Untangle the Structural and Random Zeros in Statistical Modelings

Tang, W.[a], He, H.[b*], Wang, W.J.[c] and Chen, D. G.[d]

[a] *Department of Global Biostatistics & Data Science, Tulane University School of Public Health and Tropical Medicine, New Orleans, LA70122, USA*;
[b] *Department of Epidemiology, Tulane University School of Public Health and Tropical Medicine, New Orleans, LA70122, USA*;
[c] *Brightech International, LLC, New Jersey, NJ 08873, USA*;
[d] *School of Social Work, University of North Carolina at Chapel Hill, Chapel Hill, NC27599, USA*;

Count responses with structural zeros are common in behavioral and social studies. There are considerable research focusing on zero-inflated models such as zero-inflated Poisson (ZIP) and zero-inflated Negative Binomial (ZINB) models for such zero-inflated count data. However, when such variables are used as covariates or predictors, the difference between structural and random zeros is often ignored and biased estimates may be resulted. One remedy is to include an indicator of the structural zero in the model as a predictor if observed. However, structural zeros are often not observed in practice, in which case no statistical method is available to address the biasing issue. This paper is aimed to fill this methodological gap by developing parametric methods to model zero-inflated count data when used as explanatory variables based on the maximum likelihood approach. The response variable can be any type of data including continuous, binary, count or even zero-inflated count responses. Simulation studies are performed to assess the numerical performance of this new approach when sample size is small to moderate. A real data example is also used to demonstrate the application of this method.

Keywords: generalized linear models; maximum likelihood; structural zeros; zero-inflated Poisson; zero-inflated explanatory variables.

This work was supported by the NIH under grants R33 DA027521, R01GM108337, R01HD075635 and a Faculty Research Support Grant from School of Nursing, University of Rochester.

1. Introduction

Count variables recording frequencies of some specific behaviors during a period of time, such as days of alcohol consumption or number of unprotected sexual activities in the past month, are common in behavioral and social studies. It is important, both conceptually and methodologically, to pay close attention to *structural zeros* in such count variables. Structural zeros refer to zero responses by those subjects whose count response will always be zero, in contrast to random (or sampling) zeros that occur to subjects whose count responses can be greater than zero, but appear to be zero due to sampling variability. For example, in HIV-AIDS prevention research, the count of unprotected vaginal sex is commonly used to measure the risk of HIV/AIDS. Subjects who are always, or become,

*Corresponding author. Email: hhe2@tulane.edu

continually abstinent from unprotected sex in a given time period form a *non-risk* group as defined by structural zeros in their count outcomes, while the remaining subjects constitute an *at-risk* group with their count outcomes consisting of random zeros or a positive number of episodes of unprotected sex. Such a partition of the study population is not only supported by the excess number of zeros observed in real studies, but is also conceptually needed to serve as a basis for valid inference.

The main issue in modeling count data with structural zeros is that structural zeros are often not observed. In fact, structural zeros may be latent and not observable, so the issue cannot be solved by refining the study design. For example, in toxicological studies, long-term exposure to food-borne toxins is often estimated using short-term food intake measures. Zeros in the measures may be structural or random simply due to the variability in their food intake from day to day. It is typically impossible, from the survey, to separate these two types of zeroes. In such cases, Appropriate statistical methods are needed to address the issue. The issue has been acknowledged and dealt with when the zero-inflated count data is treated as dependent, or a response variable in literature, see for example, [2, 4–6, 8, 12–14, 18, 20]. **Zero-inflated models such as zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) models have been developped and also successfully applied to various fields in biomedical health fields such as HIV-AIDS, cancer, nursing, and health care outcome research, as well as non-health fields such as zoology, econometric, manufacturing and traffic accident modeling** [1, 3, 7, 9, 10, 15–17, 19, 21–24]. However, the statistical problem and associated implications when such a count outcome is used as an explanatory variable has received far less attention in literature. In such cases, count variables are typically treated as continuous predictors in regression models, with no effort to distinguish structural zeros from their random counterparts. This practice is adopted mainly for modeling convenience, which in many studies does not reflect the realistic association of variables involved. For example, as illustrated in a study on alcohol research [11], a structural zero in drinking outcomes represents an individual who abstains from drinking, while a random zero corresponds to a drinker who happens not to drink during a period of time. Thus, the structural and random zeros represent two distinct groups of subjects with different psychosocial outcomes. Indeed, ignoring the differences between structural and random zeros and simply using the count variable as a predictor may yield biased inferences and uninterpretable findings [11].

To tease out the distinctive effects of structural and random zeros on the response of interest, we can include an indicator of structural zeros in the model (in addition to the count variable itself). This approach requires that the structural zeros are observed, such as alcohol abstainers in alcohol research. **However, as indicated above, structural zeros are often latent and are not directly observable. This paper is aimed at filling the methodological gap by developing a new approach to model the distinctive effects of structural and random zeros as predictors in regression analysis, in the situations where the structural are not observed. Our method relies on modeling structural zeros by zero-inflated models and may be potentially applied to a broad range of fields as mentioned above.**

2. Models for Count Predictors with Structural Zeros

2.1 *Problems from Structural Zeros*

Given a sample of n subjects, let y_i denote the response of interest and x_i a zero-inflated count predictor from the ith subject ($1 \leq i \leq n$). Suppose that the structural zero in x_i

measures some personal trait and the random zero and positive count assesses the level of activities of some behavior of interest such as alcohol use. Further, we assume there are some other covariates, collectively denoted by $\mathbf{z}_i = (z_{i1}, \ldots, z_{ip})^\top$.

Let r_i be an indicator of structural zero of x_i, i.e., $r_i = 1$ if x_i is a structural zero and $r_i = 0$ otherwise. In studies where the structural zeros are observed, one may simply add the indicator r_i of structural zero as an additional predictor in the model to address the differential effects between random and structural zeros. However, in many studies r_i is latent as it is only partially observed; for subjects with $x_i > 0$, $r_i = 0$, however, r_i is unknown for subjects with $x_i = 0$.

The latent indicator r_i partitions the study sample (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 $r_i = 0$. Since the trait in many studies is often a risk factor, we refer to the first group as the non-risk subgroup, while the second as the at-risk subgroup.

If we do not distinguish between structural and random zeros, we may apply the generalized linear models (GLMs) to model the association between the explanatory variables including the predictor of interest x_i and the covariates $\mathbf{z}_i$, and the outcome, as follows:

$$y_i \mid x_i, \mathbf{z}_i \sim \text{i.d. } f_i, \quad \mu_i = E\left(y_i \mid x_i, \mathbf{z}_i\right) = h(\alpha x_i + \mathbf{z}_i^\top \beta) \tag{1}$$

where i.d. denotes independently distributed, f denotes some distribution such as Poisson and h is the inverse of some link function such as the log function [21]. For example, if y_i is continuous, we may use the following linear model:

$$y_i \mid x_i, \mathbf{z}_i \sim \text{i.d. } N\left(\mu_i, \sigma^2\right), \quad \mu_i = E\left(y_i \mid x_i, \mathbf{z}_i\right) = \alpha x_i + \mathbf{z}_i^\top \beta, \quad 1 \le i \le n, \tag{2}$$

where $N\left(\mu, \sigma^2\right)$ denotes a normal distribution with mean μ and variance σ^2. Note that $\mathbf{z}_i$ includes a covariate with the constant value 1 so the models above contain the intercept term.

However, as mentioned in Section 1, when a count variable 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 in (1) and (2). For example, if x_i is a drinking outcome such as days of heaving drinking, the difference between a subject with $r_i = 1$ and $r_i = 0$ is substantial. If $x_i = 0$ is a random zero, the coefficient of x_i represents the differential effect of drinking on the response y_i within the drinker subgroup when the drinking outcome changes from 0 to 1. If $x_i = 0$ represents a structural zero, such a difference speaks to the effect of the trait of drinking on the response y_i. When only including x_i as in (1), the coefficient of x_i has a dubious interpretation. Thus, the model in (1) is flawed and must be revised to tease out such conceptually distinctive effects of structural and random zeros.

Now consider the following GLM:

$$y_i \mid x_i, r_i, \mathbf{z}_i \sim \text{i.d. } f, \quad E\left(y_i \mid x_i, r_i, \mathbf{z}_i\right) = h(\alpha_1 x_i + \alpha_2 r_i + \mathbf{z}_i^\top \beta), \quad 1 \le i \le n. \tag{3}$$

The above is identical to (1), except for the additional indicator of structural zeros in the set of explanatory variables. Under the refined model in (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.

