University of North Carolina at Chapel Hill

The University of North Carolina at Chapel Hill Department of Biostatistics Technical Report Series

Year 2014 *Paper* 43

A Marginalized Zero-Inflated Negative Binomial Regression Model with Overall Exposure Effects

John S. Preisser[*] Kalyan Das[†]

D. Leann Long[‡] John W. Stamm[**]

[*]University of North Carolina at Chapel Hill, jpreisse@bios.unc.edu

[†]University of Calcutta, kalyanstat@gmail.com

[‡]West Virginia University, dllong@hsc.wvu.edu

[**]University of North Carolina at Chapel Hill, john_stamm@unc.edu

A Marginalized Zero-Inflated Negative Binomial Regression Model with Overall Exposure Effects

John S. Preisser, Kalyan Das, D. Leann Long, and John W. Stamm

Abstract

The zero-inflated negative binomial regression model (ZINB) is often employed in diverse fields such as dentistry, health care utilization, highway safety, and medicine, to examine relationships between exposures of interest and overdispersed count outcomes exhibiting many zeros. The regression coefficients of ZINB have latent class interpretations for a susceptible subpopulation at risk for the disease/condition under study with counts generated from a negative binomial distribution and for a non-susceptible subpopulation that provides only zero counts. The ZINB parameters, however, are not well-suited for estimating overall exposure effects, specifically, in quantifying the effect of an explanatory variable in the overall mixture population. In this paper, a marginalized zero-inflated negative binomial regression (MZINB) model for independent responses is proposed to model the population marginal mean count directly, providing straightforward inference for overall exposure effects based on maximum likelihood estimation. Through simulation studies, the performance of MZINB with respect to test size is compared to marginalized zero-inflated Poisson, Poisson, and negative binomial regression. The MZINB model is applied to data from a randomized clinical trial of three toothpaste formulations to prevent incident dental caries in a large population of Scottish schoolchildren.

A Marginalized Zero-Inflated Negative Binomial Regression Model with Overall Exposure Effects

John S. Preisser[1], Kalyan Das[2], D. Leann Long[3], John W. Stamm[4]

[1]Department of Biostatistics, University of North Carolina at Chapel Hill

[2]Department of Statistics, University of Calcutta, Kolkata, India

[3]Department of Biostatistics, West Virginia University, Morgantown, WV

[4]Department of Dental Ecology, University of North Carolina at Chapel Hill

Abstract

The zero-inflated negative binomial regression model (ZINB) is often employed in diverse fields such as dentistry, health care utilization, highway safety, and medicine, to examine relationships between exposures of interest and overdispersed count outcomes exhibiting many zeros. The regression coefficients of ZINB have latent class interpretations for a susceptible subpopulation at risk for the disease/condition under study with counts generated from a negative binomial distribution and for a non-susceptible subpopulation that provides only zero counts. The ZINB parameters, however, are not well-suited for estimating overall exposure effects, specifically, in quantifying the effect of an explanatory variable in the overall mixture population. In this paper, a marginalized zero-inflated negative binomial regression (MZINB) model for independent responses is proposed to model the population marginal mean count directly, providing straightforward inference for overall exposure effects based on maximum likelihood estimation. Through simulation studies, the performance of MZINB with respect to test size is compared to marginalized zero-inflated Poisson, Poisson, and negative binomial regression. The MZINB model is applied to data from a randomized clinical trial of three toothpaste formulations to prevent incident dental caries in a large population of Scottish schoolchildren.

Keywords: caries prevention, excess zeros, marginalized models, over-dispersion, zero-inflation.

1 Introduction

Zero-inflated count regression models are widely used to analyze count data that include many zeros in such diverse fields as dentistry, health care utilization, highway safety, and medicine. Historically, the negative binomial distribution has been used for characterizing dental caries counts owing to the fact that they are routinely over-dispersed relative to the Poisson distribution [5]. As populations have become healthier over time, reported distributions of dental caries indices such as the number of decayed, missing or filled teeth (DMFT) or surfaces (DFMS) have been increasingly characterized by a preponderance of zero counts in proportions greater than expected under either the Poisson or negative binomial distributions. To account for these “excess zeros”, zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) regression models have been increasingly applied to caries indices [2, 16, 18]. Other applications of ZINB models include modeling lesion counts in the analysis of magnetic resonance imaging for multiple sclerosis [3], mosquito counts in a study of Malaria [10], and vaccine adverse event count data [20]. ZINB models are applicable when there is interest in a model for latent classes corresponding to a susceptible subpopulation at risk for the disease/condition under study with counts generated from a negative binomial distribution and a non-susceptible subpopulation that provides only zero counts.

While the ZINB model regression coefficients have latent class interpretations for these two subpopulations, researchers in the health sciences sometimes seek to make inference on the marginal mean of the mixture population. In dentistry, for example, zero-inflated count models are not well-suited for generating inference concerning overall effects of risk factors and treatments on caries count outcomes that are often of primary interest. Albert *et. al.* [1] proposed a causal inference estimator for overall effects of a binary variable in a zero-inflated count model as the average within-subject difference of the marginal means under exposed and unexposed conditions. Their estimator may give results that are not generalizable to populations with other configurations of explanatory variables and its extension to continuous exposures is not straightforward. With a similar goal for overall effect estimation, Long *et. al.* [13] proposed a marginalized zero-inflated Poisson (MZIP) regression model with maximum likelihood estimation in the framework of the ZIP model. Instead of modeling the Poisson mean in the at-risk latent class, the marginal mean is modeled directly as a function of covariates in conjunction with a logistic model for the probability of an excess zero. As with (log-link) Poisson regression, exponentiating regression coefficients from the marginal mean component of the MZIP model gives incidence density ratios for the overall effects of exposures and covariates on the count outcome.

