Published in final edited form as:

Stat Med. 2015 October 30; 34(24): 3235–3245. doi:10.1002/sim.6560.

On Performance of Parametric and Distribution-free Models for Zero-inflated and Over-dispersed Count Responses

W. Tang¹, N. Lu³, T. Chen¹, W. Wang¹, D. Gunzler⁴, Y. Han¹, and X.M. Tu^{1,2,5}

¹Department of Biostatistics and Computational Biology, University of Rochester

²Department of Psychiatry, University of Rochester

³Department of Management, Harbin Institute of Technology, Harbin, P.R.China

⁴Center for Health Care Research & Policy, Case Western Reserve University

⁵Center of Excellence for Suicide Prevention, Canandaigua VA Medical Center

Summary

Zero-inflated Poisson (ZIP) and negative binomial (ZINB) models are widely used to model zero-inflated count responses. These models extend the Poisson and Negative Binomial (NB) to address excessive zeros in the count response. By adding a degenerate distribution centered at 0 and interpreting it as describing a non-risk group in the population, the ZIP (ZINB) models a two-component population mixture. As in applications of Poisson and NB, the key difference between ZIP and ZINB is the allowance for overdispersion by the ZINB in its NB component in modeling the count response for the at-risk group. Overdispersion arising in practice too often does not follow the NB and applications of ZINB to such data yields invalid inference. If sources of overdispersion are known, other parametric models may be used to directly model the overdispersion. Such models too are subject to assumed distributions. Further, this approach may not be applicable if information about the sources of overdispersion is unavailable.

In this paper, we propose a distribution-free alternative and compare its performance with these popular parametric models as well as a moment-based approach proposed by Yu et al. [*Statistics in Medicine* 2013; **32**: 2390-2405]. Like the generalized estimating equations (GEE), the proposed approach requires no elaborate distribution assumptions. Compared with the approach of Yu et al., it is more robust to overdispersed zero-inflated responses. We illustrate our approach with both simulated and real study data.

Keywords

functional response models; generalized estimating equations; zero-inflated Poisson; zero-inflated Poisson with random effects; zero-inflated negative binomial; population mixtures

1 Introduction

Zero-inflated Poisson (ZIP) and Negative Binomial (ZINB) models are widely used to model zero-inflated count responses [1–12]. Unlike standard count responses, which are typically modeled by the Poisson or Negative Binomial (NB) distributions, zero-inflated count responses define a two-component mixture consisting of a degenerate distribution

centered at 0 and a distribution for count responses such as the Poisson. The former (latter) is often called the “at-risk” (“non-risk”) group. Thus, if a study population includes a group of subjects who are not at risk for the phenomenon of interest such as heart attack, then the number of heart attacks experienced by each subject is a zero-inflated count outcome. The zero from the non-risk group is referred to as “structural” zero and the proportion of such structural zeros is the mixing probability of the two-component mixture distribution such as ZIP or ZINB.

In practice it may happen that there is a lack of sufficient zeros in a count response as modeled by the Poisson (NB), in which case zero-deflated models may be used. Both zero-inflated and zero-deflated models can be studied under the framework of zero-altered models [13, 14]. However, we focus on zero-inflated cases in this paper, since in the zero-deflated model all zeros are from a single group and there is no distinction across such zeros, such as random vs. structural as in the zero-inflated model. Note that when zeros all come from a single group, regardless of inflated or deflated zeros, hurdle models with either a truncated Poisson or truncated negative-binomial model for the positive (i.e., non-zero) outcome is also a common approach [15, 16]. Under this approach, one models the zero and positive components of the mixture separately. This approach can also be applied to zero-inflated data, if the structural zero is observed, i.e., the at-risk group is known. In this paper, we focus on zero-inflated models when structural zeros are not observed.

Like the difference between Poisson and NB, the only difference between ZIP and ZINB is the additional dispersion parameter in ZINB to account for overdispersion in the count response from the at-risk group [17–20]. Although ZINB provides more robust inference than ZIP, it still is a parametric model and may yield incorrect inference, if the overdispersion is not described by the NB [21, 22]. For example, if a normal random effect is used in ZIP to model a zero-inflated response, the resulting overdispersion no longer follows NB and ZINB becomes inappropriate for modeling the overdispersed zero-inflated count response. In some cases, even if the source of overdispersion is known, parametric models may not apply. For example, in a multi-center study on reducing risks for HIV/STD infection for male drug users (see Section 5 for more details about this study), overdispersion is likely due to site-induced data clustering. However, it is not possible to control for site effects using random effects, since information about sites of the study is unavailable from the public released dataset.

