{\mathbf{x}_i'\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i'\boldsymbol{\gamma}}} \tag{1.13}$$

The incidence density ratio (IDR_i) or the ratio of overall means corresponding to a one unit increase in the j^{th} exposure variable, x_{ij}, is (Long et al., 2014),

$$IDR_i = \frac{E(y_i|x_{ij}=c+1, \tilde{\mathbf{x}}_i' = \tilde{\mathbf{x}}_i')}{E(y_i|x_{ij}=c, \tilde{\mathbf{x}}_i' = \tilde{\mathbf{x}}_i')} = e^{\beta_j}\frac{1+\exp(c\gamma_j + \tilde{\mathbf{x}}_i'\tilde{\boldsymbol{\gamma}})}{1+\exp((c+1)\gamma_j + \tilde{\mathbf{x}}_i'\tilde{\boldsymbol{\gamma}})}, \tag{1.14}$$

where $\tilde{\mathbf{x}}_i$ is the vector of covariates without x_{ij}, c is a possible value of x_{ij} and $\tilde{\boldsymbol{\gamma}}$ is the vector of parameters in the logit model corresponding to $\tilde{\mathbf{x}}_i$ (Preisser et al., 2012; Long et al., 2014). When $\gamma_j \neq 0$ in equation (1.14), the estimate of IDR_i changes as the values of the covariates in $\tilde{\mathbf{x}}_i$ change. In other words, ZIP regression parameters do not allow the estimation of an overall constant incident density ratio when the exposure variable of interest is included in the logit model (Long et al., 2014).

In the literature, several approaches have been proposed for the estimation of overall effects of explanatory variables on population-level parameters. While many of these methods involve fitting traditional zero-inflated models and then using the estimates to describe the parameters of interest, more recent approaches specify regression coefficients directly for the marginal mean.

1.5.1 Estimation Based on Latent Coefficients

In the analysis of zero-inflated data, population-wide parameters have traditionally been estimated by exploiting latent coefficients obtained from ZIP, ZINB and similar models. Acknowledging the inadequacy of ZIP coefficients in determining changes in marginal mean

defect counts as levels of manufacturing settings change, Lambert (1992) estimates the overall population mean at a level of a categorical covariate by averaging the model estimated means across all design points sharing the specific level of the covariate. This way, comparisons are made among levels of a covariate with regard to the overall mean of manufacturing defects. Although the method can be employed for problems where all involved predictor variables are categorical, it may not be appropriate when the ZIP model includes one or more continuous covariates. In a further attempt to characterize the overall population mean, Böhning et al.(1999) propose large sample methods to construct $(1-\alpha)100\%$ confidence intervals for the marginal mean, ν, as $\bar{Y} \pm z_{(1-\frac{\alpha}{2})}\sqrt{\frac{Var(Y)}{n}}$, where $\bar{Y}$ is the observed mean count, $Var(Y)$ is the variance, n is the sample size and $z_{(1-\frac{\alpha}{2})}$ is the $1-\frac{\alpha}{2}$ quantile of the standard normal distribution.

Albert et al.(2014) propose two methods of assessing overall population exposure effects in clinical trials and observational studies using estimates from zero-inflated beta-binomial and negative binomial models. The first method, called average predicted value (APV), allows to estimate differences and ratios of the marginal means for exposed and non-exposed subjects. However, this approach is not directly applicable when the exposure variables are continuous and calculation of variances may not be straightforward even for binary covariates. Another limitation of the method is that distributions need to be assumed for the covariates. Although average exposure effects can be estimated using empirical distributions of the explanatory variables, the approach may not be generalizable to populations with other configurations of covariates (Preisser et al., 2016). The second method proposed by Albert et al.(2014) specifies the log link, instead of the logit, to model the probability of excess zeros in zero-inflated negative binomial and beta-binomial regressions. The use of the log link for ψ allows for the estimation of ratios of means for the exposed and non-exposed groups that are independent of any of the covariates. However, the log link may not be appropriate to model ψ, since it may give predicted values greater than 1.

1.5.2 Marginalized Models

To estimate directly interpretable regression parameters for marginal means of ZIP distributed counts, Long et al.(2014) propose marginalized zero-inflated Poisson (MZIP) models, where regression parameters are specified for the overall population mean as well as for the probability of being an excess zero. Preisser et al.(2016) extend MZIP models to handle counts with extra-Poisson dispersion in addition to zero-inflation, by using the ZINB likelihood function. As in MZIP, the marginalized zero-inflated negative binomial (MZINB) regression provides coefficients for the effects of covariates on the marginal means as well as for the excess zero probabilities.

Let Y_i be a random variable having a ZIP distribution with marginal mean ν_i and excess zero probability ψ_i. The MZIP model relates ν_i and ψ_i with covariates as (Long et al., 2014)

$$\begin{aligned} logit(\psi_i) &= \mathbf{z}_i'\boldsymbol{\gamma} \\ log(\nu_i) &= \mathbf{x}_i'\boldsymbol{\alpha}, \end{aligned} \tag{1.15}$$

In (1.15), $\mathbf{z}_i$ and $\mathbf{x}_i$ are $q \times 1$ and $p \times 1$ vectors of covariates, and the parameters in $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, ..., \gamma_q)'$ have the same interpretation as in standard ZIP models. Unlike ZIP models, however, parameters $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_p)'$ describe heterogeneity in the overall population mean, instead of the mean count for subjects in the 'susceptible' latent class. Since the mean μ_i of the Poisson part of ZIP and the overall mean ν_i are related by $\nu_i = (1-\psi_i)\mu_i = e^{\mathbf{x}_i'\alpha}$, to find the MZIP likelihood, Long et al.(2014) replace μ_i by $\frac{\nu_i}{1-\psi_i}$ in the ZIP likelihood function. Thus, for n independent subjects, the log-likelihood function for MZIP models is written as

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

The corresponding score equations are (Long et al., 2014),

$$\begin{aligned}\frac{\partial l(\boldsymbol{\alpha},\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} &= \sum_{i=1}^{n}\Big[\frac{I(y_i=0)\psi_i(1-\psi_i)^{-1}(e^{\nu_i(1-\psi_i)^{-1}}-\nu_i)}{\psi_i(1-\psi_i)^{-1}e^{\nu_i(1-\psi_i)^{-1}}+1} \\ &\quad + \psi_i(y_i-1)-I(y_i>0)\psi_i(1-\psi_i)^{-1}\nu_i\Big]\mathbf{z}_i' \\ \frac{\partial l(\boldsymbol{\alpha},\boldsymbol{\gamma})}{\partial \boldsymbol{\alpha}} &= -\sum_{i=1}^{n}\Big[\frac{I(y_i=0)\nu_i(1-\psi_i)^{-1}}{\psi_i(1-\psi_i)^{-1}e^{\nu_i(1-\psi_i)^{-1}}+1}-(y_i-\nu_i(1-\psi_i)^{-1})I(y_i>0)\Big]\mathbf{x}_i'\end{aligned} \tag{1.16}$$

Long et al.(2014) employ quasi-Newton optimization methods to obtain parameter estimates. 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, robust standard errors are estimated.

For MZINB models, in addition to the standard regression parameter specifications for ψ_i and ν_i as in (1.15), Preisser et al (2016) model ψ_i by using shared parameters from the linear predictor of ν_i as

$$\begin{aligned} logit(\psi_i) &= \gamma_0+\gamma_1(\mathbf{x}_i'\boldsymbol{\alpha}) \\ log(\nu_i) &= \mathbf{x}_i'\boldsymbol{\alpha}, \end{aligned} \tag{1.17}$$

where γ_0 and γ_1 are scalar parameters.

1.6 Missing Data

In the absence of straightforward methods and software to analyze incomplete data, modeling is often done by deleting all cases with missing values on any of the variables (Ibrahim et al., 2005). However, this approach, known as complete case (CC) analysis, is valid only when the probability of missingness is independent of any observed and unobserved data. Even when CC analysis is valid, estimates can be inefficient if too many observations are missing (Ibrahim et al., 2005). Other ad-hoc methods of handling missing data include filling in plausible values for the missing observations, available case analysis, dummy variable

adjustments, and variable deletion (Allison, 2002). The use of such methods, however, may result in biased estimates, reduced efficiency and model mis-specifications (Allison, 2002).

Over the past few decades, much attention has been given to missing data methods for a wide range of models. In general, such methods work under certain assumptions about the dependence of the missingness mechanism on observed and missing values of relevant variables. Based on the nature of missingness, Little and Rubin (2002) group missing data into three categories: missing completely at random (MCAR), missing at random (MAR) and not missing at random (NMAR). Under MCAR, missingness is independent of any observed or unobserved information, and the MAR assumption holds when missingness is independent of any unobserved data. NMAR has the weakest assumptions among the three categories, and assumes that the probability of missingness is dependent on missing data. In maximum likelihood estimation, when data are MAR and the model of interest and missingness mechanism have separate parameters, missingness is ignorable, meaning that estimation can be done without modeling the missing data mechanism (Ibrahim et al., 2005). However, NMAR data require specification of a model for the missingness mechanism as part of the estimation process (Ibrahim et al., 1999, 2005). Maximum likelihood methods for missing data often estimate model parameters either by directly maximizing the observed data likelihood or by using the expectation-maximization algorithm on a convenient complete data likelihood function (Allison, 2002). However, since computing and maximizing the observed data likelihoods is often difficult, many of maximum likelihood based missing data methods rely on the EM algorithm and related approaches.

1.6.1 EM Algorithm and Monte Carlo EM Methods

The EM algorithm (Dempster, Laird and Rubin, 1977) is a two-step iterative method of estimation widely used in missing data problems as well as in situations where direct maximizations of likelihoods are difficult. EM works by first constructing a complete data likelihood and then iteratively applying the expectation and the maximization steps until

convergence is attained. While the expectation or E-step of EM computes the expected value of the complete data log-likelihood conditional on the observed data and current parameter values, the maximization or M-step maximizes the expected log-likelihood. In situations where the E-step is difficult to compute, the Monte Carlo EM (MCEM) algorithm (Wei and Tanner, 1990) may be employed to estimate the log-likelihood numerically. Ibrahim et al.(1999) apply the method for missing covariates in parametric models by using samples obtained from the Gibbs sampler with adaptive rejection sampling (ARS) algorithm (Gilks and Wild, 1992). Following Ibrahim et al.(1999, 2005), we review the applications of EM and MCEM methods for missing covariate problems in count models. In the following discussions, the outcome variable is assumed to be fully observed, but covariates can have missing values for some of the the study subjects.

Suppose that $\mathbf{y} = (y_1, y_2, \ldots, y_n)'$ is a vector of independent count outcomes from n subjects. For the i^{th} subject, let $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})'$ be a $p \times 1$ vector of covariates. Because covariates are partially missing for some subjects, Ibrahim et al.(1999, 2005) write the covariate vector $\mathbf{x}_i$ as $\mathbf{x}_i = (\mathbf{x}_i^{obs}, \mathbf{x}_i^{mis})$, with $\mathbf{x}_i^{obs}$ and $\mathbf{x}_i^{mis}$ representing the observed and the missing parts of $\mathbf{x}_i$, respectively. Using these notations, the observed data vector for the i^{th} subject is $(y_i, \mathbf{x}_i^{obs}, \mathbf{r}_i)$, where $\mathbf{r}_i = (r_{i1}, r_{i2}, \ldots, r_{ip})$ is a vector of missingness indicators for components of $\mathbf{x}_i$, defined by,

$$r_{ij} = \begin{cases} 1, & \text{if the } j^{th} \text{ component of } \mathbf{x}_i \text{ is observed.} \\ 0, & \text{otherwise.} \end{cases} \tag{1.18}$$

Under MAR, the conditional distribution of $\mathbf{r}_i$ given the data is a function only of the observed information and is independent of the missing data. Thus,

$$Pr(\mathbf{r}_i | y_i, \mathbf{x}_i^{obs}, \mathbf{x}_i^{mis}, \boldsymbol{\phi}) \propto Pr(r_i | y_i, \mathbf{x}_i^{obs}, \boldsymbol{\phi}), \tag{1.19}$$

where $\boldsymbol{\phi}$ is a vector of parameters. In addition, if $\boldsymbol{\phi}$ is distinct from the parameters in the

joint distribution of $(y_i, \mathbf{x}_i)$, missingness is ignorable and estimation can be done based on the likelihood L from the outcome and the covariates, where L is often written as a product of the conditional distribution of the outcome given the covariates and the joint distribution of the covariates as (Ibrahim et al., 1999)

$$\begin{aligned} L(\boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\gamma} | \mathbf{y}, \mathbf{x}^{obs}, \mathbf{x}^{mis}) &= \prod_{i=1}^{n} Pr(y_i | \mathbf{x}_i^{obs}, \mathbf{x}_i^{mis}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) Pr(\mathbf{x}_i^{mis} | \mathbf{x}_i^{obs}, \boldsymbol{\xi}) \\ &= \prod_{i=1}^{n} L_i(\boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\gamma} | y_i, \mathbf{x}_i^{obs}, \mathbf{x}_i^{mis}), \end{aligned} \tag{1.20}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ are parameter of the model that are of primary interest, $\boldsymbol{\xi}$ is a vector of parameters in the joint distribution of the missing covariates. Note that the conditional distributions $Pr(\mathbf{x}_i^{mis} | \mathbf{x}_i^{obs}, \boldsymbol{\xi})$ are used in (1.20) since the joint distribution of the covariates is proportional to the distribution of the missing covariates conditional on the observed (Ibrahim et al., 1999, 2005). From (1.20), the complete data log-likelihood $\ell(\boldsymbol{\theta} | \mathbf{y}, \mathbf{x}^{obs}, \mathbf{x}^{mis})$ can be written as

$$\begin{aligned} \ell(\boldsymbol{\theta} | \mathbf{y}, \mathbf{x}^{obs}, \mathbf{x}^{mis}) &= \sum_{i=1}^{n} \ell(\boldsymbol{\eta} | y_i; \mathbf{x}_i^{obs}, \mathbf{x}_i^{mis}) + \sum_{i=1}^{n} \ell(\boldsymbol{\xi} | \mathbf{x}_i^{mis}; \mathbf{x}_i^{obs}) \\ &= \sum_{i=1}^{n} \ell_i(\boldsymbol{\theta} | y_i, \mathbf{x}_i^{obs}, \mathbf{x}_i^{mis}) \end{aligned} \tag{1.21}$$

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

The observed data log-likelihood, based on which estimation is normally done, is obtained by integrating (or summing) $\ell(\boldsymbol{\theta} | \mathbf{y}, \mathbf{x}^{obs}, \mathbf{x}^{mis})$ over the domain of the missing covariates. Such integrals or summations are often difficult to evaluate and estimation is typically done using the EM algorithm. In the E-step, EM estimates the expected value of the complete data log-likelihood conditional on current parameter estimates and the observed data, and maximizes the expected log-likelihood. If the vector of parameter estimates at iteration

t is $\boldsymbol{\theta}^{(t)}$, at the $(t+1)^{th}$ iteration, the E step of EM computes,

$$Q_i(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = E\Big(\ell_c(\boldsymbol{\theta}|y_i, \mathbf{x}_i^{obs}, \mathbf{x}_i^{mis})|y_i, \mathbf{x}_i^{obs}, \boldsymbol{\theta}^{(t)}\Big). \tag{1.22}$$

In (1.22), ℓ_c is the log-likelihood from the complete data. The M-step of EM then maximizes $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \sum_{i=1}^{n} Q_i(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ to obtain the parameter estimates at iteration $t+1$, and the process continues until convergence. Values of $\boldsymbol{\theta}$ obtained at convergence are maximum likelihood estimates and the corresponding covariance matrix is commonly obtained using the method of Louise (1982).

For problems where a direct evaluation of the E-step is difficult, Monte Carlo EM methods estimate the expected log-likelihood numerically. At iteration t+1, the MCEM approach generates Monte-Carlo samples of size, say s, from the conditional distribution of the missing covariates and estimates $Q_i(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ in equation (1.22) by (Ibrahim et al., 1999),

$$Q_i(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \frac{1}{s}\sum_{j=1}^{s} \ell(\boldsymbol{\theta}|y_i, \mathbf{d}_{ij}, \mathbf{x}_i^{obs}) \tag{1.23}$$

where $\mathbf{d}_{i1}, \mathbf{d}_{i2}, \ldots \text{and}, \mathbf{d}_{is}$ are vectors of Monte-Carlo samples from the conditional distributions of the missing covariates. Ibrahim et al.(1999) generate Monte Carlo samples using adaptive rejection algorithm with Gibbs sampling for problems where the conditional distributions $Pr(y_i|\mathbf{x}_i^{obs}, \boldsymbol{\eta}^{(t)})$ and $Pr(\mathbf{x}_i^{mis}|\mathbf{x}_i^{obs}, \boldsymbol{\xi}^{(t)})$ are log-concave.

CHAPTER 2: MARGINALIZED MIXTURE MODELS FOR COUNT DATA FROM MULTIPLE SOURCE POPULATIONS

2.1 Introduction

The analysis of data from populations with unexplained heterogeneity presents special challenges to researchers. When count data arise from mixtures of unobserved populations, models based on standard probability distributions are often inadequate to explain observed variability (Wedel and DeSarbo, 1995; Frühwirth-Schnatter, 2005). For example, in dental caries research and many other areas, proportions of observations with zero counts are often higher than expected under the Poisson or negative binomial distributions and regression models based on these distributions may result in biased estimates and poor predictions. To account for such excess zeros, Mullahy (1986) and Lambert (1992) proposed zero-inflated Poisson (ZIP) regression. ZIP models, which employ two-component mixture distributions, hypothesize that observed counts arise from one of two latent classes within the source population: one class provides only zeros and the other produces both zero and non-zero values. However, the assumption of a model based on 'at-risk' and 'not-at-risk' latent classes may not be appropriate in some settings or may provide an inadequate fit (Preisser et al., 2012). To model counts from heterogeneous populations, Wang et al.(1996) proposed multi-component Poisson mixture distributions and their approach has been extended to other finite mixtures of non-degenerate count distributions. Despite the flexibility of finite mixtures for describing highly dispersed count data, parameters from standard mixture regression models are not directly applicable to making inferences about the overall effects of covariates on marginal means of count outcomes (Preisser et al., 2012; Albert et al., 2014). Even with the application of indirect methods of parameter estimation such as the use of

post-modeling transformations, there are many instances where latent class model formulations fail to fully explain relationships between covariates and population-wide parameters.