In extending the MZIP to the negative binomial case, this article draws on the marginalized model literature. Heagerty [6] proposed marginalized multilevel models for correlated binary data, which directly model the marginal means whose regression parameters are specified in a likelihood that links the marginal model with a flexibly specified conditional model with random effects. Lee *et al.* [11] proposed marginalized negative binomial hurdle models to analyze clustered data with excess zeros, marginalizing over the random effects. These methods for regression of correlated outcomes combine the desire for population average interpretations with the convenience of estimation with a likelihood function. In a comparatively simple implementation of the principle of marginalization, the marginalized models approach was adapted in the ZIP model by Long *et. al.* [13] in order to achieve population-wide parameter interpretations for independent count responses with many zeros. Instead of integrating (averaging) over mixtures of distributions defined by random effects, their approach marginalizes over the Poisson and degenerate components of the two-part ZIP model to obtain overall effects.

Extending the model of Long *et. al.* [13], this article introduces a marginalized zero-inflated negative binomial model (MZINB) to model the population mean count directly, allowing straightforward inference for overall exposure effects that accounts for both excess zeros and overdispersion. Section 2 reviews traditional ZIP and ZINB models, and the marginalized zero-inflated Poisson regression model (MZIP) [13], section 3 introduces the MZINB, section 4 presents a simulation study, section 5 presents an application to a randomized clinical trial comparing three toothpaste formulations to prevent dental caries incidence and section 6 draws conclusions.

2 Zero-inflated Count Data Regression Models

2.1 Zero-inflated Poisson and negative binomial models

In ZIP and ZINB models, caries counts arise from a mixture of two latent (i.e., unobserved) classes of subjects [9, 15]. The first model part (defined by a Bernoulli 0/1 process) selects subject i with probability ψ_i to be considered not-at-risk for caries where, conditional on being a member in this class, subject i has a zero count with probability one. The second model part provides a caries count with probability 1- ψ_i from a stochastic distribution with mean μ_i for the response counts in an "at-risk" class of subjects. Specifically, let Y_i be a random variable for the i-th individual's caries count. The zero-inflated count distribution of the i-th individual's caries counts, Y_i, is:

$$\begin{aligned} P(Y_i = 0) &= \psi_i + (1 - \psi_i) g(0|\theta_i) \\ P(Y_i = y_i) &= (1 - \psi_i) g(y_i|\theta_i), \quad y_i > 0, \end{aligned} \tag{1}$$

where $g(y_i|\theta_i)$ is the either the Poisson probability function with $\theta_i = \mu_i$ so that (1) is the ZIP probability distribution, or the negative binomial with $\theta_i = (\mu_i, \alpha)$ where α is the overdispersion parameter giving the ZINB distribution. Thus, the individuals with zeros include both fixed zeros from not-at-risk subjects and random zeros from at-risk subjects with group membership being unknown. While the latent classes are sometimes of intrinsic interest, their mixture has been viewed in the dental caries literature as a convenient construct that leads to a statistical distribution for caries counts that accounts for excess zeros [16, 18].

The joint distribution across all n individuals in the sample based on equation (1) is

$$f(\mathbf{y}|\boldsymbol{\psi}, \boldsymbol{\mu}) = \prod_{y_i=0} \Big[\Big(\frac{\psi_i}{1-\psi_i} + g(0|\mu_i)\Big)(1-\psi_i)\Big] \prod_{y_i>0} \Big[(1-\psi_i) g(y_i|\mu_i)\Big] \tag{2}$$

where $\mathbf{y} = (y_1, \ldots, y_n)$, $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_n)$ and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$. Lambert [9] and Mullahy [15] proposed the model

$$\text{logit}(\psi_i) = \mathbf{Z}_i' \boldsymbol{\gamma} \qquad \text{and} \qquad \log(\mu_i) = \mathbf{X}_i' \boldsymbol{\lambda} \tag{3}$$

where $\mathbf{Z}_i$ and $\mathbf{X}_i$ are the covariate vectors for the i-th individual for excess zeros and Poisson (or negative binomial) processes, respectively. Insertion of equation (3) into equation (2) produces the likelihood function, say $L_{zip}(\boldsymbol{\gamma}, \boldsymbol{\lambda}|\mathbf{y})$. In the traditional ZIP or ZINB model in equation (3), $\boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$ have latent class interpretations: γ_j, the element of $\boldsymbol{\gamma}$ corresponding to the j-th covariate, is the multiplicative increase in the log-odds of being an excess zero due to a unit increase in the covariate (z_{ij}) and λ_j is the multiplicative increase in the log-incidence caries rate due to a unit increase in the covariate x_{ij} in the susceptible population. In practice, models often have specification $\mathbf{Z}_i = \mathbf{X}_i$. Occasionally, $\mathbf{Z}_i$ consists of a subset of the covariates in $\mathbf{X}_i$.

