A Statistical Inference of Regression Model for Longitudinal Counting Data with Zero Expansion Covariates

Jingrui Mu[1]

Supervisor: Ye Peng[1]

[1] School of Statistics, University of International Business and Economics, Beijing 100029, China

Abstract

Counting data is an important class of data, which often appears in medicine, biological, and finance areas. When there are excessive zeros in counting data, it's possible that there is a zero-inflated phenomenon in the count data. And zero-inflated count models such as zero-inflated Poisson (ZIP) and zero-inflated Negative Binomial (ZINB) can be applied to address this issue when such zero-inflated counting data are used as response variable. 2018, Peng Y proposed a GEE-type approach to untangle structural and random zeros in predictors for cross-section data. This paper extended this GEE-type approach to zero-inflation count longitudinal data. In this paper, a generalized estimating equation (GEE)-type mixture model is proposed to jointly model the response of interest and thezero-inflated predictors. The estimates of parameters have been acquired by constructing estimating equations. In simulation study, this paper used Monte Carlo method to analyze three different classes of response data and got the results of estimators to verify the reliability of this method.

Keywords: Generalized Estimating Equation (GEE); zero-inflated counting data; structural zeros; random zeros

1 A GEE-type Mixture Model

Suppose that there is a sample of n independent subjects and there are l assessment times, and let y_{it} denote the response of interest and x_{it} denote a zero-inflated count predictor from time t for the i th subject $(i = 1, \ldots, n)$ and $\mathbf{y}_i = (y_{i1}, \ldots, z_{il})^\top$ and $\mathbf{x}_i = (x_{i1}, \ldots, x_{il})^\top$ be the outcome and the explanatory variables across the time points for the i th subject $(i = 1, \ldots, n)$. In practice, the structural zeros in x_{it} usually measure some personal trait and the random zeros and positive counts assess levels of activities of some behavior of interest such as alcohol drinking. Since the trait in many applications is often a risk factor, we refer to the group of subjects with structural zeros as the non-risk subgroup, and the others as the at-risk subgroup. In addition, we assume that there is a p-dimensional vector of covariates to be adjusted for and denoted by $\mathbf{z}_{it} = (z_{it1}, \ldots, z_{itp})^\top$ and $\mathbf{z}_i = (z_{i1}, \ldots, z_{il})^\top$.

If we do not distinguish the structural zeros from random zeros, we may apply a generalized linear model (GLM) to model the association between the response y_{it} and the zero-inflated predictor x_{it}, controlling for covariates $\mathbf{z}_{it}$, as follows:

$$y_{it} \mid x_{it}, \mathbf{z}_{it} \sim \text{i.d. } f, \quad E\left(y_{it} \mid x_{it}, \mathbf{z}_{it}\right) = g(\alpha_1 x_{it} + \mathbf{z}_{it}^\top \beta) \tag{1.1}$$

where i.d. denotes independently distributed, f denotes some distribution functions such as a Poisson distribution, g is a known link function such as a log function, and α_1 and β are the regression parameters. One may include a constant value 1 in $\mathbf{z}_{it}$ so that the intercept term is included in β as well. For example, if y_{it} is count data from a Poisson distribution, one can choose an exponential function for g. Then (1.1) becomes

$$y_{it} \mid x_{it}, \mathbf{z}_{it} \sim \text{i.d. Poisson}\left(\mu_{it}\right), \quad \mu_{it} = E\left(y_{it} \mid x_{it}, \mathbf{z}_{it}\right) = \exp(\alpha_1 x_{it} + \mathbf{z}_{it}^{\top}\beta), \quad i = 1, \ldots, n. \tag{1.2}$$