If r_i is observed, (3) is a regular GLM and commonly used inference tools such as maximum likelihood can be applied for inferences about the model parameters. When

r_i is unobserved as in most real studies, (3) cannot be estimated using such standard methods. Next we discuss how to make inferences about (3) in the latter case.

2.2 *A Mixture Model*

We construct a model with a zero-inflated count predictor under the generalized linear regression model framework. Our mixture model consists of two components, one for modeling the outcome y and the other for modeling the zero-inflated count predictor.

Main Model: This component pertains to the model of primary interest. Given x_i, $\mathbf{z}_i$ and r_i, the outcome y_i follows some parametric distribution indexed by parameter vector $\alpha = (\alpha_1, \alpha_2, \beta)$:

$$y_i \mid x_i, r_i, \mathbf{z}_i \sim \text{i.d. } f, \quad \mu_i = E(y_i \mid x_i, \mathbf{z}_i, r_i) = g(\alpha_1 x_i + \alpha_2 r_i + \mathbf{z}_i^\top \beta), \quad 1 \le i \le n. \tag{4}$$

The link function $g(\cdot)$ can be specified depending on the type of the outcome y. For example, if y_i is continuous and normally distributed, we can choose the identity link function, then model (4) becomes

$$y_i \mid x_i, r_i, \mathbf{z}_i \sim \text{i.d. } N\left(\mu_i, \sigma^2\right), \quad \mu_i = \alpha_1 x_i + \alpha_2 r_i + \mathbf{z}_i^\top \beta, \quad 1 \le i \le n. \tag{5}$$

Inclusion of the indicator r_i for the risk, as a predictor in the Main Model, enables us to model the differential effects between structural and random zeros. There are two effects associated with the trait in the Main model. One is for the difference between structural zeros and random zeros, say the trait effect, measured by α_2. The other is the dosage effect of the count predictor for the at-risk subgroup, measured by α_1. The coefficient α_1 measures the change in y_i per unit increase in x_i within the at-risk group, which is the effect of the severity of the risk factor on the response for subjects who have such risk factor. Without including r_i, two effects are mixed together and hence may potentially provide biased and misleading conclusions. Our model can tease apart the two effects and hence can provide a more comprehensive relationship between the outcome and the trait.

Auxiliary Zero-inflated Model: This component models the zero-inflated predictor x_i. Because of the inflation of zeros from the non-risk subgroup, we model the count variable x_i with some zero-inflated count response models. For example, we may assume that x_i follows a popular ZIP distribution with the probability of being structural zero ρ_i and the Poisson mean μ_i, i.e., ZIP(ρ_i, μ_i). The Auxiliary ZIP model, ZIP$_x$, models both the structural zero and the Poisson count in x_i. We assume that $\mathbf{u}_i$ is a set of predictors for both ρ_i and μ_i. Although ρ_i and μ_i may depend on different sets of predictors, for notational brevity we assume a common set $\mathbf{u}_i$, which includes all the predictors for both components, but with different coefficients γ_1 and γ_2, i.e.:

$$x_i \mid \mathbf{u}_i \sim \text{i.d. ZIP}(\rho_i,\ \mu_i), \quad \rho_i = h_1(\mathbf{u}_i^T \gamma_1), \quad \mu_i = h_2(\mathbf{u}_i^T \gamma_2), \qquad 1 \le i \le n, \tag{6}$$

where $h_1(\cdot)$ and $h_2(\cdot)$ are the link functions for the structural zero component and the Poisson component. The predictors $\mathbf{u}_i$ may be different from or overlap with $\mathbf{z}_i$ in the Main Model (4). Other commonly used zero-inflated count response models such as zero-inflated negative binomial (ZINB) may also be adopted for the Auxiliary zero-inflated model.

The validity of the Main Model and the Auxiliary Model is given by the following assumptions:

Assumption A: Conditional Independence. Given $\mathbf{u}_i$, we assume that x_i and r_i are independent of $\mathbf{z}_i$, i.e.,

$$(x_i, r_i) \perp \mathbf{z}_i \mid \mathbf{u}_i.$$

This assumption implies that x_i and r_i may depend on the covariates $\mathbf{z}_i$, but the dependence is only through the predictors $\mathbf{u}_i$. This condition can be satisfied by including additional predictors from $\mathbf{z}_i$ in (4), as needed for the conditional independence, into $\mathbf{u}_i$ in (6) for the zero-inflated model of x_i.

Assumption B: Comprehensiveness of the Main Model. Given the predictors $x_i, \mathbf{z}_i, r_i$, y_i is independent of $\mathbf{u}_i$, i.e.,

$$y_i \perp \mathbf{u}_i \mid x_i, \mathbf{z}_i, r_i.$$

The assumption implies that y_i may depend on $\mathbf{u}_i$, but the dependence is only through $x_i, \mathbf{z}_i$ and r_i. This condition can always be satisfied by including additional predictors from $\mathbf{u}_i$ in (6) into $\mathbf{z}_i$ in (4).The comprehensiveness here means that all the information about y_i carried by or contained in $\mathbf{u}_i$ is captured by $x_i, \mathbf{z}_i$ and r_i through the Main Model.

In practice, we may choose a set of covariates $\mathbf{u}_i$ and $\mathbf{z}_i$ based on the subject matter of the study. As long as important predictors for the outcome y_i and the count x_i are included, the two assumptions should approximately true.

The proposed mixture model can be applied to different types of responses in the Main Model including continuous, categorical, count, and survival data and different models such as ZIP and ZINB for zero-inflated count data x_i in the Auxiliary Model. We discussed the linear regression model for the continuous response in (5). Below we illustrate the approach with some other common response variables for the Main Model.

2.2.1 Models for categorical responses

When y_i is binary, we may consider modeling the response in the Main Model using the following logistic regression:

$$\begin{aligned} & y_i \mid x_i, r_i, \mathbf{z}_i \sim \text{i.d. Bern}\left(\mu_i\right), \\ & \mu_i = E\left(y_i \mid x_i, \mathbf{z}_i, r_i\right) = \text{logit}^{-1}\left(\alpha_1 x_i + \alpha_2 r_i + \mathbf{z}_i^\top \beta\right), \quad 1 \le i \le n, \end{aligned} \tag{7}$$

where Bern(μ) denotes a Bernoulli with mean μ and $\text{logit}^{-1}(\cdot)$ denotes the inverse of the logit link. Alternatively, we may apply the probit, complementary log-log, or other commonly used link functions for the binary response y_i. Further, we can readily extend (7) to nominal or ordinal responses using the cumulative logistic or generalized logit models in [21].

2.2.2 Models for count responses

When y_i is a count response, Poisson and negative binomial (NB) regression models may be applied. For example, under a log-linear Poisson regression we may assume:

$$\begin{aligned} & y_i \mid x_i, r_i, \mathbf{z}_i \sim \text{i.d. Poisson}\left(\mu_i\right), \quad \log(u_i) = \alpha_1 x_i + \alpha_2 r_i + \mathbf{z}_i^\top \beta \\ & \mu_i = E(y_i \mid x_i, \mathbf{z}_i, r_i) = \exp(\alpha_1 x_i + \alpha_2 r_i + \mathbf{z}_i^\top \beta), \quad 1 \le i \le n, \end{aligned} \tag{8}$$

where Poisson(μ) denotes a Poisson with mean μ.

2.2.3 Models for zero-inflated count responses

If y_i is a zero-inflated count response itself, we may apply ZIP or ZINB to model the data and to account for structural zeros. For example, if using ZIP, we can apply a logistic model for the structural zero and loglinear model for the Poisson component of the response y_i in (4) as:

$$
\begin{aligned}
& y_i \mid x_i, \mathbf{z}_i, r_i \sim \text{i.d. ZIP}(\rho_i(\mathbf{v}_i; \theta_2), \mu_i(\mathbf{v}_i; \theta_2)), \quad 1 \le i \le n. \qquad (9) \\
& \rho_i(\mathbf{v}_i; \theta_1) = \text{logit}^{-1}\left(\alpha_1 x_i + \alpha_2 r_i + \mathbf{z}_i^\top \beta\right), \quad \mu_i(\mathbf{v}_i; \theta_2) = \exp\left(\alpha_1' x_i + \alpha_2' r_i + \mathbf{z}_i^\top \beta'\right) \\
& \theta_1 = \left(\alpha_1, \alpha_2, \beta^\top\right)^\top, \quad \theta_2 = \left(\alpha_1', \alpha_2', \beta'^\top\right)^\top.
\end{aligned}
$$

Note that as in the case of x_i, we assume a common set of explanatory variables $\mathbf{v}_i = \left(x_i, r_i, \mathbf{z}_i^\top\right)^\top$, but different coefficients θ_1 and θ_2 for the logistic and Poisson components of the ZIP.

