Published in final edited form as:

Stat Methods Med Res. 2019 December ; 28(12): 3683–3696. doi:10.1177/0962280218812228.

A GEE-type approach to untangle structural and random zeros in predictors

Peng Ye^{1,2}, Wan Tang³, Jiang He², Hua He²

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

²Department of Epidemiology, Tulane University School of Public Health and Tropical Medicine, New Orleans, LA, USA

³Department of Global Biostatistics and Data Science, Tulane University School of Public Health and Tropical Medicine, New Orleans, LA, USA

Abstract

Count outcomes with excessive zeros are common in behavioral and social studies, and zero-inflated count models such as zero-inflated Poisson (ZIP) and zero-inflated Negative Binomial (ZINB) can be applied when such zero-inflated count data are used as response variable. However, when the zero-inflated count data are used as predictors, ignoring the difference of structural and random zeros can result in biased estimates. In this paper, a generalized estimating equation (GEE)-type mixture model is proposed to jointly model the response of interest and the zero-inflated count predictors. Simulation studies show that the proposed method performs well for practical settings and is more robust for model misspecification than the likelihood-based approach. A case study is also provided for illustration.

Keywords

Generalized estimating equations; mixture model; structural zeros; zero-inflated explanatory variables; zero-inflated Poisson

1 Introduction

Zero-inflation due to the existence of structural zeros is common for count data. In such data, there are two types of zeros, structural zeros and random zeros. Structural zeros refer to zero observations from those subjects whose count responses are always zero, in contrast to random or sampling zeros that occur for subjects whose count response can be greater than zero, but are reported as zero due to sampling variability. For example, in HIV-AIDS prevention research, sexual behavior is a risk factor for HIV/AIDS. The number of

Corresponding author: Hua He, Department of Epidemiology, Tulane University School of Public Health and Tropical Medicine 1440 Canal Street, Suite 2014, New Orleans 70112-2703, LA, USA., hhe2@tulane.edu.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Supplemental Material

Supplemental material for this article is available online.

unprotected sexual occurrences is usually collected. The zero counts in the number of unprotected sexual occurrences for subjects with lifetime celibacy or sexual problems can be defined by structural zeros, while zero counts from those sexually active individuals who happen to have no sex during a given time period can be treated as random zeros. In this example, the structural zeros and random zeros represent two distinct groups of subjects with different psychosocial nature, and the former (latter) group is often called the ‘non-risk’ (‘at-risk’) group. The importance of distinguishing structural zeros from random zeros when studying such zero-inflated count data has been well recognized in various disciplines, from biomedical and psychosocial studies to social sciences including sociology, economics, business, and politics.^{1–4}

When zero-inflated count variables are treated as the response, a number of statistical models have been proposed to address the structural zero issue.^{5–14} Among these models, the zero-inflated Poisson (ZIP) model has been widely used to deal with the issue of structural zeros in the zero-inflated count data.^{15–18} When there is an overdispersion issue arising from the at-risk group, the zero-inflated negative-binomial (ZINB) model^{19–22} is typically suggested instead. However, the issue of structural zeros has received little attention when such zero-inflated count data are treated as a predictor/explanatory variable. In many applications, the zero-inflated count predictors are just treated as continuous predictors, with no effort to distinguish structural zeros from their random counterparts. This approach is widely used in practice due to modeling convenience. One can also include the indicator variable for zeros as a predictor to model the difference between subjects with zero counts and positive counts; however, under such an approach, the differences between structural and random zeros are still ignored. Ignoring the differences between structural and random zeros may fail to describe the realistic relationships between zero-inflated count predictors and the response of interest. For example, in a study on alcohol research, it has been shown that ignoring the differences between structural and random zeros by simply using the count variable as a continuous predictor would yield biased inferences.²³ In recent years, several statistical methods have been proposed to address the structural zero issue in zero-inflated count predictors. For example, He et al.²³ teased out the differential effects of structural and random zeros of zero-inflated count predictors by adding an indicator of structural zeros into generalized linear regression models, in which the structural zeros are required to be observed. Tang et al.²⁴ constructed a likelihood-based mixture regression model for a zero-inflated count predictor by jointly modeling the outcome of interest and the zero-inflated count predictor and employed maximum likelihood method (MLE) to untangle the effects of structural zeros from random zeros. However, the MLE approach relies on a parametric distribution assumption on both the response variable and the zero-inflated explanatory variables, and thus may yield biased estimates if either of these distribution assumptions is not satisfied. In this paper, we propose a semi-parametric GEE-type mixture model to simultaneously model the outcome and the zero-inflated count predictor to address the structural zero issue in predictors. The proposed GEE-type approach relaxes the distribution assumptions on the outcome and the zero-inflated explanatory variable, but only specifies a functional form between the outcome and the explanatory variable; hence, it is expected to yield robust estimates for model parameters.

The rest of the article is organized as follows. In Section 2, we first give a brief review of the likelihood-based mixture model, and then propose a GEE-type mixture model to address the structural zero issue in predictors. The asymptotic properties of the GEE-type estimates are also presented in Section 2. In Section 3, simulation studies are conducted to evaluate the performance of the GEE-type estimates and compare the results with the estimates based on the likelihood-based mixture model. A real data example is provided in Section 4, and discussion and concluding remarks are given in Section 5.

2 A GEE-type mixture model

Suppose that there is a sample of n independent subjects, and let y_i denote the response of interest and x_i denote a zero-inflated count predictor for the i th subject ($i = 1, \dots, n$). In practice, the structural zeros in x_i 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 $\mathbf{z}_i = (z_{i1}, \dots, z_{ip})^\top$.

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_i and the zero-inflated predictor x_i , controlling for covariates $\mathbf{z}_i$ as follows

$$y_i | x_i, \mathbf{z}_i \sim \text{i.d. } f, \quad E(y_i | x_i, \mathbf{z}_i) = g(\alpha_1 x_i + \mathbf{z}_i^\top \beta) \quad (1)$$

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 $\mathbf{z}_i$ so that the intercept term is included in β as well. For example, if y_i is count data from a Poisson distribution, one can choose an exponential function for g . Then equation (1) becomes

$$y_i | x_i, \mathbf{z}_i \sim \text{i.d. Poisson}(\mu_i), \quad \mu_i = E(y_i | x_i, \mathbf{z}_i) = \exp(\alpha_1 x_i + \mathbf{z}_i^\top \beta), \quad i = 1, \dots, n \quad (2)$$

However, as discussed in literature,^{23,24} when a count predictor x_i 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 equations (1) and (2). Let $r_i = 1$ if x_i 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 $r_i = 1$ and the other comprising of the remaining subjects with $r_i = 0$. For example, if x_i is an alcohol drinking count variable such as days of alcohol drinking, the difference between a subject with $r_i = 1$ and $r_i = 0$ 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 equation

(1), if $x_j = 0$ is a random zero, the coefficient α_1 of x_j represents the effect of drinking on the response y_j within the drinker subgroup when the drinking outcome changes from 0 to 1; while if $x_j = 0$ represents a structural zero, such a difference speaks to the effect of the trait of drinking on the response y_j . When x_j is included in the model as in equation (1), the coefficient of x_j has a dubious interpretation. Thus equation (1) is flawed and must be revised to tease out such conceptually distinctive effects of structural and random zeros.

