NIH Public Access
Author Manuscript
Comput Stat Data Anal. Author manuscript; available in PMC 2016 May 01.

Published in final edited form as:
Comput Stat Data Anal. 2015 May 1; 85: 54–66. doi:10.1016/j.csda.2014.11.014.

GEE type inference for clustered zero-inflated negative binomial regression with application to dental caries

Maiying Kong[a,*], **Sheng Xu**[a], **Steven M Levy**[b], and **Somnath Datta**[a]

[a]Department of Bioinformatics and Biostatistics, University of Louisville, Louisville, KY 40202, USA

[b]Department of Preventive & Community Dentistry and Department of Epidemiology, University of Iowa, Iowa City, IA 52242, USA

Abstract

Use of zero-inflated count data models is common in applications where the number of zero counts exceeds that predicted from a traditional count data model such as Poisson or negative binomial. When count data exhibiting inflated zero counts are correlated among subjects, a natural approach will be to fit a marginal model with the help of generalized estimating equations (GEE) that can incorporate subject-to-subject correlations. A GEE based zero-inflated negative binomial (ZINB) model is proposed to fit clustered counts with excessive zeros. However, the corresponding sandwich variance estimator appears to underestimate the true variance. The theoretical reasons for its failure are explained and a correction under additional modeling assumptions is offered. In addition, a clustered resampling (bootstrap) procedure is proposed to estimate the variance and it is shown that the bootstrap procedure captures the correct variance under no additional model assumptions. Utility of this marginal GEE based ZINB model over two other competing models has been assessed using a thorough simulation study. The resulting inference procedure is applied to study the association between the dental caries and fluoride exposures using a dataset extracted from the Iowa Fluoride Study. A number of risk factors of clinical significance are reliably identified using the proposed model.

Keywords

Zero-inflated models; Generalized estimating equations (GEE); Sandwich variance estimate; Bootstrap; Iowa Fluoride Study

1. Introduction

Statistical methods for longitudinal/clustered data have been developed in the past two decades. The longitudinal/clustered data have the feature that the observations from the

© 2014 Elsevier B.V. All rights reserved.

*Corresponding author. Tel: +1 502 852 3988. Fax: +1 502 852 3291. maiying.kong@louisville.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

same subject (or cluster) are correlated. Failure to consider the correlations of within-subject (or within-cluster) observations could result in invalid statistical inferences (Hedeker and Gibbons, 2006; Fitzmaurice et al., 2011). Many statistical methods have been developed for longitudinal/clustered data. Among them, the marginal models, i.e., the generalized estimating equations (GEEs), have been applied to characterize the relationships between responses and covariates at the population level (Hedeker and Gibbons, 2006; Fitzmaurice et al., 2011). The GEE model is based on the first and second moments of the response variable (Albert et al., 1988; Liang and Zeger, 1986). It has been shown that the GEE estimators are consistent and asymptotically normally distributed as long as the first moment is correctly specified. As a result, the GEE technique has been widely applied to longitudinal/clustered data analyses. In particular, the GEE model has been applied to correlated count data (Zeileis et al., 2008), where the log-link function and the quasi-Poisson first and second moments are assumed for each individual observation. In the case where the counts for zeros are above and beyond the number of sampling zeros expected by a quasi-Poisson distribution, the standard Poisson GEE model is no longer sufficient.

Even for independent count data, zero-inflated negative binomial (ZINB) and zero-inflated Poisson models have been developed to model excessive zero counts in the data (Zeileis et al., 2008; Mwalili et al., 2008). Zero-inflated regression models consist of two regression models: a logistic (or probit) regression model component for zero inflation and a count model component following a generalized linear model. The logistic regression models the probability of excess zeros in terms of available covariates. The count model relates the mean of the counts with available covariates using the framework of a generalized linear model when the response is not from the distribution degenerated at zero. As mentioned earlier, in many applications, data come in clusters where the counts from a cluster are correlated. The standard zero-inflated model cannot capture the correlation of the observations within a cluster. Ignoring the correlations among the observations could result in smaller variance estimates, thus leading to invalid statistical inferences. Although models for correlated zero-inflated count data have been proposed in the literature (Dobble and Welsh, 2001; Hall and Zhang, 2004), these methods have not specifically considered the case of a ZINB marginal.

In the current paper, we consider a ZINB model for correlated counts and the related inference. In order to distinguish it from the standard ZINB model, where the counts are independent, we refer to this model as the GEE-based ZINB model (or GEE.ZINB for short). This model allows for specifying a working correlation matrix along with the marginal relationship. We follow Hall and Zhang's (2004) adaptation of a general algorithm, called the expectation solution (ES) algorithm, to estimate the model parameters. However, when it came to estimating the variance of our parameter estimators, the sandwich formula given by Hall and Zhang (2004) seemed not to account for the variability due to the introduction of the latent variables which indicate whether each response variable is from a zero-degenerated distribution or from the count model. Following the work of Satten and Datta (2000), we introduce a correction term which substantially improves the situation in several simulation scenarios. This correction term is based on estimation of quantities that require additional modeling assumptions which could be difficult to verify and only hold approximately. We finally settled for a non-parametric bootstrap-based variance estimator

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

(Efron, 1979; Davison and Hinkley, 1997) which does not require any additional assumptions. Furthermore, it provided accurate answers in all settings.

We apply our GEE.ZINB model to analyze the caries scores at age five for children participating in the Iowa Fluoride Study (Levy et. al, 2001). For comparison, we also apply the standard ZINB model that does not account for correlation and a GEE based zero-inflated Poisson (GEE.ZIP) model that incorporates both the correlation and the zero inflation. We show GEE.ZINB is better than ZINB and is similar to GEE.ZIP. Both GEE.ZINB and GEE.ZIP identify a clinically relevant risk factor for dental caries in children. Also noteworthy is that the estimated coefficient of the extra zero component of the GEE.ZINB model was statistically significant indicating that a negative binomial model would have been inadequate for these data. Next, we carry out extensive simulation studies with realistic parameter choices guided by the real data.

The rest of the paper is organized as follows. In Section 2, we present the details of the proposed models, the estimation of the parameters, and their variance estimators. In Section 3, we report our data analysis for the dental study. Simulation results are reported in Section 4. The paper concludes with a discussion in Section 5.

2. The model and statistical inference

2.1. GEE-based ZINB model and estimation of parameters

A GEE-based ZINB model (GEE.ZINB) is developed to handle correlated/clustered count data, where the counts of zeros are above and beyond the number of sampling zeros expected from a NB distribution. Without loss of generality, let us denote by Y_{ij}, the number of counts for the j^{th} subject within the i^{th} cluster ($i = 1, \cdots, N; j = 1, \ldots, n_i$). The response variable Y_{ij} ($i = 1, \cdots, N; j = 1, \cdots, n_i$) follows a mixture of a degenerated distribution at zero with mixing probability of p_{ij} and a NB distribution with mean λ_{ij} with mixing probability $1 - p_{ij}$. The probability mass function of a NB distributed random variable W_{ij} can be written as

$$Pr(W_{ij}=w_{ij})=f_{NB}(w_{ij}|\lambda_{ij},\tau)=\frac{\Gamma(w_{ij}+\frac{1}{\tau})}{w_{ij}!\Gamma(\frac{1}{\tau})}\left(\frac{1}{1+\tau\lambda_{ij}}\right)^{\frac{1}{\tau}}\left(\frac{\tau\lambda_{ij}}{1+\tau\lambda_{ij}}\right)^{w_{ij}}, \quad (1)$$

where $w_{ij} = 0,1, \cdots$; and $\tau\,(\tau>0)$ is a shape parameter that quantifies the amount of over-dispersion. The mean and variance of W_{ij} are given by $E(W_{ij}|\lambda_{ij}, \tau) = \lambda_{ij}$ and $Var(W_{ij}|\lambda_{ij},\tau)=\lambda_{ij}+\tau\lambda_{ij}^2$, respectively. Unless $\tau = 0$, the variance is always larger than the mean λ_{ij}. Thus, the NB model adds a quadratic term $\tau\lambda_{ij}^2$ to the variance of Poisson to account for the extra-Poisson variation or over-dispersion (Wan et al., 2012). Considering Y_{ij} as the mixture of a zero-degenerated distribution with a probability p_{ij} and a NB model random variable W_{ij} with a probability $1 - p_{ij}$, the probability distribution of the response variable Y_{ij} can be written as:

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

$$Pr(Y_{ij}=y)=\begin{cases} p_{ij}+(1-p_{ij})\,Pr(W_{ij}=0) & if\ y=0, \\ (1-p_{ij})\,Pr(W_{ij}=y) & if\ y\geq 1, \end{cases}$$
$$=\begin{cases} p_{ij}(1-p_{ij})\left(\frac{1}{1+\tau\lambda_{ij}}\right)^{\frac{1}{\tau}} & if\ y=0, \\ (1-p_{ij})\frac{\Gamma(y+\frac{1}{\tau})}{y!\Gamma(\frac{1}{\tau})}\left(\frac{1}{1+\tau\lambda_{ij}}\right)^{\frac{1}{\tau}}\left(\frac{\tau\lambda_{ij}}{1+\tau\lambda_{ij}}\right)^{y} & if\ y\geq 1. \end{cases} \quad (2)$$

It is easy to check that the mean and variance of Y_{ij} are given by

$$E\{Y_{ij}\}=(1-p_{ij})\lambda_{ij}, \quad (3)$$

and

$$Var\{Y_{ij}\}=(1-p_{ij})\lambda_{ij}(1+\tau\lambda_{ij}+p_{ij}\lambda_{ij}). \quad (4)$$

When all observations are independent, the ZINB has been developed to model the zero-inflated data. In a ZINB model, both p_{ij} and λ_{ij} are modeled as functions of explanatory variables. The log link function is used to relate λ_{ij} to the explanatory variables (say, x_{ij}), and the logit link function is used to relate p_{ij} to the explanatory variable (say, z_{ij}). The predictors (say, x_{ij}) for λ_{ij} can be different from the predictors (say, z_{ij}) for p_{ij}. Let us assume that $\log(\lambda_{ij})=x_{ij}^T\beta$ and $\mathrm{logit}(p_{ij})=z_{ij}^T\gamma$ $(i=1,2,\cdots,N; j=1,\cdots,n_i)$. Thus, the mean of Y_{ij}, $E\{Y_{ij}\} = \mu_{ij} = (1 - p_{ij})\lambda_{ij}$, depends on the parameters β and γ. To account for the correlations for the observations within the same cluster, a correlation matrix, say $R_i(\alpha)$ for i^{th} cluster is introduced, and β and γ can be obtained by applying the following generalized estimating equations (GEE) (Hall and Zhang, 2004):

$$\sum_{i=1}^{N}\begin{pmatrix} \frac{\partial\mu_i^T}{\partial\beta} \\ \frac{\partial\mu_i^T}{\partial\gamma} \end{pmatrix} V_i^{-1}(\alpha,\beta,\gamma,\tau)(Y_i-\mu_i)=0. \quad (5)$$

Here $Y_i = (Y_{i1}, \cdots, Y_{in_i})^T$, $\mu_i = (\mu_{i1}, \cdots \mu_{in_i})^T$, $V_i(\alpha,\beta,\gamma,\tau)= Var(Y_i)=A_i^{\frac{1}{2}}R_i(\alpha)A_i^{\frac{1}{2}}$ with $A_i = Diag\{Var(Y_{ij})|_{j=1,\cdots,n_i}\}$. This direct application of GEE to the clustered zero-inflated models may not be identifiable because β and γ are typically confounded (e.g., share information) in equation (5) (Hall and Zhang, 2004). It has been recommended to estimate β and γ in two separate equations by introducing latent variables u_{ij} ($i = 1, \cdots, N; j = 1, \cdots, n_i$), which indicate whether the random variable Y_{ij} is from a zero-degenerated distribution or NB distribution.

Let $u_{ij} = 0$ if $Y_{ij} \sim f_{NB}(\lambda_{ij}, \tau)$, and $u_{ij} = 1$ if Y_{ij} is from a zero-degenerated distribution. We have $\Pr(u_{ij} = 1) = p_{ij}$. The GEE for γ can be written as:

$$\sum_{i=1}^{N}\frac{\partial p_i^T}{\partial\gamma}\{V_{\gamma i}\}^{-1}(u_i-p_i)=0. \quad (6)$$

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Here $u_i = (u_{i1}, \cdots, u_{in_i})^T$, $p_i^T = (p_{i1}, \cdots, p_{in_i})$, and $p_{ij}(\gamma) = \frac{\exp(z_{ij}^T\gamma)}{1+\exp(z_{ij}^T\gamma)}$ implying $\frac{\partial p_{ij}}{\partial \gamma} = z_{ij}\frac{\exp(z_{ij}^T\gamma)}{(1+\exp(z_{ij}^T\gamma))^2}$; $V_{\gamma i} = A_i^{\frac{1}{2}} R_{1i}(\alpha_1) A_i^{\frac{1}{2}}$ is the variance matrix for u_i, where $A_i = Diag\{p_{i1}(1 - p_{i1}), p_{i2}(1 - p_{i2}), \cdots, p_{in_i}(1 - p_{in_i})\}$ with the j^{th} entry being the variance of u_{ij}, and $R_{1i}(\alpha_1)$ is the correlation matrix for u_i. Similarly, the GEE for β can be written as

$$\sum_{i=1}^{N} \frac{\partial \lambda_i^T}{\partial \beta}\{V_{\beta i}\}^{-1} Diag(1-u_i)(y_i-\lambda_i)=0 \quad (7)$$

Here $\lambda_i^T = (\lambda_{i1}, \cdots, \lambda_{in_i})$, where $\log(\lambda_{ij}) = x_{ij}^T\beta$, which implies that $\lambda_{ij} = \exp(x_{ij}^T\beta)$ and $\frac{\partial \lambda_{ij}}{\partial \beta} = x_{ij}\exp(x_{ij}^T\beta)$. As before, $V_{\beta i}\left(=D_i^{1/2} R_{2i}(\alpha_2) D_i^{1/2}\right)$ is the variance matrix for Y_i, where $D_i = Diag\{\lambda_{i1}(1 + \tau\lambda_{i1}), \lambda_{i2}(1 + \tau\lambda_{i2}), \cdots, \lambda_{in_i}(1 + \tau\lambda_{in_i})\}$ and $R_{2i}(\alpha_2)$ is the correlation matrix for Y_i. The diagonal matrix $Diag\{1 - u_i\} = Diag(1 - u_{i1}, \cdots, 1 - u_{in_i})$ in equation (7) indicates that only Y_{ij} from the NB distribution (*i.e.*, $u_{ij} = 0$) contributes the estimating equation (7). Given u_{ij} ($i = 1, \cdots, N$; $j = 1, \cdots, n_i$), β and γ are estimated by the Fisher-scoring method, which is an iterative algorithm for solving the estimating equations such as (6) and (7). However, since u_{ij} ($i = 1, \cdots, N$; $j = 1, \cdots, n_i$) are latent (unobserved) variables, the solutions are not directly usable. Instead the expectation-solution algorithm (Hall and Zhang, 2004; Rosen et al., 2000) can be applied, where in each iteration, u_{ij} in (6) and (7) is replaced by the conditional mean of u_{ij} given Y_{ij} and the current estimates for γ, β and τ.

Suppose that, in the b^{th} iteration, the estimates for γ, β, and τ are $\gamma^{(b)}$, $\beta^{(b)}$, and $\tau^{(b)}$, then the conditional mean of u_{ij} can be obtained as

$$\begin{aligned} u_{ij}^{(b)} &= \Pr(u_{ij}=1|y, \gamma^{(b)}, \beta^{(b)}, \tau^{(b)}) \\ &= \{\frac{\Pr(u_{ij}=1, y_{ij}=0|\gamma^{(b)}, \beta^{(b)}, \tau^{(b)})}{\Pr(y_{ij}=0|\gamma^{(b)}, \beta^{(b)}, \tau^{(b)})}\}^{1_{\{y_{ij}=0\}}} \\ &= \{1+\frac{(1-p_{ij}^{(b)})(\frac{1}{1+\tau^{(b)}\lambda_{ij}^{(b)}})^{\frac{1}{\tau^{(b)}}}}{p_{ij}^{(b)}}\}^{-1} 1_{\{y_{ij}=0\}}, \end{aligned} \quad (8)$$

where $p_{ij}^{(b)} = \frac{\exp(z_{ij}^T\gamma^{(b)})}{1+\exp(z_{ij}^T\gamma^{(b)})}$ and $\lambda_{ij}^{(b)} = \exp(x_{ij}^T\beta^{(b)})$. Thus, combining this with (6) and (7), estimates of γ and β can be updated by the following iterated formulas:

$$\gamma^{(b+1)} = \gamma^{(b)} + \{\sum_{i=1}^{N} \frac{\partial p_i^T}{\partial \gamma}\{V_{\gamma i}\}^{-1}\frac{\partial p_i}{\partial \gamma^T}\}^{-1} \sum_{i=1}^{N} \frac{\partial p_i^T}{\partial \gamma}\{V_{\gamma i}\}^{-1}\left(u_i^{(b)} - p_i^{(b)}\right)|_{(\gamma,\alpha_1)=(\gamma^{(b)},\alpha_1^{(b)})}, \quad (9)$$