Distribution-free, or semi-parametric, models such as the popular generalized estimating equations (GEE) provide a robust alternative for addressing overdispersed count responses. By modeling only the mean response, the GEE provides valid inference for a much wider class of data distributions. However, GEE does not apply to the current context, since the mean response alone does not provide sufficient information to identify all the parameters in a mixture model such as ZIP. By modeling both the first- and second-order moments, Yu et al. [12] developed an approach to extend the GEE to zero-inflated count responses. Although improving robustness of ZIP, their approach may not provide protection against overdispersion because the second-order moment is specified based on the ZIP.

In this paper, we propose a new approach to address this key limitation of Yu's approach to provide robust inference in the presence of overdispersion. In Section 2, we give an overview of ZIP and ZINB, followed by a discussion of the functional response model (FRM) and its application to the current setting. In Section 3, we discuss inference for the FRM-based model. In Section 4, we illustrate the proposed approach with both real and simulated study data. In Section 5, we give our concluding remarks.

2 Functional Response Models for Count Responses

We start with a brief review of ZIP and ZINB.

2.1 Models for Structural Zeros

Let y_i be a zero-inflated count response and $\mathbf{x}_i$ a vector of explanatory variables. Let $\mathbf{u}_i$ and $\mathbf{v}_i$ be two subsets of $\mathbf{x}_i$, which may overlap one another or even be identical. The zero-inflated Poisson (ZIP) regression is defined by:

$$y_i | \mathbf{x}_i \sim \text{ZIP}(p_i, \mu_i), \quad \text{logit}(p_i) = \mathbf{u}_i^T \beta_u, \quad \log(\mu_i) = \mathbf{v}_i^T \beta_v, \quad 1 \leq i \leq n, \quad (1)$$

where $\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$ is the logit link and $\text{ZIP}(p, \mu)$ denotes the ZIP distribution defined by:

$$f_{\text{ZIP}}(y|p, \mu) = \begin{cases} pf_0(0) + (1-p)f_p(0|\mu) & \text{if } y=0 \\ (1-p)f_p(y|\mu) & \text{if } y>0 \end{cases}, \quad (2)$$

with $f_0(y)$ denoting a degenerate distribution centered at 0 and $f_p(y|\mu)$ the distribution function of $\text{Poisson}(\mu)$ with mean μ . In (2), the Poisson probability at 0, $f_p(0|\mu)$, is modified by $pf_0(0) + (1-p)f_p(0|\mu)$ with $pf_0(0) = p$ accounting for structural zeros. For example, if there is a proportion of p subjects who never have unprotected sex (non-risk group) and the number of unprotected sexual occasions for the remaining $1-p$ proportion of subjects engaged in unprotected sex (at-risk group) follows a $\text{Poisson}(\mu)$ with mean μ , then the count of unprotected sexual occasions for the whole population follows the $\text{ZIP}(p, \mu)$.

Under $\text{ZIP}(p, \mu)$, the mean and variance of the Poisson component are both equal to μ . In many studies, the count response from the at-risk group may be overdispersed, causing invalid inference when applying ZIP to such data. By mixing the zero-centered degenerate distribution with NB, we obtain the ZINB:

$$y_i | \mathbf{x}_i \sim \text{ZINB}(p_i, \mu_i, \tau), \quad \text{logit}(p_i) = \mathbf{u}_i^T \beta_u, \quad \log(\mu_i) = \mathbf{v}_i^T \beta_v, \quad 1 \leq i \leq n,$$

where $\text{ZINB}(p, \mu, \tau)$ is a ZINB distribution defined by:

$$f_{\text{ZINB}}(y|p, \mu, \tau) = \begin{cases} pf_0(0) + (1-p)f_{\text{NB}}(0|\mu, \tau) & \text{if } y=0 \\ (1-p)f_{\text{NB}}(y|\mu, \tau) & \text{if } y>0 \end{cases}.$$

with $f_{NB}(y | \mu, \tau)$ denoting the distribution function of NB:

$$f_{NB}(y | \mu, \tau) = \frac{\Gamma(y + \tau)}{y! \Gamma(\tau)} \left(\frac{1}{1 + \mu/\tau} \right)^y \left(\frac{\mu/\tau}{1 + \mu/\tau} \right)^\tau, \quad \tau > 0, \quad y = 0, 1, 2, \dots \quad (3)$$

As in ZIP, p is the mixing probability of the ZINB mixture.