While the importance of the marginal mean as a target of inference in the analysis of finite mixtures of counts is well established (Lambert, 1992; Böhning et al., 1999; Preisser et al., 2012; Albert et al., 2014), marginally-specified mean models for finite mixtures of count distributions have more recently been proposed. Within a ZIP likelihood framework, Long et al.(2014) proposed marginalized zero-inflated Poisson (MZIP) regression, which specifies a two-part model for counts with a set of regression coefficients for the marginal mean and, to complete model specification, a second set of regression coefficients for the latent parameter defining membership in the 'excess-zero' class. The marginalized zero-inflated negative binomial (MZINB) model (Preisser et al., 2016) extended the MZIP model to zero-inflated negative binomial (ZINB) distributions. Todem et al.(2016) described a general representation of two-part marginalized mean count models including distributions for bounded counts, e.g., the zero-inflated beta binomial distribution. All these marginalized models assume that the count outcomes follow two-component mixtures consisting of a standard count distribution with a point-mass at zero. Data-generating mechanisms based on mixtures of non-degenerate count distributions could provide better fits in the class of marginalized mixture models for count data.

In this article, we seek to expand the class of marginalized mixture models for zero-inflated and other heterogeneous count data to allow for greater model choice with maximum likelihood estimation, when there is interest in evaluating the effects of exposures on the overall mean count. For counts with excess zeros, we extend the degenerate component of traditional zero-inflated models to standard count distributions for more flexible modeling of the marginal mean. Our motivation comes from a double-blind caries incidence trial conducted between 1988 and 1992 in Lanarkshire, Scotland, to compare the anti-caries efficacy of three toothpaste formulations in children. In this trial, a total of 4294 children ages 11-12 were randomized to either sodium fluoride or sodium monofluorophosphate or

the combination of sodium fluoride and sodium trimetaphosphate (Stephen et al., 1994; Preisser et al., 2013). The outcome variable of interest was the number of new decayed, missing and filled surfaces (DMFS) and dental exams were performed at baseline and after 1, 2 and 3 years. Because the DMFS counts exhibit many zeros, Poisson or negative binomial regression is not appropriate to model the counts. We consider marginalized, two-component finite mixture models to obtain direct inference about the relationship between toothpaste formulation and the marginal mean caries count in the trial population. Section 2.2 reviews zero-inflated mixture distributions and marginalized zero-inflated models, while Section 2.3 briefly discusses traditional finite mixture models. Section 2.4 presents two different two-component marginalized mixture models involving non-degenerate distributions. Simulation studies and an application of the proposed models are discussed in Sections 2.5 & 2.6 respectively. Concluding remarks follow in Section 2.7.

2.2 Models for Zero-inflated Data

2.2.1 Zero-inflated Poisson and Negative Binomial Models

Traditional zero-inflated models assume that counts arise from a two-component mixture of a standard count distribution with a distribution degenerate at zero. Under such models, counts are generated either from a 'non-susceptible' or 'perfect' state that always gives zeros, or from a 'susceptible', 'imperfect' state that produces both zero and positive counts according to a standard count data distribution (Lambert, 1992; Long et al.,2014; Preisser et al., 2012). Lambert (1992) introduce the zero-inflated Poisson regression and applied it for modeling defects in manufacturing processes, where defects are assumed coming from a 'perfect' state with a probability π or an 'imperfect' state with a probability $1-\pi$. While counts from the 'perfect', 'no-defect' state are always zero, those from the 'imperfect' state follow a Poisson distribution. The probability mass function of a random variable having a

zero-inflated Poisson or negative binomial distribution can be written as

$$Pr(Y_i = k) = \pi_i I(k = 0) + (1 - \pi_i) g(k|\boldsymbol{\theta}_i), k = 0, 1, 2, ..., \tag{2.24}$$

where the mixing parameter π_i is interpreted as the probability of a count being from the 'non-susceptible' or 'not-at-risk' latent class, $I(T)$ is an indicator variable taking 1 when T is true, and 0 when T is false; g is a Poisson or negative binomial mass function, and $\boldsymbol{\theta}_i$ is the vector of parameters in g. When g is the Poisson mass function, $\boldsymbol{\theta}_i$ is equal to the mean μ_i of the distribution, and for a negative binomial probability mass function g, $\boldsymbol{\theta}_i = (\mu_i, \alpha)$, where μ_i is the mean of the distribution and ϕ is the dispersion parameter. In this paper we will use the following parameterization for the probability mass function of a negative binomial distribution with mean μ and dispersion parameter α.

$$f(y|\mu, \alpha) = \frac{\Gamma(y + \alpha)}{y!\Gamma(\alpha)} \left(\frac{\alpha}{\alpha + \mu}\right)^{\alpha} \left(\frac{\mu}{\alpha + \mu}\right)^{y}, \text{ where } y = 0, 1, \ldots. \tag{2.25}$$

In zero-inflated models, regression parameters are specified for the mixing probability π_i and the mean of the assumed standard distribution μ_i, by using the logit and the log links as in equation (3) of Preisser et al.(2016), as

$$logit(\pi_i) = \mathbf{z}_i'\boldsymbol{\gamma} \quad \text{and} \quad log(\mu_i) = \mathbf{x}_i'\boldsymbol{\xi}, \tag{2.26}$$

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 $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_q)'$ and $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_p)'$ are regression parameters.

For n independent observations, the ZIP likelihood function is

$$L(\boldsymbol{\xi}, \boldsymbol{\gamma}|\mathbf{y}) = \prod_{i=1}^{n} \{1 + e^{(\mathbf{z}_i'\boldsymbol{\gamma})}\}^{-1} \left\{e^{(\mathbf{z}_i'\boldsymbol{\gamma})} + e^{-\exp(\mathbf{x}_i'\boldsymbol{\xi})}\right\}^{I(y_i=0)} \left\{\frac{e^{-\exp(\mathbf{x}_i'\boldsymbol{\xi})} e^{\mathbf{x}_i'\boldsymbol{\xi} y_i}}{y_i!}\right\}^{I(y_i>0)} \tag{2.27}$$

The corresponding likelihood function for the ZINB model can be written as

$$\begin{aligned} L(\boldsymbol{\xi}, \boldsymbol{\gamma}|\mathbf{y}) &= \prod_{i=1}^{n} \{1 + e^{(\mathbf{z}_i'\boldsymbol{\gamma})}\}^{-1} \left\{ e^{(\mathbf{z}_i'\boldsymbol{\gamma})} + \left(\frac{\alpha}{\alpha + e^{\mathbf{x}_i'\boldsymbol{\xi})}} \right)^{\alpha} \right\}^{I(y_i=0)} \\ &\times \prod_{i=1}^{n} \left\{ \frac{\Gamma(y_i + \alpha)}{y_i! \Gamma(\alpha)} \left(\frac{\alpha}{\alpha + e^{\mathbf{x}_i'\boldsymbol{\xi})}} \right)^{\alpha} \left(\frac{e^{\mathbf{x}_i'\boldsymbol{\xi}}}{\alpha + e^{\mathbf{x}_i'\boldsymbol{\xi})}} \right)^{y_i} \right\}^{I(y_i>0)} \end{aligned} \tag{2.28}$$

Since interpretations of parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\xi}$ in ZIP and ZINB models apply to the two latent subpopulations, they do not directly describe the overall population mean. Although the overall mean, $E(Y_i) = \nu_i$, for i^{th} subject could be estimated from such models by

$$\nu_i = \frac{e^{\mathbf{x}_i'\boldsymbol{\xi}}}{1 + e^{\mathbf{z}_i'\boldsymbol{\gamma}}} \tag{2.29}$$

and transformations such as the delta method could be applied to estimate the corresponding variance, it is not always easy to understand the behavior of ν_i. In particular, determining the effects of an exposure variable on incidence density ratios is challenging especially when the linear predictor for the mixing proportions contain some of the covariates in the Poisson mean model (Long et al., 2014).

2.2.2 Marginalized ZIP and ZINB Models

To estimate the overall effects of covariates on the population mean, marginalized zero-inflated Poisson (Long et al., 2014) and marginalized zero-inflated negative binomial (Preisser et al., 2016) models specify parameters for the marginal mean $\nu_i = E(y_i) = (1-\pi_i)\mu_i$ and the probability of being an excess zero (i.e., π_i) as

$$log(\nu_i) = \mathbf{x}_i'\boldsymbol{\beta} \quad \text{and} \quad logit(\pi_i) = \mathbf{z}_i'\boldsymbol{\gamma}, \tag{2.30}$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)$ is a vector of regression parameters for ν_i, and the parameters in $\boldsymbol{\gamma}$ have the same latent class interpretations as in ZIP and ZINB. The MZIP and MZINB

likelihood functions are obtained by replacing μ_i by $\frac{\nu_i}{1-\pi_i}$ in the ZIP and ZINB likelihoods, respectively.

2.3 Finite Mixture Models

Finite mixture distributions have been used to model counts obtained from heterogeneous populations (Wang et al., 1996; Morgan et al., 2014; Schlattmann et al., 2009). In the finite mixture model, the source population is assumed to be a partition of latent subpopulations; with a probability π_{ij}, the count random variable Y_i corresponding to the i^{th} individual takes a value from the j^{th} subpopulation according to a distribution specific to the subpopulation. An m component mixture distribution can be defined as (Wedel and DeSarbo, 1995; Frühwirth-Schnatter, 2005)

$$Pr(Y_i = y_i|\pi, \boldsymbol{\theta}_{ij}) = \sum_{j=1}^{m} \pi_j f_j(y_i|\boldsymbol{\theta}_{ij}), \tag{2.31}$$

where the components f_1, f_2, ..., f_m are probability mass functions of known distributions, $\boldsymbol{\theta}_{ij}$ is the vector of parameters in f_j, and $\boldsymbol{\pi} = (\pi_1, \pi_2..., \pi_m)'$ is a vector of mixing probabilities with $0 \leq \pi_j \leq 1$ and $\sum_{j=1}^{m} \pi_j = 1$. While the mixture distribution for zero-inflated counts in equation (2.24) allows mixing probabilities (i.e., π_i) to vary across individuals, conventional finite mixture models assume a constant probability, π_j, corresponding to the j^{th} subpopulation and impose heterogeneity through $f_j(y_i|\boldsymbol{\theta}_{ij})$.

The Poisson mixture distribution, where

$$f_j(y_i|\mu_{ij}) = \frac{e^{-\mu_{ij}}\mu_{ij}^{y_i}}{y_i!}$$

with μ_{ij} being the mean of the j^{th} component distribution, is a popular finite mixture model for count data. In finite Poisson mixture regression, the mean μ_{ij} is modeled as a function of covariates using the log link. Wang et al.. (1996) discuss that such models are identifiable

for full rank design matrices. While finite mixture models enable flexible modeling of counts from heterogeneous populations, their parameters have latent class interpretations. Such coefficients do not enable one to make direct inferences of the effects of covariates on the overall population mean (Roeder et al., 1999; Min and Agresti, 2005).

2.4 Marginalized Finite Mixture Models

In this section we propose methods of estimating regression parameters for the overall population mean of zero-inflated and other types of heterogeneous counts by employing non-degenerate mixture distributions. With the aim of expanding the pool of marginalized models for such counts, we consider data generating mechanisms based on mixtures of two Poissons (Pois-Pois) and a negative binomial and a Poisson (NB-Pois) distributions.

2.4.1 Models

The probability mass function (pmf) of a random variable with a Pois-Pois mixture distribution can be written as

$$f(y_i|\pi, \mu_{1i}, \mu_{2i}) = \pi f_{P1}(y_i|\mu_{1i}) + (1-\pi) f_{P2}(y_i|\mu_{2i}), \tag{2.32}$$

where π is a mixing probability,and f_{P1} and f_{P2} are Poisson mass functions with corresponding mean parameters μ_{1i} and μ_{2i}. Similarly, a NP-Pois random variable has a pmf given by,

$$f(y_i|\pi_i, \mu_{1i}, \mu_{2i}, \alpha) = \pi f_P(y_i|\mu_{1i}) + (1-\pi) f_{NB}(y_i|\mu_{2i}, \alpha). \tag{2.33}$$

In (2.33), f_P is a Poisson pmf with mean parameter μ_{1i} and f_{NB} a negative binomial pmf with mean and dispersion parameters μ_{2i} and α, respectively. The marginal mean, ν_i, of a random variable Y_i having either of the two mixture distributions can be written as

$$\nu_i = \pi\mu_{1i} + (1-\pi)\mu_{2i}. \tag{2.34}$$

Solving for μ_{2i} in equation (2.34) gives

$$\mu_{2i} = \frac{\nu_i - \pi\mu_{1i}}{1-\pi}. \tag{2.35}$$

To estimate a model for ν_i, the likelihood functions of Pois-Pois and NB-Pois mixture models can be written as functions of ν_i using equation (2.35) and replacing μ_{2i} by a linear function of the marginal mean. Thus, marginalized Poisson-Poisson (MPois-Pois) and marginalized NB-Poisson (MNB-Pois) models defined immediately below can be estimated utilizing the pmfs in equations (2.36) and (2.37), respectively.

$$f_{MPP}(y_i|\pi,\mu_{1i},\nu_i) = \pi\frac{e^{-\mu_{1i}}\mu_{1i}^{y_i}}{y_i!} + (1-\pi)\frac{e^{-\frac{\nu_i-\pi\mu_{1i}}{1-\pi}}\left[\frac{\nu_i-\pi\mu_{1i}}{1-\pi}\right]^{y_i}}{y_i!} \tag{2.36}$$

$$\begin{aligned} f_{NBP}(y_i|\pi,\alpha,\mu_{1i},\nu_i) = {} & \pi\frac{e^{-\mu_{1i}}\mu_{1i}^{y_i}}{y_i!} \\ & + (1-\pi)\frac{\Gamma(y_i+\alpha)}{y_i!\Gamma(\alpha)}\left(\frac{\alpha}{\alpha+\frac{\nu_i-\pi\mu_{1i}}{1-\pi}}\right)^{\alpha}\left(\frac{\frac{\nu_i-\pi\mu_{1i}}{1-\pi}}{\alpha+\frac{\nu_i-\pi\mu_{1i}}{1-\pi}}\right)^{y_i} \end{aligned} \tag{2.37}$$

The MPois-Pois model is defined through the specification of generalized linear models for the relationship of covariates to ν_i and μ_{1i}. Given a $p \times n$ design matrix $\mathbf{X}$, a model for ν_i is specified as

$$log(\nu_i) = \mathbf{x}_i'\boldsymbol{\beta}, \tag{2.38}$$

where $\mathbf{x}_i$ is the i^{th} column of $\mathbf{X}$ and $\boldsymbol{\beta}$ is a $p \times 1$ vector of parameters. Although π, μ_{1i} and α are considered nuisances that are not of study interest, these parameters need to be modeled to facilitate maximum likelihood estimation of regression coefficients in the marginal mean model. The logarithm of μ_{1i} is modeled by using a linear predictor that involves covariates of interest as in standard finite mixture Poisson models. The nuisance

parameter π is modeled as a constant using the logit link. Thus, the complete marginalized Pois-Pois (MPois-Pois) model can be written as in equation (2.39).

$$\begin{aligned} log(\nu_i) &= \mathbf{x}_i'\boldsymbol{\beta} \\ log(\mu_{1i}) &= \mathbf{z}_i'\boldsymbol{\xi} \\ logit(\pi) &= \rho, \end{aligned} \tag{2.39}$$

where $\mathbf{x}_i$ and $\mathbf{z}_i$ are vectors of covariates, $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$ are vectors of regression coefficients, and $-\infty < \rho < \infty$ is a constant.

Marginalized NB-Pois models require estimation of the dispersion parameter (i.e., α) in addition to the regression coefficients in equation (2.39). We specify a model for α as

$$log(\alpha) = -\tau. \tag{2.40}$$

The link functions in equations (2.39) and (2.40) correspond to $\nu_i > 0$, $\mu_{1i} > 0$, $0 < \pi < 1$ and $\alpha > 0$. For n independent count random variables $Y_1, Y_2, ..., Y_n$ with corresponding realizations $y_1, y_2, ..., y_n$, the likelihood function for MPois-Pois models is given by (2.41).

$$L(\rho, \boldsymbol{\beta}, \boldsymbol{\xi}|\mathbf{y}) = \prod_{i=0}^{n} \frac{1}{(1+e^{\rho})y_i!} \left\{ e^{\rho} \exp(-e^{\mathbf{z}_i'\boldsymbol{\xi}}) e^{\mathbf{z}_i'\boldsymbol{\xi} y_i} + e^{-\eta(\rho,\boldsymbol{\beta},\boldsymbol{\xi};\mathbf{x}_i,\mathbf{z}_i)} \eta(\rho, \boldsymbol{\beta}, \boldsymbol{\xi}; \mathbf{x}_i, \mathbf{z}_i)^{y_i} \right\}, \tag{2.41}$$

with

$$\eta(\rho, \boldsymbol{\beta}, \boldsymbol{\xi}; \mathbf{x}_i, \mathbf{z}_i) = e^{\mathbf{x}_i'\boldsymbol{\beta}}(1+e^{\rho}) - e^{\rho} e^{\mathbf{z}_i'\boldsymbol{\xi}}. \tag{2.42}$$

Similarly, the likelihood function for marginalized NB-Pois (MNB-Pois) models can be specified as

$$L(\rho,\tau,\boldsymbol{\beta},\boldsymbol{\xi}|\mathbf{y}) = \prod_{i=0}^{n} \frac{\Gamma(y_i+e^{-\tau})}{(1+e^{\rho})\Gamma(y_i+1)\Gamma(e^{-\tau})}\left(\frac{e^{-\tau}}{e^{-\tau}+\eta(\rho,\boldsymbol{\beta},\boldsymbol{\xi};\mathbf{x}_i,\mathbf{z}_i)}\right)^{e^{-\tau}} \times \prod_{i=0}^{n}\left(\frac{\eta(\rho,\boldsymbol{\beta},\boldsymbol{\xi};\mathbf{x}_i,\mathbf{z}_i)}{e^{-\tau}+\eta(\rho,\boldsymbol{\beta},\boldsymbol{\xi};\mathbf{x}_i,\mathbf{z}_i)}\right)^{y_i} + \prod_{i=0}^{n}\frac{e^{\rho}\exp(-e^{\mathbf{z}_i'\boldsymbol{\xi}})e^{\mathbf{z}_i'\boldsymbol{\xi}y_i}}{(1+e^{\rho})y_i!}, \tag{2.43}$$

where $\eta(\rho,\boldsymbol{\beta},\boldsymbol{\xi};\mathbf{x}_i,\mathbf{z}_i)$ has the same interpretation as in equation (2.42).

2.4.2 Estimation

With carefully chosen starting parameter values, regression coefficients in MPois-Pois and MNB-Pois models can be estimated by the use of quasi-Newton optimization. While MZIP or MZINB (Long et al., 2014; Preisser et al.,2016) model estimates can be used as starting values of coefficients in the marginal mean model (i.e., the $\boldsymbol{\beta}$s), starting values for coefficients in the latent parameter models (i.e.,π, μ_i, and α) may be obtained from two-component Poisson-Poisson and negative binomial-Poisson models. Following Ramaswamy et al.(1994) and Leisch (2004), we employ EM algorithm to find starting values for parameters ρ, $\boldsymbol{\xi}$ and τ in MNB-Pois models. The same approach can be applied for MPois-Pois models.