In addition to the common types of response variables, this approach can be readily adapted to other response variables.

3. Statistical Inference

3.1 *Likelihood Function*

Since the indicator variable r_i is only partially observed, inferences cannot be made just based on the Main Model (4). Under Assumptions A and B, we can apply maximum likelihood for inference. For a subject with $x_i > 0$, note that $r_i = 0$, so the likelihood is:

$$
\begin{aligned}
L_{(x_i>0)} &= f(y_i, x_i, \mathbf{z}_i, \mathbf{u}_i) = f(y_i, x_i, \mathbf{z}_i, \mathbf{u}_i, r_i = 0) \\
&= f(y_i \mid x_i, \mathbf{z}_i, \mathbf{u}_i, r_i = 0) \Pr(x_i \mid \mathbf{z}_i, \mathbf{u}_i, r_i = 0) \Pr(r_i = 0 \mid \mathbf{z}_i, \mathbf{u}_i) f(\mathbf{z}_i, \mathbf{u}_i) \\
&= f(y_i \mid x_i, \mathbf{z}_i, r_i = 0) \Pr(x_i \mid \mathbf{u}_i, r_i = 0) \Pr(r_i = 0 \mid \mathbf{u}_i) f(\mathbf{z}_i, \mathbf{u}_i). \qquad (10)
\end{aligned}
$$

Here we use $f()$ as a generic notation for the (joint) likelihood for the variables in the parenthesis. So it will be the density function for continuous variables and mass probabilities for discrete and categorical variables.

For a subject with $x_i = 0$, since r_i is unknown, the likelihood can be expressed as:

$$
\begin{aligned}
L_{(x_i=0)} &= f(y_i, x_i = 0, \mathbf{z}_i, \mathbf{u}_i) = f(y_i, x_i = 0, \mathbf{z}_i, \mathbf{u}_i, r_i = 0) + f(y_i, x_i = 0, \mathbf{z}_i, \mathbf{u}_i, r_i = 1) \\
&= f(y_i \mid x_i = 0, \mathbf{z}_i, \mathbf{u}_i, r_i = 0) \Pr(x_i = 0, r_i = 0 \mid \mathbf{u}_i) f(\mathbf{z}_i, \mathbf{u}_i) \\
&\quad + f(y_i \mid \mathbf{z}_i, \mathbf{u}_i, r_i = 1) \Pr(x_i = 0, r_i = 1 \mid \mathbf{u}_i) f(\mathbf{z}_i, \mathbf{u}_i) \\
&= f(\mathbf{z}_i, \mathbf{u}_i) \left\{ f(y_i \mid x_i = 0, \mathbf{z}_i, r_i = 0) \Pr(x_i = 0 \mid r_i = 0, \mathbf{u}_i) \Pr(r_i = 0 \mid \mathbf{u}_i) \right. \\
&\quad \left. + f(y_i \mid x_i, \mathbf{z}_i, r_i = 1) \Pr(r_i = 1 \mid \mathbf{u}_i) \right\}. \qquad (11)
\end{aligned}
$$

In the above likelihood, $f(y_i \mid x_i, \mathbf{z}_i, r_i)$ can be computed based on (4), while $\Pr(x_i \mid \mathbf{u}_i, r_i)$

and $\Pr(r_i \mid \mathbf{u}_i)$ are provided by (6). For example, under (5) for a continuous y_i, we have:

$$f(y_i \mid x_i, \mathbf{z}_i, r_i) = \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{[y_i - \left(\alpha_1 x_i + \alpha_2 r_i + \mathbf{z}_i^\top \beta\right)]^2}{2\sigma^2} \right\}.$$

Under (7) for a binary y_i,

$$\Pr(y_i \mid x_i, \mathbf{z}_i, r_i) = \left[\frac{\exp\left(\alpha_1 x_i + \alpha_2 r_i + \mathbf{z}_i^\top \beta\right)}{1 + \exp\left(\alpha_1 x_i + \alpha_2 r_i + \mathbf{z}_i^\top \beta\right)} \right]^{y_i} \left[\frac{\exp\left(\alpha_1 x_i + \alpha_2 r_i + \mathbf{z}_i^\top \beta\right)}{1 + \exp\left(\alpha_1 x_i + \alpha_2 r_i + \mathbf{z}_i^\top \beta\right)} \right]^{1-y_i}$$

Under (6), if we assume:

$$h_1 = \log it^{-1}(\mathbf{u}_i^T \gamma_1), \quad h_2 = \log^{-1}(\mathbf{u}_i^T \gamma_2), \quad 1 \le i \le n,$$

then we have

$$\Pr(r_i = 1 \mid \mathbf{u}_i) = \rho_i(\mathbf{u}_i \gamma_1) = \frac{\exp\left(\mathbf{u}_i \gamma_1\right)}{1 + \exp\left(\mathbf{u}_i \gamma_1\right)},$$
$$\Pr(x_i \mid \mathbf{u}_i, r_i = 0) = \frac{\exp\left(-\exp\left(\mathbf{u}_i \gamma_2\right)\right) \exp\left(x_i \mathbf{u}_i \gamma_2\right)}{x_i!}.$$

By substituting $f(y_i \mid x_i, \mathbf{z}_i, r_i)$, $\Pr(r_i \mid \mathbf{u}_i)$ and $\Pr(x_i \mid \mathbf{u}_i, r_i = 0)$ into the likelihood functions $L_{(x_i>0)}$ and $L_{(x_i=0)}$ in (10) and (11), we can apply maximum likelihood methods for inferences about the parameters.

Note that as in standard regression analysis, the likelihood for each subject contains the joint distribution of $\mathbf{z}_i$ and $\mathbf{u}_i$. However, since $f(\mathbf{z}_i, \mathbf{u}_i)$ contains no parameter of primary interest, it provides no contribution to the score equations and thus can be factored out from the likelihood function.

Because we are mainly interested in the differential effects of structural zeros in the main model, we naturally adopted the "pattern mixture model" approach, which involves a formulation of $f(y_i, x_i, r_i, \mathbf{z}_i, \mathbf{u}_i) = f(y_i|x_i, \mathbf{z}_i, \mathbf{u}_i, r_i) \Pr(x_i|r_i, \mathbf{z}_i, \mathbf{u}_i) f(r_i, \mathbf{z}_i, \mathbf{u}_i)$**. However, we can also formulate the model following a "selection model" scheme. In the "selection model" scheme, the model would involve a selecting distribution** $\Pr(r_i|x_i, \mathbf{z}_i, \mathbf{u}_i)$**, and the likelihood is factored as** $f(y_i, x_i, r_i, \mathbf{z}_i, \mathbf{u}_i) = f(y_i|x_i, \mathbf{z}_i, \mathbf{u}_i, r_i) \Pr(r_i|x_i, \mathbf{z}_i, \mathbf{u}_i) f(x_i, \mathbf{z}_i, \mathbf{u}_i)$. Thus the likelihood for a subject with $x_i > 0$ (hence, $r_i = 0$) will be

$$f(y_i, x_i, \mathbf{z}_i, \mathbf{u}_i) = f(y_i|x_i, \mathbf{z}_i, \mathbf{u}_i, r_i = 0) \Pr(r_i = 0|x_i, \mathbf{z}_i, \mathbf{u}_i) f(x_i, \mathbf{z}_i, \mathbf{u}_i),$$

and the likelihood for a subject with $x_i = 0$ (hence, r_i can be 0 or 1) will be

$$f(y_i, x_i, \mathbf{z}_i, \mathbf{u}_i) = [f(y_i|x_i, \mathbf{z}_i, \mathbf{u}_i, r_i = 0) \Pr(r_i = 0|x_i, \mathbf{z}_i, \mathbf{u}_i)$$

$$+ f(y_i|x_i, \mathbf{z}_i, \mathbf{u}_i, r_i = 1) \Pr(r_i = 1|x_i, \mathbf{z}_i, \mathbf{u}_i)] f(x_i, \mathbf{z}_i, \mathbf{u}_i).$$

Under this formulation, the distribution $f(x_i, \mathbf{z}_i, \mathbf{u}_i)$ **does not need to be specified if a model (say logistic) is specified for** r_i **and** r_i **is observed. However, since the structural zeros are unobserved, a (logistic) model** r_i **would not be**

identifiable. We may rely on the zero-inflated models to model the structural zeros, i.e., use the auxiliary model to model r_i, and the pattern mixture approach we adopted above would be a natural choice in the situation.