However, when a count predictor x_{it} has structural zeros, the conceptual difference between structural and random zeros carries quite a significant implication for the interpretation of the coefficient α_1 in (1.1) and (1.2). Let $r_{it} = 1$ if x_{it} is a structural zero and $r_i = 0$ otherwise. The indicator r_i partitions the study population into two distinctive subgroups, with one consisting of all subjects corresponding to $r_{it} = 1$ and the other comprising of the remaining subjects with $r_{it} = 0$. For example, if x_{it} is an alcohol drinking count variable such as days of alcohol drinking, the difference between a subject with $r_{it} = 1$ and $r_{it} = 0$ is substantial as the former represents subjects who are abstinent of alcohol drinking and the latter represents subjects who are at-risk for alcohol drinking. The two subgroups of subjects may have very different relationship with the outcome. In (1.1), if $x_{it} = 0$ is a random zero, the coefficient α_1 of x_{it} represents the effect of drinking on the response y_{it} within the drinker subgroup when the drinking outcome changes from 0 to 1; while if $x_{it} = 0$ represents a structural zero, such a difference speaks to the effect of the trait of drinking on the response y_{it}. When x_{it} is included in the model as in (1.1), the coefficient of x_{it} has a dubious interpretation. Thus (1.1) is flawed and must be revised to tease out such conceptually distinctive effects of structural and random zeros.

1.1 A GEE-type Mixture Model

Let r_{it} be an indicator of structural zeros, with value 1 for a structural zero and 0 otherwise. Similar to the likelihood-based mixture model, the GEE-type mixture model consists of two models, a main model for the outcome variable, and an auxiliary model for the zero-inflated count predictor.

Main Model: Based on GLM framework, the main model is constructed to model the conditional mean of the outcome given the zero-inflated predictor and covariates.

$$E\left(y_{it} \mid x_{it}, r_{it}, \mathbf{z}_{it}\right) = g(\alpha_1 x_{it} + \alpha_2 r_{it} + \mathbf{z}_{it}^{\top}\beta), \quad i = 1, \ldots, n. \tag{1.3}$$

Auxiliary Zero-inflated Model: For the zero-inflated predictor x_{it}, we need to model both the probability of structural zeros ρ_{it} and the count mean μ_{it} by logit and loglinear regression models. Let $\mathbf{w}_{it}$ be a set of predictors for both ρ_{it} and μ_{it}. The auxiliary zero-inflated model is given by

$$\operatorname{logit}(\Pr(r_{it} = 1 \mid \mathbf{w}_{it})) = \operatorname{logit}(\rho_{it}) = \mathbf{w}_{it}^{T}\gamma_1, \quad \log(E(x_{it} \mid r_{it} = 0, \mathbf{w}_{it})) = \log(\mu_{it}) = \mathbf{w}_{it}^{T}\gamma_2, \quad i = 1, \ldots, n. \tag{1.4}$$

To estimate the parameters in the main model (1.3), Let $\alpha = (\alpha_1, \alpha_2, \beta)^T$.

Define

$$S_{1it}^{\alpha} = I(x_{it} = 0)[y_{it} - E(y_{it}|x_{it} = 0, \mathbf{z}_{it}, \mathbf{w}_{it})],$$
$$S_{2it}^{\alpha} = I(x_{it} > 0)[y_{it} - E(y_{it}|x_{it} > 0, \mathbf{z}_{it}, \mathbf{w}_{it})],$$

and $S_{it}^{\alpha} = (S_{1it}^{\alpha}, S_{2it}^{\alpha})$.

Let $\delta_{it} = \Pr(r_{it} = 1 \mid x_{it} = 0, \mathbf{w}_{it})$, based on the main model, and replace r_{it}s in S_{it}^{α} with their conditional mean, yielding

$$\begin{aligned} S_{1it}^{\alpha} &= I\left(x_{it} = 0\right)\left[y_{it} - (1 - \delta_{it})g(\mathbf{z}_{it}^{T}\beta) - \delta_{it}g(\alpha_2 + \mathbf{z}_{it}^{T}\beta)\right], \\ S_{2it}^{\alpha} &= I\left(x_{it} > 0\right)\left[y_{it} - g(\alpha_1 x_{it} + \mathbf{z}_{it}^{T}\beta)\right]. \end{aligned} \tag{1.5}$$

The following estimating equation

$$\mathbf{Q}_n^\alpha(\alpha) = \sum_{i=1}^{n} D_i^\alpha (V_i^\alpha)^{-1} S_i^\alpha = \mathbf{0}\mathbf{S}_i^\alpha = (S_{i1}^\alpha, \ldots, S_{il}^\alpha)^\top \mathbf{D}_i^\alpha = (D_{i1}^\alpha, \ldots, D_{il}^\alpha)^\top \tag{1.6}$$

can be used to estimate α, where $D_{it}^\alpha = \frac{\partial S_{it}^\alpha}{\partial \alpha}$ and $V_{it}^\alpha = Var\left(S_{it}^\alpha | \mathbf{z}_{it}, \mathbf{w}_{it}\right)$.