Under ZINB, the conditional variance of the count response for the at-risk group is

$Var(y_i | \mathbf{x}_i) = \mu_i \left(1 + \frac{\mu_i}{\tau} \right)$, which is larger than the conditional mean, $E(y_i | \mathbf{x}_i) = \mu_i$. However, the overdispersion under ZINB follows a specific form, which may not fit overdispersed count responses arising in practice. To improve robustness of inference, an alternative approach is to use moment-based models. For example, in the absence of structural zeros, we may model the mean, or first-order moment, as:

$$E(y_i | \mathbf{x}_i) = \mu_i, \quad \log(\mu_i) = \exp(\mathbf{x}_i^T \beta), \quad 1 \leq i \leq n. \quad (4)$$

Then inference based on the above is valid regardless of whether y_i given $\mathbf{x}_i$ follows Poisson, NB, or any distribution so long as (4) is the correct model for the conditional mean $E(y_i | \mathbf{x}_i)$ [18, 23]. Unfortunately, if we simply model the mean of ZIP (or ZINB since it has the same mean as ZIP) as:

$$E(y_i | \mathbf{x}_i) = (1 - p_i) \mu_i, \quad \text{logit}(p_i) = \mathbf{u}_i^T \beta_u, \text{logit}(\mu_i) = \mathbf{v}_i^T \beta_v, \quad 1 \leq i \leq n. \quad (5)$$

we would not be able to estimate β_u and β_v , since the mean alone is not sufficient to identify these parameters.

Yu et al. [12] added the second-order moment, $E(y_i^2 | \mathbf{x}_i)$, based on the ZIP to solve the identifiability issue. Although applicable to a wider class of data distributions than ZIP, this approach may not be robust against overdispersed y_i from the at-risk group. Below, we discuss an alternative to identify β_u and β_v without modeling the second-order moment. We start with a brief overview of the functional response models upon which the new approach is based.

2.2 Functional Response Models

Consider a class of distribution-free regression models defined by:

$$E[\mathbf{f}(\mathbf{y}_{i_1}, \dots, \mathbf{y}_{i_q}) | \mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_q}] = \mathbf{h}(\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_q}; \theta), \quad 1 \leq q, \quad 1 \leq i \leq n, \quad (6)$$

where $\mathbf{y}_i = (y_{i1}, \dots, y_{im})^T$ denotes a vector of responses from the i th subject, $\mathbf{f}$ some vector-valued function, $\mathbf{h}(\theta)$ some vector-valued smooth function (e.g. with continuous derivatives up to the second order), θ a vector of parameters, and $i_1, \dots, i_q$ represents q distinct elements from the integer set $\{1, \dots, n\}$. The *functional response models* (FRM) in (6) extend single-subject linear responses $\mathbf{y}_i$ in generalized linear models (GLM) to an arbitrary function of responses from multiple subjects. By setting $q = 1$ and $\mathbf{f}(\mathbf{y}_i) = \mathbf{y}_i$, (6) yields the class of distribution-free GLM. The FRM has been applied to a range of problems such as the zero-

inflated count response in the current context, extension of the Mann-Whitney-Wilcoxon rank sum test to causal inference and mediation analyses [12, 25, 26].

Consider the following FRM:

$$\mathbf{f}_i = \mathbf{f}(y_i) = (f_{1i}, f_{2i})^T, \quad \mathbf{h}_i = (h_{1i}, h_{2i})^T, \quad f_{1i} = I(y_i = 0), \quad f_{2i} = y_i, \quad (7)$$

$$h_{1i} = \text{logit}^{-1}(\mathbf{u}_i^T \beta_u) + \frac{\exp(-\exp(\mathbf{v}_i^T \beta_v))}{1 + \exp(\mathbf{u}_i^T \beta_u)}, \quad h_{2i} = \frac{\exp(\mathbf{v}_i^T \beta_v)}{1 + \exp(\mathbf{u}_i^T \beta_u)}.$$

In addition to y_i , the above also includes a non-linear response $I(y_i = 0)$ to identify the parameters. Unlike Yu's approach, it does not use the second-order moment, but rather a lesser-stringent constraint, $I(y_i = 0)$, to help identify model parameters, thereby potentially improving robustness against overdispersed y_i from the at-risk group. The conditional mean $\mathbf{h}_i$ given $\mathbf{u}_i$ and $\mathbf{v}_i$ in (7) is evaluated based on the ZIP in (1).