and

$$\beta^{(b+1)} = \beta^{(b)} + \{\sum_{i=1}^{N} \frac{\partial \lambda_i^T}{\partial \beta}\{V_{\beta i}\}^{-1} Diag\left(1-u_i^{(b)}\right)\frac{\partial \lambda_i}{\partial \beta^T}\}^{-1} S_\beta|_{(\beta,\tau,\alpha_2)=(\beta^{(b)},\tau^{(b)},\alpha_2^{(b)})}, \quad (10)$$

where $S_\beta = \sum_{i=1}^{N} \frac{\partial \lambda_i^T}{\partial \beta}\{V_{\beta i}\}^{-1} Diag\left(1-u_i^{(b)}\right)\left(y_i-\lambda_i^{(b)}\right)$.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Note that the parameter τ is only related to the variance function of the NB part. To estimate τ, let us introduce $\varepsilon_{ij}^2(\beta)=(Y_{ij}-\lambda_{ij}(\beta))^2$, and $\varepsilon_i^2(\beta)=(\varepsilon_{i1}^2(\beta),\cdots,\varepsilon_{in_i}^2(\beta))^T$. If β is set at the correct value, we have $E\left(\varepsilon_{ij}^2(\beta)\right)=v_{ij}(\tau)$ provided u_{ij} = 1, where $v_{ij}=\lambda_{ij}(1+\tau\lambda_{ij})$. Thus, we propose estimating τ by solving the following equations:

$$\sum_{i=1}^{N}\Gamma_i^T H_i(\varepsilon_i^2(\beta)-v_i(\tau))=0. \quad (11)$$

Here $\Gamma_i=\left(\frac{\partial(v_i(\tau))}{\partial\tau}\right)=\begin{pmatrix}\lambda_{i1}^2\\ \cdots\\ \lambda_{in_i}^2\end{pmatrix}$, $H_i = Diag(1-u_i)$ and $v_i(\tau)=(v_{i1},\cdots,v_{in_i})^T$. Given β and γ, τ can be obtained by solving equation (11), which results in:

$$\tau=\frac{\sum_{i=1}^{N}\sum_{j=1}^{n_i}\lambda_{ij}^2(1-u_{ij})(\varepsilon_{ij}^2(\beta)-\lambda_{ij})}{\sum_{i=1}^{N}\sum_{j=1}^{n_i}(1-u_{ij})\lambda_{ij}^4}=\frac{\sum_{i=1}^{N}\sum_{j=1}^{n_i}\lambda_{ij}^2(1-u_{ij})((y_{ij}-\lambda_{ij})^2-\lambda_{ij})}{\sum_{i=1}^{N}\sum_{j=1}^{n_i}(1-u_{ij})\lambda_{ij}^4} \quad (12)$$

Now we come to estimating α_1 and α_2. To estimate α_1, let us set

$$U_{\gamma ist}=\frac{(u_{is}-p_{is})(u_{it}-p_{it})}{\sqrt{p_{is}(1-p_{is})p_{it}(1-p_{it})}},$$

which has the expected value of $\rho_{\gamma ist}$, the correlation coefficient between u_{is} and u_{it}. Denote $U_{\gamma i}=(U_{\gamma i12},U_{\gamma i13},\cdots,U_{\gamma in_i-1,n_i})^T$, and $\rho_{\gamma i}(\alpha_1)=E\{U_{\gamma i}\}=(\rho_{\gamma i12},\rho_{\gamma i13},\cdots,\rho_{\gamma in_i-1,n_i})^T$. The parameter α_1 can be estimated by solving the following equation:

$$\sum_{i=1}^{N}E_{\gamma i}^T W_{\gamma i}^{-1}(U_{\gamma i}-\rho_{\gamma i}(\alpha_1))=0, \quad (13)$$

where $E_{\gamma i}=\frac{\partial\rho_{\gamma i}(\alpha_1)}{\partial\alpha_1}$, and $W_{\gamma i}$ is a working variance covariance matrix for the $U_{\gamma i}$. If we set $W_{\gamma i}$ to be the identity matrix, and assume a symmetric compound structure for $R_{\gamma i}(\alpha_1)$, an estimate of the common correlation parameter α_1 can be obtained as:

$$\hat{\alpha}_1=\frac{1}{N^*}\sum_{i=1}^{N}\sum_{s<t}\frac{(u_{is}-p_{is})(u_{it}-p_{it})}{\sqrt{p_{is}(1-p_{is})p_{it}(1-p_{it})}} \quad (14)$$

where $N^*=\sum_{i=1}^{N}\frac{n_i(n_i-1)}{2}$. Standardizing it, we get our final estimator of α_1:

$$\hat{\alpha}_1=\frac{\frac{1}{N^*}\sum_{i=1}^{N}\sum_{s<t}\frac{(u_{is}-p_{is})(u_{it}-p_{it})}{\sqrt{p_{is}(1-p_{is})p_{it}(1-p_{it})}}}{\frac{1}{N_{tot}}\sum_{i=1}^{N}\sum_{j=1}^{n_i}\frac{(u_{ij}-p_{ij})^2}{p_{ij}(1-p_{ij})}} \quad (15)$$

where $N_{tot}=\sum_{i=1}^{N}n_i$.

Similarly, to estimate α_2, we set

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

$$U_{\beta ist}=\frac{(y_{is}-\lambda_{is})(y_{it}-\lambda_{it})}{\sqrt{\lambda_{is}(1+\tau\lambda_{is})\lambda_{it}(1+\tau\lambda_{it})}}. \quad (16)$$

and $E\{U_{\beta ist}\} = \rho_{\beta ist}$, the correlation coefficient between y_{is} and y_{it} when they follow NB distributions. Denote $U_{\beta i} = (U_{\beta ist}, U_{\beta ist}, \cdots, U_{\beta in_i-1,n_i})^T$, and $\rho_{\beta i}(\alpha_1) = E\{U_{\beta i}\} = (\rho_{\beta ist}, \rho_{\beta ist}, \cdots, \rho_{\beta in_i-1,n_i})^T$ when each component of Y_i follows a NB distribution. Thus, α_2 can be estimated by the following equation:

$$\sum_{i=1}^{N} E_{\beta i}^T W_{\beta i}^{-1} H_{\beta i}\,(U_{\beta i}-\rho_{\beta i}(\alpha_2)) = 0. \quad (17)$$

Here $E_{\beta i}=\frac{\partial \rho_{\beta i}(\alpha_2)}{\partial \alpha_2}$, $W_{\beta i} \approx Cov(U_{\beta i})$, and $H_{\beta i} = Diag\{(1 - u_{i1})(1 - u_{i2}), \cdots, (1 - u_{in_i-1})(1 - u_{in_i})\}$. In the case that $W_{\beta i}$ is the identity matrix and $R_{\beta i}(\alpha_2)$ is the symmetric compound structure, α_2 could be obtained as follows:

$$\hat{\alpha}_2=\frac{1}{N^*}\sum_{i=1}^{N}\sum_{s<t}\frac{(1-u_{is})(1-u_{it})(y_{is}-\lambda_{is})(y_{it}-\lambda_{it})}{\sqrt{\lambda_{is}(1+\tau\lambda_{is})\lambda_{it}(1+\tau\lambda_{it})}}. \quad (18)$$

Here $N^*=\sum_{i=1}^{N}\sum_{s<t}(1-u_{is})(1-u_{it})$. As before, an alternative range preserving (standardized) estimate for α_2 is

$$\hat{\alpha}_2=\frac{1}{N^*}\sum_{i=1}^{N}\sum_{s<t}\frac{\frac{(1-u_{is})(1-u_{it})(y_{is}-\lambda_{is})(y_{it}-\lambda_{it})}{\sqrt{\lambda_{is}(1+\tau\lambda_{is})\lambda_{it}(1+\tau\lambda_{it})}}}{\frac{1}{N_{tot}}\sum_{i=1}^{N}\sum_{j=1}^{n_i}\frac{(1-u_{ij})^2(y_{ij}-\lambda_{ij})^2}{\lambda_{ij}(1+\tau\lambda_{ij})}}, \quad (19)$$

where $N_{tot}=\sum_{i=1}^{N}\sum_{j=1}^{n_i}(1-u_{ij})^2$. In this paper, the range preserving estimates for α_1 and α_2 given by (15) and (19) are used.

To obtain the final parameter estimates for β, γ, τ, α_1 and α_2, an iterative method is required to iterate between estimating β and γ (given the current estimate of τ, α_1 and α_2) as the solution of equation (6) and (7), and estimating τ, α_1 and α_2 (given the current estimate of β and γ) as the solution of (11), (13) and (17) until convergence.

2.2. Variance estimation

Liang and Zeger (1986) showed that the GEE estimators are consistent and asymptotic normal for any choice of working correlation matrix, provided that the regression model for the mean response has been correctly specified. However, the standard GEE model does not involve any latent variables. When the latent variables do exist, the expectation-solution algorithm replaces the latent variables by their conditional mean given the response variable and current estimates of regression parameters. However, ignoring the variation due to the replacement of each latent variable with its conditional mean can result in smaller estimated variance as compared to the true variance of the regression parameter estimators.