2.1 Review of likelihood-based mixture model

Tang et al.²⁴ considered a likelihood-based mixture model to model the distinctive effects of structural and random zeros. The likelihood-based mixture model consists of two components, one for modeling the response y and the other for modeling the zero-inflated count predictor x .

2.1.1 Main Model—In the settings where the structural zeros are observed, we may simply add the indicator r_i of a structural zero as an additional predictor to address the effects of structural zeros. Then we can specify the following GLM for y

$$y_i | x_i, r_i, \mathbf{z}_i \sim \text{i.d. } f, E(y_i | x_i, r_i, \mathbf{z}_i) = g(\alpha_1 x_i + \alpha_2 r_i + \mathbf{z}_i^T \beta), i = 1, \dots, n \quad (3)$$

The above model is identical to equation (1), except for an additional indicator of structural zeros in the set of explanatory variables. Under the refined model (3), the effects of traits on the response are explained by α_2 , while the effects of the level of activities of the behavior are indicated by α_1 . Thus model (3) can tease apart the two effects and provide a more comprehensive relationship between the outcome and the trait. If all quantities in model (3) are observed, the standard MLE can be employed to estimate model parameters. However, in many studies, we can not observe the structural zeros, which makes r_i not fully observed; that is, r_i is unknown for subjects with $x_i = 0$.

2.1.2 Auxiliary zero-inflated model—For the zero-inflated predictor x_i , we can model it by some commonly used zero-inflated count models. Here we assume x_i follows the ZIP model with the probability of structural zero ρ_i and the Poisson mean μ_i i.e. $x_i \sim \text{ZIP}(\rho_i, \mu_i)$.

We also let $\mathbf{w}_i$ be a set of predictors for both ρ_i and μ_i . The ZIP model indexed by a parameter vector $\gamma = (\gamma_1, \gamma_2)^T$ is given by

$$x_i | \mathbf{w}_i \sim \text{i.d. ZIP}(\rho_i, \mu_i), \text{logit}(\rho_i) = \mathbf{w}_i^T \gamma_1, \log(\mu_i) = \mathbf{w}_i^T \gamma_2, i = 1, \dots, n \quad (4)$$

Note that although ρ_i and μ_i may depend on different sets of predictors, we assume a common set $\mathbf{w}_i$ for notational brevity, which includes all the predictors for both components. On the other hand, $\mathbf{w}_i$ may be different from or overlap with $\mathbf{z}_i$ in the main model (3). The purpose of the auxiliary model is to model r_i . If the structural zeros are observed, regression models such as the logistic model can be applied. However, since the structural zeros are

unobserved, we need the zero-inflated model to address the structural zeros. The ZIP model is flexible, and other zero-inflated count models can also be easily accommodated.

To ensure the validity of the main model and the auxiliary model, we assume the following two conditions hold.

Assumption A. Conditional Independence. Given $\mathbf{w}_i$, x_i and r_i are independent of $\mathbf{z}_i$, i.e.

$$(x_i, r_i) \perp \mathbf{z}_i | \mathbf{w}_i$$

This assumption implies that x_i and r_i may be associated with $\mathbf{z}_i$, but the association is only through $\mathbf{w}_i$. This condition can be easily satisfied by including additional predictors from $\mathbf{w}_i$ in equation (3), as needed for the conditional independence, into $\mathbf{w}_i$ in equation (4).

Assumption B. Comprehensiveness of the main model. Given the predictors x_i , $\mathbf{z}_i$, r_i , the response y_i is independent of $\mathbf{w}_i$, i.e.

$$y_i \perp \mathbf{w}_i | x_i, \mathbf{z}_i, r_i$$

which implies that y_i may depend on $\mathbf{w}_i$, but the dependence is only through x_i , $\mathbf{z}_i$ and r_i . This condition can always be met through including additional predictors from $\mathbf{w}_i$ in equation (4) into $\mathbf{z}_i$ in equation (3). The comprehensiveness here means that all the information on y_i carried by or contained in $\mathbf{w}_i$ is captured by x_i , $\mathbf{z}_i$ and r_i through model equation (3). In practice, the selection of $\mathbf{w}_i$ and $\mathbf{z}_i$ is based on the subject matter of the study. As long as pertinent predictors for the outcome y_i and the count x_i are included, the two assumptions would be approximately true.

When the Assumptions A and B are satisfied, the MLE method can be applied to estimate the parameters in equations (3) and (4), and make inferences about the relationship between the zero-inflated count predictor and the outcome.

The likelihood-based mixture model proposed by Tang et al.²⁴ depends on the distribution assumptions for both the outcome and zero-inflated count predictors. However, in many applications, knowledge regarding the exact distribution is limited and the corresponding distributions may be misspecified. In such cases, the MLE approach may yield biased estimates. To relax the distribution assumption, we propose the following GEE-type mixture model to address the structural zero issue in count predictors.

2.2 A GEE-type mixture model

Let r_i 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.

2.2.1 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, that is

$$E(y_i | x_i, r_i, \mathbf{z}_i) = g(\alpha_1 x_i + \alpha_2 r_i + \mathbf{z}_i^T \beta), \quad i = 1, \dots, n \quad (5)$$

2.2.2 Auxiliary zero-inflated model—For the zero-inflated predictor x_i , we need to model both the probability of structural zeros ρ_i and the count mean μ_i by logit and loglinear regression models. Let $\mathbf{w}_i$ be a set of predictors for both ρ_i and μ_i . The auxiliary zero-inflated model is given by

$$\text{logit}(\Pr(r_i = 1 | \mathbf{w}_i)) = \text{logit}(\rho_i) = \mathbf{w}_i^T \gamma_1, \log(E(x_i | r_i = 0, \mathbf{w}_i)) = \log(\mu_i) = \mathbf{w}_i^T \gamma_2, \quad i = 1, \dots, n$$

(6)

To estimate the parameters in the main model (5), let $\alpha = (\alpha_1, \alpha_2, \beta)^T$. Define

$$S_{1i}^\alpha = I(x_i = 0)[y_i - E(y_i | x_i = 0, \mathbf{z}_i, \mathbf{w}_i)], \\ S_{2i}^\alpha = I(x_i > 0)[y_i - E(y_i | x_i > 0, \mathbf{z}_i, \mathbf{w}_i)]$$

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

Let $\delta_i = \Pr(r_i = 1 | x_i = 0, \mathbf{w}_i)$, based on the main model, and replace r_i in S_i^α with their conditional mean, yielding

$$S_{1i}^\alpha = I(x_i = 0)[y_i - (1 - \delta_i)g(\mathbf{z}_i^T \beta) - \delta_i g(\alpha_2 + \mathbf{z}_i^T \beta)], \quad (7) \\ S_{2i}^\alpha = I(x_i > 0)[y_i - g(\alpha_1 x_i + \mathbf{z}_i^T \beta)]$$

The following estimating equation

$$\mathbf{Q}_n^\alpha(\alpha) = \sum_{i=1}^n D_i^\alpha (V_i^\alpha)^{-1} S_i^\alpha = \mathbf{0} \quad (8)$$

can be used to estimate α , where $D_i^\alpha = \frac{\partial S_i^\alpha}{\partial \alpha}$ and $V_i^\alpha = \text{Var}(S_i^\alpha | \mathbf{z}_i, \mathbf{w}_i)$.