If δ_{it} is known or can be easily estimated given that all r_{it} are observed, the α can be estimated based on the equation (1.6). However, in most cases, r_{it} for some subjects are not observed or are unobservable, and equation (1.6) does not provide enough information to estimate α. Hence we need the auxiliary model to provide us additional information to estimate α.

Since $\delta_{it} = \Pr(r_{it} = 1 \mid x_{it} = 0, \mathbf{w}_{it}) = \Pr(r_{it} = 1, x_{it} = 0 \mid \mathbf{w}_{it}) / \Pr(x_{it} = 0 \mid \mathbf{w}_{it}) = \rho_{it}/(\rho_{it} + \Pr(x_{it} = 0, r_{it} = 0 \mid \mathbf{w}_{it}))$ involves γ_1 and γ_2 in (1.4), we need to construct estimating equations to estimate γ_1 and γ_2. Since the conditional probability of random zeros $\Pr(x_{it} = 0, r_{it} = 0 \mid \mathbf{w}_{it})$ cannot be uniquely identified based on the models in (1.4), we further assume that the zero-inflated count predictor x_{it} follows a zero-inflated Poisson distribution, i.e., $x_{it} \mid \mathbf{w}_{it} \sim$ i.d. ZIP$(\rho_{it},\ \mu_{it})$. Under the assumption, δ_{it} can be expressed as $\frac{\rho_{it}}{\rho_{it} + (1-\rho_{it})\exp(-\mu_{it})}$ and can be identified by the models in (1.4). Let $\gamma = (\gamma_1, \gamma_2)^T$, and define

$$\begin{aligned} S_{1it}^\gamma &= I(x_{it} = 0) - E[I(x_{it} = 0 | \mathbf{w}_{it})], \\ S_{2it}^\gamma &= I\left(x_{it} > 0\right)\left[x_{it} - E(x_{it} | x_{it} > 0, \mathbf{w}_{it})\right]. \end{aligned}$$

Under the models in (1.4), we further get

$$
\begin{aligned}
S_{1it}^{\gamma} &= I(x_{it}=0) - \rho_{it} - (1-\rho_{it})\exp(-\mu_{it}), \\
S_{2it}^{\gamma} &= I\left(x_{it}>0\right)\left[x_{it} - \mu_{it}/(1-\exp(-\mu_{it}))\right],
\end{aligned} \tag{1.7}
$$

and $S_{it}^{\gamma} = (S_{1it}^{\gamma}, S_{2it}^{\alpha})$. The following estimating equation

$$
\mathbf{Q}_n^{\gamma}\left(\gamma\right) = \sum_{i=1}^{n} D_i^{\gamma}(V_i^{\gamma})^{-1}S_i^{\gamma} = \mathbf{0}\mathbf{S}_i^{\gamma} = (S_{i1}^{\gamma},\ldots,S_{il}^{\gamma})^{\top}\,\mathbf{D}_i^{\gamma} = (D_{i1}^{\gamma},\ldots,D_{il}^{\gamma})^{\top} \tag{1.8}
$$

can be used to estimate γ, where $D_{it}^{\gamma} = \frac{\partial S_{it}^{\gamma}}{\partial \gamma}$ and $V_{it}^{\gamma} = Var\left(S_{it}^{\gamma}|\mathbf{w}_{it}\right)$.

Please note that although we assume a zero-inflated Poisson model for x_{it} to define $S_{1it}^{\gamma}, S_{2it}^{\gamma}$ in (1.8), we do not use all the information about the distribution, but only the mean. In this sense, our method does not rely on a full specification of the distribution for the auxiliary model. Given that we do not assume any distribution for the main model, and that for the auxiliary model, we only use the information about the mean of the distribution, the proposed method is a GEE-type approach, and it is expected to be more robust than the likelihood-based model.