2.4.3 Algorithm for Finding Starting Values of Parameters

Consider a random variable Y_i that takes a value y_i according to the two-component NP-Pois mixture model in equation (2.33). Latent class regression coefficients can be specified

for parameters π, μ_{1i}, μ_{2i} and α as

$$\begin{aligned} log(\mu_{1i}) &= \mathbf{z}_i'\boldsymbol{\gamma} \\ log(\mu_{2i}) &= \mathbf{x}_i'\boldsymbol{\zeta} \\ \pi &= \pi \\ log(\alpha) &= -\tau, \end{aligned} \tag{2.44}$$

where $\boldsymbol{\zeta}$ is a vector of parameters and all the other parameters and variables are as described in equations (2.39) and (2.40). In line with standard mixture models (Ramaswamy et al., 1994; and Leisch, 2004), the logit link is not used to model π in equation (2.44); once π is estimated, a starting value for ρ in the marginal mean model can be obtained by setting $\rho = \text{logit}(\pi)$.

As a complete data likelihood function is needed to implement EM algorithm, we define an indicator variable U_i corresponding to the i^{th} subject as (Ramaswamy et al., 1994; and Leisch, 2004)

$$U_i = \begin{cases} 1, & \text{if subject } i \text{ belongs to subpopulation 1} \\ 0, & \text{if subject } i \text{ belongs to subpopulation 2} \end{cases} \tag{2.45}$$

Thus, U_i has a Bernoulli distribution with parameter π.

$$Pr(U_i = u_i|\pi) = \pi^{u_i}(1-\pi)^{1-u_i}, u_i = 0, 1.$$

The random variable (Y_i, U_i) contains an observed outcome Y_i and a missing variable U_i, and the contribution of (Y_i, U_i) to the complete data likelihood is given by,

$$\begin{aligned} L_{ic}(\pi, \boldsymbol{\gamma}, \boldsymbol{\zeta}, \tau|u_i, y_i, \mathbf{x}_i, \mathbf{z}_i) &= Pr(Y_i = y_i|\boldsymbol{\gamma}, \boldsymbol{\zeta}, \tau, \mathbf{x}_i, \mathbf{z}_i; U_i = u_i)Pr(U_i = u_i|\pi) \\ &= \Big[\pi f_P(y_i|\boldsymbol{\gamma}, \mathbf{z}_i)\Big]^{u_i}\Big[(1-\pi)f_{NB}(y_i|\boldsymbol{\zeta}, \tau, \mathbf{x}_i)\Big]^{1-u_i} \end{aligned} \tag{2.46}$$

The likelihood function L_c from n independent counts is the product of each likelihood in equation (2.46). That is (Ramaswamy et al., 1994),

$$L_c(\pi, \boldsymbol{\gamma}, \boldsymbol{\zeta}, \tau | \mathbf{u}, \mathbf{y}, \mathbf{x}, \mathbf{z}) = \prod_{i=0}^{n} \Big[\pi f_P(y_i|\boldsymbol{\gamma}, \mathbf{z}_i)\Big]^{u_i} \Big[(1-\pi) f_{NB}(y_i|\boldsymbol{\zeta}, \tau, \mathbf{x}_i)\Big]^{1-u_i} \tag{2.47}$$

The log-likelihood function is given by

$$\begin{aligned} \ell_c(\pi, \boldsymbol{\gamma}, \boldsymbol{\zeta}, \tau | \mathbf{u}, \mathbf{y}, \mathbf{x}, \mathbf{z}) = &\sum_{i=0}^{n} \Big[u_i \, logit(\pi) + log(1-\pi)\Big] + \sum_{i=0}^{n} u_i \, log(f_P(y_i|\boldsymbol{\gamma}, \mathbf{z}_i)) \\ &+ \sum_{i=0}^{n} \Big[(1-u_i) log(f_{NB}(y_i|\boldsymbol{\zeta}, \tau, \mathbf{x}_i))\Big] \end{aligned} \tag{2.48}$$

Given initial parameter values $\boldsymbol{\theta}^{(0)} = (\pi^{(0)}, \boldsymbol{\gamma}^{(0)}, \boldsymbol{\zeta}^{(0)}, \tau^{(0)})$, the E step of EM computes the expected value of ℓ_c conditional on the observed variables and $\boldsymbol{\theta}^{(0)}$.

$$\begin{aligned} E(\ell_c(\pi, \boldsymbol{\gamma}, \boldsymbol{\zeta}, \tau | \mathbf{u}, \mathbf{y}, \mathbf{x}, \mathbf{z}) | \boldsymbol{\theta}^{(0)}, \mathbf{y}, \mathbf{x}, \mathbf{z})) = &\sum_{i=0}^{n} \Big[E(u_i|\boldsymbol{\theta}^{(0)}, y_i, \mathbf{x}_i, \mathbf{z}_i) \, logit(\pi) + log(1-\pi)\Big] \\ &+ \sum_{i=0}^{n} E(u_i|\boldsymbol{\theta}^{(0)}, y_i, \mathbf{x}_i, \mathbf{z}_i) \, log(f_P(y_i|\boldsymbol{\gamma}, \mathbf{z}_i)) \\ &+ \sum_{i=0}^{n} \Big[log(f_{NB}(y_i|\boldsymbol{\zeta}, \tau, \mathbf{x}_i))(1 - E(u_i|\boldsymbol{\theta}^{(0)}, y_i, \mathbf{x}_i, \mathbf{z}_i))\Big] \end{aligned} \tag{2.49}$$

It can be shown that (Ramaswamy et al., 1994)

$$\begin{aligned} E(u_i|\boldsymbol{\theta}^{(0)}, y_i, \mathbf{x}, \mathbf{z}) &= \frac{\pi^{(0)} f_P(y_i|\boldsymbol{\gamma}, \mathbf{z}_i)}{\pi^{(0)} f_P(y_i|\boldsymbol{\gamma}, \mathbf{z}_i) + (1-\pi^{(0)}) f_{NB}(y_i|\boldsymbol{\zeta}, \tau, \mathbf{x}_i)} \\ &\equiv P_i^{(0)} \end{aligned} \tag{2.50}$$

Thus, the M step maximizes,

$$
\begin{aligned}
E(\ell_c(\pi, \boldsymbol{\beta}, \boldsymbol{\zeta}, \tau | \mathbf{u}, \mathbf{y}, \mathbf{x}, \mathbf{z}) | \boldsymbol{\theta}^{(0)}, \mathbf{y}, \mathbf{x}, \mathbf{z})) &= \sum_{i=0}^{n} \left[P_i^{(0)} \, logit(\pi) + log(1-\pi) \right] \\
&+ \sum_{i=0}^{n} P_i^{(0)} \, log(f_P(y_i | \boldsymbol{\gamma}, \mathbf{z}_i)) \\
&+ \sum_{i=0}^{n} \left[log(f_{NB}(y_i | \boldsymbol{\zeta}, \tau, \mathbf{x}_i))(1 - P_i^{(0)}) \right] \\
&= \ell_\pi + \ell_{\boldsymbol{\gamma}} + \ell_{(\boldsymbol{\zeta}, \tau)}
\end{aligned} \tag{2.51}
$$

To obtain the next estimates in the M step, the three components ℓ_π, $\ell_{\boldsymbol{\gamma}}$ and $\ell_{(\boldsymbol{\zeta},\tau)}$ of the expected log-likelihood in (2.51), can be optimized separately. Maximizing ℓ_π with respect to π gives (Ramaswamy et al., 1994)

$$
\pi^{(1)} = \sum_{i=0}^{n} \frac{P_i^{(0)}}{n}.
$$

The remaining two components of the expected log-likelihood (i.e., $\ell_{\boldsymbol{\gamma}}$ and $\ell_{(\boldsymbol{\zeta},\tau)}$) correspond to weighted log-likelihoods of generalized linear models and estimation can be performed separately to obtain the next set of parameters $\boldsymbol{\gamma}^{(1)}$, $\boldsymbol{\zeta}^{(1)}$ and $\tau^{(1)}$. Utilizing the parameters ($\pi^{(1)}$,$\boldsymbol{\beta}^{(1)}$,$\boldsymbol{\zeta}^{(1)}$, $\tau^{(1)}$) estimated in the first step, EM again computes and optimizes the expected log-likelihood and continues iterations between the two steps until convergence. The NB-Poisson mixture model estimates of π,$\boldsymbol{\gamma}$ and τ at convergence are then employed as starting values for parameters $\rho = \text{logit}(\pi)$, $\boldsymbol{\xi}$ and τ respectively, in the MNB-Pois model.

2.5 Simulation Study

Simulation studies were performed to examine the properties of MPois-Pois and MNB-Pois models for various sample sizes. Counts with Pois-Pois and NB-Pois mixture distributions were generated from the probability mass functions in equation (2.36) and (2.37),

where π, μ_{1i}, ν_i and α are determined from

$$\begin{aligned} log(\nu_i) &= \mathbf{x}_i'\boldsymbol{\beta} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} \\ log(\mu_{1i}) &= \mathbf{z}_i'\boldsymbol{\xi} = \xi_0 + \xi_1 x_{1i} + \xi_2 x_{2i} + \xi_3 x_{3i} \\ logit(\pi) &= \rho, \\ log(\alpha) &= -\tau \end{aligned} \tag{2.52}$$

with $\mathbf{x}_i = \mathbf{z}_i$ and $x_{1i} \sim$ Poisson(2)/3, $x_{2i} \sim \exp(1)$, $x_{3i} \sim$ Benoulli(0.4), $\beta_0 = 1.5$, $\beta_1 = -0.1$, $\beta_2 = -0.2$, $\beta_3 = 0.5$, $\xi_0 = 1.5$, $\xi_1 = -0.5$, $\xi_2 = -0.5$, $\xi_3 = 1$, $\rho = -0.4$ and $\tau = -0.5$. Using these specifications, samples of sizes 100, 200, 500 and 1000 were generated corresponding to marginalized Pois-Pois and NB-Pois models. Four marginalized models, namely, MPois-Pois, MNB-Pois, MZIP and MZINB models were then fitted to the data, where each simulation was repeated 10,000 times. To estimate Type I error rates of testing $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$, all the simulations were repeated by generating data using $\beta_1 = 0$, but keeping all the remaining parameter and covariate values the same as described previously. For each of the four models, the Type I error rates were calculated as the proportion of 10,000 models that converged and estimated a p-value from two-sided Wald tests of less than 0.05 for β_1.

Table 2.1 shows that for all sample sizes (i.e., 100, 200, 500 and 1000), estimates of β_1, β_2 and β_3 from the MPois-Pois model have low biases when the true model is MPois-Pois, and that the biases tend to decrease when the sample sizes increase. In these simulations, the MNB-Pois, MZIP and MZINB models also have low biases. From Table 2.2, it can be seen that the MPois-Pois model estimates Type I error rates for β_1 close to 0.05, but that MNB-Pois, MZIP and MZINB models tend to over-estimate the error rates when the true model is MPois-Pois. For such data, the MPois-Pois model estimated coverages of 95% confidence intervals for β_1, β_2 and β_3 are in general close to the nominal value, particularly when the sample sizes are 200, 500 and 1000 (Table 2.3). In the simulations, over 96% of

MNB-Pois models converged, but convergence rates for the remaining marginalized models range from 88.0% to 90.2% for MNB-Pois, from 75.9% to 98.4% for MZIP, and from 72.0% to 96.6% for the MZINB models.

When the data are generated from MNB-Pois models, Table 2.5 shows that the MNB-Pois model gives low percent relative median biases for β_1, β_2 and β_3, and the biases appear to decrease as sample sizes increase. The corresponding estimates from the MZINB model also have low biases, but those from MPois-Pois and MZIP models are generally higher. In addition, the performance of the true MNB-Pois model with regard to Type I error rates (for β_1) and coverages of 95% confidence intervals (for β_1, β_2 and β_3) is superior to the other three marginalized models (Tables 2.6 and 2.7, respectively) for larger sample sizes. Overall, the simulation results indicate that when the true model is MPois-Pois or MNB-Pois, the model estimates parameters with small biases, Type I errors close to the assumed rate and coverages of 95% confidence intervals near 95% for large sample sizes.

2.6 Application to a Caries Incidence Trial

The methods described in this paper were applied to the Lanarkshire caries incidence trial briefly discussed in Section 2.1. A total of 4294 children ages 11-12 were randomized to either sodium fluoride (NaF), sodium fluoride plus sodium trimetaphosphate (NaFTMP) or sodium monofluorophosphate (SMFP) and dental exams were performed at baseline and after 1, 2 and 3 years. The analysis was based on 3412 children followed up until year 2 and the response variable of interest was the number of new decayed, missing and filled surfaces (DMFS). In addition to treatment allocation, baseline caries (bc: 1= high, 0 = low), baseline calculus (calc:1=yes, 0= no) and the interaction of the two (bc_calc) were considered as explanatory variables. High baseline caries values correspond to at least one decayed, missing or filled interior tooth or premolar, and a baseline calculus value of '1' refers to the existence of calcified deposits on the teeth formed by the continuous presence of dental plaque (Stephen et al., 1994; Preisser et al., 2013). An important feature of the

data is the large number of zero counts in the outcome variable, as 658 (19.28 %) of the 3412 children had zero DMFS counts (Figure 2.1). Since the number of zeros is much higher than what is expected under standard count probability mass functions (such as the Poisson and negative binomial), regression models based on these distributions may provide biased estimates and poor predictions. Marginalized models, however, account for zero-inflation and enable the estimation of treatment effects on DMFS counts in the overall population.

We applied each of the two mixture distributions discussed in this article (i.e., Pois-Pois and NB-Pois mixtures) to model the marginal mean of DMFS. In each model, the marginal mean ν_i of DMFS and the mean parameter in a Poisson part of Pois-Pois and NB-Pois mixtures (i.e, μ_{1i}) are related to the explanatory variables of interest as follows.

$$\begin{aligned} log(\nu_i) &= \beta_0 + \beta_1 bc_i + \beta_2 calc_i + \beta_3 bc\_calc_i + \beta_4 NaF_i + \beta_5 NaFTMP_i \\ log(\mu_{1i}) &= \xi_0 + \xi_1 bc_i + \xi_2 calc_i + \xi_3 bc\_calc_i \end{aligned} \tag{2.53}$$

where bc_i is baseline caries from the i^{th} child, $calc_i$ is baseline calculus, bc_calc_i is the interaction of bc_i and $calc_i$, $NaF_i = 1$ if the child was given sodium fluoride, and $NaFTMP_i = 1$ if the child was randomized to the NaFTMP group with children in the SMFP group making up the reference treatment category.

To model the mixing probability π and the reciprocal of the dispersion parameter α (for the NB-Pois model), only intercepts were specified using the logit and the negative log links, respectively.

$$\begin{aligned} logit(\pi) &= \rho \\ log(\alpha) &= -\tau. \end{aligned} \tag{2.54}$$

For comparisons, MZIP and MZINB models were also fitted to the data by employing the same covariates as in equation (2.53) to model the marginal mean and the probability

of excess zeros.

Table 2.9 summarizes the estimated log-likelihood and AIC values from the the four marginalized models together with incidence rate ratios for the NaF and NaFTMP groups relative to the SMFP group. The estimated regression coefficients and standard errors for the marginal mean part of each of the four marginalized models are presented in Table 2.10. Based on the AIC criteria, the MNB-Pois (AIC=17192.9) provides the best fit to the data compared to the other three models. The MZINB model has the next lowest AIC value and appears to give a good prediction of observed DMFS proportions as the MNB-Pois model (Figure 2.2).

Based on the best-fitting model (i.e., MNB-Pois), the estimated incidence density ratio of a child in the NaF group is 0.942 CI (0.874, 1.015), relative to children with the same baseline status of caries and calculus who were assigned to SMFP. The corresponding incidence density ratio for children in the NaFTMP group is 0.970 CI (0.884, 1.063). Thus, children in the NaF and NaFTMP groups had a decrease in the marginal mean DMFS count by 5.5% and 3.0%, respectively, compared to children with the same baseline characteristics who were assigned to the SMFP group. However, the associations are not significant since the confidence intervals of the two incidence density ratios include 1.

2.7 Discussion

We proposed two-component mixture distributions to model marginal means of counts generated from heterogeneous populations. To estimate the effects of exposure variables on the overall population mean count, we specify regression parameters directly to the mean and perform estimation using maximum likelihood methods. The resulting model parameters have straightforward interpretations in describing exposure effects on the marginal mean. The two proposed mixture distributions generalize the ZIP and ZINB distributions and can be applied to a wide range of overdispersed outcomes. For zero-inflated counts, the

proposed method expands the family of two-part marginalized regression models by providing alternatives to MZIP and MZINB regression. The merit of each model in the larger class of alternative marginalized models is then judged based on goodness of fit considerations. Because our main interest is in modeling marginal means of counts, model parameters that are not of our primary interest are allowed to depend on covariates or none whatsoever, to complete specification of the likelihood function. This provides for model parsimony as needed while allowing all the relevant covariates to be estimated in the overall mean model.

Simulations indicate that when the true model is specified, each of the proposed marginalized mixture model provides low biases, Type I errors and confidence interval coverages close to the nominal levels. The models were also applied to a randomized trial aimed at comparing the anti-caries efficacy of three toothpaste formulations in children ages 11-12. Since the counts in this trial (i.e., number of decayed and filled tooth surfaces) are zero for a large proportion of children, traditional count models such as Poisson regression do not fit the data sufficiently. Conventional zero-inflated models fitted to the data also have limitations in that the estimated parameters are interpreted in terms of latent classes representing children 'at-risk' and 'not-at-risk' for dental caries. Parameter estimates from two-part marginalized count models are directly interpretable and are also easily employed to compute incidence density ratios for the exposure effect of the main exposure variables and the other covariates. The models are compared with each other and with MZIP and MZINB model fits based on the AIC goodness of fit criteria. Comparisons of the new models with each other and with MZIP and MZINB models show that the MNB-Pois model has the best fit, as evidenced by a smaller AIC value. The proposed marginalized mixture modeling framework provides a wide range of alternatives to estimate exposure effects on marginal means of counts generated from heterogeneous populations. The methods are straightforward and can be implemented in most statistical softwares. Future research could extend the marginalized mixture models to allow the mixing probabilities to depend on covariates as well as to accommodate longitudinal data.

Table 2.1: Percent relative median biases of estimates of β_1, β_2 and β_3 from marginalized mixture models fitted to data generated from the MPois-Pois model with 10,000 replications.