Let us stack the estimating equations (6) and (7) to write $\sum_{i=1}^{N} S_i(u_i, y_i|\theta)=0$, where

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

$$S_i(u_i, y_i|\theta)=\begin{pmatrix} \frac{\partial p_i^T}{\partial \gamma}\{V_{\gamma i}\}^{-1}(u_i-p_i) \\ \frac{\partial \lambda_i^T}{\partial \beta}\{V_{\beta i}\}^{-1}Diag(1-u_i)(y_i-\lambda_i) \end{pmatrix}. \quad (20)$$

When the latent variables are replaced by their conditional means, the estimating equations become $\sum_{i=1}^{N} S_i(y_i|\theta)=0$, where

$$S_i(y_i|\theta)=E\{S_i(u_i, y_i|\theta)|y_i\}=\begin{pmatrix} \frac{\partial p_i^T}{\partial \gamma}\{V_{\gamma i}\}^{-1}(E\{u_i|y_i\}-p_i) \\ \frac{\partial \lambda_i^T}{\partial \beta}\{V_{\beta i}\}^{-1}Diag(1-E\{u_i|y_i\})(y_i-\lambda_i) \end{pmatrix}. \quad (21)$$

The asymptotic covariance for the estimator of θ is given by the following sandwich form

$$\hat{V}ar\begin{pmatrix} \hat{\gamma} \\ \hat{\beta} \end{pmatrix}=\hat{B}^{-1}\hat{M}\hat{B}^{-1}, \quad (22)$$

where $\hat{B}=\sum_{i=1}^{N}\frac{\partial S_i(y_i|\theta)}{\partial \theta}$, and

$$\hat{M}=\sum_{i=1}^{N}\begin{pmatrix} \frac{\partial \hat{p}_i^T}{\partial \hat{\gamma}}\{\hat{V}_{\gamma i}\}^{-1}(\hat{u}_i-\hat{p}_i) \\ \frac{\partial \hat{\lambda}_i^T}{\partial \hat{\beta}}\{\hat{V}_{\beta i}\}^{-1}Diag(1-\hat{u}_i)(y_i-\hat{\lambda}_i) \end{pmatrix}\begin{pmatrix} \frac{\partial \hat{p}_i^T}{\partial \hat{\gamma}}\{\hat{V}_{\gamma i}\}^{-1}(\hat{u}_i-\hat{p}_i) \\ \frac{\partial \hat{\lambda}_i^T}{\partial \hat{\beta}}\{\hat{V}_{\beta i}\}^{-1}Diag(1-\hat{u}_i)(y_i-\hat{\lambda}_i) \end{pmatrix}^T.$$

Here the parameters for equation (22) have been replaced by their final estimates. In order to obtain the Hessian matrix $\frac{\partial S_i(y_i|\theta)}{\partial \theta}$, Satten and Data (2000) provided the following formula for the marginal Hessian matrix, where $F_\theta(u_i|y_i)$ is the conditional CDF of u_i given y_i:

$$\frac{\partial S_i(y_i|\theta)}{\partial \theta^T}=\int\left\{\frac{\partial S_i(u_i,y_i|\theta)}{\partial \theta^T}+(S_i(u_i,y_i|\theta)-S_i(y_i|\theta))\,(S_i(u_i,y_i|\theta)-S_i(y_i|\theta))^T\right\}dF_\theta(u_i|y_i).$$

In the present context, the first term turns out to be

$$\int\left\{\frac{\partial S_i(u_i,y_i|\theta)}{\partial \theta^T}\right\}dF_\theta(u_i|y_i)=\begin{pmatrix} -\frac{\partial p_i^T}{\partial \gamma}\{V_{\gamma i}\}^{-1}\left(\frac{\partial p_i}{\partial \gamma^T}\right) & 0 \\ 0 & -\frac{\partial \lambda_i^T}{\partial \beta}\{V_{\beta i}\}^{-1}Diag(1-u_i)\frac{\partial \lambda_i}{\partial \beta^T} \end{pmatrix}=B_{1i}, \quad (23)$$

and the second term turns out to be

$$\begin{aligned} &\int\left\{(S_i(u_i,y_i|\theta)-S_i(y_i|\theta))\,(S_i(u_i,y_i|\theta)-S_i(y_i|\theta))^T\right\}dF_\theta(u_i|y_i) \\ &=\begin{pmatrix} \frac{\partial p_i^T}{\partial \gamma}\{V_{\gamma i}\}^{-1} \\ -\frac{\partial \lambda_i^T}{\partial \beta}\{V_{\beta i}\}^{-1}Diag(y_i-\lambda_i) \end{pmatrix}Var\{u_i|y_i\}\begin{pmatrix} \frac{\partial p_i^T}{\partial \gamma}\{V_{\gamma i}\}^{-1} \\ -\frac{\partial \lambda_i^T}{\partial \beta}\{V_{\beta i}\}^{-1}Diag(y_i-\lambda_i) \end{pmatrix}^T=B_{2i}, \end{aligned} \quad (24)$$

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

where $Var\{u_i|y_i\} = \int (u_i - E\{u_i|y_i\})(u_i - E\{u_i|y_i\})^T dF_\theta(u_i|y_i)$, the conditional variance of u_i given y_i. Recall that u_{ij} is a latent variable, indicating whether the observation y_{ij} is from zero-degenerated distribution, the conditional mean $E(u_{ij}|y_{ij})$ is estimated by $u_{ij}^{(b)}$ as defined in equation (8), and the conditional variance of $Var(u_{ij}|y_{ij})$ is estimated by $u_{ij}^{(b)}\left(1-u_{ij}^{(b)}\right)$. We assume the conditional correlation matrix for u_i given y_i is the same as its marginal correlation matrix, say $R_{\gamma i}$. Thus, $Var\{u_{ij}|y_{ij}\}$ in equation (24) could be estimated by $\hat{var}\{u_i|y_i\} = Diag\left(u_{ij}^{(b)}\left(1-u_{ij}^{(b)}\right)\right)^{1/2}\{R_{\hat{\gamma}i}\}Diag\left(u_{ij}^{(b)}\left(1-u_{ij}^{(b)}\right)\right)^{1/2}$. Thus, our sandwich estimator the variance of $\hat{\theta}$ is $\hat{B^{-1}M}\hat{B^{-1}} = (\hat{B_1} + B_2)^{-1}M(\hat{B_1} + B_2)^{-1}$, where $B_1 = \Sigma B_{1i}$, $B_2 = \Sigma B_{2i}$, and the estimates are obtained by replacing all quantities with the corresponding estimates at convergence.

It should be noted that the proposed sandwich estimate $(\hat{B_1} + \hat{B_2})^{-1}M(\hat{B_1} + \hat{B_2})^{-1}$ accounts for the variation of the latent variables, which is captured in $\hat{B_2}$. We expect the performance of the proposed sandwich estimator to be better than the traditional sandwich estimator (Hall and Zhang, 2004) that ignores the variation from the latent variable and has the form of $\hat{B_1}^{-1}M\hat{B_1}^{-1}$. However, when the counts are correlated, neither can the conditional distribution of y_i given u_i, nor the distribution of u_i be exactly specified from the marginal model. Thus, the proposed sandwich estimate can still fall short of the true variance when the counts are highly correlated; we have observed this in our simulation studies in Section 4.

To remedy the situation, we also consider a non-parametric alternative. We adapt the non-parametric bootstrap technology (Efron, 1979; Davison and Hinkley, 1997) to the clustered data setup to obtain this variance estimates. To do the bootstrap resampling, we draw a random sample (with replacement) of clusters from the collection of all clusters in the observed dataset. Note that resampling this way preserves the correlation structure that was present in the original sample. We refit the GEE.ZINB model to the resampled clustered data. This process is repeated 200 times, and the variance for each variable is calculated by the empirical variance of the 200 sets of estimated parameters. Note that covariances between parameter estimators can also be calculated this way. We show in Section 4 that these resulting variance estimators capture the correct variances in all simulation settings. However, first we apply our model to infer the risk factors for childhood caries and to guide our simulation setup.

3. An application of GEE.ZINB inference to Iowa Fluoride Study data

We illustrate our method using a dataset extracted from the Iowa Fluoride Study (Levy et al., 2001). The goal of our analysis will be to identify risk factors for childhood caries experience (CE). Dental caries, also known as tooth decay or cavities, is an infection that causes demineralization and destruction of the dental hard tissues. The Iowa Fluoride Study (http://www1.dentistry.uiowa.edu/preventive-fluoride-study) is an ongoing study of a cohort of Iowa children (mostly white, non-Hispanic, and middle or upper class socio-economic status) that began in 1992. Substantial dietary data, including data on fluoride intake, were obtained longitudinally through periodic questionnaires sent to parents every 1.5–6 months.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Surface-specific assessments of dental fluorosis and caries were made through dental examinations at the ages of 5 years, 9 years, 13 years and 17 years.