With no assumed parametric model such as ZIP, inference cannot proceed using maximum likelihood. Next we discuss an adaptation of the generalized estimating equations (GEE) to the current setting to provide inference for this FRM.

3 Distribution-free Inference

For the FRM in (7), let $\theta = (\beta_u^T, \beta_v^T)^T$ and

$$D_i = \frac{\partial}{\partial \theta} \mathbf{h}_i, \quad S_i = \mathbf{f}_i - \mathbf{h}_i, \quad V_i = \begin{pmatrix} \text{Var}(f_{1i} | \mathbf{u}_i, \mathbf{v}_i) & \text{Cov}((f_{1i}, f_{2i}) | \mathbf{u}_i, \mathbf{v}_i) \\ \text{Cov}((f_{1i}, f_{2i}) | \mathbf{u}_i, \mathbf{v}_i) & \text{Var}(f_{2i} | \mathbf{u}_i, \mathbf{v}_i) \end{pmatrix}. \quad (8)$$

Under the ZIP model in (1), the elements of V_i above are readily evaluated (see Appendix A). Thus, along with (7), the quantities D_i , V_i and S_i in (8) are well defined. We estimate θ by solving the following set of generalized estimating equations (GEE):

$$\mathbf{w}_n(\theta) = \sum_{i=1}^n D_i V_i^{-1} S_i = \mathbf{0}. \quad (9)$$

The above extends the GEE beyond linear [23] or quadratic responses [27] to general functions of responses of the FRM.

Under (7), the GEE estimate $\hat{\theta}$ of θ obtained as the solution to (9) is consistent and asymptotically normal (see Appendix B for a sketch of the proof):

$$\sqrt{n}(\hat{\theta} - \theta) \rightarrow_d N(0, \Sigma_\theta), \quad B = E(D_i^T V_i^{-1} D_i), \quad \Sigma_\theta = B^{-1} E(D_i V_i^{-1} S_i S_i^T V_i^{-1} D_i^T) B^{-T}, \quad (10)$$

where $\rightarrow_d$ denotes convergence in distribution [18]. Unlike maximum likelihood estimates (MLE), the asymptotic results above do not require that y_i (given $\mathbf{u}_i$ and $\mathbf{v}_i$) follow the ZIP in

(1). A consistent estimate of Σ_θ is obtained by substituting moment estimates in place of the respective parameters:

$$\hat{\Sigma}_\theta = \hat{B}^{-1} \left(\frac{1}{n-1} \sum_{i=1}^n \hat{D}_i \hat{V}_i^{-1} \hat{S}_i \hat{S}_i^T \hat{V}_i^{-1} \hat{D}_i^T \right) \hat{B}^{-T}, \quad \hat{B} = \frac{1}{n-1} \sum_{i=1}^n \hat{D}_i \hat{V}_i^{-1} \hat{D}_i^T,$$

where $\hat{B}_i, \hat{D}_i, \hat{S}_i$ and $\hat{V}_i$ denote the corresponding quantities with θ replaced by $\hat{\theta}$.

Note that our approach is based on some moments of y_i . Thus, it is not specific to zero-inflated outcomes and is applicable to more general zero-altered models as well with appropriately defined f_{1i} and f_{2i} .

4 Simulation Study

We first investigate the performance of the approach by comparing it with the parametric ZIP and ZINB as well as Yu's method by simulation. All simulations are performed with a Monte Carlo (MC) sample of $M = 1,000$, with a statistical significance level at $\alpha = 0.05$.

4.1 Absence of Overdispersion under ZIP

We first simulate data from ZIP and then fit ZIP, ZINB, Yu's method and the proposed approach. In this case, ZIP is the optimal model and its maximum likelihood estimates are most efficient (asymptotically). By comparing ZIP with others, we are able to assess potential loss of power for the other methods.