Sample Size	Parameter	MPois-Pois	MNB-Pois	MZIP	MZINB
100	β_1	-2.04	0.56	1.40	0.97
	β_2	0.08	1.54	-3.11	-3.45
	β_3	-0.70	-0.33	-0.61	-0.74
200	β_1	-0.68	1.34	1.70	1.89
	β_2	-0.69	0.62	-2.64	-2.65
	β_3	-0.29	0.06	-0.43	-0.41
500	β_1	-0.87	0.07	-0.36	-1.18
	β_2	0.11	0.78	-1.51	-1.44
	β_3	-0.14	0.19	-0.16	-0.11
1000	β_1	-0.40	0.43	-0.37	-0.64
	β_2	0.27	0.88	-1.43	-0.91
	β_3	0.06	0.22	-0.08	-0.07

Table 2.2: Type I error rates for the estimate of β_1 from marginalized models fitted to data generated from the MPois-Pois model with 10,000 replications.

Sample Size	MPois-Pois	MNB-Pois	MZIP	MZINB
100	0.068	0.073	0.102	0.070
200	0.067	0.069	0.106	0.072
500	0.060	0.065	0.112	0.073
1000	0.054	0.061	0.112	0.066

Table 2.3: Coverages of 95% confidence intervals for estimates of β_1, β_2 and β_3 from marginalized models fitted to data generated from the MPois-Pois model with 10,000 replications.

Sample Size	Parameter	MPois-Pois	MNB-Pois	MZIP	MZINB
100	β_1	93.7	93.8	91.3	93.4
	β_2	93.2	92.9	90.9	92.8
	β_3	95.2	95.1	92.9	94.7
200	β_1	94.1	94.1	91.2	93.8
	β_2	93.3	93.2	90.9	92.9
	β_3	95.1	95.2	92.6	94.9
500	β_1	94.1	93.9	90.7	93.5
	β_2	94.4	93.9	90.5	93.1
	β_3	94.9	94.9	92.0	94.8
1000	β_1	94.7	94.4	90.9	93.9
	β_2	94.7	93.8	90.8	93.1
	β_3	95.0	94.9	92.1	95.0

Table 2.4: Percentages of converged marginalized models fitted to data generated from the MPois-Pois model with 10,000 replications.

Sample Size	MPois-Pois	MNB-Pois	MZIP	MZINB
100	96.2	88.2	75.9	72.0
200	97.2	90.2	87.0	82.6
500	98.3	90.0	95.0	94.2
1000	99.3	88.0	98.4	94.6

Table 2.5: Percent relative median biases of estimates of β_1, β_2 and β_3 from marginalized mixture models fitted to data generated from the MNB-Pois model with 10,000 replications.

Sample Size	Parameter	MPois-Pois	MNB-Pois	MZIP	MZINB
100	β_1	6.80	11.95	23.51	13.72
	β_2	4.00	4.44	7.95	1.89
	β_3	-4.35	-0.25	1.40	0.88
200	β_1	-14.85	4.57	20.12	7.41
	β_2	-1.12	2.02	5.11	0.07
	β_3	-5.44	0.33	1.49	0.36
500	β_1	-29.97	-0.75	11.79	0.73
	β_2	-3.90	0.62	2.81	0.14
	β_3	-7.66	0.46	1.52	0.61
1000	β_1	-34.68	0.00	10.34	2.39
	β_2	-4.75	0.87	2.63	0.39
	β_3	-10.13	-0.19	0.97	-0.01

Table 2.6: Type I error rates for the estimate of β_1 from marginalized models fitted to data generated from the MNB-Pois model with 10,000 replications.

Sample Size	MPois-Pois	MNB-Pois	MZIP	MZINB
100	0.262	0.103	0.271	0.079
200	0.255	0.064	0.272	0.073
500	0.232	0.053	0.273	0.074
1000	0.240	0.049	0.273	0.072

Table 2.7: Coverages of 95% confidence intervals for estimates of β_1, β_2 and β_3 from marginalized models fitted to data generated from the MNB-Pois model with 10,000 replications.

Sample Size	Parameter	MPois-Pois	MNB-Pois	MZIP	MZINB
100	β_1	76.9	89.7	77.4	92.4
	β_2	77.8	89.6	79.6	91.8
	β_3	83.0	92.0	79.4	93.7
200	β_1	78.1	93.0	77.6	92.3
	β_2	78.9	92.7	79.1	91.8
	β_3	83.9	93.5	80.0	94.0
500	β_1	78.1	94.2	77.0	92.2
	β_2	80.8	94.5	78.6	91.3
	β_3	80.2	94.5	79.7	93.9
1000	β_1	76.2	95.0	77.5	93.1
	β_2	81.5	95.0	78.9	91.6
	β_3	71.6	95.3	80.7	94.6

Table 2.8: Percentages of converged marginalized models fitted to data generated from the MNB-Pois model with 10,000 replications.

Sample Size	MPois-Pois	MNB-Pois	MZIP	MZINB
100	92.0	91.0	97.4	85.3
200	96.8	96.9	99.7	87.3
500	97.7	99.8	100.0	90.3
1000	99.4	100.0	100.0	91.4

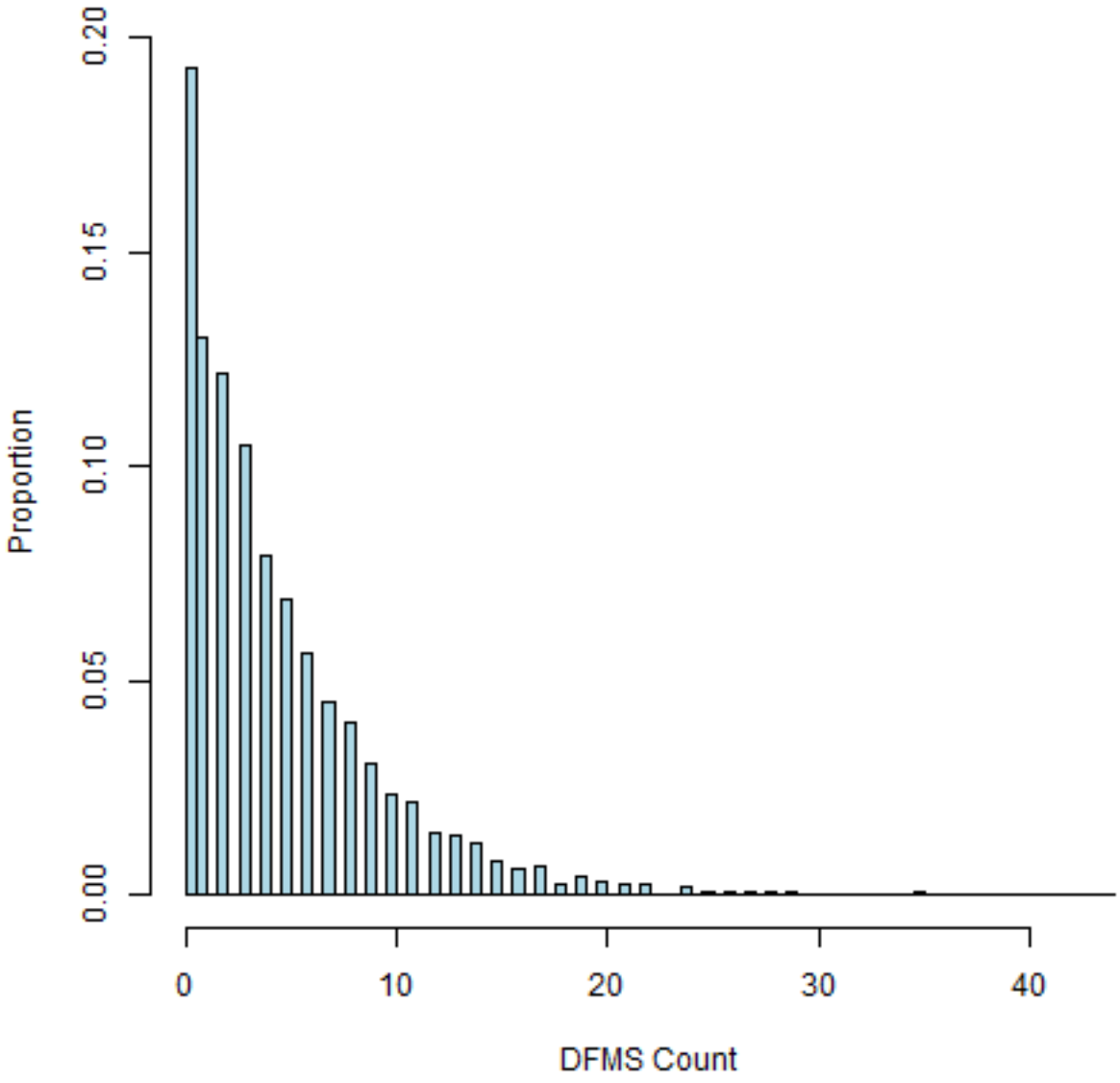

Figure 2.1: Distribution of DFMS counts after 2 years for 3412 children ages 11-12 participating in the Lanarkshire trial.

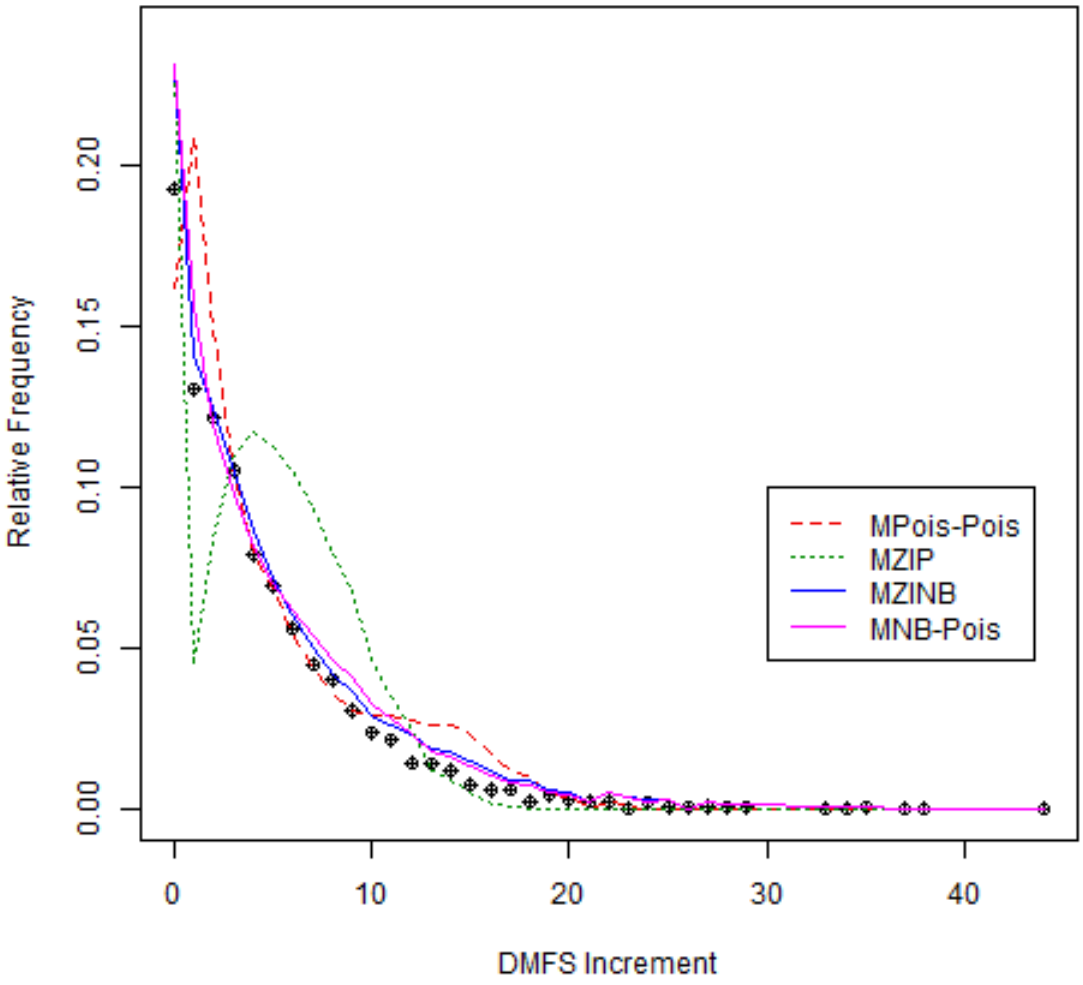

Figure 2.2: Predicted and observed proportions of DMFS count increments after 2 years in the Lanarkshire trial.

Table 2.9: Estimated log-likelihood, AIC and incidence density ratios (95% CI) comparing NaF and NaFTMP with SMFP in the Lanakshire trial, based on four marginalized models.

			IDR (95% CI)	
Model	-2Log-lik.	AIC	NaF	NaFTMP
MPois-Pois	18074.2	18096.2	0.989 (0.964, 1.015)	1.008 (0.977, 1.039)
MNB-Pois	17168.9	17192.9	0.942 (0.874, 1.015)	0.970 (0.884, 1.063)
MZIP	20413.4	20433.4	0.933 (0.900, 0.967)	0.939 (0.898, 0.981)
MZINB	17190.1	17212.1	0.948 (0.880, 1.021)	0.977 (0.892, 1.070)

Table 2.10: Marginal mean model Estimates and standard errors from MPois-Pois, MNB-Pois, MZIP and MZINB models for the Lanarkshire caries trial.

	MPois-Pois		MNB-Pois		MZIP		MZINB	
Variable	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
Marginal mean model								
Intercept	1.228	0.026	1.190	0.037	1.200	0.024	1.187	0.036
bc	0.683	0.030	0.784	0.040	0.784	0.026	0.783	0.040
calc	-0.226	0.042	-0.148	0.052	-0.151	0.038	-0.151	0.050
bc_calc	-0.067	0.062	-0.112	0.079	-0.108	0.053	-0.109	0.080
NaF	-0.011	0.013	-0.060	0.038	-0.069	0.018	-0.053	0.038
NaFTMP	0.008	0.016	-0.031	0.047	-0.063	0.022	-0.023	0.046
Latent class mean model					*Zero-inflation model*			
Intercept	2.041	0.027	-1.831	0.758	-1.124	0.064	-1.938	0.154
bc	0.529	0.031	3.886	0.777	-1.315	0.131	-2.237	0.722
calc	-0.225	0.042	0.568	0.806	0.032	0.112	-0.137	0.263
bc_calc	-0.042	0.061	-0.610	0.831	0.216	0.259	-0.087	1.917
Mixing probability and dispersion parameter model estimates								
ρ	-0.751	0.056	-1.863	0.175				
τ			-0.327	0.041			-0.047	0.055

CHAPTER 3: MARGINALIZED BIVARIATE ZERO-INFLATED POISSON REGRESSION

3.1 Introduction

Counts with excess zeros are often encountered in health research and many other areas. While much of the literature on zero-inflated counts is focused on univariate outcomes, studies sometimes involve two or more correlated and zero-inflated count variables (Divaris et al.2012; Li et al., 1999; Wang et al., 2003). When two dependent count outcomes take higher proportions of zeros than expected under standard bivariate count distributions, modeling requires accounting for zero-inflation and the dependence between the outcomes. To model counts of occupational injuries, Wang et al.(2003) employ a mixture of a bivariate Poisson distribution with a point mass at (0,0), and estimate parameters using the expectation maximization (EM) algorithm. Li et al.(1999) propose several component mixture distributions for multivariate zero-inflated counts and describe their applications to defects in manufacturing processes. Based on similar distributions, Mujumdar and Gries (2010) employ Bayesian approaches to model bivariate plant count data with excess zeros, and Arab et al.(2012) apply semi-parametric methods to model species abundances. Other works on bivariate zero-inflated count models include Yang, Das and Majumdar (2016), Cheung and Lam (2005), Bermùdez and Karlis (2012), Gurmu and Elder (2008), and Walhin (2001).

A common feature of existing models for bivariate zero-inflated counts is that, when covariates are included in model building processes, regression coefficients are specified for latent parameters representing unobserved subpopulations within the sampled population. When interest is to make inferences about the overall population, such coefficients do not have direct interpretations in describing the effects of covariates on the marginal mean

vector in the population. Thus, one has to rely on indirect methods such as the use of post-modeling transformations, to estimate effects of covariates on the marginal means. In addition to the difficulty to compute relevant variances, these methods sometimes fail to fully explain relationships between covariates and population-wide parameters such as incidence density ratios (Preisser et al., 2012; Albert et al., 2014, Long et al., 2014).

For univariate zero-inflated counts, the importance of methods yielding marginally interpretable parameters has long been recognized (Lambert, 1992; Böhning et al., 1999; Preisser et al., 2012; Albert et al., 2014); lately, marginally specified mean models for zero-inflated counts have been promulgated (Long et al., 2014; Preisser et al., 2016; Todem et al., 2016). Based on the framework of ZIP model likelihood function, Long et al.(2014) propose a maximum likelihood method of estimating regression coefficients for the marginal means of counts with excess zeros. Instead of modeling the latent class mean parameter in the Poisson part of ZIP, the marginalized zero-inflated Poisson (MZIP) model specifies regression parameters directly to the overall mean and estimates an additional set of coefficients for the probability of being excess zero. Marginalized zero-inflated negative binomial models (Preisser et al., 2016) extend the MZIP model to counts having zero-inflated negative binomial distributions, where the marginal mean and the probability of being excess zero are modeled by using shared as well as distinct regression parameters. 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.

In this paper, we propose a marginalized model for bivariate zero-inflated counts that provides directly interpretable regression parameters for the marginal means of the two correlated outcomes in the overall population. As in standard Poisson regression, the model relates the marginal mean of each outcome variable with a linear predictor through the log link function, but it also specifies parameters for the underlying mixing distribution of the latent subpopulations. The resulting estimates can be directly used in explaining the effects of exposure variables on the means of the outcomes in the overall population and in

estimating other population-wide parameters such as incidence density ratios. We illustrate the method by using simulations and in the evaluation of the caries preventive effects of a school-based weekly fluoride mouthrinse (FMR) program among North Carolina (NC) schoolchildren.

This article is organized as follows. Section 3.2 reviews bivariate distributions for zero-inflated counts and Section 3.3 discusses marginalized models for such counts. Simulation studies and an application of the proposed model are presented in Sections 3.4 & 3.5, respectively. We conclude with a discussion in Section 3.6.