For this illustration, we focus on the caries data at the age of 5 years from the Iowa Fluoride Study. There was a small fraction of children with missing covariates; we assume that such missingness was completely at random and these children were dropped from our analysis. We score each dental surface with a value of 0, 1 or 2 depending on the caries severity, and these scores were then added to get overall caries experience scores (CESs) for each tooth of each child. Note that, in the general terminology of Section 2, each child is a "cluster" and the CES for various teeth for a child are the "subjects" in a cluster.

We analyze the CES treating them as count data. Figure 1A shows the marginal histogram of CES (averaged over all teeth and children in our sample), and Figure 1B shows the marginal histogram of CES excluding zeros. It is clear that the zeros dominate in the dataset, and the excessive zeros provide evidence of zero inflation. This was more formally validated in our subsequent analysis.

Since the CES for different teeth within a child are potentially correlated, a GEE-based model to analyze whether the dental caries is associated with different exposures and treatments seems appropriate. The potentially useful covariates from a clinical standpoint included daily fluoride intake (*AUCmgF0_5yrs*), daily soda pop intake (*AUCSodaOz0_5yrs*), average of tooth brushing frequencies (*ToothBrushingFreqPerDayAvg*), the average proportion of times a dental visit was conducted with each individual point assessing the previous 6 months (*DentalVisitPast6moAvg*), the average proportion of times a professional dental fluoride treatment was received with each individual point assessing the previous 6 months (*FluorideTreatmentPast6moAvg*), and the average home tap water fluoride level for all the returned questionnaires (*HomeFluorideppmAvg*). We also included gender and the exact age as covariate in our models. After removing the observations with missing covariates, the study subset included 8189 data values (CES) grouped into 414 clusters (children) in total. The minimum size of the clusters (*i.e.*, number of teeth per child) was 13, and the maximum size was 20. Indeed 90% of clusters had size of 20.

We have fitted a GEE.ZINB model to these data. A compound symmetric correlation structure is assumed for each of the two components (i.e., count component and zero-inflation component) of the model and the two correlation coefficients are estimated using formulas (15) and (19), respectively. The parameter estimates, their various estimated standard errors and p-values from two-sided Wald tests are reported in Table 1. We use the sandwich variance estimator $(\hat{B_1} + \hat{B_2})^{-1} M (\hat{B_1} + \hat{B_2})^{-1}$ that accounts for the variability due to estimating the latent variables *u*. The estimated S.E. and the resulting p-values are listed under the column "S.E. (SW)" and the next column, respectively. For comparison, we also report the results both from using the sandwich variance estimator $\hat{B_1}^{-1} M \hat{B_1}^{-1}$ as prescribed by Hall and Zhang (2004) and the non-parametric cluster bootstrap-based variance estimators. These values are indicated by "S.E. (HZ)" and "S.E. (BS)", respectively (see Table 1A).

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

For the sake of comparison, we also fit a GEE based zero-inflated Poisson (GEE.ZIP) model and an independent ZINB model to these data. The implementation of GEE.ZIP is similar to GEE.ZINB, where the parameter τ in the variance component of NB is set zero. The ZINB is implemented using the existing R package *pscl* (http://cran.rproject.org/web/packages/pscl/pscl.pdf; also see Zeileis et al., 2008). The initial values for regression parameters β and γ in the GEE.ZINB and GEE.ZIP were those obtained from ZINB. These results obtained from GEE.ZIP and ZINB models are reported in Table 1B and 1C respectively. As can be seen from Table 1, the magnitude and the direction (i.e., the sign of the estimates) of the estimates based on the three different models are similar. The correlation coefficient for the observations within the same child was 0.165 for the count data component in the GEE.ZINB model. The small correlation coefficient indicates that the point estimates from ZINB will be similar to those from GEE.ZINB for this count data part, which indeed happens to be the case to a large extent. The *DentalExamAge* effect is positive and declared to be statistically significant by all three models, which makes clinical sense since waiting longer for the dental examination increases the severity of CE. Both ZINB and GEE.ZINB models indicate that *AUCmgFO_5yrs* is negatively associated with CES, which also makes clinical sense since fluoride acts to prevent or slow subsequent caries development.

For the zero inflation part, the GEE.ZINB and GEE.ZIP are quite similar, while the GEE.ZINB and ZINB model inferences differ somewhat. In particular, the variables *DentalExamAge* and *ToothBrushingFreqPerDayAvg* are both deemed to be statistically significant (at $p<0.05$) by the ZINB model, but not by the GEE.ZINB model using the proposed sandwich estimate of variance. While the signs of the corresponding parameter estimates make good clinical sense, the ZINB model estimates (including the variance estimates) are based on the maximum likelihood estimation theory for independent data. Consequently, the estimated standard errors are incorrect, since there was a modest within-cluster correlation of 0.266 for the zero-inflated part of the data. The estimated standard error for ZINB could be underestimating the true standard error and, hence, the p-values could be too optimistic. We have verified this in the simulation studies presented in Section 4. Comparing the three sets of standard error estimates for the GEE.ZINB model in Table 1, we see that our sandwich estimate lies between that using the HZ recipe and the bootstrap-based estimate in all cases. The extensive simulation studies performed and reported in the next section show that the bootstrap version is most accurate. On the other hand, the HZ estimates could severely underestimate the true values, leading to optimistic p-values and potentially false positive results.

4. Simulation

We conducted an extensive simulation study to evaluate the performance of the estimators obtained from GEE.ZINB and ZINB under the general setting where both within-cluster correlation and zero inflation are present in the data. While the parameter choices in the simulation studies were guided by the data analysis section, we varied the amount of within-cluster correlation as high, medium and absent.

Data for the count model were generated from the following log-linear model:

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

$$\begin{aligned}\log(\lambda_{ij}) = &-2.713 - 0.018\,Gender + 0.696\,DentalExamAge - 0.856\,AUCmgF0\_5yrs \\ &+ 0.072\,AUCSodaOz0\_5yrs - 0.023\,ToothBrushingFreqPerDayAvg \\ &+ 0.740\,DentalVisitPast6moAvg + 0.955\,FluorideTreatmentPast6moAvg \\ &- 0.062\,HomeFluoridepmAvg,\end{aligned} \tag{25}$$

and the probability p_{ij} of a data value coming from the zero component was controlled via the following logistic regression model:

$$\begin{aligned}\text{logit}(p_{ij}) = &3.599 + 0.201\,Gender - 0.389\,DentalExamAge + 0.392\,AUCmgF0\_5yrs \\ &- 0.010\,AUCSodaOz0\_5yrs + 0.503\,ToothBrushingFreqPerDayAvg \\ &+ 0.825\,DentalVisitPast6moAvg - 0.875\,FluorideTreatmentPast6moAvg \\ &+ 0.046\,HomeFluoridepmAvg.\end{aligned} \tag{26}$$

Note that just like the real data, the covariates used in the model were cluster (child) level covariates. We used the same number of clusters (=414) and the corresponding covariates as in our dental dataset analyzed in the previous section. However, for simplicity we take the cluster size to assume a constant value of 20. To generate the indicators u_{ij} of zero inflation of subjects in each cluster, we start from a sample of 20 correlated normal variates, each with zero mean and unit variance, and with a constant pairwise correlation coefficient, say 3, and convert them into 20 correlated binary random variables using appropriate transformations while maintaining the marginal success probabilities equal to (26).

Meanwhile, we also generated 20 correlated negative binomial random variables for each subject via quantile-probability transformations starting from another (independent) set of 20 normal variates, each with zero mean and unit variance, and with a constant pairwise correlation coefficient of α. We ensured that marginally these NB variates have means specified by the log-linear model (25) and dispersion τ=0.347. Note that, for simplicity, we have used the same correlation coefficient α in both parts of the model. However, since the nonlinear transformations were different, the magnitude of the induced correlation in two pieces of the generated data was different. If for a subject, u_{ij} =1, we set the corresponding CES equal to zero; otherwise, we assign the corresponding NB variate to its CES.

A total of 1000 datasets were generated for each of three scenarios determined by amount of correlation. For each dataset, parameter estimates and their estimated standard errors were computed from GEE.ZINB and ZINB respectively. These values were averaged over the 1000 runs and reported along with the empirical standard deviations of the parameter estimates. The main results are presented in Table 2.