To increase the efficiency of the estimate of α, we estimate α and γ simultaneously by constructing estimating equations for both α and γ. Let $S_i = (S_i^{\alpha}, S_i^{\gamma})$, $\theta = (\alpha, \gamma)$. and $\mathbf{S}_i\alpha = (S_{i1}\alpha, \ldots, S_{il}\alpha)^{\top}$ and $\mathbf{S}_i^{\gamma} = (S_{i1}^{\gamma}, \ldots, S_{il}^{\gamma})^{\top}$. We define the following generalized estimating equation to estimate θ:

$$
\mathbf{U}_n\left(\theta\right) = \sum_{i=1}^{n} D_i V_i^{-1} S_i = \mathbf{0}, \mathbf{S}_i = (S_{i1},\ldots,S_{il})^{\top}\,\mathbf{D}_i = (D_{i1},\ldots,D_{il})^{\top} \tag{1.9}
$$

where $D_{it} = \frac{\partial S_{it}}{\partial \theta}$

Based on (1.5) and (1.7), we have

$$D_{it} = \begin{pmatrix} 0 & \frac{\partial S^{\alpha}_{2it}}{\partial \alpha_1} & 0 & 0 \\ \frac{\partial S^{\alpha}_{1it}}{\partial \alpha_2} & 0 & 0 & 0 \\ \frac{\partial S^{\alpha}_{1it}}{\partial \beta} & \frac{\partial S^{\alpha}_{2it}}{\partial \beta} & 0 & 0 \\ \frac{\partial S^{\alpha}_{1it}}{\partial \gamma_1} & 0 & \frac{\partial S^{\gamma}_{1it}}{\partial \gamma_1} & \frac{\partial S^{\gamma}_{2it}}{\partial \gamma_1} \\ \frac{\partial S^{\alpha}_{1it}}{\partial \gamma_2} & 0 & \frac{\partial S^{\gamma}_{1it}}{\partial \gamma_2} & \frac{\partial S^{\gamma}_{2it}}{\partial \gamma_2} \end{pmatrix}$$

Next, we calculate the variance matrix V_i.

$$V_i = A_i * R(\alpha) * A_i, A_i = diag_t(A_{it})$$

where

$$A_{it} = \begin{pmatrix} Var\left(S^{\alpha}_{1it}\right) & Cov\left(S^{\alpha}_{1it}, S^{\alpha}_{2it}\right) & Cov\left(S^{\alpha}_{1it}, S^{\gamma}_{1it}\right) & Cov\left(S^{\alpha}_{1it}, S^{\gamma}_{2it}\right) \\ Cov\left(S^{\alpha}_{1it}, S^{\alpha}_{2it}\right) & Var\left(S^{\alpha}_{2it}\right) & Cov\left(S^{\alpha}_{2it}, S^{\gamma}_{1it}\right) & Cov\left(S^{\alpha}_{2it}, S^{\gamma}_{2it}\right) \\ Cov\left(S^{\alpha}_{1it}, S^{\gamma}_{1it}\right) & Cov\left(S^{\alpha}_{2it}, S^{\gamma}_{1it}\right) & Var\left(S^{\gamma}_{1it}\right) & Cov\left(S^{\gamma}_{1it}, S^{\gamma}_{2it}\right) \\ Cov\left(S^{\alpha}_{1it}, S^{\gamma}_{2it}\right) & Cov\left(S^{\alpha}_{2it}, S^{\gamma}_{2it}\right) & Cov\left(S^{\gamma}_{1it}, S^{\gamma}_{2it}\right) & Var\left(S^{\gamma}_{2it}\right) \end{pmatrix},$$