3.2 *Hypothesis Testing*

As discussed above, there are two effects associated with the trait. One is for the trait effect, the difference between structural and random zeros, measured by α_2, and the other is for the dosage effect of the count predictor on the outcome for the at-risk subgroup, measured by α_1, in the Main Model. We may test them separately using common hypothesis testing techniques such as the Wald test and the likelihood ratio test. For an overall testing of whether the trait is associated with the outcome, a linear composite hypothesis of $\alpha_1 = \alpha_2 = 0$ needs to be tested.

3.3 *Selection of Initial Values*

Due to the complexity of the mixture model, we generally do not obtain closed-form ML estimates (MLEs) of the parameters. Numerical optimization is needed to find the MLEs, such as the popular Newton-Raphson (NR) method. In using the Newton-Raphson method, it is important to start with good initial values in order for the iterations to converge to the global maximum of the likelihood function. The following strategies can be used for setting initial values to achieve this objective as well as to speed up convergence for the Newton-Raphson method.

We first estimate the initial values of the parameters in (6) for the count predictor x_i, and then estimate the initial values for the parameters in (4). More specifically, we follow a two-step procedure to obtain initial values:

Step 1. Initial values for the regression parameters γ_1 and γ_2 in the Auxiliary Model in (6), as well as the marginal probability of structural zeros ρ_x and the marginal Poisson mean μ_x, for the count predictor x_i.

(a) Estimate the initial value of μ_x and ρ_x. We fit a ZIP model for x_i with intercept only. The estimated probability of structural zeros and the Poisson mean then serve as the initial values of μ_x and ρ_x, denote as μ_{x_I} and ρ_{x_I}, respectively.

(b) Estimate the initial value of γ_1 and γ_2. We fit a ZIP model for x_i with predictors $\mathbf{u}_i$. The estimated coefficients from the logistic regression for the structural zeros are then used as the initial value of γ_1, denote as γ_{1_I}, while the coefficients from the loglinear model serve as the initial value of γ_2, denoted as γ_{2_I}.

Step 2. Initial values for the parameters for the Main Model of y in (4). The difficulties of model estimation lie in the fact that structural zeros are not observed. However, since subjects with positive values of x_i are not structural zeros, i.e., $r_i = 0$ for these subjects. Thus we may apply regular regression methods to this subsample to obtain initial estimates of all the parameters except for the coefficient of r_i.

For example, if we have a regression model on y and

$$y_i \mid x_i, r_i, z_i \sim \text{i.d. } N\left(\mu_i, \sigma^2\right), \quad \mu_i = c_0 + c_x x_i + c_r r_i + c_z z_i.$$

we can apply the model

$$y_i \mid x_i, z_i \sim \text{i.d. } N\left(\mu_i, \sigma^2\right), \quad \mu_i = c_0 + c_x x_i + c_z z_i, \tag{12}$$

to the subjects with $x_i > 0$ to obtain the initial values for c_{0_I}, c_{x_I}, c_{z_I} and σ_I. We can then set the initial value of c_r based on the following equations:

$$\begin{aligned}E(y \mid x=0) &= \Pr(x \text{ is random zero})(c_{0_I} + c_{x_I} E(x \mid x=0)) \\ &\quad + \Pr(x \text{ is structural zero}) \cdot (c_{0_I} + c_r + c_{x_I} E(x \mid x=0)) \\ &= (c_{0_I} + c_{x_I} E(x \mid x=0)) + c_r \Pr(x \text{ is structural zero}) \\ &= (c_{0_I} + c_{x_I} E(x \mid x=0)) + c_r \frac{\rho_{x_I}}{\rho_{x_I} + (1-\rho_{x_I}) \cdot e^{-\mu_{x_I}}},\end{aligned}$$

so the initial value of c_r can be obtained by:

$$c_{r_I} = \frac{[E(y_i \mid x_i = 0) - c_{0_I}]\,[\rho_{x_I} + (1-\rho_{x_I}) \cdot e^{-\mu_{x_I}}]}{\rho_{x_I}}. \tag{13}$$

The initial value for c_r is obtained by comparing the mean response outcome for all the subjects with zero counts in x, including both structural and random zeros, to the intercept estimated from the subjects with $x_i > 0$ in (12). The choices of initial values depend on the models we use, and we will give some examples in the simulation section. Since ignoring the difference between structural and random zeros means the coefficient involving r is zero, thus it is also reasonable to use 0 as initial values for c_r.

4. Simulation Studies

4.1 *Simulation Setup*

We use simulation studies to examine the performance of the proposed method when modeling zero-inflated outcomes as predictors in regression analysis. We assume that the predictor $x \sim ZIP(\rho_x, \mu_x)$, a ZIP with ρ_x denoting the probability of structural zeros in the logistic component and μ_x denoting the mean of a count response in the Poisson component of the ZIP. A larger ρ_x means more structural zeros, while a larger μ_x indicates a smaller proportion of random zeros in the simulated data.

The predictor x is generated based on the following Auxiliary Model:

$$\begin{aligned} x \sim \text{i.d. ZIP}(\rho_x, \mu_x), \quad z_1 \sim \text{i.d. } N\left(0, \sigma_{z_1}^2\right), \\ \log it(\rho_x) = a_0 + a_1 z_1, \quad \log(\mu_x) = b_0 + b_1 z_1. \end{aligned} \tag{14}$$

The values of a_0, a_1, b_0 and b_1 control the amount of structural and random zeros in the predictor. We consider four different types of response: continuous, binary, Poisson and zero inflated Poisson y. To investigate the performance of the proposed method under different conditions for each type of outcomes, we consider three scenarios: a) when the structural zeros have effect on the outcome y, and the Main Model (4) correctly specifies the effect; b) when the structural zeros do have an effect on y, but the Main Model (4) is misspecified by not including the effect of structural zeros in the model, i.e., the difference between structural zeros and random zeros is ignored; c) when the structural zeros don't have effect on y, but the Main Model does include an effect of the structural zeros.

In all simulations, a Monte Carlo (MC) sample size of 1,000 is used for the models. We summarize results of model estimates by reporting point and variance estimates (both model-based obtained from the asymptotic theory and empirical estimates from MC replications), as well as the coverage probability of confidence intervals (probability whether the true value is covered by the confidence interval).

4.2 *Continuous Response Y*

For a continuous y, the association of y with x, z and r based on (5) is specified as follows:

$$y \mid x, r, z \sim \text{i.d. } N\left(\mu, \sigma_y^2\right), \ z \sim \text{i.d. } N\left(0, \sigma_z^2\right),$$
$$\mu = c_0 + c_x x + c_r r + c_z z, \tag{15}$$

and x is generated based on (14). For the simulation studies, we set $\sigma_y^2 = \sigma_z^2 = \sigma_{z_1}^2 = 1$, $c_0 = -1$, $c_x = c_z = 1$, $a_0 = b_0 = 0.5$, and $a_1 = b_1 = 1$. To see if the effect of structural zeros on the response has any impact on the estimates of other parameters such as c_x and c_z, we consider $c_r = 1$ and $c_r = 0$. When $c_r = 1$, we consider both the true Main Model (15) and a misspecified Main Model by excluding c_r in the model fitting, i.e., we fit a model on y as $y = c_0 + c_x x + c_z z$. The sample sizes considered were 200, 500 and 1000. As described above, we set the initial values of the parameters based on the discussion in Section 3.3. By applying the model (12) to the subsample with $x > 0$, we obtained the initial values for c_x and c_z, and the initial value c_r is set based on (13).