4.1 Comparison of estimators from GEE.ZINB and ZINB

Based on the simulation results, we conclude that (i) the biases of the estimators based on GEE.ZINB and ZINB are similar, and all the estimated coefficients in the GEE.ZINB are asymptotically normally distributed (see Q-Q plots in Figure 2); (ii) when the correlation of within-subject counts is high (Table 2), the ratios of the sandwich SE and the true SE of regression parameter estimators for the count model lie between 0.691 and 0.798 for GEE.ZINB. The same ratios for ZINB are between 0.339 and 0.378, indicating that, while

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

both large sample estimators (SW for GEE.ZINB and inverse Fisher information for ZINB) underestimate the true variance, the degree of underestimation for ZINB is more severe; (iii) for the zero inflation model in GEE.ZINB, the ratio of the estimated standard error from the proposed sandwich estimator and the true SE is close to 1 regardless of the magnitude of the within cluster correlation, for that reason we only showed the simulation results for the case of high correlation for zero inflation model in Table 2; (iv) for ZINB model, the ratios of the estimated SE and the true SE are about 0.35 under high correlation (see Table 2), 0.50~0.70 under medium correlation (see Table S1 in Supplementary materials), and 1 under independence (see Table S2 in Supplementary materials), indicating that the standard ZINB inference may not be suitable for clustered data.

4.2 Comparisons of the three variance estimators for GEE.ZINB

It is noticed that the proposed sandwich variance estimator for GEE.ZINB is underestimating the true variance for the count model part, although we have accounted for the variation of the latent variables, and the degree of underestimation is more severe when the correlation is high. A plausible explanation for its less than perfect behavior is that estimation of this additional part B_2 is not possible without making some crude approximation (e.g., the conditional correlation matrix for u_i given y_i is the same as the marginal correlation matrix). There does not seem to be a natural (e.g., model-free) closed form estimator for this part. It should be noted, however, that this estimator is still preferable over the HZ estimator in all cases. The estimator based on the non-parametric bootstrap (each based on a BS replication of size 200) seems to be most accurate in all cases, as shown by the ratio comparison with the true values.

The proposed sandwich variance estimator provides correct estimates for the zero-inflation part of the model regardless of the magnitude of the correlation. In real data analysis, it may be safer to use the bootstrap method for valid statistical inference for the GEE.ZINB, even though it is computationally a bit more intensive and does not offer a fixed answer due to resampling.

5. Discussion

We have developed a GEE-based ZINB model and associated inference procedures for analyzing count data that are correlated and zero-inflated. Utility of this model over existing models and software has been demonstrated in analyzing a dental dataset relating caries experience with clinical covariates. For the dataset extracted from the Iowa Fluoride Study, we found that dental examination age and daily fluoride intake were significant predictors of caries experience at age five. These findings may be limited to the population under the Iowa Fluoride Study and not generalizable to the entire US population. However, the underlying statistical methodology could be useful in analyzing other clustered count datasets encountered in different disciplines of biomedical and other scientific research. The associated R code (available under supplementary materials) will provide a useful data analysis tool to practicing statisticians and data analysts.

Another important purpose of this paper is to caution potential users about the danger of using incorrect variance formulas in making statistical inference. While sandwich variance

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

formulas related to estimating equations are regarded as robust against potential model misspecification, they are not without pitfalls. In case of complex model fitting procedures, a naïve sandwich variance formula could underestimate the true variance by failing to properly account for all sources of estimation errors. This, in turn, could inflate the power of test of significance of an effect, leading to false positive results. When in doubt, non-parametric procedures such as the bootstrap should be used. Of course, for developing a proper resampling procedure, certain model structures should be respected. In particular for clustered data, resampling entire clusters of observation is warranted, rather than subject-level resampling.

As stated earlier, we have compared the GEE.ZINB method developed here with the independent ZINB where an existing R-package was used to obtain the necessary inference for the later method. In our simulation study, we noticed that this method of estimation encountered convergence problems more often as compared to the GEE.ZINB method. In terms of the variance estimates, it may be possible to create a robust variance estimate for the ZINB likelihood scores that is valid under a clustered data setting. These issues may be investigated elsewhere.

The GEE.ZINB is a marginal model and is suitable for population level inferences. The multilevel zero-inflated NB and Poisson models (Lim et al., 2013; Moghimbeigi et al., 2008; Yau et al., 2003) have been proposed to analyze clustered count data with extra zeros. However, the regression coefficients in the multilevel models may not be interpreted as population level effects (Fitzmaurice et al., 2011). The comparisons between the multilevel ZINB models and GEE based ZINB may deserve further investigation. In addition, Bayesian zero-inflated Poisson (ZIP) model (Xia et al., 2014) has been proposed recently, and the estimation and influence diagnostics for ZINB (Garay et al., 2011) and Bayesian ZIP (Xia et al., 2014) have been studied. How to extend the estimation and influence diagnostics to GEE based ZINB model may be investigated.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported by NIH grants 1R03DE020839-01A1, 5R03DE020839-02, 1R03DE022538-01, and 5R03DE022538-02. We thank Barbara Broffitt for helpful discussions. We thank the two referees for a number of useful comments leading to an improved manuscript.

References

Albert P, Zeger SL, Liang KY. Models for longitudinal data: a generalized estimating. Biometrics. 1988; 44:1049–1060. [PubMed: 3233245]

Davison, AC.; Hinkley, DV. Bootstrap Methods and Their Application. Cambridge; New York: 1997.

Dobble MJ, Welsh AH. Modeling correlated zero-inflated count data. Australian and New Zealand Journal of Statistics. 2001; 43:431–444.

Efron B. Bootstrap methods: Another look at the jackknife. Annals of Statistics. 1979; 7:1–26.

Fitzmaurice, GM.; Laird, NM.; Ware, JH. Applied Longitudinal Analysis. 2. Wiley; New York: 2011.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Garay AM, Ortega EM, Lachos VH. On estimation and influence diagnostics for zero-inflated negative binomial regression models. Computational Statistics and Data Analysis. 2011; 55:1304–1318.

Hall DB, Zhang Z. Marginal models for zero inflated clustered data. Statistical Modelling. 2004; 4:161–180.

Hedeker, D.; Gibbons, RD. Longitudinal Data Analysis. Wiley; New York: 2006.

Levy SM, Warren JJ, Davis CS, Kirchner HL, Kanellis MJ, Wefel JS. Patterns of fluoride intake from birth to 36 months. Journal of Public Health Dentistry. 2001; 61:70–77. [PubMed: 11474917]

Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika. 1986; 73:13–22.

Lim HK, Song J, Jung BC. Score tests for zero-inflation and overdispersion in two-level count data. Computational Statistics and Data Analysis. 2013; 61:67–82.

Moghimbeigi A, Eshraghian MR, Mohammad K, Mcardle B. Multilevel zero-inflated negative binomial regression modeling for over-dispersed count data with extra zeros. Journal of Applied Statistics. 2008; 35:1193–1202.

Mwalili SM, Lesaffre E, Belgium L, Declerck D. The zero-inflated negative binomial regression model with correction for misclassification: an example in caries research. Statistical Methods in Medical Research. 2008; 17:123–139. [PubMed: 17698937]

Rosen O, Jiang W, Tanner MA. Mixtures of marginal models. Biometrika. 2000; 87:391–404.

Satten GA, Datta S. The S-U algorithm for missing data problems. Computational Statistics. 2000; 15:243–277.

Wan, T.; Hua, H.; Xin, T. Applied Categorical and Count Data Analysis. Chapman and Hall/CRC; New York: 2012.

Xie FC, Lin JG, Wei BC. Bayesian zero-inflated generalized Poisson regression model: estimation and case influence diagnostics. Journal of Applied Statistics. 2014; 41(6):1383–1392.

Yau KKW, Wang K, Lee AH. Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros. Biometrical Journal. 2003; 45:437–452.

Zeileis A, Kleiber C, Jackman S. Regression models for count data in R. Journal of Statistical Software. 2008; 27:1–25.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

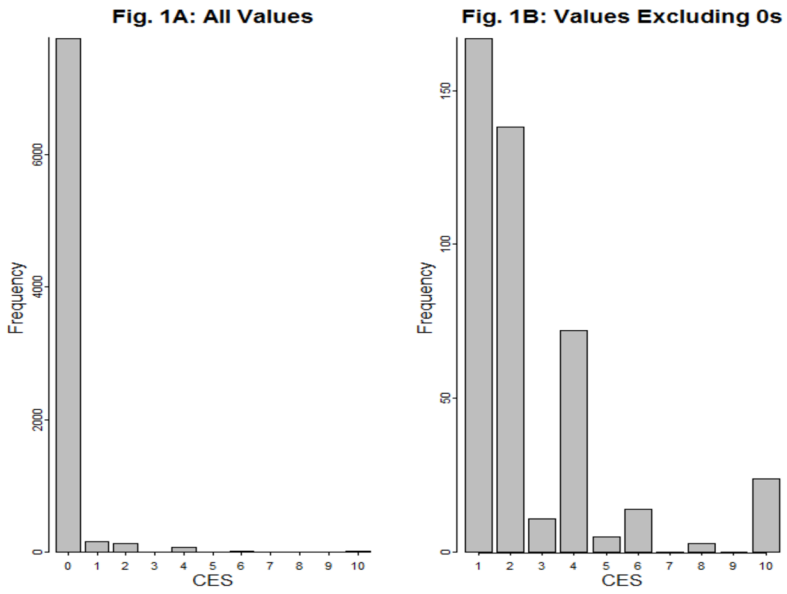

Figure 1.