If δ_i is known or can be easily estimated given that all r_i are observed, the α can be estimated based on equation (8). However, in most cases, r_i for some subjects are not observed or are unobservable, and equation (8) does not provide enough information to estimate α . Hence, we need the auxiliary model to provide us additional information to estimate α .

Since $\delta_i = \delta_i = Pr(r_i = 1 | x_i = 0, \mathbf{w}_i) = Pr(r_i = 1, x_i = 0 | \mathbf{w}_i) / Pr(x_i = 0 | \mathbf{w}_i) = \rho_i$ involves γ_1 $I(\rho_i + Pr(x_i = 0, r_i = 0, r_i = 0 | \mathbf{w}_i))$

and γ_2 in equation (6), we need to construct estimating equations to estimate γ_1 and γ_2 . Since the conditional probability of random zeros $Pr(x_i = 0, r_i = 0 | \mathbf{w}_i)$ cannot be uniquely identified based on the models in equation (6), we further assume that the zero-inflated count predictor x_i follows a zero-inflated Poisson distribution, i.e. $x_i | \mathbf{w}_i \sim \text{i.d. ZIP}(\rho_i, \mu_i)$.

Under the assumption, δ_i can be expressed as $\frac{\rho_i}{\rho_i + (1 - \rho_i)\exp(-\mu_i)}$ and can be identified by the models in equation (6). Let $\gamma = (\gamma_1, \gamma_2)^T$, and define

$$\begin{aligned} S_{1i}^\gamma &= I(x_i = 0) - E[I(x_i = 0 | \mathbf{w}_i)], \\ S_{2i}^\gamma &= I(x_i > 0)[x_i - E(x_i | x_i > 0, \mathbf{w}_i)] \end{aligned}$$

Under the models in equation (6), we further get

$$\begin{aligned} S_{1i}^\gamma &= I(x_i = 0) - \rho_i - (1 - \rho_i)\exp(-\mu_i), \\ S_{2i}^\gamma &= I(x_i > 0)[x_i - \mu_i / (1 - \exp(-\mu_i))] \end{aligned} \quad (9)$$

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

$$\mathbf{Q}_n^\gamma(\gamma) = \sum_{i=1}^n D_i^\gamma (V_i^\gamma)^{-1} S_i^\gamma = \mathbf{0} \quad (10)$$

can be used to estimate γ , where $D_i^\gamma = \frac{\partial S_i^\gamma}{\partial \gamma}$ and $V_i^\gamma = \text{Var}(S_i^\gamma | \mathbf{w}_i)$.

Please note that although we assume a zero-inflated Poisson model for x_i to define $S_{1i}^\gamma, S_{2i}^\gamma$ in equation (10), 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 $S_i = (S_i^\alpha, S_i^\gamma)$, $\theta = (\alpha, \gamma)$. We define the following generalized estimating equation to estimate θ

$$\mathbf{U}_n(\theta) = \sum_{i=1}^n D_i V_i^{-1} S_i = \mathbf{0} \quad (11)$$

where $D_i = \frac{\partial S_i}{\partial \theta}$ and $V_i = \text{Var}(S_i | \mathbf{Z}_i, \mathbf{w}_i)$.

Based on equations (7) and (9), we have

$$D_i = \begin{pmatrix} 0 & \frac{\partial S_{2i}^\alpha}{\partial \alpha_1} & 0 & 0 \\ \frac{\partial S_{1i}^\alpha}{\partial \alpha_2} & 0 & 0 & 0 \\ \frac{\partial S_{1i}^\alpha}{\partial \beta} & \frac{\partial S_{2i}^\alpha}{\partial \beta} & 0 & 0 \\ \frac{\partial S_{1i}^\alpha}{\partial \gamma_1} & 0 & \frac{\partial S_{1i}^\gamma}{\partial \gamma_1} & \frac{\partial S_{2i}^\gamma}{\partial \gamma_1} \\ \frac{\partial S_{1i}^\alpha}{\partial \gamma_2} & 0 & \frac{\partial S_{1i}^\gamma}{\partial \gamma_2} & \frac{\partial S_{2i}^\gamma}{\partial \gamma_2} \end{pmatrix}$$

with

$$\begin{aligned}
\frac{\partial S_{1i}^\alpha}{\partial \alpha_2} &= -I(x_i = 0) \delta_i \dot{g}(\alpha_2 + \mathbf{z}_i^T \beta), \\
\frac{\partial S_{1i}^\alpha}{\partial \beta} &= -I(x_i = 0) \left[(1 - \delta_i) \dot{g}(\mathbf{z}_i^T \beta) + \delta_i \dot{g}(\alpha_2 + \mathbf{z}_i^T \beta) \right] \mathbf{z}_i, \\
\frac{\partial S_{1i}^\alpha}{\partial \gamma_1} &= I(x_i = 0) \left[g(\mathbf{z}_i^T \beta) - g(\alpha_2 + \mathbf{z}_i^T \beta) \right] \frac{\exp(-\mu_i)}{[\rho_i + (1 - \rho_i) \exp(-\mu_i)]^2} \frac{\partial \rho_i}{\partial \gamma_1}, \\
\frac{\partial S_{1i}^\alpha}{\partial \gamma_2} &= I(x_i = 0) \left[g(\mathbf{z}_i^T \beta) - g(\alpha_2 + \mathbf{z}_i^T \beta) \right] \delta_i (1 - \delta_i) \frac{\partial \mu_i}{\partial \gamma_2}, \\
\frac{\partial S_{2i}^\alpha}{\partial \alpha_1} &= -I(x_i > 0) \dot{g}(\alpha_1 x_i + \mathbf{z}_i^T \beta) x_i \\
\frac{\partial S_{2i}^\alpha}{\partial \beta} &= -I(x_i > 0) \dot{g}(\alpha_1 x_i + \mathbf{z}_i^T \beta) \mathbf{z}_i \\
\frac{\partial S_{1i}^\gamma}{\partial \gamma_1} &= -(1 - \exp(-\mu_i)) \frac{\partial \rho_i}{\partial \gamma_1} \\
\frac{\partial S_{1i}^\gamma}{\partial \gamma_2} &= (1 - \rho_i) \exp(-\mu_i) \frac{\partial \mu_i}{\partial \gamma_2} \\
\frac{\partial S_{2i}^\gamma}{\partial \gamma_1} &= 0 \\
\frac{\partial S_{2i}^\gamma}{\partial \gamma_2} &= -I(x_i > 0) \frac{1 - \exp(-\mu_i) - \mu_i \exp(-\mu_i)}{(1 - \exp(-\mu_i))^2} \frac{\partial \mu_i}{\partial \gamma_2}
\end{aligned}$$

and $\dot{g}(u) = \frac{\partial g(u)}{\partial u}$.