*** Table 1 goes about here ***

Shown in Table 1 and tables S1-S2 (in the supporting web material) are the averages of the estimates for both the Main and Auxiliary Model. In Table 1, the Main Model (15) is correctly specified, while in Table S1, the Main Model is misspecified. The results for $c_r = 0$ are provided in Table S2 as the supporting web material. As shown in Table 1 and S2, the estimates for both the Main Model and the Auxiliary Model are very close to the true values, the coverage probabilities are also very close to 95%, and the asymptotic variances are very close to the empirical variances. Table 1 and Table S2 also show that structural zeros do not have much impact on the estimates of other parameters such as c_x and c_z, as long as the Main Model is correctly specified. But when the Main Model is misspecified by not including the structural zeros in the model, as shown in Table S1, the estimates of c_x are quite biased, and the coverage probabilities are very low. The misspecification of the Main Model does not have a big impact on the estimates of c_z. Therefore, when the structural zeros do have effect on the outcome y, a model failing to include the structural zeros of the count variable x can't capture the true association between x and y, but the associations between the outcome y and other covariates z may not be affected much by the misspecification. The estimate of the intercept c_0 is biased.

4.3 *Binary Response Y*

For a binary outcome y, we simulate the data from a GLM for the Main Model with a logit link as follows:

$$y \mid x, r, z \sim \text{i.d. Bern}\,(p), \quad \text{logit}\,(p) = c_0 + c_x x + c_r r + c_z z. \tag{16}$$

The explanatory variables x and z are simulated the same way as in the continuous case. The values of the parameters are set to be the same as in the continuous case. A MC sample of 1000 replications is simulated for each of the sample sizes 200, 500 and 1000 using the same parameter values as in the case of a continuous y. The initial values for a_0, b_0, a_1 and b_1 are again determined by Section 3.3. For the initial values of c_x and c_z, we apply a logistic regression model to the subset of subjects with $x > 0$, i.e., c_{x_I} and c_{z_I} are estimated based on:

$$E(y \mid x, z) = \text{logit}^{-1}\left(c_0 + c_x x + c_z z\right), \quad x > 0. \tag{17}$$

After obtaining the initial values of c_x and c_z by applying Step 2 in Section 3.3, the initial value of c_r is estimated by:

$$c_{r_I} = \log \frac{A}{1-A} - c_{0_I} - c_{x_I} E(x \mid x = 0),$$

where

$$A = \frac{\Pr(y = 1 \mid x = 0) - (1 - \rho_{x_I}) e^{-\mu_{x_I}} \log it^{-1} \left[c_{0_I} - c_{x_I} E(x \mid x = 0) \right]}{\rho_{x_I}}.$$

*** Table 2 goes about here ***

The simulation results are summarized in Table 2, S3 and S4. As shown in Table 2 and S4, when the Main Model (16) is correctly specified, all the estimates are very good, even for a relatively small sample size. As the sample size increases from 200 to 1000, the point estimates are closer to the true value. When the Main Model (16) is misspecified, as shown in Table S3, the estimates of c_x become quite biased, although other parameters except for c_0 are all estimated quite well.

4.4 *Poisson Count Response Y*

For a Poisson count variable y, we generate y from a GLM with a log function as follows:

$$y \mid x, r, z \sim \text{i.d. Poisson}(\mu), \quad \mu = \exp(c_0 + c_x x + c_r r + c_z z). \tag{18}$$

With the same set of values of the parameters as in the continuous case, we simulate 1,000 MC samples from each of the three sample sizes considered.

The initial values of the estimates of μ_x and ρ_x are determined by the same algorithm as in the previous cases. In order to obtain a proper initial value of c_0, c_x and c_z, we fit the following Poisson to the subsample with $x > 0$:

$$y \mid x, z \sim \text{i.d. Poisson}(\mu), \quad \mu = \exp(c_{0_I} + c_{x_I} x + c_{z_I} z), \quad x > 0,$$

with the initial values $\mu_{x_I}, \rho_{x_I}, c_{0_I}, c_{x_I}$ and c_{z_I}. We estimate an initial value of c_r using the following estimating equations:

$$\begin{aligned} E(y \mid x = 0) &= \Pr(x \text{ is random zero}) \cdot e^{c_{0_I} + c_{z_I} * E(z|x=0)} \\ &\quad + \Pr(x \text{ is structural zero}) \cdot e^{c_{0_I} + c_r + c_{z_I} * E(z|x=0)}, \\ c_r &= \log \left(\frac{E(y \mid x = 0) - (1 - \rho_{x_I}) e^{-\mu_{x_I}} \cdot e^{c_{0_I} + c_{z_I} * E(z|x=0)}}{\rho_{x_I} \cdot e^{c_{0_I} + c_{z_I} * E(z|x=0)}} \right) \end{aligned}$$

The simulation results are summarized in Tables S5, S6 and S7 as the supporting web material. Similar to the continuous and binary cases, all estimates are quite close to the true values when the Main Model (18) is correctly specified. But when the Main Model is misspecified, as shown in Table S6, the estimates are biased and the coverage probabilities are very small as well. Again, misspecification of the Main Model does not have much impact on the Auxiliary Model.

4.5 *Zero-inflated Poisson Response Y*

Finally, we consider a zero-inflated count response y generated from the following ZIP model:

$$\begin{aligned} & y \mid \mathbf{v} \sim \text{i.d. ZIP}(\rho(\mathbf{v};\theta_2), \mu(\mathbf{v};\theta_2)), \\ & \rho = \text{logit}^{-1}\left(c_0 + c_x x + c_r r + c_z z\right), \quad \mu = \exp\left(c_0' + c_x' x + c_r' r + c_z' z\right), \end{aligned} \tag{19}$$

where $\mathbf{v} = (x, r, z)^\top$. We set $c_0' = c_0 = -1$ and $c_x' = c_x = c_r' = c_r = c_z' = c_z = 1$. Since the latent nature of ZIP requires a larger sample size to obtain reliable estimates, especially within the context of a latent x following another ZIP, we consider bigger sample sizes 500, 1000 and 1500 for each case.

Again, we determine the initial values for estimating μ_x and p_x as discussed in Section 3.3. For the initial values of c_0, c_x, c_z of the logistic component and c_0', c_x', c_z' of the loglinear component of the ZIP for y, we apply the following models for the subsample with $x > 0$:

$$\begin{aligned} & y \mid x, z \sim \text{i.d. ZIP}(\rho(x, z; \eta_1), \mu(x, z; \eta_2)), \\ & \nu = \text{logit}^{-1}\left(c_0 + c_x x + c_z z\right), \quad \log(\mu) = c_0' + c_x' x + c_z' z, \\ & \eta_1 = (c_0, c_x, c_z)^\top, \quad \eta_2 = \left(c_0', c_x', c_z'\right)^\top. \end{aligned}$$

Due to the complexity of the model, we set $c_r = c_r' = 0$ as the initial values for estimating c_r and c_r'.

*** Table 3 goes about here ***

Shown in Tables 3, S8 and S9 are the simulation results. When the Main Model (19) is correctly specified, as shown in Table 3 and S9, the estimates from both the log-linear Poisson component and the logistic zero-inflated component of the Main Model are very good. The point estimates are close to the true values and the coverage probabilities are close to 95%. The asymptotic variances are also quite close to their corresponding empirical counterparts. But when the Main Model is misspecified, as shown in Table S8, the estimates from both components of the Main Model are biased, especially for the estimates of c_{xp} and c_{xb}. This indicates that when the outcome follows a ZIP model and has zero-inflated count predictor x, if the difference between the structural and random zeros in the predictor is ignored, the Main Model can detect neither the true relationship between y and x, nor the associations between y and other covariates. Comparing to the two components of the Main Model, the estimates of c_{zb} in the zero-inflated component are relatively better than the estimates of c_{zp} in the Poisson component. The misspecification of the Main Model do not have much impact on the performance of the Auxiliary Model.

5. Real Data Analysis

5.1 *The Data*

We now use the 2009-2010 National Health and Nutrition Examination Survey (NHANES) study discussed in [11] as an illustrative example of a real study application. The NHANES is a survey research program conducted by the National Center for Health Statistics to assess the health and nutritional status of people in the United States.

A brief introduction of the study and a more detailed description of the NHANES data can be found in [11]. In NHANCE study, alcohol use is measured by the number of days of alcohol consumption (DAD) in a week, while depressive symptoms are assessed by the Patient Health Questionnaire (PHQ-9). As discussed in [11], both DAD and PHQ-9 have excessive zeros in their distributions. By fitting a zero-inflated Poisson (ZIP) model, we revealed that the DAD outcome has excessive zeros and the structural zeros in DAD were as high as 30%. Note that for illustrative purpose, in all these analysis and the following analysis, we ignore the complex survey study design of NHANES and did not incorporate the sampling weight in the analysis.