The marginal histogram of caries experience score (CES) summarized over all teeth and children in our sample (Fig. 1A), and the marginal histogram of CES excluding zero counts summarized over all teeth and children in our sample (Fig. 1B).

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

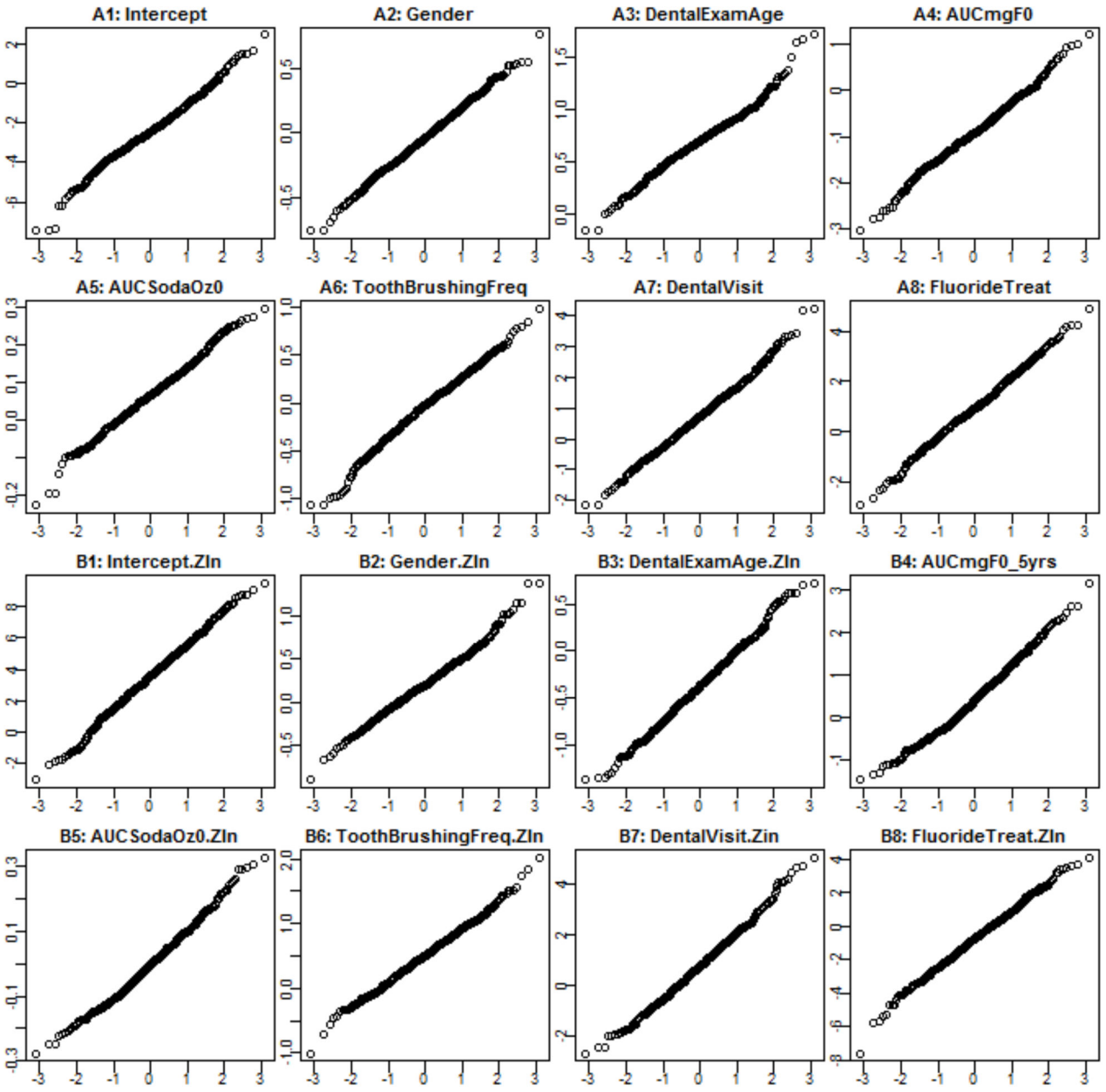

Figure 2.

The Q-Q plots for the estimated parameters for count model (Panels A1–A8) and for the zero-inflated model (Panels B1–B8) in the GEE-based ZINB model, each plot is based on the estimated parameters from 1000 simulated datasets. The x-axis shows the theoretical quantiles of standard normal distribution, and the y-axis is the quantiles of the 1000 estimated parameters.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Table 1

The analysis results for the effects of exposures, treatments and other covariates on caries experience (CE) at age five from the Iowa Fluoride Study dataset based on three different models: the GEE-based ZINB model (see the results under the title "A: GEE.ZINB Model"), the GEE-based zero inflated Poisson model (see the results under the title "B. GEE.ZIP Model"), and the ZINB model (see the results under the title "C: ZINB Model"). Here SW indicates our proposed Sandwich estimate, HZ indicates Hall & Zhang sandwich estimates, BS indicates Bootstrap estimates.

A: GEE.ZINB Model

	GEE.ZINB count component						
	Estimate	**S.E. (HZ)**	**p-value**	**S.E. (SW)**	**p-value**	**S.E. (BS)**	**p-value**
Intercept	−2.713	0.795	0.001	0.923	0.003	1.263	0.032
Gender	−0.018	0.131	0.892	0.160	0.911	0.228	0.938
Dental ExamAge	**0.696**	**0.144**	**<0.001**	**0.166**	**<0.001**	**0.219**	**0.001**
AUCmgF0_5yrs	**−0.856**	**0.197**	**<0.001**	**0.254**	**0.001**	**0.435**	**0.049**
AUCSodaOz0_5yrs	0.072	0.043	0.093	0.054	0.181	0.073	0.327
ToothBrushingFreqPerDayAvg	−0.023	0.119	0.848	0.152	0.881	0.255	0.929
Dental VisitPast6moAvg	0.740	0.487	0.129	0.571	0.195	0.692	0.285
FluorideTreatmentPast6moAvg	0.955	0.620	0.123	0.759	0.208	1.017	0.348
HomeFluorideppmAvg	−0.062	0.104	0.554	0.125	0.622	0.343	0.857
τ (or 1/τ for ZINB)	0.327						
Correlation α_2	0.165						
	GEE.ZINB zero inflation component						
Intercept	**3.599**	**1.346**	**0.007**	**1.613**	**0.026**	**1.813**	**0.047**
Gender	0.201	0.163	0.218	0.219	0.359	0.243	0.408
DentalExamAge	−0.389	0.251	0.121	0.295	0.188	0.344	0.259
AUCmgF0_5yrs	0.392	0.42	0.351	0.639	0.540	0.726	0.589
AUCSodaOz0_5yrs	−0.010	0.049	0.840	0.066	0.879	0.073	0.891
ToothBrushingFreqPerDayAvg	0.503	**0.212**	**0.018**	0.298	0.092	0.329	0.127
Dental VisitPast6moAvg	0.825	0.693	0.234	0.928	0.374	0.923	0.372
FluorideTreatmentPast6moAvg	−0.875	0.866	0.312	1.127	0.438	1.151	0.447
HomeFluorideppmAvg	0.046	0.144	0.747	0.188	0.805	0.318	0.884
Correlation α_1	0.266						