where

$$\begin{aligned} &Var\left(S^{\alpha}_{1i}\right) = \Pr(x_i = 0|\mathbf{w}_i) Var\left(y_i|x_i = 0, \mathbf{z}_i, \mathbf{w}_i\right), \\ &Var\left(S^{\alpha}_{2i}\right) = \Pr(x_i > 0|\mathbf{w}_i) Var\left(y_i|x_i > 0, \mathbf{z}_i, \mathbf{w}_i\right), \\ &Var\left(S^{\gamma}_{1i}\right) = \left(\rho_i + (1-\rho_i)\exp(-\mu_i)\right)\left((1-\rho_i)(1-\exp(-\mu_i))\right), \\ &Var\left(S^{\gamma}_{2i}\right) = (1-\rho_i)(1-\exp(-\mu_i))\left[\frac{\mu_i(1+\mu_i)}{1-\exp(-\mu_i)} - \left(\frac{\mu_i}{1-\exp(-\mu_i)}\right)^2\right], \\ &Cov\left(S^{\alpha}_{1i}, S^{\alpha}_{2i}\right) = 0,\ Cov\left(S^{\alpha}_{1i}, S^{\gamma}_{2i}\right) = 0,\ Cov\left(S^{\alpha}_{1i}, S^{\gamma}_{1i}\right) = E\left(S^{\alpha}_{1i} S^{\gamma}_{1i}\right), \\ &Cov\left(S^{\gamma}_{1i}, S^{\gamma}_{2i}\right) = 0,\ Cov\left(S^{\alpha}_{2i}, S^{\gamma}_{1i}\right) = 0,\ Cov\left(S^{\alpha}_{2i}, S^{\gamma}_{2i}\right) = E\left(S^{\alpha}_{2i} S^{\gamma}_{2i}\right). \end{aligned}$$

For calculating the variance V_i, this paper assumed these two assumptions A and B. **Assumption A:** Conditional Independence. Given $\mathbf{w}_{it}$, x_{it} and r_{it} are independent of $\mathbf{z}_{it}$, i.e.,

$$(x_{it}, r_{it}) \perp \mathbf{z}_{it} \mid \mathbf{w}_{it}.$$

This assumption implies that x_{it} and r_{it} may be associated with $\mathbf{z}_{it}$, but the association is only through $\mathbf{w}_{it}$. This condition can be easily satisfied by including additional predictors from $\mathbf{z}_{it}$ in (**??**)**,** as needed for the conditional independence, into $\mathbf{w}_{it}$ in (**??**).

Assumption B: Comprehensiveness of the main model. Given the predictors $x_{it}, \mathbf{z}_{it}, r_{it}$, the response y_{it} is independent of $\mathbf{w}_{it}$**,** i.e.,

$$y_{it} \perp \mathbf{w}_{it} \mid x_{it}, \mathbf{z}_{it}, r_{it},$$

which implies that y_{it} may depend on $\mathbf{w}_{it}$, but the dependence is only through $x_{it}, \mathbf{z}_{it}$ and r_{it}. Above these assumptions, we can calculate V_i as follows:

$$Cov\left(S_{1i}^{\alpha}\,, S_{2i}^{\alpha}\right) = 0,\ Cov\left(S_{1i}^{\alpha}, S_{2i}^{\gamma}\right) = 0,\ Cov\left(S_{1i}^{\alpha}, S_{1i}^{\gamma}\right) = E\left(S_{1i}^{\alpha} S_{1i}^{\gamma}\right),$$
$$Cov\left(S_{1i}^{\gamma}\,, S_{2i}^{\gamma}\right) = 0,\ Cov\left(S_{2i}^{\alpha}, S_{1i}^{\gamma}\right) = 0,\ Cov\left(S_{2i}^{\alpha}, S_{2i}^{\gamma}\right) = E\left(S_{2i}^{\alpha} S_{2i}^{\gamma}\right).$$

Let $\hat{\theta} = (\hat{\alpha}, \hat{\gamma})$ be the estimator of $\theta = (\alpha, \gamma)$ by solving the generalized estimating equation (1.9). Since there is no closed-form for the estimates, numeric solutions can be obtained easily through the popular Newton-Raphson (NR) method. Under (1.5) and (1.7), $\hat{\theta}$ is consistent and asymptotically normal.

Asymptotic Results:

a). The GEE estimator $\hat{\theta}$ *is consistent and* $\sqrt{n}\left(\hat{\theta} - \theta\right)$ *is asymptotically normally distributed with mean zero and covariance matrix:*

$$\Sigma_\theta = A^{-1} E\left[(D_i V_i^{-1} S_i)(D_i V_i^{-1} S_i)^T\right] A^{-T}, \quad A = E(D_i V_i^{-1} D_i^T).$$