Next, we calculate the variance matrix $V_i = \text{Var}(S_i | \mathbf{z}_i, \mathbf{w}_i)$, i.e.,

$$V_i = \text{Var}(S_i | \mathbf{z}_i, \mathbf{w}_i) = \begin{pmatrix} \text{Var}(S_{1i}^\alpha) & \text{Cov}(S_{1i}^\alpha, S_{2i}^\alpha) & \text{Cov}(S_{1i}^\alpha, S_{1i}^\gamma) & \text{Cov}(S_{1i}^\alpha, S_{2i}^\gamma) \\ \text{Cov}(S_{1i}^\alpha, S_{2i}^\alpha) & \text{Var}(S_{2i}^\alpha) & \text{Cov}(S_{2i}^\alpha, S_{1i}^\gamma) & \text{Cov}(S_{2i}^\alpha, S_{2i}^\gamma) \\ \text{Cov}(S_{1i}^\alpha, S_{1i}^\gamma) & \text{Cov}(S_{2i}^\alpha, S_{1i}^\gamma) & \text{Var}(S_{1i}^\gamma) & \text{Cov}(S_{1i}^\gamma, S_{2i}^\gamma) \\ \text{Cov}(S_{1i}^\alpha, S_{2i}^\gamma) & \text{Cov}(S_{2i}^\alpha, S_{2i}^\gamma) & \text{Cov}(S_{1i}^\gamma, S_{2i}^\gamma) & \text{Var}(S_{2i}^\gamma) \end{pmatrix}$$

where

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

The above calculation of the variance V_i is under Assumptions A and B.

Let $\hat{\theta} = (\hat{\alpha}, \hat{\gamma})$ be the estimator of $\theta = (\alpha, \gamma)$ by solving the generalized estimating equation (11). Since there is no closed-form for the estimates, numeric solutions can be obtained easily through the popular Newton-Raphson (NR) method. Under equations (7) and (9), $\hat{\theta}$ is consistent and asymptotically normal distributed (see Appendix, Supplementary material for a sketch of the proof).

2.2.3 Asymptotic results—a). The GEE estimator $\hat{\theta}$ is consistent and $\sqrt{n}(\hat{\theta} - \theta)$ is asymptotically normally distributed with mean zero and covariance matrix

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

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

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

b). The estimator $\hat{\alpha}$ of α for the main model (5) is consistent and $\sqrt{n}(\hat{\alpha} - \alpha)$ is asymptotically normally distributed with mean zero and covariance matrix $\Sigma_{\alpha} = B^{-1}[\Psi + \Phi]B^{-T}$, where

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

A consistent estimate of the asymptotic variance Σ_{α} can be obtained by substituting consistent estimates of the respective quantities, i.e.

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

The $\hat{V}_i^{\alpha}$ and $\hat{D}_i^{\alpha}$ 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 .

Please note that V_i^{γ} in equation (10) can be any invertible matrix function of $\mathbf{w}_i$, although the commonly used $V_i^{\gamma} = \text{Var}(S_i^{\gamma}|\mathbf{w}_i)$ provides the most efficient estimates when the data does follow the ZIP model. However, regardless of the choice of function of $\mathbf{w}_i$ for V_i^{γ} , inference based on equation (10) is valid as long as the specification in equation (6) is true.

3 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.²⁴ 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_i as well as the associated indicator for the structural zero r_i is generated from the following ZIP model

$$\begin{aligned} x_i | w_i &\sim ZIP(\rho_i, \mu_i), w_i \sim \text{Uniform}(0, 1), \\ \text{logit}(\rho_i) &= \gamma_{10} + \gamma_{11}w_i, \log(\mu_i) = \gamma_{20} + \gamma_{21}w_i \end{aligned} \quad (12)$$

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 $\gamma_{10} = -1$, $\gamma_{11} = 0$, $\gamma_{20} = 1$, and $\gamma_{21} = -0.5$. In this case, the proportion of structural zeros in x is around 27%.

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

$$y_i = \alpha_1 x_i + \alpha_2 r_i + \mathbf{z}_i^T \beta + e_i \quad (13)$$

where $e_i \sim N(0, 1)$.

Binary response Y : We simulate a binary response y_i based on the following GLM with logit link function

$$y_i | x_i, r_i, \mathbf{z}_i \sim \text{Bernoulli}(p_i), \text{logit}(p_i) = \alpha_1 x_i + \alpha_2 r_i + \mathbf{z}_i^T \beta \quad (14)$$

Poisson response Y : The Poisson response y_i is generated through the following GLM with a log link function

$$y_i | x_i, r_i, \mathbf{z}_i \sim \text{Poisson}(\mu_i), \log(\mu_i) = \alpha_1 x_i + \alpha_2 r_i + \mathbf{z}_i^T \beta \quad (15)$$

Zero-inflated Poisson response Y : We consider a zero-inflated count response y_i based on the following ZIP model

$$y_i | x_i, r_i, \mathbf{z}_i \sim \text{i.d.} \text{ ZIP}(p_i, \mu_i), \quad (16)$$

$$\text{logit}(p_i) = c_0, \log(\mu_i) = \alpha_1 x_i + \alpha_2 r_i + \mathbf{z}_i^T \beta$$

For all of the above data scenarios, we set $\alpha_1 = 0.2$, $\alpha_2 = 0.5$, $\beta = (\beta_0, \beta_1) = (-1, 1)^T$ and $c_0 = -1$. The performance of the GEE-type method and the likelihood-based method is evaluated for different sample sizes 200, 500 and 1000. Due to the latent nature of the structural zeros, the ZIP requires a larger sample size to obtain reliable estimates, especially within the context of a zero-inflated x following another ZIP, we consider larger sample sizes 500, 1000 and 1500 for each scenario. All simulations are performed with 1000 Monte Carlo replicates.

3.1 Both models for X and Y are correctly specified

For data $(y_i, x_i, r_i, \mathbf{z}_i)$ generated based on equations (12) and (13), (14), (15) or (16), the auxiliary zero-inflated model for X is fitted by

$$\text{logit} (Pr(r_i = 1 | w_i)) = \gamma_{10} + \gamma_{11}w_i, \log(E(x_i | r_i = 0, w_i)) = \gamma_{20} + \gamma_{21}w_i \quad (17)$$

and the main model for Y is fitted by

$$E(y_i | x_i, r_i, \mathbf{z}_i) = g(\alpha_1 x_i + \alpha_2 r_i + \mathbf{z}_i^T \beta) \quad (18)$$

with identify, logit and log link function $g(\cdot)$ for continuous, binary, and Poisson responses, respectively. The main model for a ZIP response is fitted with logit and log-link function for the zero-inflated component and Poisson component, respectively.