3.2 Zero-inflated Bivariate Poisson Models

To model multivariate zero-inflated counts, Li et al.(1999) propose mixtures of m-dimensional distributions. For the bivariate case, they assume that a zero-inflated random variable (Y_1, Y_2) arises either from a distribution degenerate at $(0,0)$, from a bivariate Poisson, or from a bivariate distribution with one component degenerate at 0 and the other component having a standard Poisson mass function. That is,

$$(Y_1, Y_2) \sim \begin{cases} (0,0), \text{with probability } p_0 \\ (\text{Poisson}(\lambda_1), 0), \text{with probability } p_1 \\ (0, \text{Poisson}(\lambda_2)), \text{with probability } p_2 \\ \text{Bivariate Poisson}(\lambda_{10}, \lambda_{20}, \lambda_{00}), \text{with probability } p_3, \end{cases} \tag{3.55}$$

where $p_k \geq 0$, $k = 0, 1, 2, 3$, $\sum_{k=0}^{3} p_k = 1$, and λ_1, λ_2, λ_{10}, λ_{20}, $\lambda_{00} > 0$. The probability mass function of the random variable (Y_1, Y_2) is given by,

$$\begin{aligned}
P_{00} &= p_0 + p_1 \exp(-\lambda_1) + p_2 \exp(-\lambda_2) + p_3 \exp(-\lambda) \\
P_{10} &= \frac{p_1 \lambda_1^{y_1} \exp(-\lambda_1) + p_3 \lambda_{10}^{y_1} \exp(-\lambda)}{y_1!} \\
P_{01} &= \frac{p_2 \lambda_2^{y_2} \exp(-\lambda_2) + p_3 \lambda_{20}^{y_2} \exp(-\lambda)}{y_2!} \\
P_{11} &= \sum_{j=0}^{\min(y_1, y_2)} \frac{\lambda_{10}^{y_1 - j} \lambda_{20}^{y_2 - j} \lambda_{00}^{j}}{(y_1 - j)!(y_2 - j)! j!} p_3 \exp(-\lambda),
\end{aligned} \tag{3.56}$$

where $P_{00} = \Pr(Y_1 = 0, Y_2 = 0)$, $P_{10} = \Pr(Y_1 = y_1, Y_2 = 0)$, $P_{01} = \Pr(Y_1 = 0, Y_2 = y_2)$, $P_{11} = \Pr(Y_1 = y_1, Y_2 = y_2)$, $\lambda = \lambda_{00} + \lambda_{10} + \lambda_{20}$, $y_1 > 0$ and $y_2 > 0$.

The zero-inflated bivariate Poisson distribution in (3.56) reduces to the standard bivariate Poisson model for $p_0 = p_1 = p_2 = 0$. For three Poisson random variables W_1, W_2 and W_0 with respective means λ_{10}, λ_{20} and λ_{00}, if $Y_1 = W_1 + W_0$ and $Y_2 = W_2 + W_0$, then (Y_1, Y_2) is distributed as Bivariate Poisson$(\lambda_{10}, \lambda_{20}, \lambda_{00})$. In addition, Y_1 and Y_2 are marginally distributed as Poisson$(\lambda_{10} + \lambda_{00})$ and Poisson$(\lambda_{20} + \lambda_{00})$, respectively. In a similar fashion, when $\lambda_1 = \lambda_{10} + \lambda_{00}$ and $\lambda_2 = \lambda_{20} + \lambda_{00}$ in equation (3.55), the marginal distributions of Y_1 and Y_2 become univariate ZIP (Li et al., 1999). That is,

$$\Pr(Y_t = k) = \begin{cases} (1 - p_t - p_3) + (p_t + p_3) \exp(-\lambda_t), & k = 0 \\ (p_t + p_3) \frac{\exp(-\lambda_t) \lambda_t^k}{k!}, & k = 1, 2, \ldots \end{cases} \tag{3.57}$$

where $t = 1, 2$. In this article, we consider the case where $\lambda_1 = \lambda_{10} + \lambda_{00}$ and $\lambda_2 = \lambda_{20} + \lambda_{00}$. Li et al.(1999) employ directional grid search methods (Powell, 1964) to obtain maximum likelihood estimates of model parameters by using method of moment estimates as initial values. Majundar and Gries (2010) describe a Bayesian approach in conjunction with the EM algorithm to estimate parameters of bivariate zero-inflated regression models, where

they express Y_1 and Y_2 using latent variables as

$$Y_1 = (1 - Z_0)(1 - Z_2)(W_0 + W_1) \tag{3.58}$$
$$Y_2 = (1 - Z_0)(1 - Z_1)(W_0 + W_2),$$

where $W_t \sim \text{Poisson}(\lambda_{t0})$, $t = 0, 1, 2$ and $(Z_0, Z_1, Z_2, Z_3) \sim \text{Multinomial}(1, (p_0, p_1, p_2, p_3))$ with the parameters λ_{00}, λ_{10}, λ_{20}, p_0, p_1, p_2 and p_3 as defined in equation (3.55). In addition, W_0, W_1, W_2 and (Z_0, Z_1, Z_2, Z_3) are assumed independent of each other.

When covariates are used to model bivariate zero-inflated Poisson counts, linear predictors are specified for the mean parameters and the mixing probabilities, for example, as $\log(\lambda_{10i}) = \mathbf{x}'_{1i}\boldsymbol{\alpha}_1$, $\log(\lambda_{20i}) = \mathbf{x}'_{2i}\boldsymbol{\alpha}_2$, $\log(\lambda_{00i}) = \mathbf{x}'_{3i}\boldsymbol{\alpha}_3$, $\log(p_{0i}/p_{3i}) = \mathbf{x}'_{4i}\boldsymbol{\gamma}_0$, $\log(p_{1i}/p_{3i}) = \mathbf{x}'_{5i}\boldsymbol{\gamma}_1$ and $\log(p_{2i}/p_{3i}) = \mathbf{x}'_{6i}\boldsymbol{\gamma}_2$, where $\mathbf{x}_{1i}$, ..., $\mathbf{x}_{6i}$ are vectors of covariates from the i^{th} individual, and $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$, $\boldsymbol{\alpha}_3$, $\boldsymbol{\gamma}_0$, $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ are vectors of parameters (Li et al., 1999; Majundar and Gries 2010). Because the model parameters have latent class interpretations, one has to employ post-modeling transformations to estimate the effects of covariates on the overall population means $\nu_{1i} = E(Y_{1i})$ and $\nu_{2i} = E(Y_{2i})$. The marginal means and the model parameters can be related by

$$\nu_{1i} = (p_{1i} + p_{3i})(\lambda_{00i} + \lambda_{10i}) = \frac{(e^{\mathbf{x}'_{1i}\boldsymbol{\alpha}_1} + e^{\mathbf{x}'_{3i}\boldsymbol{\alpha}_3})(1 + e^{\mathbf{x}'_{5i}\boldsymbol{\gamma}_1})}{1 + e^{\mathbf{x}'_{4i}\boldsymbol{\gamma}_0} + e^{\mathbf{x}'_{5i}\boldsymbol{\gamma}_1} + e^{\mathbf{x}'_{6i}\boldsymbol{\gamma}_2}} \tag{3.59}$$

$$\nu_{2i} = (p_{2i} + p_{3i})(\lambda_{00i} + \lambda_{20i}) = \frac{(e^{\mathbf{x}'_{2i}\boldsymbol{\alpha}_2} + e^{\mathbf{x}'_{3i}\boldsymbol{\alpha}_3})(1 + e^{\mathbf{x}'_{6i}\boldsymbol{\gamma}_2})}{1 + e^{\mathbf{x}'_{4i}\boldsymbol{\gamma}_0} + e^{\mathbf{x}'_{5i}\boldsymbol{\gamma}_1} + e^{\mathbf{x}'_{6i}\boldsymbol{\gamma}_2}}$$

Although ν_{1i} and ν_{2i} could be estimated at fixed covariate values by using equation (3.59), the quantification of the relationship between covariates and the marginal means with appropriate variance estimates may be difficult in practice. In addition, when interest is in determining whether the effects of an exposure variable on ν_{1i} or ν_{2i} are homogeneous across levels of covariates, existing bivariate zero-inflated models usually do not provide

the desired estimates as in the case of traditional zero-inflated models for univariate counts (Long et al., 2014).

3.3 Marginalized Zero-inflated Bivariate Poisson Models

Our primary interest is in modeling the marginal means (ν_{1i}, ν_{2i}) as functions of covariates, while also estimating the nuisance parameters for model completion. For univariate zero-inflated Poisson outcomes, a similar marginalized model is previously discussed in Long et al.(2014) and Preisser et al.(2016). Using (3.59) and eliminating nuisance parameters λ_{10i} and λ_{20i}, the probabilities in equation (3.56) can be written as functions of the marginal means and latent parameters λ_{00i} and (p_{0i}, p_{1i}, p_{2i}). If $\mathbf{x}_{1i}$, $\mathbf{x}_{2i}$ and $\mathbf{x}_{3i}$ are vectors of covariates from the i^{th} individual, we specify regression parameters for ν_{1i}, ν_{2i}, and λ_{00i} as

$$\begin{aligned} log(\nu_{1i}) &= \mathbf{x}'_{1i}\boldsymbol{\beta}_1 \\ log(\nu_{2i}) &= \mathbf{x}'_{2i}\boldsymbol{\beta}_2 \\ log(\lambda_{00i}) &= \mathbf{x}'_{3i}\boldsymbol{\zeta}, \end{aligned} \tag{3.60}$$

where $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ and $\boldsymbol{\zeta}$ are vectors of parameters. To model the multinomial probabilities p_{0i}, p_{1i}, p_{2i} and $p_{3i} = 1 - p_{0i} - p_{1i} - p_{2i}$, multicategory logit models are employed as follows.

$$\begin{aligned} log\Big(\frac{p_{0i}}{p_{3i}}\Big) &= \mathbf{u}'_{1i}\boldsymbol{\gamma}_0 \\ log\Big(\frac{p_{1i}}{p_{3i}}\Big) &= \mathbf{u}'_{2i}\boldsymbol{\gamma}_1 \\ log\Big(\frac{p_{2i}}{p_{3i}}\Big) &= \mathbf{u}'_{3i}\boldsymbol{\gamma}_2, \end{aligned} \tag{3.61}$$

where, $\mathbf{u}'_{1i}$, $\mathbf{u}'_{2i}$ and $\mathbf{u}'_{3i}$ are vectors of covariates and $\boldsymbol{\gamma}_0$, $\boldsymbol{\gamma}_1$, and $\boldsymbol{\gamma}_2$ are vectors of parameters. Using equations (3.56), (3.60) and (3.61) together with the relations $\lambda_{1i} = \lambda_{10i} + \lambda_{00i}$ and

$\lambda_{2i} = \lambda_{20i} + \lambda_{00i}$, the log-likelihood function from n subjects can be written as

$$\begin{aligned}\ell(\boldsymbol{\theta}|Y_1, Y_2, X) = &\sum_{i=1}^{n} I(Y_{1i} = 0, Y_{1i} = 0)\log(P_{00i}) + \sum_{i=1}^{n} I(Y_{1i} > 0, Y_{1i} = 0)\log(P_{10i}) \\ &+ \sum_{i=1}^{n} I(Y_{1i} = 0, Y_{1i} > 0)\log(P_{01i}) + \sum_{i=1}^{n} I(Y_{1i} > 0, Y_{1i} > 0)\log(P_{11i}).\end{aligned} \tag{3.62}$$

In (3.62), $\boldsymbol{\theta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\zeta}, \boldsymbol{\gamma}_0, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)$, X is the design matrix, $P_{00i} = \Pr(Y_{1i} = 0, Y_{2i} = 0)$, $P_{10i} = \Pr(Y_1 = y_{1i}, Y_2 = 0)$, $P_{01i} = \Pr(Y_{1i} = 0, Y_{2i} = y_{2i})$ and $P_{11i} = \Pr(Y_{1i} = y_{1i}, Y_{2i} = y_{2i})$ with $y_{1i} > 0$, $y_{2i} > 0$. The maximum likelihood estimates of the parameters satisfy

$$(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \hat{\boldsymbol{\zeta}}, \hat{\boldsymbol{\gamma}}_0, \hat{\boldsymbol{\gamma}}_1, \hat{\boldsymbol{\gamma}}_2) = \operatorname{argmax} \ell(\boldsymbol{\theta}|Y_1, Y_2, X).$$

With a proper choice of starting values, we perform parameter estimation employing quasi-Newton algorithms. Starting values for $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ may be obtained from separate MZIP models fitted to Y_1 and Y_2 respectively, and estimates from the bivariate zero-inflated models discussed in Section 3.2 may be used as starting values for $\boldsymbol{\zeta}$, $\boldsymbol{\gamma}_0$, $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$.

3.4 Simulation Study

To evaluate the performance of the marginalized bivariate zero-inflated Poisson (MBZIP) model in finite samples, simulation studies were performed for various sample sizes. Let (Y_{1i}, Y_{2i}) be a zero-inflated bivariate outcome and x_{1i}, x_{2i} and x_{3i} be covariates from the i^{th}

subject. Data were generated from the model,

$$\begin{aligned}
\log(\nu_{1i}) &= \beta_{10} + \beta_{11}x_{1i} + \beta_{12}x_{2i} + \beta_{13}x_{3i} = \mathbf{x}_i'\boldsymbol{\beta}_1 \\
\log(\nu_{2i}) &= \beta_{20} + \beta_{21}x_{1i} + \beta_{22}x_{2i} + \beta_{23}x_{3i} = \mathbf{x}_i'\boldsymbol{\beta}_2 \\
\log(\lambda_{00i}) &= \zeta_0 \\
\log(p_{0i}/p_{3i}) &= \gamma_{00} \\
\log(p_{1i}/p_{3i}) &= \gamma_{10} + \gamma_{11}x_{1i} \\
\log(p_{2i}/p_{3i}) &= \gamma_{20} + \gamma_{21}x_{1i} + +\gamma_{22}x_{2i},
\end{aligned} \tag{3.63}$$

where $x_{1i} \sim N(1,1)$, $x_{2i} \sim$ Binomial(1, 0.4), $x_{3i} \sim$ Exponential(2), (β_{10}, β_{11}, β_{12}, β_{13}) = (1, –0.2, 0.3, 0.2), (β_{20}, β_{21}, β_{22}, β_{23}) = (1, 0.2, 0.3, 0.4), ζ_0 = 0.5, (γ_{00}, γ_{10}, γ_{11}, γ_{20}, γ_{21}, γ_{22}) = (–0.25, –0.2, –0.3, –0.4, 0.2, –0.2) and $p_{3i} = 1 - p_{0i} - p_{1i} - p_{2i}$. To generate the bivariate outcome, first the latent parameters λ_{10i} and λ_{20i} were estimated from equations (3.59) and (3.63), and Y_{1i} and Y_{2i} were determined from

$$\begin{aligned}
Y_{1i} &= (1 - Z_{0i})(1 - Z_{2i})(W_{0i} + W_{1i}) \\
Y_{2i} &= (1 - Z_{0i})(1 - Z_{1i})(W_{0i} + W_{2i}),
\end{aligned} \tag{3.64}$$

where $(Z_{0i}, Z_{1i}, Z_{2i}, Z_{3i}) \sim$ Multinomial$(1, (p_{0i}, p_{1i}, p_{2i}, p_{3i}))$, $W_{0i} \sim$ Poisson(λ_{00i}), $W_{1i} \sim$ Poisson(λ_{10i}) and $W_{2i} \sim$ Poisson(λ_{20i}) are independent random variables.

Using these specifications, 10000 samples of sizes n = 100, 200, 500 and 1000 were generated, and the MBZIP model was fitted for each replication. In the quasi-Newton optimization, starting values for the marginal mean model parameters were obtained from univariate MZIP models fitted separately for the two outcomes. Additionally, estimates from the bivariate zero-inflated model of Li et al.(1999) were used as starting values for ζ_0, γ_{00}, γ_{10}, γ_{11}, γ_{20}, γ_{21} and γ_{22}. From each model, parameter estimates, the corresponding percent relative median biases, standard errors, coverages of 95% confidence intervals and

Type I error rates with the nominal value set at 0.05 for testing $H_0 : \beta_{11} = 0$ and $H_0 : \beta_{21} = 0$, were retained. In each simulation, univariate MZIP models (with covariates x_{1i} and x_{2i} in the excess zero model parts) were fitted for the two outcomes to allow comparisons of MBZIP and MZIP model performances. MBZIP model convergence rates for sample sizes n = 100, 200, 500 and 1000 were 96.0%, 99.1%, 99.9% and 100%, respectively.

Table 3.1 shows that the MBZIP model has low biases for parameters in the marginal mean models and that the biases tend to decrease as sample sizes increase. Although biases of the marginal mean parameters from the MZIP models are generally low, parameters β_{10}, β_{20}, β_{11} and β_{21} have larger biases than the corresponding MZBIP model parameters for all sample sizes. Coverages of 95% confidence intervals for the marginal mean parameters of the MBZIP and MZIP models are also close to 95%. From Table 3.2, we note that coefficients for the mixing probabilities in MBZIP model have low biases and coverages of 95% confidence intervals that are close to the nominal value; and that the estimate for λ_{00} (i.e., ζ_0) has larger biases and smaller coverage probabilities for smaller sample sizes. However, the biases and coverage probabilities for ζ_0 tend to improve as sample sizes increase resulting in a small bias and a coverage probability close to the nominal value when the sample sizes reach 1000.

Table 3.3 presents mean standard errors and Monte Carlo standard deviations of parameters in the marginal mean models of MBZIP and MZIP. For each sample size, mean standard errors and Monte Carlo standard deviations of the marginal parameters from the MBZIP model are very close to each other and they are almost identical, for sample sizes $n = 500$, and 1000. In general, mean standard errors and Monte Carlo standard deviations from the MZIP models are higher than the corresponding estimates from MBZIP models, highlighting the statistical efficiency that is gained by modeling the two correlated outcomes jointly. Regarding the nuisance parameters in the MBZIP model, Table 3.2 shows that the mean standard errors and Monte Carlo standard deviations of parameters γ_{00}, γ_{10}, γ_{11}, γ_{20}, γ_{21} and γ_{22} are very close to each other, but the mean standard errors for ζ_0 are much higher

than the corresponding Monte Carlo standard deviations for sample sizes 100 and 200. The difference between the two quantities decreases as sample sizes increase.

Two scenarios were employed to compute Type I error rates for β_{11} and β_{21} in the MBZIP model and the corresponding coefficients of x_{1i} in MZIP models for each outcome. In the first scenario, data were generated by setting $\beta_{11} = 0$ or $\beta_{21} = 0$ separately and the error rates were calculated from MBZIP and MZIP models. In the second case, the Type I error rates were calculated separately for β_{11} and β_{21}, but data were generated by setting $\beta_{11} = \beta_{21} = 0$. As can be seen from Table 3.4, Type I error rates of the marginal parameters in the MBZIP model are close to the nominal value for the two parameters and under both scenarios.