Challenges arise in model choice because the latent class interpretations of ZIP and ZINB models are not well-suited for dental caries studies where interest is in the marginal parameters of the mixture population, for example, the incident proportion of individuals with any caries, $\pi_i = P(Y_i > 0)$, and the overall mean caries increment $\nu_i = \mathrm{E}(Y_i)$ in a prospective study. In particular, the importance of $\boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$ lies in their relationship to the incidence and mean caries increment via $\pi_i = (1 - \psi_i)[1 - g(0|\theta_i)]$ and $\nu_i = (1 - \psi_i)\mu_i$ [1]. When interest is in the effects of explanatory variables on π_i, hurdle models [15, 17] provide an appropriate modeling choice. When interest is in covariate effects on the marginal mean, equation (3) with $\mathbf{Z}_i = \mathbf{X}_i$ implies $\nu_i = \exp(X_i'\boldsymbol{\lambda})/[1 + \exp(X_i'\boldsymbol{\gamma})]$. It follows that incidence density ratios defined as ratios of two such marginal means where one covariate is allowed to vary while the others are held fixed will generally depend upon $\boldsymbol{\gamma}$ so that the elements of $\boldsymbol{\lambda}$ will not have interpretations for covariate effects as incidence density ratios in the mixture population. However, overall effects may be obtained indirectly with post-modelling calculations [2, 9], including counterfactual and log-log model approaches [1]. These approaches require variance estimation for functions of multiple parameters, for which statisticians have suggested the delta method and bootstrap resampling methods. Unfortunately, such indirect methods of overall effects estimation for a binary exposure variable are sufficiently complicated to be inaccessible to many dental research analysts, and they do not have obvious extensions to continuous explanatory variables. Hence, marginalized model approaches are proposed immediately below to provide easy, direct inference for overall effects for count data with many zeros through modeling ν_i instead of μ_i.

2.2 Marginalized zero-inflated Poisson regression model

When the overall mean caries increment $\nu_i = \mathrm{E}(Y_i)$ is of primary interest, one may specify a marginalized zero-inflated count response model [13]

$$\mathrm{logit}(\psi_i) = \mathbf{Z}_i'\boldsymbol{\gamma} \qquad \text{and} \qquad \log(\nu_i) = \mathbf{X}_i'\beta. \tag{4}$$

While $\boldsymbol{\gamma}$ models the excess zeros as in the traditional ZIP model, the vector parameter of log(IDR)'s denoted by β represents the same overall effect of covariates on caries increment as in Poisson or negative binomial regression. In other words, $\exp(\beta_j)$ is the log-incidence rate ratio for caries in the overall population corresponding to a one-unit increase in the covariate x_{ij}. Adopting ideas from marginalized longitudinal data model approaches [6], β in equation (4) is estimated, accounting for excess zeros, in a maximum likelihood framework via substitution of $\mu_i = \nu_i/(1-\psi_i)$ into (2) giving

$$f(\mathbf{y}|\boldsymbol{\psi},\boldsymbol{\nu}) \;=\; \prod_{all\, y_i}(1-\psi_i)\prod_{y_i=0}\Big[\Big(\frac{\psi_i}{1-\psi_i}+g(0|\psi_i,\nu_i)\Big)\Big]\prod_{y_i>0}\Big[g(y_i|\psi_i,\nu_i)\Big]. \tag{5}$$

Long *et. al.* [13] developed procedures for a marginalized zero-inflated Poisson (MZIP) regression model with maximum likelihood estimation of $(\boldsymbol{\gamma},\beta)$ in model (4) where $g(y_i|\mu_i) = \exp(-\mu_i)\mu_i^{y_i}/y_i!$ in (2) becomes $g(y_i|\psi_i,\nu_i) = \exp[-\nu_i/(1-\psi_i)][\nu_i/(1-\psi_i)]^{y_i}/y_i!$ in (5) with simplification for $y_i = 0$. Insertion of the model equations (4) into (5) yields the MZIP log-likelihood function

$$\begin{aligned} L_{mzip}(\boldsymbol{\gamma},\beta|\mathbf{y}) \;=\; & \prod_{all\, y_i}(1+e^{\mathbf{Z}_i'\boldsymbol{\gamma}})^{-1}\prod_{y_i=0}(e^{\mathbf{Z}_i'\boldsymbol{\gamma}}+e^{-[1+\exp(\mathbf{Z}_i'\boldsymbol{\gamma})]\exp(\mathbf{X}_i'\beta)}) \\ & \times\prod_{y_i>0}\big[(1+e^{\mathbf{Z}_i'\boldsymbol{\gamma}})^{y_i}e^{\mathbf{X}_i'\beta y_i}e^{-[1+\exp(\mathbf{Z}_i'\boldsymbol{\gamma})]\exp(\mathbf{X}_i'\beta)}\big]/y_i! \end{aligned}$$

Generally, equations (3) and (4) are non-nested models and $L_{zip}(\boldsymbol{\gamma},\boldsymbol{\lambda}|\mathbf{y})$ is not the same as $L_{mzip}(\boldsymbol{\gamma},\beta|\mathbf{y})$. Instances of equivalence arise when models (3) and (4) are null (no covariates) or saturated, i.e., all covariates are categorical with all possible interactions involving them included in both model parts. Choice of ZIP versus MZIP (or ZINB versus the proposed MZINB in section 3) should depend upon the desired parameter interpretations, as given by $\boldsymbol{\lambda}$ for the latent class of susceptible persons or by β for the overall population. Model (4) is easy to fit with SAS Proc NLMIXED [13]. The next section extends MZIP to MZINB allowing for extra-Poisson variation in addition to excess zeros.