The resulting bias, ESE, ASE and CP for both the likelihood-based method and the GEE-type method based on 1000 realizations are presented in Table 1 and Tables S1 and S2 as the Supplementary material. The bias is defined as the mean of difference between the Monte Carlo estimates and the true value. The ESE is the Monte Carlo sample standard deviation of estimates across the 1000 samples; for samples with large size, this should be close to the asymptotic standard deviation. The ASE is the Monte Carlo average of the estimated asymptotic standard deviation across the 1000 samples, and thus if the normal approximation is appropriate, the ESE and the ASE should be close. The CP is the Monte Carlo coverage rate of the 95% asymptotic confidence intervals, or the proportion of samples whose 95% confidence intervals contain the corresponding true values. The ranges of the Monte Carlo sample standard error (SE) of the Bias, the Monte Carlo standard error of the asymptotic standard deviation, as well as the Monte Carlo standard error of the coverage rate are also provided as footnotes of the tables. The simulation results show that the proposed method performs equally well, in both the point estimates and the variance estimates, compared to the likelihood-based method when both models for X and Y are correctly specified. Specifically, the proposed method practically yields unbiased estimates, the estimated standard errors based on the asymptotic distribution are very close to the Monte Carlo sample standard deviations of parameter estimates, and the 95% empirical coverage probabilities are quite close to the nominal level, 95%. The performance of the proposed method improves when the sample size increases from 200 to 1000, as expected.

3.2 Only the model for X is misspecified

To investigate how the two methods perform under misspecification of the auxiliary model, we generate an over-dispersed zero-inflated predictor. Essentially, for x_i generated in equation (12), the positive x_i s are replaced by positive data generated from a negative-binomial (NB) distribution, i.e.

$$x_i | x_i > 0, w_i \sim \text{NB}(\mu_i, \tau), \log(\mu_i) = \gamma_{20} + \gamma_{21}w_i \quad (19)$$

All parameter values are the same as in the aforementioned ZIP setting, except for a new dispersion parameter τ , which is set to $\tau = 1.5$. Because $\text{NB}(\mu_x, \tau)$ converges to Poisson with the same mean μ_x as $\tau \rightarrow \infty$, selecting a relatively small τ such as $\tau = 1.5$ allows us to better assess performance of both methods under this specific type of overdispersion.

The results for sample size 1000 are presented in Table 2. For sample size 200 and 500, the results are presented in Tables S3–S5 of the Supplementary material. Based on the results, the GEE-type method produces good estimates for both the main model and the auxiliary model. The asymptotic and Monte Carlo sample standard deviation are very similar, and the CP is close to the true value of 0.95. As for the likelihood-based method, the estimates for the main model are very good and the CP is close to 0.95 as well. For the auxiliary model, even though the likelihood-based method still yields good point estimates of γ_{20} and γ_{21} , the variance is underestimated and hence the CP is below 0.95. These results are consistent with Tang et al.²⁵ Therefore, if the auxiliary model is not of interest, both methods can be applied if only the auxiliary model is misspecified.

3.3 Only the model for Y is misspecified

Because the GEE-type method only used the information on the first-order moment of response y , it provides some protection against a misspecified distribution of response y . To misspecify a model, for a continuous outcome, y_i is generated through

$$y_i = \alpha_1 x_i + \alpha_2 r_i + \mathbf{z}_i^T \beta + e_i \quad (20)$$

where $e_i \sim t(2)$, which makes the distribution of the response heavy-tailed. This case is designed to examine the robustness of the proposed GEE-type method. We estimate regression parameters of model (20) using likelihood-based method and the normal distribution assumption. For a binary response, we first employ the Copula approach to generate multivariate correlated Bernoulli random variables $y_{ik}, k = 1, \dots, 7$, based on model (14), and then set $y_i = \sum_{k=1}^7 y_{ik}$. The Copula exchangeable correlation coefficient is taken as 0.6. We also consider the likelihood-based method, which assumes that y_i follows Binomial(7, p_i), where p_i is given in equation (14). For a count response, we simulate y_i from a modified Poisson model with a random effect. Specifically, the response y_i is generated based on the normal-random-effect (NRE) Poisson model as follows

$$y_i | x_i, r_i, b_i, \mathbf{z}_i \sim \text{Poisson}(\mu_i^*), \log(\mu_i^*) = \alpha_1 x_i + \alpha_2 r_i + \mathbf{z}_i^T \beta - \frac{1}{2} \sigma_b^2 + b_i \quad (21)$$

where the random effect $b_i \sim N(0, \sigma_b^2)$ with $\sigma_b^2 = 1$. All regression parameters are set to the same values as in the aforementioned two examples. Under model (21), it is easy to show that the conditional mean and variance of y_i given $(x_i, r_i, \mathbf{z}_i)$ are

$$E(y_i | x_i, r_i, \mathbf{z}_i) = \mu_i = \exp(\alpha_1 x_i + \alpha_2 r_i + \mathbf{z}_i^T \beta), \text{Var}(y_i | x_i, r_i, \mathbf{z}_i) > E(y_i | x_i, r_i, \mathbf{z}_i)$$

Thus the NRE-Poisson model has the same mean as the corresponding Poisson model but an over-dispersed variance.

For a ZIP count response, we first generated y_i based on (16), and then replaced the positive y_i s by positive data generated from a negative-binomial (NB) distribution, i.e.

$$y_i | y_i > 0, x_i, r_i, \mathbf{z}_i \sim \text{NB}(\mu_i, \tau), \log(\mu_i) = \alpha_1 x_i + \alpha_2 r_i + \mathbf{z}_i^T \beta \quad (22)$$

All the parameter values are the same as in the aforementioned settings, except for a dispersion parameter τ , which is again set to $\tau = 1.5$.

The simulation results for sample size 1000 are presented in Table 3. The simulation results for sample sizes 200 and 500, and for the ZIP response, are given in Tables S6 and S7 as supplementary material. The results show that the likelihood-based method underestimates the variance of the estimates for all cases and yields considerable biased CPs which are far below the nominal level 0.95 when the outcome y is a Binomial-type or modified Poisson variable. However, the proposed GEE-type method produces pretty robust estimates for all model parameters. Therefore, as our primary interest focuses on the main model, the likelihood-based method should not be used when the model for the outcome y is misspecified. In this situation, the GEE-type method is recommended.