3.5 Application to a School-based Fluoride Mouthrinse Program

This analysis is aimed at estimating the caries preventive effects of a school-based fluoride mouthrinse program (FMR) on North Carolina (NC) schoolchildren, based on clinical and parent reported data from a probability sample of NC schoolchildren in grades 1 through 5. As measures of caries experiences, clinical data on counts of decayed and filled primary tooth surfaces (dmfs) and the corresponding counts of permanent tooth surfaces (DMFS) were collected. The exposure variable of interest is the number of years of participation in the FMR program (Years). While the original data involved a total of 1363 children, only 677 of them had complete outcome and covariate values. The data exhibit high proportions of zeros on both outcomes variables: out of the 677 children with complete data, 330 (48.7 %) had zero dmfs and 512 (75.6 %) had zero DMFS counts. Previously, Divaris et al.(2012) employed zero-inflated negative binomial regression to fit separate models for the dfms counts and the sum of the two outcomes (i.e., dfms + DFMS) by including the exposure variable as well as other demographic and dental care related covariates in the linear predictors. Because primary and permanent caries counts are obtained from the same child, dmfs and DMFS values are correlated (corr. coef. = 0.15, p-value < 0.0001).

We modeled the dfms and DFMS outcomes jointly by including the main exposure

variable and the same adjustment variables as in Divaris et al.(2012). The model is given by

$$\begin{aligned}
\log(\nu_{1i}) &= \beta_{10} + \beta_{11}x_{1i} + \beta_{12}x_{2i} + \beta_{13}x_{3i} + \beta_{14}x_{4i} + \beta_{15}x_{5i} + \beta_{16}x_{6i} + \beta_{17}x_{7i} + \beta_{18}x_{8i} \\
&\quad + \beta_{19}x_{9i} + \beta_{110}x_{10i} + \beta_{111}x_{11i} \\
\log(\nu_{2i}) &= \beta_{20} + \beta_{21}x_{1i} + \beta_{22}x_{2i} + \beta_{23}x_{3i} + \beta_{24}x_{4i} + \beta_{25}x_{5i} + \beta_{26}x_{6i} + \beta_{27}x_{7i} + \beta_{28}x_{8i} \\
&\quad + \beta_{29}x_{9i} + \beta_{210}x_{10i} + \beta_{211}x_{11i} \\
\log(p_{0i}/p_{3i}) &= \gamma_{10} + \gamma_{10}x_{1i} + \gamma_{10}x_{2i} + \gamma_{10}x_{3i} + \gamma_{10}x_{4i} + \gamma_{10}x_{5i} + \gamma_{10}x_{6i} + \gamma_{10}x_{7i} + \gamma_{10}x_{8i} \\
&\quad + \gamma_{10}x_{9i} + \gamma_{10}x_{10i} + \gamma_{10}x_{11i} \\
\log(p_{1i}/p_{3i}) &= \gamma_{10} \\
\log(p_{2i}/p_{3i}) &= \gamma_{20} \\
\log(\lambda_{00i}) &= \zeta_0.
\end{aligned}$$

where, ν_{1i} and ν_{2i} are the i^{th} marginal dmfs and DMFS mean counts respectively, x_{1i} is Years divided by 3, x_{2i} is an indicator of whether the child is African American, x_{3i} is a binary indicator for other non-Caucasian race, x_{4i} is the child's brushing frequency, x_{5i} is family income in $ 10,000, x_{6i} is an indicator for the availability of established dental home, x_{7i} is an indicator for whether the child had sealants, x_{8i} is an indicator for availability of dental care when needed, and x_{9i}, x_{10i} and x_{11i} are the child's age centered at the mean, its square and cubic values respectively.

Table 3.5 shows parameter estimates and standard errors of the MBZIP model and the marginal parameters of MZIP models fitted for dmfs and DMFS separately. In both parts of the bivariate model and each of the MZIP models, the exposure variable (Years) has negative estimates. Based on the MBZIP model, the estimated incidence rate ratios for the overall effect of three years participation in the fmr program on dmfs and DMFS counts are exp(−0.058) = 0.944 (95% CI: (0.83,1.08)) and exp(−0.013) = 0.987 (95% CI: (0.80,1.22)),

respectively. Thus, conditional on covariates, the mean dmfs count for a child in the overall population with three years participation in the fmr program is approximately 94.4% of the mean dmfs count of a child with zero years of participation. Similarly, on average, three years of participation in the program corresponds to a 1.3% reduction in DMFS counts in the overall population. However, because the confidence interval of each incidence rate ratio includes the value 1.0, the association between the exposure variable and dmfs or DMFS counts is not significant. Likewise, the associations between Years and the two caries counts are not significant based on estimates from the MZIP models. An advantage of a bivariate model for the two outcomes is that one can perform joint statistical tests across the two sets of regression parameters. Testing for the overall effect of Years on dmfs and DMFS counts (i.e., null hypothesis $H_0 : \beta_{11} = \beta_{21} = 0$) gives a likelihood ratio statistic of 0.772 with a p-value of 0.680, confirming a lack of significant overlall exposure effect on dmfs and DMFS counts.

3.6 Discussion

In this article, we proposed a joint marginalized model for two correlated counts with zero-inflation. The model specifies regression coefficients to the marginal means of the two outcomes and provides estimates that allow direct inferences about the overall population. Unlike traditional bivariate count models, parameters from the marginalized model have straightforward interpretations in describing the effects of explanatory variables on the marginal means of the two correlated, zero-inflated counts; and can easily be employed to determine the relationships between covariates and population-wide parameters such as incidence density ratios with appropriate variance estimates. Under the marginalized model, counts are assumed to have come from four latent classes: a 'non-susceptible' or 'perfect' state, from which both outcomes take zero-values, two "partially-susceptible" states in which one outcome takes only zeros and the other follows a Poisson distribution, and 'susceptible" class where the two outcomes take both zero and positive counts according to a bivariate Poisson distribution. Earlier approaches to model such counts utilize four component mixtures of bivariate distributions and estimate regression parameters with latent class interpretations. These parameters, however, are not suitable when interest is to make inferences about the overall population. The new model extends univariate marginalized models by accommodating two correlated outcomes, and modifies existing bivariate models for zero-inflated counts by directly estimating overall covariate effects in the population, when interest is in studying the relationships between the covariates and the two marginal means.

Our simulations show that estimates of the marginal parameters in the model have low biases with Type I error rates and coverage probabilities close to the nominal values. When the MBZIP model is correctly specified, the mean standard errors for these parameters are very close to the Monte Carlo standard deviations of their estimates. Except for the small sample estimates of one parameter, estimates of the nuisance parameters in the model

also have low biases and good Type I error and coverage properties. In the simulations, the new model provides smaller standard error estimates than marginalized zero-inflated Poisson models separately fitted for each outcome; underscoring the potential for statistical efficiency gains from modeling the two outcomes jointly. In its application to evaluate the caries preventive effects of a school-based fluoride mouthrinse program among North Carolina schoolchildren, the marginal model estimated the effects of the exposure variable and other covariates on the marginal means as well as incidence density ratios with confidence intervals. An advantage of using the marginalized bivariate model is that it allows hypothesis testing across parameters of the two outcomes. A likelihood ratio test showed that participation in the fluoride mouthrinse program was not significantly associated with caries counts in primary and permanent teeth. Except for a few cases, estimates from the model also have smaller standard errors than similar univariate models applied to each outcome.

The MBZIP model should be used with caution when extra-Poisson dispersion in addition to excess zeros is suspected. Preisser et al.(2016) showed that the univariate MZIP model gives inflated Type I error and poor coverage of 95% confidence intervals when the true model is marginalized zero-inflated negative binomial regression; similar results are expected to apply to the bivariate setting.

While we performed estimation by direct maximization of the likelihood function with carefully selected starting values, applications of Bayesian methods or the expectation-maximization algorithm could provide alternative estimation methods. Future research could also extend the model to handle three or more correlated outcomes with zero-inflation or to counts that are overdispersed in addition to zero-inflated. Another possible extension could be the modeling of repeated or longitudinal data in problems where the bivariate zero-inflated outcome is measured repeatedly for each sampling unit.

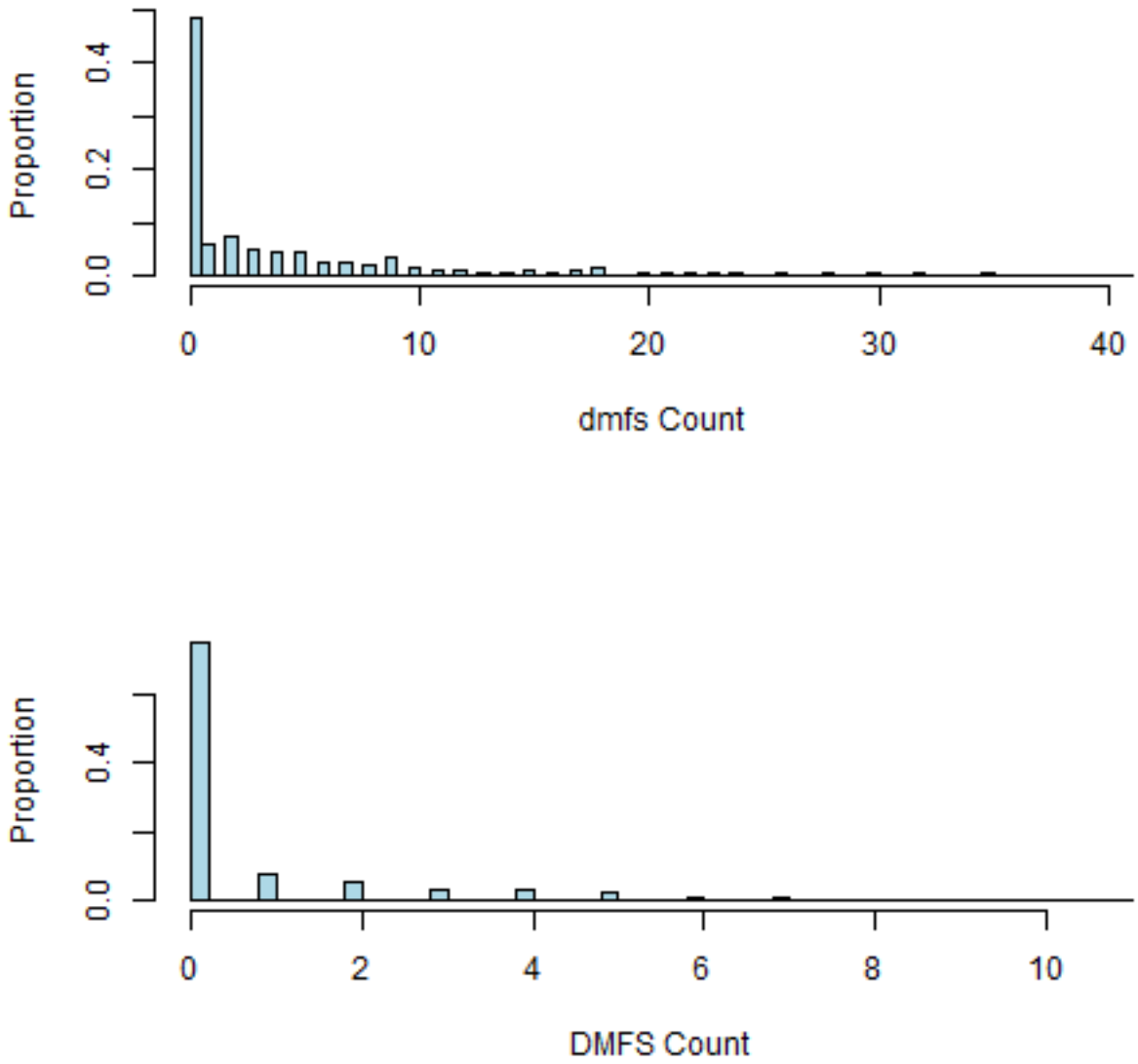

Figure 3.1: Distributions of dmfs and DFMS counts from 677 children in the NC FMR study.

Table 3.1: Percent relative median biases and coverages of 95% confidence intervals of MBZIP and MZIP model estimates based on 10,000 replications.

		MBZIP		MZIP	
Sample Size	Par.	% Rel. Med. Bias	Cov. Prob.	% Rel. Med. Bias	Cov. Prob.
100	β_{10}	-0.54	94.2	-1.12	95.2
	β_{11}	-0.29	93.6	-2.06	94.7
	β_{12}	-1.44	94.5	0.97	95.1
	β_{13}	-0.62	95.1	-0.72	95.4
	β_{20}	0.17	94.9	-0.42	94.5
	β_{21}	-0.90	94.3	1.31	93.2
	β_{22}	-0.55	94.5	-0.34	95.0
	β_{23}	-0.38	95.2	-0.07	95.6
200	β_{10}	-0.39	94.6	-0.65	94.4
	β_{11}	-0.27	94.6	-1.00	95.1
	β_{12}	1.11	94.6	1.95	95.1
	β_{13}	-0.31	94.9	-0.31	95.2
	β_{20}	-0.02	94.7	-0.42	94.4
	β_{21}	-0.15	94.6	0.17	93.9
	β_{22}	-0.40	95.1	0.44	95.4
	β_{23}	0.15	95.2	0.12	95.4
500	β_{10}	-0.09	94.9	-0.34	95.3
	β_{11}	0.18	95.1	-0.96	95.0
	β_{12}	0.03	95.2	0.74	95.0
	β_{13}	-0.06	95.0	0.08	95.1
	β_{20}	-0.05	95.4	-0.50	95.1
	β_{21}	0.00	94.9	1.70	94.8
	β_{22}	-0.41	95.0	0.37	94.8
	β_{23}	0.03	95.0	0.06	94.9
1000	β_{10}	-0.07	95.0	-0.26	95.0
	β_{11}	-0.41	95.0	-0.81	95.2
	β_{12}	-0.39	95.2	1.07	95.3
	β_{13}	-0.13	95.3	-0.02	95.5
	β_{20}	0.00	94.9	-0.38	94.8
	β_{21}	-0.12	95.2	1.22	94.7
	β_{22}	0.24	95.0	1.20	95.1
	β_{23}	-0.06	94.7	-0.07	94.8

Table 3.2: Percent relative median biases, mean standard errors, Monte Carlo standard deviations and coverages of 95% confidence intervals of nuisance parameters in the MBZIP models, based on data generated from the MBZIP model with 10,000 replications.

Sample size	Parameter	Percent rel. med. bias	Mean SE	MC SD	Cov. Prob.
100	ζ_0	27.86	42.085	3.258	82.8
	γ_{00}	4.44	0.273	0.280	95.3
	γ_{10}	2.66	0.363	0.383	95.0
	γ_{11}	6.08	0.260	0.281	93.1
	γ_{20}	5.12	0.440	0.473	93.7
	γ_{21}	4.74	0.250	0.274	92.8
	γ_{22}	-0.02	0.501	0.529	95.1
200	ζ_0	11.97	7.702	1.678	88.6
	γ_{00}	0.70	0.193	0.194	95.1
	γ_{10}	-0.66	0.251	0.259	94.7
	γ_{11}	4.37	0.178	0.187	93.9
	γ_{20}	1.00	0.307	0.316	94.5
	γ_{21}	2.27	0.173	0.180	94.0
	γ_{22}	5.92	0.347	0.358	94.6
500	ζ_0	5.04	0.533	0.464	92.5
	γ_{00}	0.92	0.121	0.122	94.9
	γ_{10}	0.90	0.156	0.156	95.4
	γ_{11}	1.94	0.110	0.111	95.0
	γ_{20}	1.62	0.192	0.195	94.7
	γ_{21}	1.15	0.108	0.111	94.3
	γ_{22}	-0.12	0.216	0.218	95.0
1000	ζ_0	1.63	0.217	0.223	93.6
	γ_{00}	-0.07	0.086	0.086	94.9
	γ_{10}	1.30	0.110	0.109	95.0
	γ_{11}	0.32	0.077	0.077	95.0
	γ_{20}	0.82	0.135	0.135	95.3
	γ_{21}	0.70	0.076	0.076	94.9
	γ_{22}	-0.84	0.152	0.151	95.2

Table 3.3: Mean standard errors and Monte Carlo standard deviations of MBZIP and MZIP model estimates, based on data generated from the MBZIP model with 10,000 replications.

		MBZIP		MZIP	
Sample Size	Parameter	Mean SE	MC SD	Mean SE	MC SD
100	β_{10}	0.164	0.171	0.190	0.194
	β_{11}	0.088	0.092	0.112	0.114
	β_{12}	0.165	0.171	0.232	0.239
	β_{13}	0.122	0.128	0.123	0.127
	β_{20}	0.157	0.161	0.182	0.187
	β_{21}	0.075	0.077	0.097	0.104
	β_{22}	0.131	0.136	0.204	0.212
	β_{23}	0.086	0.090	0.087	0.090
200	β_{10}	0.114	0.117	0.132	0.133
	β_{11}	0.061	0.063	0.078	0.078
	β_{12}	0.114	0.118	0.161	0.163
	β_{13}	0.081	0.083	0.081	0.083
	β_{20}	0.110	0.111	0.128	0.130
	β_{21}	0.052	0.053	0.068	0.071
	β_{22}	0.092	0.092	0.142	0.143
	β_{23}	0.057	0.058	0.057	0.058
500	β_{10}	0.071	0.072	0.082	0.083
	β_{11}	0.038	0.038	0.049	0.049
	β_{12}	0.072	0.072	0.101	0.102
	β_{13}	0.049	0.049	0.049	0.050
	β_{20}	0.069	0.068	0.080	0.080
	β_{21}	0.033	0.033	0.043	0.043
	β_{22}	0.057	0.058	0.089	0.090
	β_{23}	0.034	0.034	0.034	0.035
1000	β_{10}	0.050	0.050	0.058	0.058
	β_{11}	0.027	0.027	0.035	0.034
	β_{12}	0.051	0.050	0.071	0.071
	β_{13}	0.034	0.034	0.034	0.034
	β_{20}	0.049	0.049	0.057	0.057
	β_{21}	0.023	0.023	0.030	0.030
	β_{22}	0.040	0.040	0.063	0.062
	β_{23}	0.023	0.024	0.024	0.024

Table 3.4: Type I errors of β_{11} and β_{21} from MBZIP and MZIP models based on Wald type tests, based on data generated from the MBZIP model with 10,000 replications.

Sample Size	Data Generation	Parameter	MBZIP	MZIP
100	Case 1	β_{11}	0.060	0.055
		β_{21}	0.060	0.063
	Case 2	β_{11}	0.062	0.058
		β_{21}	0.058	0.063
200	Case 1	β_{11}	0.052	0.051
		β_{21}	0.053	0.058
	Case 2	β_{11}	0.053	0.052
		β_{21}	0.054	0.058
500	Case 1	β_{11}	0.048	0.051
		β_{21}	0.054	0.051
	Case 2	β_{11}	0.051	0.047
		β_{21}	0.052	0.054
1000	Case 1	β_{11}	0.050	0.047
		β_{21}	0.054	0.054
	Case 2	β_{11}	0.049	0.050
		β_{21}	0.051	0.057

**Case 1: data generated by setting $\beta_{11} = 0$ or $\beta_{21} = 0$.*
**Case 2: data generated by setting $\beta_{11} = \beta_{21} = 0$.*

Table 3.5: Parameter estimates and standard errors for the NC FMR data based on MBZIP and MZIP models.