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

	B: GEE.ZIP Model					C: ZINB Model		
	GEE.ZIP count component					ZINB count component		
	Estimate	**S.E. (HZ)**	**p-value**	**S.E. (SW)**	**p-value**	**Estimate**	**S.E.**	**p-value**
Intercept	−2.056	0.714	0.004	0.831	0.013	−3.632	0.886	<0.001
Gender	−0.016	0.123	0.896	0.151	0.916	−0.004	0.149	0.976
Dental ExamAge	**0.590**	**0.130**	**<0.001**	**0.150**	**<0.001**	**0.821**	**0.160**	**<0.001**
AUCmgF0_5yrs	**−0.749**	**0.192**	**<0.001**	**0.249**	**0.003**	**−0.906**	**0.308**	**0.003**
AUCSodaOz0_5yrs	0.063	0.042	0.138	0.053	0.236	0.078	0.060	0.194
ToothBrushingFreqPerDayAvg	−0.008	0.119	0.948	0.150	0.958	−0.058	0.201	0.774
DentalVisitPast6monthAvg	0.696	0.45	0.122	0.528	0.187	0.833	0.525	0.112
FluorideTreatmentPast6moAvg	0.655	0.552	0.236	0.650	0.314	1.234	0.727	0.090
HomeFluorideppmAvg	−0.056	0.102	0.586	0.117	0.634	−0.088	0.176	0.615
Correlation (or 1/τ for ZINB)	0.163					0.928		
	GEE.ZIP zero inflation component					ZINB zero inflation component		
Intercept	**4.093**	**1.494**	**0.006**	**1.628**	**0.012**	**2.714**	**0.778**	**<0.001**
Gender	0.201	0.189	0.288	0.218	0.358	0.210	0.128	0.100
Dental ExamAge	−0.458	0.277	0.098	0.298	0.124	**−0.287**	**0.137**	**0.037**
AUCmgF0_5yrs	0.480	0.499	0.336	0.640	0.453	0.414	0.294	0.159
AUCSodaOz0_5yrs	−0.017	0.057	0.771	0.065	0.800	−0.008	0.046	0.856
ToothBrushingFreqPerDayAvg	**0.502**	**0.246**	**0.041**	0.294	0.087	**0.525**	**0.172**	**0.002**
DentalVisitPast6monthAvg	0.746	0.803	0.353	0.925	0.420	0.891	0.513	0.082
FluorideTreatmentPast6moAvg	−1.006	0.980	0.305	1.113	0.366	−0.676	0.629	0.283
HomeFluorideppmAvg	0.052	0.163	0.749	0.186	0.780	0.011	0.135	0.935
Correlation	0.244							

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

Comput Stat Data Anal. Author manuscript; available in PMC 2016 May 01.

Table 2

Simulation results related to bias, standard error and estimated standard error of estimators in GEE.ZINB and ZINB when the inter-cluster correlation is high, and simulation results for GEE.ZINB component model for counts when the inter-cluster correlation is medium and absent. (SW= Sandwich, HZ = Hall & Zhang, BS = Bootstrap)

	GEE.ZINB Model (High inter-cluster correlation)										**ZINB Model (High inter-cluster correlation)**				
Parameter	**GEE.ZINB count component**										**ZINB count component**				
	(1) Truth	**(2) Mean**	**(3) Bias**	**(4) S.E. (True)**	**(5) S.E. (SW)**	**(6)= (5)/(4) Ratio**	**(7) S.E. (HZ)**	**(8)= (7)/(4) Ratio**	**(9) S.E. (BS)**	**(10)= (9)/(4) Ratio**	**(11) Mean**	**(12) Bias**	**(13) S.E. (True)**	**(14) S.E.**	**(15)= (14)/(13) Ratio**
Intercept	−2.713	−2.442	0.271	1.406	0.986	0.702	0.729	0.509	1.319	0.921	−2.485	0.228	1.403	0.491	0.350
Gender	−0.018	−0.017	0.001	0.235	0.187	0.798	0.142	0.599	0.225	0.949	−0.016	0.002	0.235	0.084	0.357
DentalExamAge	0.696	0.649	−0.047	0.255	0.176	0.691	0.129	0.502	0.238	0.926	0.656	−0.040	0.255	0.089	0.351
AUCmgF0_5yrs	−0.856	−0.845	0.011	0.628	0.462	0.734	0.333	0.586	0.568	1.001	−0.856	0.000	0.631	0.221	0.351
AUCSodaOz0_5yrs	0.072	0.072	0.000	0.084	0.063	0.756	0.045	0.510	0.079	0.895	0.073	0.001	0.084	0.032	0.378
ToothBrushingFreqPerDayAvg	−0.023	−0.056	−0.033	0.316	0.245	0.775	0.183	0.574	0.307	0.962	−0.057	−0.034	0.318	0.114	0.359
DentalVisitPast6moAvg	0.740	0.693	−0.047	1.039	0.826	0.795	0.578	0.568	0.949	0.933	0.695	−0.045	1.045	0.359	0.344
FluorideTreatmentPast6moAvg	0.955	0.928	−0.027	1.210	0.925	0.764	0.692	0.611	1.134	1.002	0.941	−0.014	1.220	0.418	0.343
HomeFluorideppmAvg	−0.062	−0.020	0.042	0.333	0.245	0.735	0.163	0.496	0.318	0.967	−0.024	0.038	0.336	0.114	0.339
τ (or $1/\tau$ for ZINB)	0.327	0.123		0.077									Median: 9.377; IQR (5.717, 19.67)		
	GEE.ZINB zero inflation component										ZINB zero inflation component				
Intercept	3.599	3.737	0.138	1.996	1.811	0.907	1.430	0.764	1.907	1.019	3.713	0.114	2.001	0.632	0.316
Gender	0.201	0.221	0.020	0.305	0.303	0.995	0.222	0.752	0.297	1.009	0.221	0.020	0.305	0.105	0.344
DentalExamAge	−0.389	−0.412	−0.023	0.367	0.333	0.907	0.266	0.792	0.348	1.037	−0.408	−0.019	0.369	0.114	0.309
AUCmgF0_5yrs	0.392	0.428	0.036	0.733	0.693	0.945	0.452	0.629	0.722	1.004	0.423	0.031	0.739	0.253	0.342
AUCSodaOz0_5yrs	−0.010	−0.001	0.009	0.110	0.105	0.957	0.078	0.753	0.106	1.028	−0.001	0.009	0.110	0.032	0.289
ToothBrushingFreqPerDayAvg	0.503	0.520	0.017	0.415	0.402	0.970	0.286	0.714	0.394	0.984	0.519	0.016	0.416	0.138	0.331
DentalVisitPast6moAvg	0.825	0.802	−0.023	1.279	1.293	1.010	0.941	0.732	1.265	0.983	0.804	−0.021	1.281	0.443	0.346
FluorideTreatmentPast6moAvg	−0.875	−0.894	−0.019	1.659	1.607	0.969	1.238	0.777	1.561	0.980	−0.886	−0.011	1.661	0.542	0.326
HomeFluorideppmAvg	0.046	0.122	0.076	0.385	0.338	0.878	0.232	0.648	0.407	1.137	0.095	0.049	0.392	0.126	0.322
	GEE.ZINB zero inflation component										**ZINB zero inflation component**				

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

	GEE.ZINB Model (High inter-cluster correlation)										ZINB Model (High inter-cluster correlation)				
Parameter	**GEE.ZINB count component**										**ZINB count component**				
	(1) Truth	**(2) Mean**	**(3) Bias**	**(4) S.E. (True)**	**(5) S.E. (SW)**	**(6)= (5)/(4) Ratio**	**(7) S.E. (HZ)**	**(8)= (7)/(4) Ratio**	**(9) S.E. (BS)**	**(10)= (9)/(4) Ratio**	**(11) Mean**	**(12) Bias**	**(13) S.E. (True)**	**(14) S.E.**	**(15)= (14)/(13) Ratio**
	(1) Truth	(16) Mean	(17) Bias	(18) S.E. (True)	(19) S.E. (SW)	(20)= (19)/(18) Ratio			(1) Truth	(21) Mean	(22) Bias	(23) S.E. (True)	(24) S.E. (SW)	(25)= (24)/(23) Ratio	
Intercept	−2.713	−2.534	0.179	0.801	0.554	0.691			−2.713	−2.595	0.119	0.517	0.424	0.819	
Gender	−0.018	−0.013	0.005	0.139	0.108	0.775			−0.018	−0.016	0.002	0.098	0.092	0.943	
DentalExamAge	0.696	0.672	−0.024	0.148	0.097	0.654			0.696	0.680	−0.016	0.090	0.072	0.803	
AUCmgF0_5yrs	−0.856	−0.843	0.013	0.365	0.254	0.698			−0.856	−0.836	0.020	0.238	0.226	0.951	
AUCSodaOz0_5yrs	0.072	0.068	−0.004	0.047	0.036	0.769			0.072	0.068	−0.004	0.032	0.029	0.901	
ToothBrushingFreqPerDayAvg	−0.023	−0.040	−0.017	0.196	0.144	0.735			−0.023	−0.028	−0.005	0.133	0.125	0.940	
DentalVisitPast6moAvg	0.740	0.733	−0.007	0.585	0.471	0.805			0.740	0.712	−0.028	0.430	0.396	0.922	
FluorideTreatmentPast6moAvg	0.955	0.897	−0.058	0.696	0.547	0.786			0.955	0.946	−0.009	0.483	0.446	0.925	
HomeFluorideppmAvg	−0.062	−0.057	0.005	0.181	0.125	0.692			−0.062	−0.073	−0.011	0.114	0.097	0.853	
τ	0.327	0.177			0.012				0.327	0.244			0.028		