A consistent estimator of Σ_θ *is given by* $\widehat{\Sigma}_\theta = \widehat{A}^{-1}\widehat{A}_0\widehat{A}^{-T}$, *where*

$$\widehat{A}_0 = \frac{1}{n}\sum_{i=1}^{n}(\widehat{D}_i\widehat{V}_i^{-1}\widehat{S}_i)(\widehat{D}_i\widehat{V}_i^{-1}\widehat{S}_i)^T, \quad \widehat{A} = \frac{1}{n}\sum_{i=1}^{n}\widehat{D}_i\widehat{V}_i^{-1}\widehat{D}_i^T,$$

b). The estimator $\hat{\alpha}$ *of* α *for the main model (1.3) is consistent and* $\sqrt{n}\,(\hat{\alpha}-\alpha)$ *is asymptotically normally distributed with mean zero and covariance matrix* $\Sigma_\alpha = B^{-1}\left[\Psi+\Phi\right]B^{-T}$, *where*

$$\begin{aligned}
&\Psi = E[(D_i^\alpha (V_i^\alpha)^{-1} S_i^\alpha)(D_i^\alpha (V_i^\alpha)^{-1} S_i^\alpha)^T], \qquad B = E[D_i^\alpha (V_i^\alpha)^{-1}(D_i^\alpha)^T],\\
&\Phi = CH^{-1}E[(D_i^\gamma (V_i^\gamma)^{-1} S_i^\gamma)(D_i^\gamma (V_i^\gamma)^{-1} S_i^\gamma)^T]H^{-T}C^T - G - G^T,\\
&G = E[D_i^\alpha (V_i^\alpha)^{-1} S_i^\alpha (D_i^\gamma (V_i^\gamma)^{-1} S_i^\gamma)^T H^{-T} C^T],\\
&C = E\left[\frac{\partial}{\partial\gamma^T}\left(D_i^\alpha (V_i^\alpha)^{-1} S_i^\alpha\right)\right], \qquad H = E\left[\frac{\partial}{\partial\gamma^T}(D_i^\gamma (V_i^\gamma)^{-1} S_i^\gamma)\right].
\end{aligned}$$

A consistent estimate of the asymptotic variance Σ_α *can be obtained by substituting consistent estimates of the respective quantities, i.e.,*

$$\begin{aligned}
&\widehat{\Psi} = \frac{1}{n}\sum_{i=1}^{n}(\widehat{D}_i^\alpha (\widehat{V}_i^\alpha)^{-1}\widehat{S}_i^\alpha)(\widehat{D}_i^\alpha (\widehat{V}_i^\alpha)^{-1}\widehat{S}_i^\alpha)^T, \qquad \widehat{B} = \frac{1}{n}\sum_{i=1}^{n}\widehat{D}_i^\alpha (\widehat{V}_i^\alpha)^{-1}(\widehat{D}_i^\alpha)^T,\\
&\widehat{\Phi} = \widehat{C}\widehat{H}^{-1}\frac{1}{n}\sum_{i=1}^{n}[(\widehat{D}_i^\gamma (\widehat{V}_i^\gamma)^{-1}\widehat{S}_i^\gamma)(\widehat{D}_i^\gamma (\widehat{V}_i^\gamma)^{-1}\widehat{S}_i^\gamma)^T]\widehat{H}^{-T}\widehat{C}^T - \widehat{G} - \widehat{G}^T,\\
&\widehat{G} = \frac{1}{n}\sum_{i=1}^{n}[\widehat{D}_i^\alpha (\widehat{V}_i^\alpha)^{-1}\widehat{S}_i^\alpha (\widehat{D}_i^\gamma (\widehat{V}_i^\gamma)^{-1}\widehat{S}_i^\gamma)^T \widehat{H}^{-T}\widehat{C}^T],\\
&\widehat{C} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\gamma^T}\left(D_i^\alpha (V_i^\alpha)^{-1} S_i^\alpha\right)|\gamma=\hat{\gamma},\alpha=\hat{\alpha}, \qquad \widehat{H} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\gamma^T}(D_i^\gamma (V_i^\gamma)^{-1} S_i^\gamma)|\gamma=\hat{\gamma}.
\end{aligned}$$

The $\widehat{V}_i^{\alpha}$ and $\widehat{D}_i^{\alpha}$ are estimators of V_i^{α} and D_i^{α} with $\hat{\alpha}$ and $\hat{\gamma}$ substituting in place of α and γ, respectively. The asymptotic variance Σ_{α} includes an additional term to account for the variability from estimating γ.