4 Case study

In this section, we apply the proposed approach to a randomized clinical study for teaching awareness and self-monitoring skills to indwelling urinary catheter users conducted in New York state. A total of 202 subjects were recruited and randomized to the intervention and control groups.²⁶ Two primary outcomes are whether the subjects have experienced Urinary Tract Infections (UTIs) and catheter blockages during the last two months, as well as the corresponding counts of these experiences. For illustration, we consider the outcomes at enrollment although this is a randomized longitudinal study in which each subject was measured every two months. In this example, we use count variables for both UTI and catheter blockages and examine how the catheter blockage is associated with UTI, with UTI treated as response and blockage treated as predictor, after adjusting for age, sex and the

duration of consistent catheter use (in months). Based on the test of inflated zeros in Poisson models, proposed by He et al.,²⁷ the count of UTI follows a Poisson distribution, while the count of blockage follows a ZIP distribution.

To investigate how the catheter blockage is associated with UTI, we apply the proposed main model for the Poisson response UTI (y_i), with the count of blockage x_i and the latent indicator r_i of structural zeros of blockage as predictors and age, sex and the duration of catheter use, denoted as $\mathbf{z}_i$, as covariates. For the ZIP predictor catheter blockage, we apply a ZIP auxiliary model with age, sex and duration as the predictors for both zero-inflated component and the count component. Specifically, the main model and the auxiliary model are specified as follows

$$\begin{aligned} E(y_i | x_i, r_i, z_i) &= \mu_i, \quad \text{Blockage} \sim \text{ZIP}(\rho_{xi}, \mu_{xi}), \\ \mu_i &\sim \text{Structural zero of blockage} + \text{Blockage} + \text{Age} + \text{Sex} + \text{Duration}, \\ \rho_{xi} &\sim \text{Age} + \text{Sex} + \text{Duration}, \\ \mu_{xi} &\sim \text{Age} + \text{Sex} + \text{Duration} \end{aligned} \quad (23)$$

The regression coefficient of the structural zeros of blockage represents the effect of the trait of catheter blockages on UTI, while the coefficient of blockage provides the effect of frequency of catheter blockages on UTI for subjects who are at-risk for catheter blockages. Similar to He et al.,²⁷ we have deleted three outliers that have extremely high number of catheter blockage, although more complicated methods such as Winsorizing may be applied. As a comparison, the likelihood-based method is also applied in which a Poisson regression model is assumed for UTI, and a ZIP model assumed for catheter blockage.

Shown in Table 4 are the estimates of the main model for UTI. Although some differences exist, the estimates are in general consistent for the two methods. Both methods have successfully identified significant associations between catheter blockage and UTI. The non-blockage subjects have about 1.2 less UTI (p -value < 0.01 for both methods) than subjects at-risk for blockage. However, if subjects have blockage, the number of blockages is not associated with UTI. In addition, males are associated with more UTI.

The analysis results of the auxiliary model for catheter blockage are summarized in Table 5. Males are more likely to experience catheter blockages, while longer duration of catheter use decreases the likelihood of blockage. But for subjects who have experienced blockage, their age, sex and duration of catheter use do not associate with the number of blockages.

5 Discussion

In public health and medical studies, structural zeros are common and typically not separable from random zeros. Most research on the structural zero issue focuses on the cases where the count variables are treated as responses in regression analysis. Very little attention is paid to the cases where such count variables are treated as predictors. This paper addresses the structural zero issue in zero-inflated count predictors by proposing a mixture model for both the response and predictors using a GEE-type method. Specifically, in addition to the

count variable, a latent indicator of the structural zeros in the predictor is also included in the main model for the response variable to examine the trait effect of the predictor, while the information of identifying the structural zeros is provided by an auxiliary model. The GEE-type method does not assume any distributions for the outcome and the zero-inflated count predictor but linear functions linked to the means, therefore the GEE-type method is more robust than the likelihood-based counterpart. The asymptotic properties of GEE-type estimates are developed. Simulation studies have demonstrated that the proposed methods work well and provide more robust estimation than the maximum likelihood methods when models are misspecified.

Conditional independence is assumed for both models, or equivalently, no confounders for both the main and auxiliary models, which is easily satisfied in regression analysis. In addition to the common types of response variables, the proposed method can be used in a similar manner to study other types of response variables. For example, it is interesting to apply the GEE-type method to analyze survival data with zero-inflated predictors.

In our mixture model, linear functions of explanatory variables are specified for notational brevity. More complex functions of explanatory variables may be considered utilizing piecewise linear, polynomial functions or even non-parametric form of the mean response functions. Nonparametric techniques such as local polynomial regression and B-spline approximation are suggested for parameter estimation. Although we limited our considerations to cross-sectional data, the same idea can readily be extended to longitudinal data.

In this paper, we discussed the issue of structural zeros when a zero-inflated count variable is used as a covariate. The same problems may arise in similar cases of population heterogeneity. For example, zero and one-inflation for count data were observed in Zhang et al.²⁸ and middle category inflation was observed in the ordered responses in Bagozzi and Mukherjee.²⁹ Further research is needed to address these issues, but in principle, our approach should be able to adapt to these situations.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the NIH under grants R01GM108337 and P20GM109036 and the National Natural Science Foundation of China (grant no. 11601080).

References

1. Burger M, Van Oort F and Linders G-J. On the specification of the gravity model of trade: zeros, excess zeros and zero-inflated estimation. *Spatial Economic Analys* 2009; 4: 167–190.
2. Zorn CJW. An analytic and empirical examination of zero-inflated and hurdle Poisson specifications. *Sociol Meth Res* 1998; 26: 368–400.
3. Clark DH. Can strategic interaction divert diversionary behavior? a model of us conflict propensity. *J Politics* 2003; 65: 1013–1039.