		MBZIP		MZIP	
Variable	Parameter	Estimate	SE	Estimate	SE
Marginal mean model for dmfs					
Intercept	β_{10}	1.189	0.170	1.468	0.178
Years	β_{11}	-0.058	0.068	-0.084	0.079
African Amer.	β_{12}	-0.369	0.079	-0.464	0.090
Other race	β_{13}	-0.789	0.186	-0.974	0.213
Brushing freq.	β_{14}	0.019	0.050	-0.017	0.055
Fam. income	β_{15}	-0.170	0.017	-0.213	0.019
Dental home	β_{16}	0.348	0.085	0.359	0.093
No access	β_{17}	0.371	0.063	0.426	0.071
Age	β_{18}	-0.099	0.041	-0.120	0.048
Age-sq	β_{19}	-0.013	0.013	-0.044	0.015
Age-cu	β_{110}	-0.013	0.007	-0.019	0.009
Sealants	β_{111}	0.865	0.068	0.954	0.074
Marginal mean model for DMFS					
Intercept	β_{20}	0.220	0.361	-0.253	0.473
Years	β_{21}	-0.013	0.108	-0.076	0.140
African Amer.	β_{22}	-0.243	0.158	-0.104	0.199
Other race	β_{23}	-0.548	0.291	0.529	0.273
Brushing freq.	β_{24}	-0.224	0.102	-0.290	0.142
Fam. income	β_{25}	-0.154	0.030	-0.092	0.040
Dental home	β_{26}	0.411	0.185	0.767	0.236
No access	β_{27}	0.213	0.143	0.362	0.180
Age	β_{28}	0.304	0.074	0.563	0.100
Age-sq	β_{29}	-0.232	0.047	-0.055	0.053
Age-cu	β_{210}	0.033	0.016	-0.015	0.019
Sealants	β_{211}	0.660	0.124	0.170	0.155

Table 3.6: Continued: parameter estimates and standard errors for the NC FMR data based on MBZIP and MZIP models.

		MBZIP		MZIP	
Variable	Parameter	Estimate	SE	Estimate	SE
Estimates for nuisance parameters in the MBZIP model					
	ζ_0	-0.465	0.195		
	γ_{00}	0.033	0.483		
	γ_{01}	0.038	0.148		
	γ_{02}	0.531	0.227		
	γ_{03}	1.911	0.302		
	γ_{04}	-0.069	0.124		
	γ_{05}	0.429	0.040		
	γ_{06}	-0.411	0.192		
	γ_{07}	-0.513	0.169		
	γ_{08}	0.039	0.085		
	γ_{09}	-0.021	0.032		
	γ_{010}	0.042	0.014		
	γ_{011}	-1.602	0.150		
	γ_{10}	0.247	0.140		
	γ_{20}	-0.613	0.163		

CHAPTER 4: MARGINALIZED ZERO-INFLATED POISSON MODELS WITH MISSING COVARIATES

4.1 Introduction

Counts collected in many applications often contain higher frequencies of zeros than assumed by the Poisson distribution. For example, in dental caries studies 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 with 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; Preisser et al., 2012; Long et al., 2014). 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 processes. 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 using the logit and the log links respectively.

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 (Preisser et al., 2012; Albert et al., 2014; Long 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 (Preisser et al., 2012). 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 (Long et al., 2014).

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, missing data are a common occurrence in practice. 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 unobserved data. Even when CC analysis is valid, estimates can be inefficient when too many observations are missing (Ibrahim et al., 2005). For problems 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, Chen and Lipsitz (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. Sections 4.2 and 4.3 review zero-inflated Poisson and marginalized zero-inflated Poisson models respectively. Section 4.4 describes Monte Carlo EM (MCEM) methods for MZIP models with missing covariates. Section 4.5 presents simulation studies that compare results from the proposed method with those from complete case analysis. Section 4.6 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 4.7.

4.2 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; Long et al., 2014). 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

$$Pr(Y_i = k) = \begin{cases} \psi_i + (1-\psi_i)\exp(-\mu_i), & k = 0 \\ (1-\psi_i)\frac{\exp(-\mu_i)\mu_i^k}{k!}, & k = 1, 2, ... \end{cases} \tag{4.65}$$

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 (Long et al., 2014). It can be seen from equation (4.65) that ZIP reduces to the standard Poisson model when $\psi_i = 0$. 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

$$logit(\psi_i) = \mathbf{z}_i'\boldsymbol{\gamma} \quad \text{and} \quad log(\mu_i) = \mathbf{x}_i'\boldsymbol{\beta}, \tag{4.66}$$

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 $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_q)'$ and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)'$ are regression parameters. For n independent observations, the ZIP likelihood function is

$$L(\boldsymbol{\beta}, \boldsymbol{\gamma}|\mathbf{y}) = \prod_{i=1}^{n} \{1 + e^{\mathbf{z}_i'\gamma}\}^{-1} \left\{e^{\mathbf{z}_i'\gamma} + e^{-\exp(\mathbf{x}_i'\boldsymbol{\beta})}\right\}^{I(y_i=0)} \left\{\frac{e^{-\exp(\mathbf{x}_i'\boldsymbol{\beta})}e^{\mathbf{x}_i'\boldsymbol{\beta}}}{y_i!}\right\}^{I(y_i>0)} \tag{4.67}$$

In equation (4.67), $\mathbf{y}$ is the vector of count outcomes, and $I(T)$ takes the value 1 if T is true and takes zero, otherwise. While interpretations of parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ pertain to the two latent populations, the overall, marginal mean response, $\nu_i = E(y_i|\mathbf{z}_i, \mathbf{x}_i)$, for the i^{th} subject could be estimated from the ZIP model by

$$\nu_i = \frac{e^{\mathbf{x}_i'\boldsymbol{\beta}}}{1 + e^{\mathbf{z}_i'\boldsymbol{\gamma}}}. \tag{4.68}$$

However, the quantification of the relationship between covariates and 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).

4.3 Marginalized ZIP Models

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 excess zero (i.e., ψ_i). For the i^{th} observation, MZIP relates ν_i and ψ_i with the independent variables as :

$$logit(\psi_i) = \mathbf{z}_i'\boldsymbol{\gamma} \quad \text{and} \quad log(\nu_i) = \mathbf{x}_i'\boldsymbol{\alpha}. \tag{4.69}$$

In equation (4.69), ψ_i and $\boldsymbol{\gamma}$ have the same interpretation as in ZIP, and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_p)'$ is a vector of regression parameters for ν_i having interpretations as log incidence density ratios for the entire sampled population. The mean μ_i of the Poisson part of ZIP and the overall mean ν_i are related by equation (4.68), and the MZIP likelihood function is obtained by replacing μ_i by $\frac{\nu_i}{1-\psi_i}$ in the ZIP likelihood in equation (4.67). Thus, for n independent subjects, the log-likelihood function from the marginalized ZIP model is

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

Long et al.(2014) employ quasi-Newton optimization methods for complete data to obtain parameter estimates. 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, robust standard errors are estimated.

4.4 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, Chen and Lipsitz (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 missing covariates to be log-concave, and the method of Ibrahim, Chen and Lipsitz (1999) can be applied to any settings where the log-concavity criterion is met. In the case of MZIP models, because the conditional distribution of the count outcome is not log-concave, conditional distributions of missing covariates generally fail to be 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 $\mathbf{y}' = (y_1, y_2, \ldots, y_n)$ is a vector of independent zero-inflated count outcomes from n subjects, and let $\mathbf{z}_i' = (z_{i1}, z_{i2}, \ldots, z_{iq})$ and $\mathbf{x}_i' = (x_{i1}, x_{i2}, \ldots, x_{ip})$ be the covariate vectors in the MZIP model in equation (4.69). 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 $\mathbf{z}_i = (\tilde{\mathbf{z}}_i', \mathbf{w}_i')'$ and $\mathbf{x}_i = (\tilde{\mathbf{x}}_i', \mathbf{w}_i')'$, where $\mathbf{w}_i$ represents covariates common

to $\mathbf{z}_i$ and $\mathbf{x}_i$, while $\tilde{\mathbf{z}}_i$ and $\tilde{\mathbf{x}}_i$ 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, Chen and Lipsitz (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 $\mathbf{r}_i' = (r_{i1}, r_{i2}, \ldots, r_{ik})$ is a vector of missingness indicators for the k covariates and

$$r_{ij} = \begin{cases} 1, & \text{if the } j^{th} \text{ component of } \mathbf{u}_i \text{ is observed,} \\ 0, & \text{otherwise.} \end{cases} \tag{4.70}$$

When covariate values are missing at random (MAR) (Little and Rubin, 2002), the conditional distribution of $\mathbf{r}_i$ given the data is a function only of the observed data and not depend on any missing values, i.e.,

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

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, 2005) and estimation can be done using the likelihood

$$\begin{aligned} L(\boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\gamma} | \mathbf{y}, \mathbf{u}^{obs}, \mathbf{u}^{mis}) &= \prod_{i=1}^{n} Pr(y_i | \mathbf{u}_i^{obs}, \mathbf{u}_i^{mis}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) Pr(\mathbf{u}_i^{mis} | \mathbf{u}_i^{obs}, \boldsymbol{\xi}) \\ &= \prod_{i=1}^{n} L_i(\boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\gamma} | y_i, \mathbf{u}_i^{obs}, \mathbf{u}_i^{mis}), \end{aligned} \tag{4.71}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ are the regression parameters in equation (4.69), $\boldsymbol{\xi}$ 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. Note that the conditional distributions $Pr(\mathbf{u}_i^{mis} | \mathbf{u}_i^{obs}, \boldsymbol{\xi})$ are used in equation (4.71) because the joint distribution of the

covariates is proportional to the distribution of the missing covariates conditional on the observed covariates. From equation (4.71), the complete data log-likelihood $\ell(\boldsymbol{\theta}|\mathbf{y}, \mathbf{u}^{obs}, \mathbf{u}^{mis})$ can be written as:

$$\ell(\boldsymbol{\theta}|\mathbf{y}, \mathbf{u}^{obs}, \mathbf{u}^{mis}) = \sum_{i=1}^{n} \ell(\boldsymbol{\eta}|y_i; \mathbf{u}_i^{obs}, \mathbf{u}_i^{mis}) + \sum_{i=1}^{n} \ell(\boldsymbol{\xi}|\mathbf{u}_i^{mis}; \mathbf{u}_i^{obs}) \tag{4.72}$$

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

The observed data log-likelihood is obtained by integrating (summing) $\ell(\boldsymbol{\theta}|\mathbf{y}, \mathbf{u}^{obs}, \mathbf{u}^{mis})$ 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,

$$Q_i(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = E(\ell(\boldsymbol{\theta}|y_i, \mathbf{u}_i^{obs}, \mathbf{u}_i^{mis})|y_i, \mathbf{u}_i^{obs}, \boldsymbol{\theta}^{(t)}) \tag{4.73}$$

Had the expectation in equation (4.73) been easily obtained, the M-step of EM would have maximized $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \sum_{i=1}^{n} Q_i(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ 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 $Q_i(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ 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 (Ibrahim et al., 1999),

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

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 continuous covariates, $Pr(\mathbf{u}_i^{mis}|y_i, \mathbf{u}_i^{obs}, \boldsymbol{\theta}^{(t)})$, can be written as,

$$Pr(\mathbf{u}_i^{mis}|y_i, \mathbf{u}_i^{obs}, \boldsymbol{\theta}^{(t)}) = \frac{Pr(y_i|\mathbf{u}_i^{obs}, \boldsymbol{\eta}^{(t)})Pr(\mathbf{u}_i^{mis}|\mathbf{u}_i^{obs}, \boldsymbol{\xi}^{(t)})}{\int Pr(y_i|\mathbf{u}_i^{obs}, \boldsymbol{\eta}^{(t)})Pr(\mathbf{u}_i^{mis}|\mathbf{u}_i^{obs}, \boldsymbol{\xi}^{(t)})d\mathbf{u}_i^{mis}}. \tag{4.74}$$

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 (4.74). 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 missing covariates in MZIP models.

Given the maximum likelihood estimates $\hat{\boldsymbol{\theta}}$ from MCEM, the observed information matrix $I(\hat{\boldsymbol{\theta}})$ 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

$$\begin{aligned} I(\hat{\boldsymbol{\theta}}) &= -\sum_{i=1}^{n}\frac{1}{s}\sum_{j=1}^{s}\frac{\partial^2\ell(\boldsymbol{\theta}|y_i, \mathbf{d}_{ij}, \mathbf{u}_i^{obs})}{\partial\boldsymbol{\theta}^2}\Big|_{(\boldsymbol{\theta}=\hat{\boldsymbol{\theta}})} \\ &- \sum_{i=1}^{n}\frac{1}{s}\sum_{j=1}^{s}\frac{\partial\ell(\boldsymbol{\theta}|y_i, \mathbf{d}_{ij}, \mathbf{u}_i^{obs})}{\partial\boldsymbol{\theta}}\left\{\frac{\partial\ell(\boldsymbol{\theta}|y_i, \mathbf{d}_{ij}, \mathbf{u}_i^{obs})}{\partial\boldsymbol{\theta}}\right\}'\Big|_{(\theta=\hat{\boldsymbol{\theta}})} \\ &+ \sum_{i=1}^{n}\left\{\frac{1}{s}\sum_{j=1}^{s}\frac{\partial\ell(\boldsymbol{\theta}|y_i, \mathbf{d}_{ij}, \mathbf{u}_i^{obs})}{\partial\boldsymbol{\theta}}\right\}\left\{\frac{1}{s}\sum_{j=1}^{s}\frac{\partial\ell(\boldsymbol{\theta}|y_i, \mathbf{d}_{ij}, \mathbf{u}_i^{obs})}{\partial\boldsymbol{\theta}}\right\}'\Big|_{(\boldsymbol{\theta}=\hat{\boldsymbol{\theta}})} \end{aligned} \tag{4.75}$$

Standard errors of parameter estimates are calculated by

$$\mathbf{se}(\hat{\boldsymbol{\theta}}) = \sqrt{\text{diagonal}[I(\hat{\boldsymbol{\theta}})^{-1}]}. \tag{4.76}$$

4.5 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 (4.65), with $\mu_i = \nu_i/(1-\psi_i)$ and (ψ_i, ν_i) defined by

$$\begin{aligned} logit(\psi_i) &= \gamma_0 + \gamma_1 x_{i1} + \gamma_2 x_{i2} \\ log(\nu_i) &= \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} \end{aligned} \tag{4.77}$$

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 $\sigma^2 = 1$, $x_{i1} \sim N(\omega_0 + \omega_1 x_{i2}, \kappa^2)$ 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} (i.e., x_{i1} is MAR). Denote the vector of missingness indicators 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 probability that x_{i1} is missing (i.e, $Pr(r_i = 0)$) was estimated from the logistic model

$$logit(Pr(r_i = 0)) = \phi_0 + \phi_1 y_i + \phi_2 x_{i2}, \tag{4.78}$$

with $(\phi_0, \phi_1, \phi_2) = (0.5, 1, -1)$. Linear regression was used to model the missing covariate as a function of the observed covariate.

For each of the three sample sizes, simulations were performed using 500 replications. The number of Monte-Carlo samples within each iteration of EM was 1000. 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 second set of simulations involve MZIP models with three covariates, two of which

are missing at random. Specifically, the count y_i was generated from the model

$$\begin{aligned} logit(\psi_i) &= \gamma_0 + \gamma_1 x_{i1} + \gamma_2 x_{i2} + \gamma_3 x_{i3} \\ log(\nu_i) &= \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + +\alpha_2 x_{i3}, \end{aligned} \tag{4.79}$$

with $(\gamma_0, \gamma_1, \gamma_2, \gamma_3) = (0.5, -0.5, -0.5, 0.5)$, $(\alpha_0, \alpha_1, \alpha_2, \alpha_3) = (0.5, -0.5, -0.5, 0.5)$, $x_{i3} \sim$ Exponential(λ) with $\lambda = 1$, $x_{i2} \sim N(\mu_2, \sigma_2^2)$ with $\mu_2 = 0$, $\sigma_2^2 = 1$, and $x_{i1} \sim N(\omega_0 + \omega_1 x_{i2}, \kappa^2)$ 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}. If r_{i1} and r_{i2}, take values of 1 when x_{i1} and x_{i2} are observed, missing data were generated based on the models

$$\begin{aligned} logit(Pr(r_{i1} = 0)) &= \phi_{01} + \phi_{11} y_i + \phi_{21} x_{i3} \\ logit(Pr(r_{i2} = 0)) &= \phi_{02} + \phi_{12} y_i + \phi_{22} x_{i3}. \end{aligned} \tag{4.80}$$

The missing covariates were modeled by using their true distributions and simulations were performed under two different scenarios for the missing data probabilities in equation (4.80). 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 $(\phi_{02}, \phi_{12}, \phi_{22}) = (-1, -1, -1)$.

In both simulation scenarios, the sample size was 1000 and 500 replications were used. The number of Monte-Carlo samples used within each iteration of EM was 1000. 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. Tables 1 and 2

show percent relative biases, simulation standard deviations, average standard errors of the estimated parameters, and mean squared errors (MSE) from MCEM and CC analyses. 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 tends to give estimated standard errors with small bias when the simulation standard deviation is used as the true standard deviation, whereas CC analysis underestimates the standard errors for γ_1 and γ_2. However, both methods provide estimated standard errors with little biases for the parameters in the marginal mean model, which are the parameters of primary interest.