2 Simulation Studies

Simulation studies were conducted to evaluate the performance of the proposed GEE-type mixture model, as well as to compare the performance of the new method with that of the likelihood-based counterpart [**?**]. In the simulation studies, three scenarios are considered, both models for the outcome and the zero-inflated count predictor are correctly specified, only the model for the zero-inflated count predictor is misspecified, and only the model for the outcome is misspecified. For each scenario, four types of outcomes are evaluated: continuous, binary, Poisson and zero-inflated Poisson. The outcomes, the zero-inflated count predictor as well as the covariates are generated based on the following models.

Zero-inflated count predictor X: for all the simulations, the zero-inflated predictor x_{it}, as well as the associated indicator for the structural zero r_{it}, is generated from the following ZIP model:

$$
\begin{aligned}
x_{it}|w_{it} &\sim \mathrm{ZIP}(\rho_{it}, \mu_{it}), \quad w_{it} \sim \mathrm{Uniform}\,(0,1)\,, \\
\mathrm{logit}(\rho_{it}) &= \gamma_{10} + \gamma_{11} w_{it}, \quad \log(\mu_i) = \gamma_{20} + \gamma_{21} w_{it},
\end{aligned}
\tag{2.10}
$$

Different proportions of structural zeros can be obtained by varying γ_{10} and γ_{11}, and the Poisson mean is determined by γ_{20} and γ_{21}. In our simulations, we set $\gamma_{10} = -0.9, \gamma_{11} = 0$, $\gamma_{20} = 1$, and $\gamma_{21} = -0.5$. In this case, the proportion of structural zeros in x is around 30%.

Continuous response Y: we define $\mathbf{z}_{it} = (1, w_{it})$ as the covariate vector. Then continuous outcome y_i is generated through

$$y_{it} = \alpha_1 x_{it} + \alpha_2 r_{it} + \mathbf{z}_{it}^T \beta + e_{it}, \tag{2.11}$$

where $e_{it} \sim N(0, 1)$.

Binary response Y: we simulate a binary response y_{it} based on the following GLM with logit link function:

$$y_{it} \mid x_{it}, r_{it}, \mathbf{z}_{it} \sim \text{Bernoulli}\left(p_{it}\right), \quad \text{logit}\left(p_{it}\right) = \alpha_1 x_{it} + \alpha_2 r_{it} + \mathbf{z}_{it}^T \beta. \tag{2.12}$$

Poisson response Y: the Poisson response y_{it} is generated through the following GLM with a log link function:

$$y_{it} \mid x_{it}, r_{it}, \mathbf{z}_{it} \sim \text{Poisson}\left(\mu_{it}\right), \quad \log(\mu_{it}) = \alpha_1 x_{it} + \alpha_2 r_{it} + \mathbf{z}_{it}^T \beta. \tag{2.13}$$

For different responses Y, I set all initial value α as $\alpha_1 = 0.2$, $\alpha_2 = 0.5$, $\beta = (-1, 1)^T$ and $l = 2$. I all did 500 simulations for three different sample sizes, 300, 500 and 1000, the following tables are the result for different sample sizes, where ASE represents the mean of estimated asymptotic standard deviation and CP represents the proportion of 95%. Besides, I assume the correlation matrix is identity matrix. From the following tables, we can see that the estimators are all unbiased estimators based on GEE-type mixture model.

Table 1: The estimated α values for sample size=300

Y	parameter	Bias	ESD	ASE	CP
Continuous	α_1	0.001	0.05	0.05	0.96
	α_2	0.003	0.17	0.17	0.95
	β_0	-0.008	0.17	0.17	0.95
	β_1	0.014	0.18	0.17	0.94
Binary	α_1	0.012	0.14	0.14	0.97
	α_2	0.016	0.13	0.13	0.97
	β_0	-0.033	0.13	0.13	0.96
	β_1	0.023	0.17	0.17	0.94
Poisson	α_1	-0.006	0.04	0.04	0.93
	α_2	-0.028	0.20	0.19	0.94
	β_0	0.023	0.20	0.20	0.94
	β_1	-0.023	0.20	0.20	0.94