4. Lord D, Washington SP and Ivan JN. Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. *Accident Analys Prevent* 2005; 37: 35–46.
5. Horton NJ, Bechuk JD, Jones CL, et al. Goodness-of-fit for GEE: an example with mental health service utilization. *Stat Med* 1999; 18: 213–222. [PubMed: 10028141]
6. Pardini D, White HR and Stouthamer-Loeber M. Early adolescent psychopathology as a predictor of alcohol use disorders by young adulthood. *Drug Alcohol Depend* 2007; 88: S38–S49. [PubMed: 17257781]
7. Neal DJ, Sugarman DE, Hustad JTP, et al. It's all fun and games... or is it? Collegiate sporting events and celebratory drinking. *J Studies Alcohol Drugs* 2005; 66: 291–294.
8. Hagger-Johnson G, Bewick BM, Conner M, et al. Alcohol, conscientiousness and event-level condom use. *Br J Health Psychol* 2011; 16: 828–845. [PubMed: 21988067]
9. Connor JL, Kypri K, Bell ML, et al. Alcohol outlet density, levels of drinking and alcohol-related harm in New Zealand: a national study. *J Epidemiol Commun Health* 2011; 65: 841–846.
10. Buu A, Johnson NJ, Li R, et al. New variable selection methods for zero-inflated count data with applications to the substance abuse field. *Stat Med* 2011; 30: 2326–2340. [PubMed: 21563207]
11. Fernandez AC, Wood MD, Laforge R, et al. Randomized trials of alcohol-use interventions with college students and their parents: lessons from the transitions project. *Clin Trial* 2011; 8: 205–213.
12. Cranford JA, Zucker RA, Jester JM, et al. Parental alcohol involvement and adolescent alcohol expectancies predict alcohol involvement in male adolescents. *Psychol Addict Behav* 2010; 24: 386–396. [PubMed: 20853923]
13. Hildebrandt T, McCrady B, Epstein E, et al. When should clinicians switch treatments? An application of signal detection theory to two treatments for women with alcohol use disorders. *Behav Res Ther* 2010; 48: 524–530. [PubMed: 20359693]
14. Hernandez-Avila CA, Song C, Kuo L, et al. Targeted versus daily naltrexone: secondary analysis of effects on average daily drinking. *Alcohol: Clin Exp Res* 2006; 30: 860–865. [PubMed: 16634855]
15. Hall DB. Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics* 2000; 56: 1030–1039. [PubMed: 11129458]
16. Hall DB and Zhang Z. Marginal models for zero inflated clustered data. *Stat Model* 2004; 4: 161–180.
17. Yu Q, Chen R, Tang W, et al. Distribution-free models for longitudinal count responses with overdispersion and structural zeros. *Stat Med* 2013; 32: 2390–405. [PubMed: 23239019]
18. Tang W, He H and Tu X. *Applied categorical and count data analysis*. Boca Raton, FL: Chapman & Hall/CRC, 2012.
19. Dean C and Lawless JF. Tests for detecting overdispersion in Poisson regression models. *J Am Stat Assoc* 1989; 84: 467–472.
20. Kowalski J and Tu XM. *Modern applied U-statistics*. Hoboken, NJ: Wiley-Interscience, 2008Wiley Series in Probability and Statistics.
21. Xia YL, Morrison-Beedy D, Ma JM, et al. Modeling count outcomes from HIV risk reduction interventions: a comparison of competing statistical models for count responses. *AIDS Res Treat* 2012; 25: 1–11.
22. Crits-Christoph P, Gallop R, Sadicario JS, et al. Predictors and moderators of outcomes of HIV/STD sex risk reduction interventions in substance abuse treatment programs: a pooled analysis of two randomized controlled trials. *Substance Abuse Treat Prevent Policy* 2014; 9: 3.
23. He H, Wang W, Crits-Christoph P, et al. On the implication of structural zeros as independent variables in regression analysis: applications to alcohol research. *J Data Sci* 2014; 12: 439–460. [PubMed: 28989340]
24. Tang W, He H, Wang WJ, et al. Untangle the structural and random zeros in statistical modelings. *J Appl Stat* 2018; 45: 1714–1733. [PubMed: 30906098]
25. Tang W, Lu N, Chen T, et al. On performance of parametric and distribution-free models for zero-inflated and over-dispersed count responses. *Stat Med* 2015; 34: 3235–3245. [PubMed: 26078035]

26. Wilde MH, McMahon JM, McDonald MV, et al. Self-management intervention for long-term indwelling urinary catheter users: randomized clinical trial. *Nurs Res* 2015; 64: 24–34. [PubMed: 25502058]
27. He H, Zhang H, Ye P, et al. A test of inflated zeros for poisson regression models. *Stat Meth Med Res* (in press).
28. Zhang C, Tian G-L and Ng K-W. Properties of the zero-and-one inflated poisson distribution and likelihood-based inference methods. *Stat Interface* 2016; 9: 11–32.
29. Bagozzi BE and Mukherjee B. A mixture model for middle category inflation in ordered survey responses. *Political Analys* 2012; 20: 369–386.

Table 1.

The Monte Carlo estimate of bias (Bias), Monte Carlo sample standard deviation of estimates (ESE), Monte Carlo average of estimated asymptotic standard deviations (ASE) and the Monte Carlo coverage rate of the 95% asymptotic confidence intervals (CP) for the main model based on 1000 realizations when both models for x and y are correctly specified and sample size is 1000.

Y	Parameter	GEE-type				Likelihood-based			
		Bias ^a	ESE	ASE ^b	CP ^c	Bias ^a	ESE	ASE ^b	CP ^c
Continuous	α_1	-0.001	0.03	0.03	0.96	-0.001	0.03	0.03	0.96
	α_2	-0.001	0.13	0.13	0.96	0.001	0.12	0.13	0.96
	β_0	0.006	0.11	0.11	0.95	0.004	0.10	0.11	0.96
	β_1	-0.003	0.12	0.12	0.95	-0.003	0.11	0.11	0.96
Binary	α_1	0.001	0.06	0.06	0.95	0.002	0.06	0.06	0.95
	α_2	-0.003	0.27	0.27	0.95	-0.001	0.27	0.27	0.95
	β_0	0.002	0.23	0.23	0.94	-0.000	0.23	0.23	0.94
	β_1	-0.011	0.25	0.24	0.95	-0.008	0.24	0.23	0.95
Poisson	α_1	-0.000	0.03	0.03	0.95	-0.001	0.03	0.03	0.95
	α_2	-0.001	0.13	0.12	0.95	-0.003	0.12	0.12	0.95
	β_0	0.002	0.12	0.12	0.94	0.002	0.12	0.11	0.94
	β_1	-0.001	0.12	0.12	0.95	-0.001	0.12	0.12	0.95
ZIP ^d	α_1	-0.001	0.03	0.03	0.94	-0.001	0.03	0.03	0.95
	α_2	-0.006	0.12	0.12	0.95	-0.006	0.12	0.12	0.95
	β_0	-0.006	0.13	0.13	0.95	-0.005	0.13	0.13	0.95
	β_1	0.006	0.13	0.13	0.95	0.005	0.12	0.12	0.94
	C_0	-0.028	0.15	0.15	0.96	-0.025	0.15	0.14	0.97

^aThe Monte Carlo SE for the Bias ranges from 0.00082 to 0.00854 for the GEE-type estimate and from 0.00082 to 0.00851 for the likelihood-based estimate.

^bThe Monte Carlo SE for the ASE ranges from 0.00006 to 0.00056 for the GEE-type estimate and from 0.00004 to 0.00051 for the likelihood-based estimate.

^cThe Monte Carlo SE for CP ranges from 0.0062 to 0.0075 for the GEE-type estimate and from 0.0054 to 0.0075 for the likelihood-based estimate.

^dSample size of 1500.

Table 2.

The Monte Carlo estimate of bias (Bias), Monte Carlo sample standard deviation of estimates (ESE), Monte Carlo average of estimated asymptotic standard deviations (ASE) and the Monte Carlo coverage rate of the 95% asymptotic confidence intervals (CP) for the main model based on 1000 realizations when the model for x is misspecified and sample size is 1000.