We apply the proposed approach to examine potential differential rates of depression between the at- and non-risk subgroups of alcohol use. In the proposed method, the essential component is to tease apart the effect of alcohol use, a trait of an individual, from the effect of amount of alcohol use, when modelling the relationship between alcohol use and depression. One of the unique features of the NHANES is the inclusion of the variable "NeverDrink", which measures lifetime abstinence from alcohol. This variable asks if a subject has ever used alcohol in his/her life. It is not a perfect indicator of structural zero in our context, since subjects who have used alcohol but became abstinent from it (structural zero) are not be counted as Never Drinkers. Nonetheless, the variable "NeverDrink" may serve at least as a crude benchmark to examine the performance of the proposed approach. Regarding the Main predictor DAD, we want to know if there are any demographic information to predict DAD.

5.2 *Statistical Model*

We apply the approach to model the effect of alcohol use on PHQ-9 score. For the PHQ-9 score, we applied a ZIP, with age, race, gender, education, and DAD as well as the indicator of structural zeros of DAD as the explanatory variables. Since our initial univariate analysis of DAD vs. PHQ-9 suggested a quadratic association between the two variables, a square of DAD (DAD^2) was also included as a predictor. We also consider a ZIP model for the DAD variable by including age, race, gender, education as predictors for both components of the DAD variable. So, our model (Model I) to study the effect of alcohol use on depression is specified as follows:

$$\begin{aligned} \text{PHQ-9}_i &\sim \text{ZIP}(\rho_i,\ \mu_i), \quad \text{DAD} \ \sim \text{ZIP}(\rho_{xi},\ \mu_{xi}), \qquad (20)\\ \rho_i &\sim \text{Structural zero of DAD} + \text{DAD}+\text{DAD}^2\ + \text{age}+\text{gender}+\text{race}+\text{education},\\ \mu_i &\sim \text{Structural zero of DAD} + \text{DAD}+\text{DAD}^2\ + \text{age}+\text{gender}+\text{race}+\text{education},\\ \rho_{xi} &\sim \text{age}+\text{gender}+\text{race}+\text{education},\\ \mu_{xi} &\sim \text{age}+\text{gender}+\text{race}+\text{education}, \end{aligned}$$

where ρ_{xi} is the probability for structural zeros and μ_{xi} is the Poisson mean of the DAD variable.

We apply the maximum likelihood method discussed in Section 2.2 to make inference about the parameters for the model in (20). The coefficients of the structural zeros of DAD indicate the effect of a trait of an individual for alcohol use on the depressive symptoms, while those of DAD and DAD^2 provide the effect of amount of drinking on this response for subjects in the at-risk group of alcohol use. We also apply a ZIP model to model the PHQ-9 score with exactly the same explanatory variables, except that the indicator of structural zeros of DAD in (20) is replaced by the variable "NeverDrink". The second ZIP model (Model II) does not involve the latent variable of structural zeros of DAD, providing a benchmark to assess the performance of the proposed approach. We used SAS 9.3 PROC GENMOD for the analyses with inference based on the maximum likelihood approach. Unlike Model I, Model II do not include a ZIP Auxiliary Model for the predictor of DAD.

5.3 *Results*

Due to some missing values, the actual sample size for the analysis is 5,261 (out of 5,283 subjects in the data). Shown in Table 4 are the parameter estimates for the logistic and Poisson components of the ZIP Models I (Model II) for the PHQ9 score. Both models have successfully identified significant associations between alcohol use and depression in both components. In the logistic component which models the likelihood of non-depression (structural zeros of PHQ-9 score), the non-drinkers are more likely of being non-risk for depression, or less likely of being at-risk for depression (p-value 0.0142 for Model I and <0.0001 for Model II). Among the at-risk subgroup for alcohol use, the coefficients for DAD^2 are significant in both models (p-value <0.0001 and 0.0006 for Model I and II, respectively). The negative signs of these coefficients indicate that subjects with DAD at the two ends, near 0 (few days of alcohol use) or near 7 (most days of alcohol use), are at higher risk for depressive symptoms. Based on Model I (II), subjects with 2.70 (2.04) days of any alcohol use per week are least likely to be depressed.

For the Poisson component, the non-drinkers have less depressive symptoms based on both models (p-value <0.0001 for both models). Among the subjects who are at-risk for alcohol use, the coefficients of DAD^2 are again significant in both models (p-value <0.0001 for both models). The positive signs of these coefficients indicate that subjects with DAD at the two ends near 0 and 7 have higher PHQ-9 scores. Based on Model I (II), subjects with 3.55 (2.82) days of alcohol use per week have lowest PHQ-9 scores.

The results for the Auxiliary Model of DAD are summarized in Table S10. Gender, age and education are significant predictors for both the Poisson and logistic components, older males with higher education are more likely to have more drinks and older females with lower educations are more likely to be in the non-risk group for alcohol drinking; Compared to people in other race, Mexican American and Non-Hispanic are more likely to be at-risk and also have more drinking if they are at risk for drinking.

Regarding the relationship between the alcohol drinking and depression, both models yield similar conclusions, although Model I models the latent trait of alcohol use, while Model II uses the observed measure of this trait when examining the effects of alcohol use on depression. Abstinence from alcohol or moderate alcohol consumption are protective for depression. However, there is some discrepancy in the estimates between Model I and Model II. Our estimated percentage of structural zeros (non-risk group) is 35.0%, while the percent of structural zeros based on the NeverDrink variable is only 12.0%. Since this NeverDrink variable asks if subjects have any drink in their lifetime, those who don't drink, but become abstinent from alcohol, are treated as structural zeros (non-drinkers) in the proposed approach (Model I). In contrast, such individuals are regarded as part of the at-risk subgroup in Model II. So the difference in the percent of structural zeros between the two approaches likely reflects the different interpretations of lifetime abstinence from alcohol.

6. Discussions

Zero-inflation, the observed amount of zeros is larger than that would be expected under a statistical model, is a common phenomenon in public health

and medical research, and it is often associated with the existence of structural zeros. It is important both statistically and conceptually to distinguish random and structural zeros. However, structural zeros are often latent and information about whether a zero is structural or random is often not be observed directly.

A comparatively rich literature has been focusing on statistical methodology research and their applications in addressing the structural zero issue when the count variable is treated as the response. Little attention is paid when such count variables are treated as predictors. In such cases, simply ignoring the differential effects of structural zeros in the data analyses may yield biased estimates and uninterpretable findings [11]. In this paper, we have developed statistical models to address the differential effect of structural zeros and random zeros in such explanatory variables.

The proposed approach fills a critical gap in literature to address the structural zero issues in predictors by jointly modeling the response of interest (Main Model) and the zero-inflated count predictors (Auxiliary Model). To tease out the effect of structural zeros from that of random zeros, an indicator of structural zeros, which is partially latent, is included in the Main Model to address the confounding effects of the two types of zeros. Validity of the zero-inflated model for the count predictor is critical for the application of the method.

We described four popular types of responses for the Main Model and two types of count predictors for the Auxiliary Model in this paper. In the proposed approach, we assumed conditional independence for both components, or equivalently, there is no confounder in both the Main and Auxiliary Models. Such assumptions are standard in regression analysis. The approach is easy to implement using popular statistical packages such as R and SAS. Like any mixture models, initial values are important for finding the maximum likelihood estimates. Based on our experience, the two-step procedure works quite well for selecting satisfactory initial values for computing the estimates. Also, our simulations and real data study examples have shown good performances of the approach.

Like any statistical method, the proposed approach also has some limitations. The method discussed in this paper is only applicable to cross-sectional studies. Further research is needed to extend the approach to longitudinal studies. Since it is premised upon parametric distribution assumptions, the approach lacks robustness against departures from assumed parametric models. Semiparametric approaches are needed for both cross-sectional and longitudinal studies to address such limitations.

Acknowledgements

This work was supported by the NIH under grants R33 DA027521, R01GM108337, R01HD075635 and a Faculty Research Support Grant from School of Nursing, University of Rochester.

The authors thank Professor Xin Tu and the reviewers for their comments and suggestions.

References

[1] D. BoÈhning, *Zero-inflated poisson models and ca man: A tutorial collection of evidence*, Biometrical Journal 40 (1998), pp. 833–843.

[2] A. Buu, N. Johnson, R. Li, and X. Tan, *New variable selection methods for zero-inflated count data with applications to the substance abuse field*, Statistics in medicine 30 (2011), pp. 2326–2340.