Table 2: The estimated λ values for sample size=300

Y	parameter	Bias	ESD	ASE	CP
Continuous	λ_{10}	0.014	0.25	0.25	0.95
	λ_{11}	-0.004	0.30	0.30	0.96
	λ_{20}	-0.006	0.09	0.09	0.95
	λ_{21}	0.007	0.17	0.17	0.94
Binary	λ_{10}	-0.004	0.30	0.30	0.96
	λ_{11}	-0.025	0.30	0.32	0.97
	λ_{20}	-0.009	0.10	0.10	0.95
	λ_{21}	0.006	0.19	0.19	0.94
Poisson	λ_{10}	-0.002	0.25	0.25	0.95
	λ_{11}	0.024	0.30	0.30	0.95
	λ_{20}	-0.015	0.09	0.09	0.95
	λ_{21}	0.016	0.18	0.17	0.94

Table 3: The estimated α values for sample size=500

Y	parameter	Bias	ESD	ASE	CP
Continuous	α_1	0.001	0.04	0.04	0.95
	α_2	0.004	0.13	0.13	0.96
	β_0	-0.011	0.12	0.12	0.95
	β_1	0.011	0.13	0.13	0.95
Binary	α_1	-0.005	0.10	0.10	0.94
	α_2	0.001	0.13	0.14	0.95
	β_0	-0.002	0.27	0.27	0.96
	β_1	0.002	0.27	0.28	0.95
Poisson	α_1	-0.002	0.03	0.03	0.94
	α_2	0.003	0.12	0.12	0.95
	β_0	-0.003	0.13	0.13	0.95
	β_1	0.003	0.14	0.13	0.94

Table 4: The estimated λ values for sample size=500

Y	parameter	Bias	ESD	ASE	CP
Continuous	λ_{10}	0.003	0.20	0.20	0.95
	λ_{11}	-0.012	0.37	0.37	0.96
	λ_{20}	0.004	0.07	0.07	0.95
	λ_{21}	-0.010	0.13	0.13	0.94
Binary	λ_{10}	-0.021	0.26	0.26	0.95
	λ_{11}	-0.004	0.30	0.30	0.95
	λ_{20}	-0.003	0.07	0.07	0.94
	λ_{21}	-0.007	0.14	0.15	0.96
Poisson	λ_{10}	-0.013	0.20	0.20	0.95
	λ_{11}	0.014	0.25	0.25	0.96
	λ_{20}	-0.004	0.07	0.06	0.94
	λ_{21}	-0.008	0.13	0.13	0.95

Table 5: The estimated α values for sample size=1000

Y	parameter	Bias	ESD	ASE	CP
Continuous	α_1	-0.001	0.03	0.03	0.96
	α_2	-0.004	0.09	0.09	0.96
	β_0	0.007	0.09	0.09	0.95
	β_1	-0.011	0.09	0.09	0.94
Binary	α_1	-0.004	0.09	0.09	0.95
	α_2	0.002	0.12	0.12	0.95
	β_0	-0.003	0.20	0.20	0.96
	β_1	0.001	0.21	0.21	0.95
Poisson	α_1	-0.001	0.02	0.02	0.95
	α_2	0.002	0.09	0.09	0.96
	β_0	-0.004	0.10	0.10	0.95
	β_1	0.002	0.13	0.13	0.94

Table 6: The estimated λ values for sample size=500

Y	parameter	Bias	ESD	ASE	CP
Continuous	λ_{10}	0.002	0.13	0.14	0.97
	λ_{11}	-0.018	0.25	0.25	0.96
	λ_{20}	-0.003	0.04	0.04	0.95
	λ_{21}	0.001	0.09	0.09	0.95
Binary	λ_{10}	-0.003	0.20	0.20	0.96
	λ_{11}	0.002	0.23	0.23	0.96
	λ_{20}	0.001	0.03	0.04	0.95
	λ_{21}	-0.004	0.13	0.13	0.96
Poisson	λ_{10}	-0.003	0.13	0.13	0.96
	λ_{11}	0.004	0.15	0.15	0.96
	λ_{20}	0.001	0.09	0.09	0.94
	λ_{21}	-0.0003	0.10	0.10	0.95