3 Marginalized Zero-inflated Negative Binomial Regression Model

A marginalized ZINB regression model for estimation of $(\boldsymbol{\gamma}, \beta)$ in model (4) is introduced within a likelihood framework where $\mu_i = \nu_i/(1-\psi_i)$ is substituted into the negative binomial probability function

$$g(y_i|\mu_i, \alpha) = \frac{\Gamma(y_i+\alpha)}{y_i!\Gamma(\alpha)}\Big(\frac{\alpha}{\alpha+\mu_i}\Big)^{\alpha}\Big(\frac{\mu_i}{\alpha+\mu_i}\Big)^{y_i}, \text{ where } y_i = 0, 1, \ldots$$

where $\text{var}(y_i) = \mu_i + \phi\mu_i^2$ and $\phi = 1/\alpha > 0$ so that the dependence of the ZINB density function $g(y_i|\psi_i, \nu_i, \alpha)$ on the marginal mean ν_i is made explicit. It then replaces the ZIP density function $g(y_i|\psi_i, \nu_i)$ in the likelihood expression appearing in equation (5) with simplification for $y_i = 0$. Model (4) then gives the MZINB log-likelihood function

$$\begin{aligned} L_{mzinb}(\boldsymbol{\gamma}, \beta, \alpha|\mathbf{y}) \;&=\; \prod_{all\, y_i}(1+e^{\mathbf{Z}_i'\boldsymbol{\gamma}})^{-1}\prod_{y_i=0}\Big\{e^{\mathbf{Z}_i'\boldsymbol{\gamma}} + \big[1+\frac{1}{\alpha}(1+e^{\mathbf{Z}_i'\boldsymbol{\gamma}})e^{\mathbf{X}_i'\beta}\big]^{-\alpha}\Big\} \\ &\prod_{y_i>0}\frac{\Gamma(y_i+\alpha)}{y_i!\Gamma(\alpha)}\Big[1+\frac{1}{\alpha}(1+e^{\mathbf{Z}_i'\gamma})e^{\mathbf{X}_i'\beta}\Big]^{-\alpha}\Big[\frac{(1+e^{\mathbf{Z}_i'\gamma})e^{\mathbf{X}_i'\beta}}{\alpha+(1+e^{\mathbf{Z}_i'\gamma})e^{\mathbf{X}_i'\beta}}\Big]^{y_i} \end{aligned}$$

The log-likelihood of the MZINB model is

$$\begin{aligned} l(\gamma, \beta, \alpha|\mathbf{y}) \;&=\; -\sum_i \log(1+e^{\mathbf{Z}_i'\gamma}) + \sum_{y_i=0}\log\Big\{e^{\mathbf{Z}_i'\gamma} + [1+\frac{1}{\alpha}(1+e^{\mathbf{Z}_i'\gamma})e^{\mathbf{X}_i'\beta}]^{-\alpha}\Big\} \\ &-\; \sum_{y_i>0}\log y! + \sum_{y_i>0}\sum_{j=0}^{y_i-1}\log(j+\alpha) - \sum_{y_i>0}\alpha\log\Big[1+\frac{1}{\alpha}(1+e^{\mathbf{Z}_i'\gamma})e^{\mathbf{X}_i'\beta}\Big] \\ &+\; \sum_{y_i>0}y_i\Big[\log(1+e^{\mathbf{Z}_i'\gamma}) + \mathbf{X}_i'\beta\Big] - \sum_{y_i>0}y_i\log\Big[\alpha+(1+e^{\mathbf{Z}_i'\gamma})e^{\mathbf{X}_i'\beta}\Big] \end{aligned}$$

using the relation $\frac{\Gamma(k+\alpha)}{\Gamma(\alpha)} = \prod_{j=0}^{k-1}(j+\alpha)$ for an integer k. The likelihood score equations are obtained by differentiating the log likelihood with respect to the model parameters $(\boldsymbol{\gamma}, \beta, \alpha)$. Noting that $\mu_i = e^{\mathbf{X}_i'\beta}(1+e^{\mathbf{Z}_i'\gamma})$ and defining $\theta_i = \alpha/(\alpha+\mu_i)$, the score equations are:

$$\begin{aligned}
\frac{\partial l}{\partial \beta} &= \sum_i \Big\{ I(y_i>0)\Big[y_i - \mu_i\Big(\frac{\alpha+y_i}{\alpha+\mu_i}\Big)\Big] - I(y_i=0)\frac{\theta_i^{\alpha+1}\mu_i}{e^{\mathbf{Z}_i'\gamma}+\theta_i^{\alpha}}\Big\} X_i' \\
\frac{\partial l}{\partial \alpha} &= \sum_i \Big\{ I(y_i=0)\Big[\frac{\theta_i^{\alpha}(1-\theta_i+\ln\theta_i)}{e^{\mathbf{Z}_i'\gamma}+\theta_i^{\alpha}}\Big] + I(y_i>0)\Big[\ln\theta_i - \frac{y_i-\mu_i}{\alpha+\mu_i} + \sum_{j=0}^{y_i-1}\frac{1}{j+\alpha}\Big]\Big\} \\
\frac{\partial l}{\partial \gamma} &= \sum_i \Big[(y_i-1) + I(y_i=0)\Big\{\frac{1+e^{\mathbf{Z}_i'\gamma}-\mu_i\theta_i^{\alpha+1}}{e^{\mathbf{Z}_i'\gamma}+\theta_i^{\alpha}}\Big\} - I(y_i>0)\mu_i\Big(\frac{\alpha+y_i}{\alpha+\mu_i}\Big)\Big]\psi_i Z_i' \quad (6)
\end{aligned}$$

The model-based asymptotic covariance estimator of $\hat{\boldsymbol{\zeta}}$ where $\boldsymbol{\zeta} = (\boldsymbol{\gamma}, \beta, \alpha)'$ is computed as the inverse of the Fisher information matrix $I(\boldsymbol{\zeta}) = -\mathrm{E}(\partial^2 l/\partial\boldsymbol{\zeta}\partial\boldsymbol{\zeta}')$. Estimation is by nonlinear optimization by the quasi-Newton method, which may be implemented in SAS PROC NLMIXED 9.3 as shown in Appendix 1.

4 A Small Simulation Study

Simulations were performed to examine the nominal size and power of Wald tests for marginal mean regression parameters in MZINB models and other count data models in a setting of a large clinical trial. Data were generated from MZINB models in equation (4) having

$$X_i'\beta = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3(x_{i1} \times x_{i2}) + \beta_4 x_{i4} + \beta_5 x_{i5}, \quad (7)$$

with $\beta_0 = 1.2, \beta_1 = 0.8, \beta_2 = -0.2, \beta_3 = -0.1, \beta_4 = 0$, and $\beta_5 = 0$, where x_{i1} and x_{i2} are binary covariates with 0/1 coding, $x_{i3} = x_{i1} \times x_{i2}$, and x_{i4} and x_{i5} are indicator variables for two of three treatment groups in a hypothetical clinical trial. For assessment of power $\beta_4 = -0.10$ and $\beta_5 = -0.05$. Furthermore, $Z_i = X_i$ with $\gamma_1 = -2.0, \gamma_2 = -0.2, \gamma_3 = 0, \gamma_4 = 0.4$, and $\gamma_5 = 0.5$. As shown in Table 1, six scenarios were considered for (ϕ, γ_0), where $\phi = 1/\alpha$, to examine the impact of different levels of excess zeros and varying levels of overdispersion. Design vectors $X_i, i = 1, \ldots, 3412$ from the Lanarkshire caries clinical trial example analyzed in section 5 were used for data generation. The proportion of the observations allocated in each of the four covariate groups $G = (x_{i1}, x_{i2})$ are given for treatments in the order $\{(x_{i4} = 0, x_{i5} = 0); (x_{i4} = 1, x_{i5} = 0); (x_{i4} = 0, x_{i5} = 1)\}$: $\{G = (0,0) : .162, .164, .079\}$, $\{G = (0,1) : .079, .086, .041\}$, $\{G = (1,0) : .125, .120, .063\}$ and $\{G = (1,1) : .032, .032, .016\}$.

The simulations compared several count data models where all models correctly specified the marginal mean structure, $\nu_i = \exp(X_i'\beta)$. The first set of simulations compared MZINB, MZIP, negative binomial, and Poisson regression models, the latter two approaches with empirical (sandwich) standard errors as well as model-based standard errors. The approaches using empirical standard errors were included because these approaches would be expected to maintain the nominal test size with possibly some loss of power. Type I errors of two-sided Wald tests conducted at the nominal 0.05 significance level for $H_0 : \beta_4 = 0$ vs $H_1 : \beta_4 \neq 0$ were compared. Test size was defined as the proportion of 1,000 simulation runs that reject H_0. Proc NLMIXED in SAS 9.3 was used to fit MZINB and MZIP models and Proc Genmod was used to fit negative binomial and Poisson regression models.

Table 1. Test Size for $H_0 : \beta_4 = 0$ when true model is MZINB and n=3,412.

	(ϕ, γ_0)					
model (se)	(0.5, -2)	(0.5, -1)	(0.5, 0)	(1.0, -2)	(1.0, -1)	(1.0, 0)
Percentage of 0's	18	26	42	26	34	46
Poisson	.332	.385	.468	.447	.493	.566
Poisson (emp)	.049	.047	.046	.056	.052	.048
MZIP	.267	.277	.302	.358	.359	.393
NB	.042	.028	.018	.042	.037	.029
NB (emp)	.049	.047	.048	.048	.046	.053
MZINB	.048	.051	.046	.051	.047	.048

emp = empirical (sandwich) standard error was used in the Wald test.

Table 1 shows that MZINB as well as Poisson and negative binomial regression with empirical Wald tests maintain the Type I error at the nominal 0.05 significance level. MZIP and model-based Wald tests for Poisson regression, which do not account for extra-Poisson variation, fail to maintain Type I error. Negative binomial regression is overly conservative; it rejects H_0 too infrequently. Table 2 compares power among the three approaches that maintained the nominal test size. While the power of MZINB is never surpassed by empirical Wald tests from Poisson or negative binomial regression, there is generally little difference in power between the three methods.