4.6 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 is the parent-reported number of years of participation in the FMR program (years) and the number of decayed and filled primary teeth (dfs) is an outcome variable of interest. Previous analysis was based only on 677 children who had complete covariate and outcome data. In this paper, we consider 1094 children with complete data on the outcome, race, age, and several dental care variables but with missing information on years of participation and family income. 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. Based on prior

work by Divaris et al (2012), we used linear predictors of the following form:

$$\begin{aligned} logit(\psi_i) &= \gamma_0 + \gamma_1 x_{i1} + \gamma_2 x_{i2} + \gamma_3 x_{i3} + \gamma_4 x_{i4} + \gamma_5 x_{i5} + \gamma_6 x_{i6} \\ &+ \gamma_7 x_{i7} + \gamma_8 x_{i8} + \gamma_9 x_{i9} + \gamma_{10} x_{i10} + \gamma_{11} x_{i11} \\ log(\nu_i) &= \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \alpha_3 x_{i3} + \alpha_4 x_{i4} + \alpha_5 x_{i5} + \alpha_6 x_{i6} \\ &+ \alpha_7 x_{i7} + \alpha_8 x_{i8} + \alpha_9 x_{i9} + \alpha_{10} x_{i10} + \alpha_{11} x_{i11}, \end{aligned} \tag{4.81}$$

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

$$\lambda_{i1} = \exp(\xi_{01} + \xi_{11} x_{i2} + \xi_{21} x_{i3} + \xi_{31} x_{i5} + \xi_{41} x_{i6} + \xi_{51} x_{i7} + \xi_{61} x_{i8}) \tag{4.82}$$

Similarly, income was modeled using the exponential distribution with the rate parameter

λ_{5i} depending on observed covariates as

$$\lambda_{i5} = \exp(\xi_{01} + \xi_{11}x_{i2} + \xi_{21}x_{i3} + \xi_{41}x_{i6} + \xi_{51}x_{i7} + \xi_{61}x_{i8}) \tag{4.83}$$

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 (4.84).

$$\begin{aligned} Pr(x_{i1}, x_{i5}|x_{i5}, \mathbf{x}_{oi}, \lambda_{i1}, \lambda_{i5}) &= Pr(x_{i1}|x_{i5}, \lambda_{i1}, \lambda_{i5})Pr(x_{i5}|\lambda_{i5}) \\ &= \lambda_{i1}e^{-\lambda_{i1}x_{i1}}\lambda_{i5}e^{-\lambda_{i5}x_{i5}} \end{aligned} \tag{4.84}$$

where λ_{i1} and λ_{i1} are functions of the five non-missing covariate as in equations (4.82) and (4.83).

Estimates from complete case analysis were used as starting values of the EM algorithm and $s = 500$ Monte-Carlo samples 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 variable years on income and the observed covariates in equation (4.82), and a linear regression of income on the covariates used in equation (4.83). The number of imputations was $s = 20$ and the predictive mean matching method was used to impute values.

Table 3 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 major difference between the MCEM and CC analysis is that in the zero-inflation model the two methods provide estimates of opposite signs for years and age. 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 rate 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.

4.7 Discussion

Marginalized zero inflated Poisson models allow 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 also be applied to problems 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 to model missing covariate data 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 results 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. The MCEM method was also demonstrated using real data obtained from a sample of North Carolina schoolchildren, where the resulting estimates generally had smaller standard errors than estimates obtained from CC analysis. A limitation of the proposed method is that one has to specify a distribution for the missing covariates and that 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). As a way of dealing with the problem, sensitivity analysis has been suggested to check the robustness of parameter estimates under various covariate

distributions. Multiple imputation would provide an alternative approach to missing covariates in the MZIP. In its application to the FMR data, multiple imputation gave similar results as MCEM.

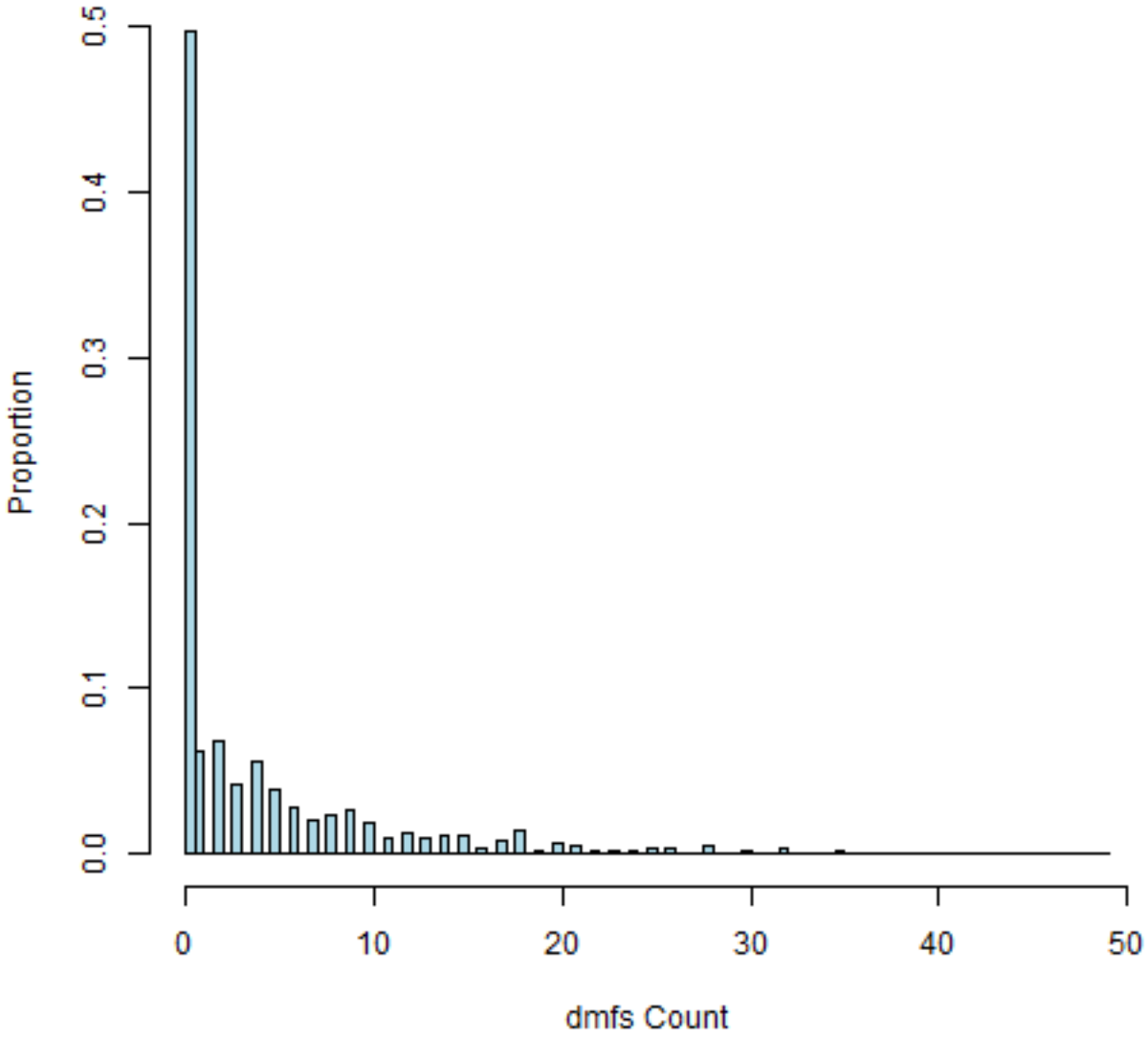

Figure 4.1: Distribution of dmfs counts from 1094 children grades 1 to 5 participating in a school-based fluoride mouthrinse program.

Table 4.1: Simulation results for scenario with two covariates, where one is potentially missing: comparison of MCEM and CC models based on 500 replications with sample sizes 250, 500 and 1000.

		MCEM				Complete Case			
S Size	Par	Percent Rel. Bias	Sim. Std	Mean SE	MSE	Percent Rel. Bias	Sim. Std	Mean SE	MSE
250	α_0	-0.734	0.185	0.184	0.034	34.594	0.171	0.165	0.149
	α_1	1.241	0.149	0.151	0.022	7.848	0.142	0.139	0.026
	α_2	0.697	0.174	0.184	0.030	25.794	0.171	0.169	0.096
	γ_0	0.215	0.232	0.232	0.054	-52.701	0.251	0.239	0.340
	γ_1	1.136	0.215	0.211	0.046	4.797	0.244	0.189	0.062
	γ_2	1.474	0.254	0.258	0.065	-26.464	0.297	0.254	0.158
500	α_0	-1.325	0.132	0.130	0.018	34.298	0.122	0.116	0.133
	α_1	-0.291	0.107	0.106	0.012	6.622	0.103	0.098	0.015
	α_2	-0.212	0.128	0.128	0.016	24.731	0.123	0.119	0.076
	γ_0	0.831	0.161	0.162	0.026	-52.337	0.173	0.167	0.304
	γ_1	1.280	0.144	0.139	0.021	4.090	0.162	0.129	0.028
	γ_2	1.428	0.173	0.172	0.030	-24.874	0.200	0.174	0.102
1000	α_0	-1.241	0.094	0.091	0.009	34.276	0.086	0.082	0.125
	α_1	-0.546	0.076	0.075	0.006	6.291	0.072	0.069	0.009
	α_2	-0.458	0.090	0.091	0.008	24.583	0.084	0.084	0.067
	γ_0	1.068	0.115	0.114	0.013	-51.964	0.119	0.118	0.284
	γ_1	1.340	0.098	0.094	0.010	4.129	0.109	0.089	0.014
	γ_2	1.410	0.118	0.118	0.014	-24.426	0.128	0.121	0.076

Table 4.2: Simulation results for scenario with three covariates, where two are potentially missing: comparison of MCEM and CC models based on 500 replications with sample size 1000 for two missing data scenarios.

	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								
α_0	0.084	0.092	0.088	0.008	-87.617	0.156	0.157	0.216
α_1	1.514	0.070	0.066	0.005	-65.266	0.084	0.090	0.113
α_2	2.177	0.073	0.072	0.005	-64.870	0.094	0.099	0.114
α_3	1.296	0.079	0.075	0.006	-38.260	0.108	0.110	0.048
γ_0	2.742	0.123	0.115	0.015	83.230	0.193	0.173	0.210
γ_1	8.946	0.100	0.088	0.012	74.417	0.108	0.101	0.150
γ_2	6.218	0.099	0.095	0.011	74.367	0.121	0.111	0.153
γ_3	2.722	0.094	0.088	0.009	47.921	0.130	0.115	0.074
Scenario 2: Mean= 30.1% of observations missing at least one covariate value								
α_0	-1.386	0.087	0.084	0.008	35.325	0.084	0.080	0.038
α_1	-1.348	0.060	0.059	0.004	8.155	0.056	0.054	0.005
α_2	0.086	0.062	0.065	0.004	9.079	0.058	0.059	0.005
α_3	-1.065	0.071	0.070	0.005	35.487	0.074	0.068	0.037
γ_0	1.164	0.110	0.106	0.012	-66.442	0.122	0.115	0.125
γ_1	1.682	0.073	0.073	0.005	2.066	0.077	0.073	0.006
γ_2	0.243	0.075	0.079	0.006	1.120	0.080	0.081	0.007
γ_3	1.264	0.080	0.080	0.006	-36.649	0.097	0.089	0.043

Table 4.3: 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

CHAPTER 5: CONCLUSION

While mixture models such as zero-inflated Poisson regression provide a flexibile platform to fit highly dispersed count data, estimates from these models do not have straightforward interpretations in describing the overall effects of explanatory variables on population-wide parameters. When interest is to make inferences about the marginal mean of the sampled population, indirect methods of parameter estimation such as the use of post-modeling transformations are often needed to make use of regression coefficients obtained from mixture models. However, these transformations may be difficult for some analysis to carry out, and may not always yield desired estimates. In the analysis of zero-inflated counts, the importance of models with marginally interpretable parameters has long been recognized (Lambert, 1992; Böhning et al., 1999; Preisser et al., 2012; Albert et al., 2014; Long et al., 2014), and the estimation of such parameters has gotten some attention in recent years. For counts with excess zeros, marginalized zero-inflated Poisson (Long et al., 2014) and negative binomial (Preisser et al., 2016) models allow for the estimation of overall exposure effects on the marginal mean in the population. Coefficients from these models have direct interpretations in describing the marginal mean, and can easily be employed to estimate incidence density ratios and other population-wide parameters.

In the second chapter of this dissertation, we proposed marginalized models for overdispersed counts based on two-component non-degenerate mixture distributions. To estimate the effects of exposure variables on the overall population mean count, we specify regression parameters directly to the marginal mean and perform maximum likelihood estimation. The models provide estimates that directly quantify the effects of exposure variables on the overall population mean, and extend the family of two-part marginalized regression models

for overdispersed count outcomes by providing alternatives to marginalized zero-inflated Poisson and negative binomial models. In addition to mixtures containing a degenerate at-zero and a Poisson or a negative binomial distributions on which existing marginalized zero-inflated models are based, the proposed method assumes other plausible mixture distributions for zero-inflated counts. Simulations indicate that when the true model is specified, each of the proposed marginalized mixture models provides smaller biases, Type I errors close to the nominal level and better confidence interval coverages compared to the other marginalized models considered. The applications of the models are demonstrated in a clinical trial aimed at comparing the anti-caries efficacy of three toothpaste formulations in children. Future research could extend the marginalized mixture models to allow the mixing probabilities to depend on covariates as well as to accommodate longitudinal data, for example, by inclusion of random effects as in Long et al.(2015).

In the third chapter of the dissertation, we developed a joint marginalized model for two correlated counts with zero-inflation. The model specifies regression coefficients to the marginal means of the two outcomes and provides estimates that allow for direct inferences about the overall population. The new model extends univariate marginalized models by accommodating two correlated outcomes, and modifies existing bivariate models for zero-inflated counts by directly estimating overall covariate effects in the population. Finite sample properties of the marginalized model estimates are examined in simulation studies. The model is further applied to dental caries data. While we performed estimation by direct maximization of the likelihood function with carefully selected starting values, applications of Bayesian methods or the EM algorithm could provide alternative estimation methods. Future research could also extend the model to handle three or more correlated outcomes with zero-inflation. Another possible extension could be the modeling of repeated or longitudinal data in problems where the bivariate zero-inflated outcome is measured repeatedly for each sampling unit.

Finally, building upon the work of Ibrahim et al.(1999), we proposed an estimation

method for marginalized zero-inflated Poisson models for problems where covariates are missing at random. The method employs Monte Carlo EM algorithms (Wei and Tanner,1990) and estimates the E step of EM based on samples generated from the conditional distributions of the missing covariates. The method was illustrated and compared with multiple imputation and complete case analysis by using simulations and dental data collected to estimate the caries preventive effects of a school-based fluoride mouthrinse program. Future research could extend the method to MZINB models or seek to handle missing response data in addition to missing covariates.

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.

Arab, A., Holan, S. H., Wikle, C. K., and Wildhaber, M. L. (2012), "Semiparametric bivariate zero-inflated Poisson models with application to studies of abundance for multiple species," *Environmetrics*, 23, 183–196.

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.

Bermúdez, L. and Karlis, D. (2012), "A finite mixture of bivariate Poisson regression models with an application to insurance ratemaking," *Computational Statistics and Data Analysis*, 56, 3988–3999.

Chen, X. D. and Fu, Y. Z. (2010), "Model selection for zero-inflated regression with missing covariates," *Computational Statistics and Data Analysis*, 136, 1360–1375.

Cheung, Y. B. and Lam, K. F. (2006), "Bivariate PoissonPoisson model of zero-inflated absenteeism data," *Statistics in Medicine*, 25, 3707–3717.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society, Series B*, 39, 1–38.

Divaris, K., Rozier, R. G., and King, R. S. (2012), "Effectiveness of a school-based fluoride mouthrinse program," *Journal of Dental Research*, 91, 282–287.

Frühwirth-Schnatter, S. (2006), *Finite Mixture and Markov Switching Models*, Springer.

Ghosh, S. K., Mukhopadhyay, P., and Lu, J. C. (2006), "Bayesian analysis of zero-inflated regression models," *Journal of Statistical Planning and Inference*, 136, 1360–1375.

Gilks, W. R., Best, N. G., and Tan, K. K. C. (1995), "Adaptive rejection Metropolis sampling within Gibbs sampling," *Journal of the Royal Statistical Society, Series C*, 44, 455–472.

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

Gurmu, S. and Elder, J. (2008), "A bivariate zero-inflated count data regression model with unrestricted correlation," *Economics Letters*, 100, 245–248.

Heagerty, P. J. (1999), "Marginally specified logistic-normal models for longitudinal binary data," *Biometrics*, 55, 688–698.

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

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

Ibrahim, J. G., Chen, M. H., and Lipsitz, S. R. (1999a), "Monte Carlo EM for missing covariates in parametric regression models," *Biometrics*, 55, 591–596.

Ibrahim, J. G., Chen, M. H., Lipsitz, S. R., and Herring, A. H. (2005), "Missing data methods for generalized linear models: A comparative review," *Journal of the American Statistical Association*, 100, 332–346.

Ibrahim, J. G., Lipsitz, S. R., 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.

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

Leisch, F. (2004), "FlexMix: A General Framework for Finite Mixture Models and Latent Class Regression in R," *Journal of Statistical Software*, 11(8).

Lewsey, J. D. and Thomson, W. M. (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, 183–189.

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.

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

Little, R. J. A. and Rubin, D. B. (2002), *Statistical analysis with missing data*, vol. Second Edition, Wiley.

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

Long, D. L., Preisser, J. S., Herring, A. H., and Golin, C. E. (2014), "A marginalized zero-inflated Poisson regression model with overall exposure effects," *Statistics in Medicine*, 33, 5151–5165.

— (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*, Wiley.

Min, Y. and Agresti, A. (2005), "Random effect models for repeated measures of zero-inflated count data," *Statistical Modelling*, 5, 1–19.

Morgan, C. J., Lenzenweger, M. F., Rubinc, D. B., and Levyd, D. L. (2014), "A hierarchical finite mixture model that accommodates zero-inflated counts,non-independence, and heterogeneity," *Statistics in Medicine*, 33, 2238–2250.

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

Mwalili, S. M., 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.

Powell, M. J. D. (1964), "An efficient method of finding the minimum of a function of several variables without calculating derivatives," *Computer Journal*, 7, 155–162.

Preisser, J. S., Das, K., Long, D. L., and Divaris, K. (2016), "A Marginalized zero-inflated negative binomial regression model with application to dental caries," *Statistics in Medicine*, 35, 1722–1735.

Preisser, J. S., Stamm, J. W., Long, D. L., 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.

Ramaswamy, V., Anderson, A. W., and DeSarbo, W. S. (1994), "A disaggregate negative binomial regression procedure for count data analysis," *Management Science*, 40(3), 405–417.

Roeder, K., Lynch, K. G., and Nagin, D. S. (1999), "Modeling uncertainty in latent class membership: A case study in criminology," *Journal of the American Statistical Association*, 94 (447), 766–776.

Schlattmann, P. (2009), *Medical Applications of Finite Mixture Models*, Springer.

Todem, D., Hsu, W. W., and Kim, K. (2012), "On the efficiency of score tests for homogeneity in two-component parametric models for discrete data," *Biometrics*, 68, 975982.

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

Walhin, J. F. (2001), "Bivariate ZIP models," *Biometrical Journal*, 43(2), 147–160.

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 and Prevention*, 35, 625–629.

Wang, P., Puterman, M., Cockburn, I., and Le, N. (1996), "Mixed Poisson regression models with covariate dependent rates," *Biometrics*, 52, 381–400.

Wedel, M. (1995), "A mixture likelihood approach for generalized linear models," *Journal of Classification*, 12, 21–55.

Wei, G. C. G. and Tanner, M. A. (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.

Yang, M., Das, K., and Majumdar, A. (2016), "Analysis of bivariate zero inflated count data with missing responses," *Journal of Multivariate Analysis*, 00, 1–12.