[3] D.A. Calsyn, M. Hatch-Maillette, S. Tross, S.R. Doyle, P. Crits-Christoph, Y.S. Song, J.M. Harrer, G. Lalos, and S.B. Berns, *Motivational and skills training hiv/sexually transmitted infection sexual risk reduction groups for men*, Journal of Substance Abuse Treatment 37 (2009), pp. 138 – 150.

[4] J. Connor, K. Kypri, M. Bell, and K. Cousins, *Alcohol outlet density, levels of drinking and alcohol-related harm in new zealand: a national study*, Journal of epidemiology and community health 65 (2011), pp. 841–846.

[5] J. Cranford, R. Zucker, J. Jester, L. Puttler, and H. Fitzgerald, *Parental alcohol involvement and adolescent alcohol expectancies predict alcohol involvement in male adolescents.*, Psychology of Addictive Behaviors; Psychology of Addictive Behaviors 24 (2010), pp. 386–396.

[6] A. Fernandez, M. Wood, R. Laforge, and J. Black, *Randomized trials of alcohol-use interventions with college students and their parents: lessons from the transitions project*, Clinical Trials 8 (2011), pp. 205–213.

[7] S. Gurmu and P.K. Trivedi, *Excess zeros in count models for recreational trips*, Journal of Business & Economic Statistics 14 (1996), pp. 469–477.

[8] G. Hagger-Johnson, B. Bewick, M. Conner, D. O'Connor, and D. Shickle, *Alcohol, conscientiousness and event-level condom use*, British journal of health psychology 16 (2011), pp. 828–845.

[9] D. Hall and Z. Zhang, *Marginal models for zero inflated clustered data*, Statistical Modelling 4 (2004), pp. 161–180.

[10] D.B. Hall, *Zero-inflated Poisson and binomial regression with random effects: A case study*, Biometrics 56 (2000), pp. 1030–1039.

[11] H. He, W. Wang, P. Crits-Christoph, R. Gallop, W. Tang, D. Chen, and X. Tu, *On the implication of structural zeros as independent variables in regression analysis: applicatoin to alcohol research*, Journal of Data Science 12 (2014), pp. 439–460.

[12] C. Hernandez-Avila, C. Song, L. Kuo, H. Tennen, S. Armeli, and H. Kranzler, *Targeted versus daily naltrexone: secondary analysis of effects on average daily drinking*, Alcoholism: Clinical and Experimental Research 30 (2006), pp. 860–865.

[13] T. Hildebrandt, B. McCrady, E. Epstein, S. Cook, and N. Jensen, *When should clinicians switch treatments? an application of signal detection theory to two treatments for women with alcohol use disorders*, Behaviour research and therapy 48 (2010), pp. 524–530.

[14] N. Horton, J. Bebchuk, C. Jones, S. Lipsitz, P. Catalano, G. Zahner, and G. Fitzmaurice, *Goodness-of-fit for GEE: An example with mental health service utilization*, Statistics in Medicine 18 (1999), pp. 213–222.

[15] K. Hur, D. Hedeker, W. Henderson, S. Khuri, and J. Daley, *Modeling clustered count data with excess zeros in health care outcomes research*, Health Services and Outcomes Research Methodology 3 (2002), pp. 5–20.

[16] D. Lambert, *Zero-inflated poisson regression, with an application to defects in manufacturing*, Technometrics 34 (1992), pp. 1–14.

[17] S.P. Miaou, *The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions*, Accident Analysis & Prevention 26 (1994), pp. 471–482.

[18] D. Neal, D. Sugarman, J. Hustad, C. Caska, and K. Carey, *It's all fun and games... or is it? Collegiate sporting events and celebratory drinking*, Journal of Studies on Alcohol and Drugs 66 (2005), pp. 291–294.

[19] I. Ozmen and F. Famoye, *Count regression models with an application to zoological data containing structural zeros*, Journal of Data Science 5 (2007), pp. 491–502.

[20] D. Pardini, H. White, and M. Stouthamer-Loeber, *Early adolescent psychopathology as a predictor of alcohol use disorders by young adulthood*, Drug and alcohol dependence 88 (2007), pp. S38–S49.

[21] W. Tang, H. He, and X. Tu, *Applied Categorical and Count Data Analysis*, Chapman & Hall/CRC, Boca Raton, 2012.

[22] A.H. Welsh, R.B. Cunningham, C. Donnelly, and D.B. Lindenmayer, *Modelling the abundance of rare species: statistical models for counts with extra zeros*, Ecological Modelling 88 (1996), pp. 297–308.

[23] M.H. Wilde, H.F. Crean, J.M. McMahon, M.V. McDonald, W. Tang, J. Brasch, E. Fairbanks, S. Shah, and F. Zhang, *Testing a model of self-management of fluid intake in community-residing long-term indwelling urinary catheter users*, Nursing research 65 (2016), pp. 97–106.

[24] Q. Yu, R. Chen, W. Tang, H. He, R. Gallop, P. Crits-Christoph, J. Hu, and X. Tu, *Distribution-free models for longitudinal count responses with overdispersion and structural zeros*, Statistics in medicine 32 (2012), pp. 2390–2405.

Appendix

See the Web-based Supplementary Materials.

Table 1. Mean estimated parameters (mean asymptotic variance, simulation variance) and the coverage probabilities (CP) (%) over 1000 realizations for the continuous response. The variances are in 10^{-3} for c_x and c_z, 10^{-1} for a_1, and 10^{-2} for the other estimates.

The estimates of the parameters for the Main Model:

N	$c_0 = -1$ Est.	CP	$c_x = 1$ Est.	CP	$c_r = 1$ Est.	CP	$c_z = 1$ Est.	CP
200	-1.000 (3.34 3.38)	93.1	1.000 (4.50 4.23)	95.0	0.998 (4.69 4.76)	95.3	1.003 (5.37 5.47)	94.0
500	-0.999 (1.28 1.31)	93.8	1.000 (1.56 1.63)	94.1	1.001 (1.81 1.83)	94.5	1.000 (2.15 2.25)	94.6
1000	-1.001 (0.62 0.61)	95.0	1.000 (0.73 0.76)	94.8	1.000 (0.89 0.88)	94.3	0.998 (1.07 1.04)	94.3

The estimates of the parameters for the Auxiliary Model:

N	$a_0 = 0.5$ Est.	CP	$a_1 = 1$ Est.	CP	$b_0 = 0.5$ Est.	CP	$b_1 = 1$ Est.	CP
200	0.478 (5.54 5.77)	95.5	1.062 (0.97 1.18)	93.8	0.488 (1.50 1.58)	94.4	1.011 (1.60 1.65)	94.7
500	0.492 (2.05 2.10)	95.0	1.016 (0.33 0.32)	96.3	0.494 (0.58 0.59)	95.4	1.003 (0.56 0.54)	96.9
1000	0.495 (1.01 1.02)	95.1	1.014 (0.16 0.18)	94.2	0.499 (0.29 0.27)	95.5	1.001 (0.26 0.27)	94.9

Table 2. Mean estimated parameters (mean asymptotic variance, simulation variance) and the coverage probabilities (CP) (%) over 1000 realizations for the binary response. The variances are in 10^{-2} for c_z, a_0, b_0 and b_1, 10^{-1} for the other estimates.

The estimates of the parameters for the Main Model:

	$c_0 = -1$		$c_x = 1$		$c_r = 1$		$c_z = 1$	
N	Est.	CP	Est.	CP	Est.	CP	Est.	CP
200	-1.118 (4.03 4.37)	95.4	1.101 (1.65 1.78)	96.2	1.129 (5.25 5.91)	94.3	1.045 (4.34 4.27)	95.5
500	-1.048 (1.33 1.39)	94.9	1.040 (0.52 0.56)	95.1	1.058 (1.80 1.90)	95.3	1.022 (1.62 1.69)	94.7
1000	-1.022 (0.62 0.65)	94.8	1.019 (0.24 0.25)	96.0	1.020 (0.86 0.86)	96.3	1.007 (0.79 0.81)	94.2

The estimates of the parameters for the Auxiliary Model:

	$a_0 = 0.5$		$a_1 = 1$		$b_0 = 0.5$		$b_1 = 1$	
N	Est.	CP	Est.	CP	Est.	CP	Est.	CP
200	0.468 (6.90 7.06)	95.8	1.081 (1.27 1.54)	93.9	0.488 (1.68 1.71)	94.4	1.013 (1.75 1.80)	94.5
500	0.488 (2.56 2.55)	94.7	1.023 (0.44 0.41)	96.5	0.493 (0.66 0.66)	95.4	1.004 (0.62 0.58)	96.2
1000	0.494 (1.25 1.29)	93.2	1.015 (0.21 0.23)	94.0	0.499 (0.33 0.32)	95.0	1.001 (0.29 0.31)	95.1

Table 3. Mean estimated parameters (mean asymptotic variance, simulation variance) and the coverage probabilities (CP) (%) over 1000 realizations for the ZIP response. The variances are in 10^{-2} for c_{zp}, a_0, b_0 and b_1, 10^{-1} for the other estimates except for c_{rb} and c_{0b}.