Y	Parameter	GEE-type				Likelihood-based			
		Bias ^a	ESE	ASE ^b	CP ^c	Bias ^a	ESE	ASE ^b	CP ^c
Continuous	α_1	0.000	0.02	0.02	0.95	0.000	0.02	0.02	0.96
	α_2	-0.001	0.12	0.12	0.96	-0.001	0.12	0.12	0.95
	β_0	-0.001	0.10	0.10	0.96	-0.001	0.10	0.09	0.95
	β_1	0.006	0.12	0.12	0.95	0.004	0.11	0.11	0.96
Binary	α_1	0.004	0.05	0.05	0.94	0.005	0.05	0.05	0.94
	α_2	0.016	0.25	0.24	0.94	0.018	0.25	0.24	0.94
	β_0	-0.014	0.21	0.21	0.95	-0.018	0.21	0.20	0.95
	β_1	0.014	0.25	0.24	0.94	0.018	0.23	0.23	0.97
Poisson	α_1	0.000	0.02	0.02	0.94	0.000	0.02	0.02	0.95
	α_2	0.003	0.11	0.11	0.94	0.001	0.10	0.10	0.95
	β_0	-0.005	0.10	0.10	0.95	-0.004	0.10	0.10	0.96
	β_1	0.001	0.12	0.12	0.95	0.001	0.11	0.11	0.96
ZIP ^d	α_1	0.000	0.02	0.02	0.93	0.000	0.02	0.02	0.94
	α_2	-0.001	0.11	0.11	0.94	-0.001	0.11	0.11	0.95
	β_0	-0.006	0.12	0.11	0.94	-0.006	0.11	0.11	0.94
	β_1	0.004	0.13	0.12	0.95	0.004	0.12	0.12	0.95
	C_0	-0.012	0.15	0.15	0.95	-0.012	0.15	0.14	0.94

^aThe Monte Carlo SE for the Bias ranges from 0.00054 to 0.00781 for the GEE-type estimate and from 0.00054 to 0.00781 for the likelihood-based estimate.

^bThe Monte Carlo SE for the ASE ranges from 0.00007 to 0.00052 for the GEE-type estimate and from 0.00004 to 0.00048 for the likelihood-based estimate.

^cThe Monte Carlo SE for CP ranges from 0.0062 to 0.0081 for the GEE-type estimate and from 0.0054 to 0.0075 for the likelihood-based estimate.

^dSample size of 1500.

Table 3.

The Monte Carlo estimate of bias (Bias), Monte Carlo sample standard deviation of estimates (ESE), Monte Carlo average of estimated asymptotic standard deviations (ASE) and the Monte Carlo coverage rate of the 95% asymptotic confidence intervals (CP) for the main model based on 1000 realizations when the model for y is misspecified and sample size is 1000.

Y	Parameter	GEE-type				Likelihood-based			
		Bias ^a	ESE	ASE ^b	CP ^c	Bias ^a	ESE	ASE ^b	CP ^c
Continuous	α_1	-0.002	0.11	0.09	0.95	-0.006	0.18	0.09	0.93
	α_2	-0.004	0.45	0.40	0.95	-0.028	0.70	0.38	0.94
	β_0	0.009	0.39	0.33	0.96	0.012	0.89	0.34	0.93
	β_1	-0.005	0.40	0.35	0.95	0.001	0.67	0.38	0.94
Binomial	α_1	0.002	0.04	0.04	0.95	0.140	0.04	0.02	0.00
	α_2	0.006	0.18	0.19	0.96	1.980	0.27	0.11	0.00
	β_0	-0.006	0.16	0.16	0.95	-0.475	0.15	0.08	0.01
	β_1	0.001	0.17	0.17	0.96	0.087	0.19	0.10	0.66
Poisson	α_1	-0.002	0.05	0.05	0.93	0.053	0.05	0.03	0.42
	α_2	-0.006	0.20	0.20	0.96	0.516	0.43	0.15	0.26
	β_0	-0.001	0.19	0.19	0.95	-0.277	0.24	0.12	0.38
	β_1	-0.002	0.20	0.20	0.95	0.139	0.24	0.13	0.63
ZIP ^d	α_1	-0.001	0.03	0.03	0.93	0.003	0.03	0.03	0.90
	α_2	-0.001	0.14	0.13	0.94	0.023	0.14	0.12	0.91
	β_0	-0.009	0.14	0.14	0.94	-0.025	0.14	0.13	0.92
	β_1	0.008	0.14	0.13	0.96	0.012	0.13	0.12	0.93
	C_0	-0.029	0.17	0.16	0.95	-0.032	0.16	0.15	0.94

^aThe Monte Carlo SE for the Bias ranges from 0.00108 to 0.01423 for the GEE-type estimate and from 0.00i04 to 0.02824 for the likelihood-based estimate.

^bThe Monte Carlo SE for the ASE ranges from 0.00007 to 0.03572 for the GEE-type estimate and from 0.00002 to 0.02063 for the likelihood-based estimate.

^cThe Monte Carlo SE for CP ranges from 0.0062 to 0.0081 for the GEE-type estimate and from 0.0000 to 0.01561 for the likelihood-based estimate.

^dSample size of 1500.

Table 4.

Comparison of estimates (Estimate), standard errors (Std Err) and p-values from the Poisson main model of UTI for the Urinary Catheter Study.

Covariate	GEE-type			Likelihood-based		
	Estimate	Std Err	<i>p</i> -Value	Estimate	Std Err	<i>p</i> -Value
Intercept	0.6482	0.6665	0.3308	0.4436	0.5909	0.4528
Blockage	−0.0840	0.0795	0.2909	−0.0998	0.1226	0.4157
<i>r</i>	−1.2133	0.3804	0.0014	−1.1481	0.4123	0.0054
Age	−0.0169	0.0090	0.0593	−0.0130	0.0072	0.0697
Male	0.8363	0.3011	0.0055	0.5893	0.2590	0.0229
Duration	−0.0036	0.0025	0.1501	−0.0024	0.0016	0.1209

Note: *r* is the latent indicator of structural zeros of catheter blockage.

Table 5.

Comparison of estimates (Estimate), standard errors (Std Err) and p -values from the ZIP auxiliary model of catheter blockage for the Urinary Catheter Study.

Covariate	GEE-type			Likelihood-based		
	Estimate	Std Err	p-Value	Estimate	Std Err	p-Value
The estimates of the Poisson component						
Intercept	1.5503	0.4260	0.0003	1.2699	0.3821	0.0009
Age	-0.0132	0.0092	0.1503	-0.0084	0.0064	0.1912
Male	0.0795	0.3841	0.8360	0.0400	0.2413	0.8682
Duration	-0.0009	0.0022	0.6976	-0.0006	0.0011	0.5584
The estimates of the zero-inflated component:						
Intercept	0.8983	0.8411	0.2855	0.8503	0.7249	0.2408
Age	0.0023	0.0125	0.8557	0.0038	0.0109	0.7258
Male	0.8645	0.4120	0.0359	0.8914	0.3978	0.0250
Duration	-0.0046	0.0026	0.0789	-0.0053	0.0023	0.0209