Table 2. Power for $H_0 : \beta_4 = 0$ when true model is MZINB and n=3,412.

	(ϕ, γ_0)					
model (se)	(0.5, -2)	(0.5, -1)	(0.5, 0)	(1.0, -2)	(1.0, -1)	(1.0, 0)
Percentage of 0's	18	26	42	26	34	46
Poisson (emp)	0.78	0.73	0.55	0.55	0.49	0.40
NB (emp)	0.79	0.70	0.46	0.57	0.49	0.36
MZINB	0.79	0.73	0.56	0.58	0.49	0.40

Percentage of simulation replicates that rejected H_0 out of 1,000 runs.

Traditional ZIP and ZINB models are intentionally excluded from the simulation study because β_4 and β_5 do not exist in these models. In contrast to the "average" treatment effects that could be obtained from ZIP and ZINB from post-modeling calculations [1], the marginalized model parameters β_4 and β_5 are treatment effects that are homogeneous across the levels of the other covariates in the model. The distinctive role of MZINB relative to ZINB is illustrated in the next section with emphasis on interpretations of exposure and treatment effects in a clinical trial.

5 Applications to a Lanarkshire Caries trial

5.1 Zero-inflated negative binomial regression model

The traditional ZINB model is applied to a three year double-blind caries incidence trial in Lanarkshire, Scotland, 1988-92, that randomized 4,294 children aged 11-12 years to three active toothpaste formulations. Dental examinations were conducted at baseline, and after 1, 2 and 3 years. The authors of the original study compared anticaries efficacy of three toothpaste formulations after three years follow-up (N=3517, 82%): sodium fluoride (NaF, n=1370), sodium fluoride plus sodium trimetaphosphate (NaF+TMP, n=680) and sodium monofluorophosphate (SMFP, n=1362). The analysis of this article compares toothpastes with respect to mean caries increment (*DMFS* or number of decayed, missing, and filled surfaces) after two years (N = 3412; 79% follow-up), while adjusting for baseline caries status and calculus and considering that 19% of counts are zero. Initially, a traditional ZINB model is considered where the probability of an excess zero caries count is

$$\text{logit}(\psi_i) = \gamma_0 + \gamma_1 x_{i1}^{bc} + \gamma_2 x_{i2}^{calc} + \gamma_3 (x_{i1}^{bc} \times x_{i2}^{calc}) + \gamma_4 x_{i4}^{Naf} + \gamma_5 x_{i5}^{Naftmp}$$

and the mean caries count μ_i of the at-risk class of children is

$$\log(\mu_i) = \lambda_0 + \lambda_1 x_{i1}^{bc} + \lambda_2 x_{i2}^{calc} + \lambda_3 (x_{i1}^{bc} \times x_{i2}^{calc}) + \lambda_4 x_{i4}^{Naf} + \lambda_5 x_{i5}^{Naftmp}$$

where x_{i1} is a 0/1 indicator for baseline caries status (1=high vs. 0=low) and x_{i2} is the presence of calculus formed by dental plaque (1=yes vs 0=no). High baseline caries refers to having at least one decayed, missing, or filled anterior tooth or premolar [21]. Additionally, the treatment variable x_{i4} is an indicator variable for NaF toothpaste, x_{i5} is an indicator for NaF+TMP, and SMFP is the reference toothpaste.

Table 3. ZINB results for two-year DMFS increments

	Neg. Bin. Process		logit(Excess zero)	
Variable	est	se	est	se
Intercept	1.316	0.039***	-2.07	0.23
Baseline Caries (high)	0.662	0.042***	-2.41	0.86**
Calculus	-0.170	0.054**	-0.17	0.27
Baseline Caries*calculus	-0.093	0.081	—	—
Naf	-0.043	0.040	0.23	0.26
Naftmp	-0.014	0.049	0.19	0.32

$\hat{\phi} = 0.629$ (se = 0.036), where $\phi = 1/\alpha$.
$*p < 0.05, **p < 0.01, ***p < 0.001$.

Results obtained from Proc GENMOD in SAS v. 9.3 using the "dist=ZINB" option on the model statement are shown in Table 3. The estimated incidence rate ratio (IRR) for the at-risk class comparing Naf to SMFP $\exp(\gamma_4)$ is $\exp(-0.043) = 0.96$, with 95% CI (0.89, 1.04). As the confidence interval contains 1.0, the result is not statistically significant at the 0.05 significance level. Thus, with the other covariates held fixed, the mean increment μ_i for NaF in the at-risk class of children is approximately 96% the mean increment of SMFP. Similarly, the estimated IRR for the at-risk class comparing NaF+TMP to SMFP $\exp(\gamma_5)$ is $\exp(-0.014) = 0.99$, with 95% CI (0.90, 1.09). Note that the interaction of x_{i1} and x_{i2} was left out of excess zero model as its estimate was unstable.