The estimates of the parameters for the poisson component of the Main Model:

	$c_{0p} = -1$		$c_{xp} = 1$		$c_{rp} = 1$		$c_{zp} = 1$	
N	Est.	CP	Est.	CP	Est.	CP	Est.	CP
200	-1.078 (3.29 3.86)	91.3	1.024 (0.99 1.27)	93.9	1.031 (3.53 4.01)	93.2	1.029 (2.86 3.18)	93.0
500	-1.031 (0.93 0.90)	94.3	1.016 (0.22 0.21)	94.3	1.012 (0.98 0.96)	94.3	1.009 (0.95 0.99)	94.5
1000	-1.011 (0.39 0.41)	94.4	1.003 (0.07 0.08)	93.9	1.000 (0.41 0.45)	93.6	1.004 (0.43 0.46)	93.7

The estimates of the parameters for the zero-inflated component of the Main Model:

	$c_{0b} = -1$		$c_{xb} = 1$		$c_{rb} = 1$		$c_{zb} = 1$	
N	Est.	CP	Est.	CP	Est.	CP	Est.	CP
200	-1.412 (1.80 1.78)	95.5	1.234 (4.75 5.15)	96.8	1.254 (1.65 1.71)	95.7	1.216 (2.92 3.33)	95.1
500	-1.194 (0.49 0.48)	96.2	1.113 (1.17 1.17)	95.4	1.140 (0.47 0.49)	96.4	1.085 (0.79 0.82)	94.8
1000	-1.076 (0.21 0.21)	94.8	1.051 (0.49 0.51)	95.4	1.038 (0.21 0.21)	95.1	1.036 (0.36 0.34)	96.0

The estimates of the parameters for the Auxiliary Model:

	$a_0 = 0.5$		$a_1 = 1$		$b_0 = 0.5$		$b_1 = 1$	
N	Est.	CP	Est.	CP	Est.	CP	Est.	CP
200	0.464 (6.89 6.78)	95.4	1.102 (1.28 1.50)	94.4	0.486 (1.67 1.68)	95.1	1.014 (1.79 1.8)	94.5
500	0.490 (2.52 2.49)	95.3	1.033 (0.43 0.41)	96.4	0.493 (0.65 0.65)	95.7	1.002 (0.64 0.6)	96.5
1000	0.494 (1.23 1.32)	94.2	1.024 (0.21 0.24)	93.3	0.498 (0.32 0.31)	95.5	0.998 (0.3 0.33)	94.8

Table 4. Mean estimated parameters (mean asymptotic variance, simulation variance) and the coverage probabilities (CP) (%) over 1000 realizations for the ZIP response when the Main Model is misspecified. The variances are in 10^{-2} for c_{zp}, a_0, b_0 and b_1, 10^{-1} for the other estimates except for c_{0b}.

The estimates of the parameters for the poisson component of the Main Model:

	$c_{0p}=-1$		$c_{xp}=1$		$c_{zp}=1$	
N	Est.	CP	Est.	CP	Est.	CP
200	-0.214 (0.38 0.47)	4.3	0.609 (0.38 0.70)	29.5	0.969 (2.46 3.35)	90.6
500	-0.211 (0.14 0.16)	0.0	0.676 (0.08 0.17)	2.9	0.947 (0.82 1.15)	85.0
1000	-0.210 (0.07 0.09)	0.0	0.699 (0.03 0.08)	0.1	0.946 (0.38 0.58)	78.1

The estimates of the parameters for the zero-inflated component of the Main Model:

	$c_{0b}=-1$		$c_{xb}=1$		$c_{zb}=1$	
N	Est.	CP	Est.	CP	Est.	CP
200	-0.230 (0.24 0.26)	52.0	0.730 (1.67 1.36)	79.4	1.038 (2.07 2.47)	92.5
500	-0.162 (0.08 0.07)	16.9	0.704 (0.33 0.24)	56.7	0.945 (0.63 0.67)	92.7
1000	-0.149 (0.04 0.04)	1.9	0.701 (0.16 0.11)	33.2	0.916 (0.30 0.30)	90.1

The estimates of the parameters for the Auxiliary Model:

	$a_0=0.5$		$a_1=1$		$b_0=0.5$		$b_1=1$	
N	Est.	CP	Est.	CP	Est.	CP	Est.	CP
200	0.468 (7.33 7.47)	94.9	1.095 (13.70 1.60)	94.3	0.486 (1.75 1.82)	95.1	1.013 (1.86 1.92)	93.9
500	0.488 (2.72 2.74)	95.3	1.036 (4.80 0.40)	96.6	0.493 (0.68 0.68)	95.5	1.003 (0.66 0.61)	96.6
1000	0.493 (1.33 1.42)	93.7	1.025 (2.30 0.30)	93.5	0.498 (0.34 0.33)	94.8	0.998 (0.32 0.35)	94.1

Table 5. Comparison of estimates (Estimate), standard errors (Std Err) and p-values (P-value) from the ZIP models of depresson for the 2009-2010 NHANES study

The estimates of the poisson component of the depresson:

Parameter		Model I			Model II		
		Estimate	Std Err	P-value	Estimate	Std Err	P-value
Intercept		2.7791	0.0635	<.0001	2.3457	0.0498	<.0001
NeverDrink	Yes vs. No	-1.3195	0.0267	<.0001	-0.1265	0.0256	<.0001
DAD		-0.5246	0.0177	<.0001	-0.1316	0.0145	<.0001
DAD^2		0.0739	0.0028	<.0001	0.0233	0.0025	<.0001
Gender	Male vs. Female	-0.1822	0.0197	<.0001	-0.1895	0.0165	<.0001
AGE		0.0007	0.0006	0.2540	-0.0032	0.0005	<.0001
Race/Ethnicity	Mexican American	-0.1724	0.0521	0.0009	-0.0954	0.0407	0.0192
	Other Hispanic	-0.0260	0.0537	0.6287	-0.0336	0.0425	0.4288
	Non-Hispanic White	-0.1571	0.0479	0.0010	-0.0889	0.0376	0.0181
	Non-Hispanic Black	-0.0212	0.0507	0.6756	0.0090	0.0401	0.8228
	NOther Race	0.0000	0.0000	.	0.0000	0.0000	.
Education		-0.1320	0.0081	<.0001	-0.1202	0.0068	<.0001

The estimates of the zero-inflated component of the depression:

Parameter		Model I			Model II		
		Estimate	Std Err	P-value	Estimate	Std Err	P-value
Intercept		-2.1316	0.2322	<.0001	-1.9935	0.2016	<.0001
NeverDrink	Yes vs. No	0.3959	0.1615	0.0142	0.4755	0.0970	<.0001
DAD		0.4050	0.1026	0.0001	0.1602	0.0591	0.0067
DAD^2		-0.0751	0.0168	<.0001	-0.0392	0.0114	0.0006
Gender	Male vs. Female	0.5152	0.0677	<.0001	0.5684	0.0645	<.0001
AGE		0.0159	0.0020	<.0001	0.0165	0.0018	<.0001
Race/Ethnicity	Mexican American	0.0626	0.1708	0.7138	0.0786	0.1597	0.6224
	Other Hispanic	-0.1192	0.1808	0.5097	-0.0971	0.1695	0.6012
	Non-Hispanic White	-0.2984	0.1580	0.0589	-0.2402	0.1473	0.1028
	Non-Hispanic Black	0.0114	0.1675	0.9457	0.0488	0.1564	0.7549
	NOther Race	0.0000	0.0000	.	0.0000	0.0000	.
Education		0.0184	0.0280	0.5095	0.0415	0.0262	0.1136