5.2 Marginalized zero-inflated negative binomial regression model

Next, the MZINB model is fitted to the data in order to model the marginal mean caries increment after two years ν_i directly:

$$\text{logit}(\psi_i) = \gamma_0 + \gamma_1 x_{i1}^{bc} + \gamma_2 x_{i2}^{calc} + \gamma_3 (x_{i1}^{bc} \times x_{i2}^{calc}) + \gamma_4 x_{i4}^{Naf} + \gamma_5 x_{i5}^{Naftmp}$$

$$\log(\nu_i) = \beta_0 + \beta_1 x_{i1}^{bc} + \beta_2 x_{i2}^{calc} + \beta_3(x_{i1}^{bc} \times x_{i2}^{calc}) + \beta_4 x_{i4}^{Naf} + \beta_5 x_{i5}^{Naftmp}$$

where β_k are overall log IRRs as in negative binomial regression. Results in Table 4 for MZINB (see Appendix 1 for the SAS code) show that the incidence rate ratio (IRR) for the overall population comparing Naf to SMFP $\exp(\beta_4)$ is estimated as $\exp(-0.060) = 0.94$ with 95% CI (0.87, 1.01). Thus, all other covariates fixed, the mean increment ν_i for NaF in the overall population of children is approximately 94% the mean increment of SMFP. However, there are no statistically significant treatment differences since the IRR is not significantly different than 1. Similarly, the incidence rate ratio (IRR) for the overall population comparing Naf+TMP to SMFP $\exp(\beta_5)$ is estimated as $\exp(-0.034) = 0.97$ with 95% CI (0.88, 1.06). Table 4 additionally reports the results of negative binomial regression, which gives similar results as MZINB for these data.

Table 4. NB and MZINB results for two-year DMFS increments

	Negative Binomial		MZINB model			
	Marginal Mean		Marginal Mean		logit(Excess zero)	
Variable	est	se	est	se	est	se
Intercept	1.188	0.036***	1.191	0.036***	-2.24	0.28
Baseline Caries (high)	0.783	0.041***	0.785	0.040***	-2.87	1.45*
Calculus	-0.151	0.050**	-0.151	0.051**	-0.19	0.27
Baseline Caries*calculus	-0.109	0.084	-0.110	0.080	—	—
Naf	-0.056	0.039	-0.060	0.038	0.41	0.30
Naftmp	-0.022	0.048	-0.034	0.047	0.54	0.36

For MZINB, $\hat{\phi} = 0.637$ (se = 0.038); for NB, $\hat{\phi} = 0.805$ (se = 0.028).
$*p < 0.05, **p < 0.01, ***p < 0.001$.

6 Conclusion

A marginalized zero-inflated negative binomial regression model was proposed for population-averaged inference of count data with many zeros. The MZINB procedure directly models the marginal means of mixtures of two discrete distributions, one consisting of negative binomial counts and the other of structural zeros. This model formulation offers meaningful statements about an exposure effect on an entire population in contrast to the traditional ZINB model whose regression parameters have interpretations for unobservable latent classes. Whereas an average effect of an exposure in a population can be determined with additional computations following the fit of a traditional ZINB model, MZINB provides direct estimates of a homogeneous exposure effect that does not require post-modeling computations. Indeed, the marginal exposure effects of MZINB are given by log incidence rate ratios that have the same interpretations as in negative binomial or Poisson regression. The logistic model part for excess zeros in MZINB is of secondary interest as its role is to provide adjustment for overdispersion due to excess zeros.

In a simulation study of negative binomial generated counts with extra zeros, empirical Wald tests based on comparatively simple models such as Poisson and negative binomial regression had power nearly as great as that of Wald tests from MZINB models that were based on the model-based variance estimator, while maintaining the nominal Type I error rates. The relatively strong performance of the simpler models may be due to the large sample size considered or that all the covariates were categorical. A previous simulation study found that the MZIP had considerably greater power than empirical Wald tests from Poisson regression for smaller sample sizes and log-normally distributed continuous covariates [13]. Further simulation studies involving MZINB will be pursued in the future.

Despite the increasing popularity of the ZINB model in health-related fields, the idea of latent class effects can be troublesome for many investigators to communicate, often yielding misleading or incorrect statements. For example, Preisser *et al.* [18] found that many dental caries researchers interpreted the negative binomial regression parameters of the ZINB model with respect to the overall caries incidence, rather than the correct model-based interpretation relating to caries incidence within the *at-risk* population. This pattern of misinterpretation suggests that investigators when genuinely interested in marginal inference for count data may choose ZINB models simply because of goodness-of-fit considerations for data with many zeros. Unless, post-modeling calculations are performed to obtain exposure effects on the marginal mean count, the research analyst may unwittingly alter the target of inference. Use of the MZINB model maintains the marginal mean as the target of inference through direct modeling while accounting for excess zeros.

Generally, the research goal should guide the identification of a class of models that can address the question of interest; only when considering competing models within the identified class should goodness-of-fit considerations prevail. This approach to model selection based on collaboration between investigators and biostatistical scientists discourages purely empirical model fitting. Indeed, such exercises often reveal ZINB, MZINB, and negative binomial hurdle models of comparable complexity to have similar goodness-of-fit statistics [20]. The marginalized ZINB model belongs to a different model class than the traditional ZINB model and so choosing between them should be based on the research question. Future research could extend the MZINB model to accommodate random effects or develop inferential methods for marginal mean regression models based on other finite mixture distributions.

Acknowledgments

The authors thank Unilever Oral Care for the caries data. John Preisser acknowledges support from NC TraCS NIH UL1 TR001111-01. Part of this work was done while Kalyan Das was Pranab K. Sen Distinguished Visiting Professor in the Department of Biostatistics, University of North Carolina in 2013.

References

1. Albert J, Wang W, Nelson S. Estimating overall exposure effects for zero-inflated regression models with application to dental caries. *Statistical Methods in Medical Research.* 2014; **23**, 257-278.

2. Bohning D, Dietz E, Schlattmann P, Mendonca L, Kirchner U. The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. *J. R. Statist. Soc A* 1999; **162**, 195-209.

3. Francois M, Peter C, Gordon F. Dealing with excess of zeros in the statistical analysis of magnetic resonance imaging lesion count in multiple sclerosis. *Pharmaceutical Statistics* 2012; **11**, 417-424.

4. Gilthorpe MS. Modelling count data with excessive zeros: the need for class prediction in zero-inflated models and the issue of data generation in choosing between zero-inflated and generic mixture models for dental caries data. *Statistics in Medicine* 2009; **28**, 3539-3553.

5. Grainger RM and Reid DBW. Distribution of dental caries in children. *Journal of Dental Research* 1954; **33**, 613-623.

6. Heagerty PJ. Marginally specified logistic-normal models for longitudinal binary data. *Biometrics* 1999; **55**, 688-698.

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

8. Jansakul N, Hinde JP. Score test for extra-zero models in zero-inflated negative binomial models. Communications in Statistics-Simulation and Computation 2009; **38**, 92-108.

9. Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* 1992; **34**, 114.

10. Lawal BH. Zero-inflated count regression models with applications to some examples. *Qual Quant* 2012; **46**, 19-38.

11. Lee K, Joo Y, Song JJ, Harper DW. Analysis of zero-inflated clustered count data: a marginalized model approach. *Computational Statistics and Data Analysis* 2011; **55**, 824-837.

12. Lewsey J, Thomson W. 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* 2004; **32**, 183-189.

13. Long DL, Preisser JS, Herring AH, Golin CE. A marginalized zero-inflated Poisson regression model with overall exposure effects. *Statistics in Medicine* 2014; published online: 14 Sep 2014; DOI: 10.1002/sim.6293.

14. Min, Y. and Agresti, A. Random effects models for repeated measures of zero-inflated count data. *Statistical Modelling* 2005; **5**, 1-19.

15. Mullahy J. Specification and testing of some modified count data models. *Journal of Econometrics* 1986; **33**, 341-365

16. Mwalili SM, Lesaffre E and Declerck D. The zero-inflated negative binomial regression model with correction for misclassification: an example in caries research. *Stat Methods Med Res* 2008; **17**, 123-139.

17. Preisser JS, Das K, Benecha H, Stamm JW. Logistic regression for dichotomized counts. *Statistical Methods in Medical Research.* (published online May 26, 2014). DOI: 10.1177/0962280214536893

18. Preisser JS, Stamm JW, Long DL, Kincade M. Review and recommendations for zero-inflated count regression modeling of dental caries indices in epidemiological studies. *Caries Research* 2012, **46**, 413-423. PMC 3424072

19. Ridout M, Hinde J, Demetrio CGB. A Score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. *Biometrics* 2001; **57**, 219-223.

20. Rose CE, Martin SW, Wannemuehler KA, Plikaytis BD. On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data. J. Biopharmaceutical Statist. 2006; **16**, 463-481.

21. Stephen, KW, Chestnutt, IG, Jacobson, APM, McCall, DR, Chesters, RK, Huntingdon, E, Schäfer F, 1994. The effect of NaF and SMFP toothpastes on three-year caries increments in adolescents. *Int Dent J* **44**:287-295.

Appendix

SAS code for MZINB

The following SAS code is used to fit the MZINB model to the caries trial data whose results are shown in Table 4. The “parms” statement is used to specify starting values for the parameters b0, . . . , b5, which are the components of β; estimates from negative binomial regression model are used. Starting values may be specified for a0, a1, a2, a4 and a5, which are the components of γ; estimates from the MZIP model are used. The outcome variable is “dmfs”.

```
proc nlmixed data= all qpoints=15;
parms a0=-2.07 a1=-2.40 a2=-0.17          a4= 0.23 a5=0.19 phi=0.81
      b0= 1.19 b1= 0.78 b2=-0.15 b3=-0.11 b4=-0.06 b5=-0.02;
linpinfl = a0 + a1*bc + a2*calc  + a4*naf + a5*naftmp;
psi = 1/(1+exp(-linpinfl));
nu = exp(b0 + b1*bc + b2*calc + b3*bc_calc + b4*naf + b5*naftmp);
mu = nu/(1-psi);
alpha = 1/phi;
theta = 1/(1+(mu/alpha));
if dmfs=0 then loglike =log(psi + (1-psi)*(theta**alpha));
else loglike = log(1-psi) + lgamma(dmfs+alpha) - lgamma(alpha)
      + dmfs*log(1-theta)+alpha*log(theta) - lgamma(dmfs+1);
model dmfs ~ general(loglike);
estimate ’naf vs smfp, rate ratio’ exp(b4);
estimate ’naftmp vs smfp, rate ratio’ exp(b5);
run;
```