We consider a single explanatory variable x_i from a normal and simulated y_i given x_i from the following ZIP:

$$y_i | x_i \sim \text{ZIP}(p_i, \mu_i), \quad \text{logit}(p_i) = \beta_{u0}, \quad \log(\mu_i) = \beta_0 + x_i \beta_1, \quad x_i \sim N(\mu_x, \sigma_x^2). \quad (11)$$

We set $\mu_x = 0.2$, $\sigma_x^2 = 0.2$ and $\beta_0 = 6$, $\beta_1 = 0.5$. Also, we set $\beta_{u0} = -1$, i.e., $p_i = \exp(\beta_{u0}) = 0.37$ so that around 37% of simulated data were structural zeros. For each sample simulated, we fit each of the four different models to the data, repeat this process M times to obtain M sets of parameter estimates and compute empirical standard errors from the M sets of estimates. Since all four models provide consistent estimates, difference only occur in standard errors. Biased inference results when asymptotic standard errors differ from their empirical counterparts. To further illustrate the ramification of differences between the two types of standard error, we also compute and report (empirical) type I error rates based on the MC replications. For brevity, we only report two-sided type I errors for testing the null hypothesis, $H_0: \beta_1 = 0.5$, which is the percent of times the null is rejected at the nominal level $\alpha = 0.05$.

Shown in Table 1 are the averaged estimates of θ , asymptotic standard errors over the M MC replicates and empirical standard errors based on the M sets of parameter estimates, with the study sample size $n = 500$. As expected, in all cases, the estimated parameters were quite close to the true values and the asymptotic standard errors also matched up their empirical

counterparts quite well. The type I error rates for testing the null $H_0: \beta_1 = 0.5$ were also quite close to the nominal value $\alpha = 0.05$ across the board.

As expected, between ZIP and ZINB, the standard errors were quite close to each other for all parameter estimates, although ZINB did have slightly larger standard errors (both asymptotic and empirical) than ZIP. The two distribution-free models also had nearly identical standard errors for the estimated parameters. By comparing the standard errors between the ZIP and the two distribution-free models, there was no evidence of power loss. Thus, the two distribution-free models performed remarkably well in this simulation setting.

4.2 Overdispersion under ZINB

In this study, we replace the ZIP in (11) by the ZINB below:

$$y_i | x_i \sim ZINB(p_i, \mu_i, \tau), \quad \text{logit}(p_i) = \beta_{u0}, \quad \log(\mu_i) = \beta_0 + x_i \beta_1, \quad x_i \sim N(\mu_x, \sigma_x^2). \quad (12)$$

We use the same parameter values as in the ZIP setting above, except for the new dispersion parameter τ , which is set to $\tau = 1.5$. Thus, this simulation setting assesses the robustness of the different methods in the presence of overdispersed count response from the at-risk group. Since $ZINB(p_i, \mu_i, \tau)$ converges to $ZIP(p_i, \mu_i)$ as $\tau \rightarrow \infty$, selecting a relatively small τ such as $\tau = 1.5$ allows us to better assess performance of ZIP and the distribution-free models under this specific type of overdispersion.

Shown in Table 2 are the averaged estimates of θ and asymptotic standard errors over the M replicates, along with empirical standard errors based on the M sets of parameter estimates. As seen, all but Yu's method yielded parameter estimates that were quite close to the true values. Overdispersion seemed to have quite a dramatic effect on Yu's method, even changing the sign of the estimate of β_{u0} for the logistic component of the model. Note that Yu's method did provide correct estimates (not shown) for large τ such as $\tau = 1,000$. Thus the constraints imposed by the Poisson on both the first and second moments in Yu's method seems to have a more effect on its estimates than the Poisson does on its MLE.

Although the ZIP provided good parameter estimates, its asymptotic standard errors underestimated the true variability of the estimates of β_0 and β_1 of the Poisson component by more than 15 times, causing a highly inflated type I error rate. The proposed method remained robust, with closely matched-up asymptotic and empirical standard errors. What is particularly interesting is that this distribution-free model yielded nearly identical results (asymptotic and empirical standard errors and type I errors) as the ZINB, showing again no loss of power as compared to the fully metric ZINB.

4.3 Overdispersion under Normal Random Effect

We simulate data in this case from a modified ZIP with a random effect in the mean of its Poisson component to create overdispersion for the count response from the at-risk group. Specifically, the response y_i under this normal-random-effect ZIP (NRE-ZIP) is modeled according to:

$$y_i|x_i, b_i \stackrel{i.i.d.}{\sim} NRE - ZIP(\tilde{\mu}_i), \quad 1 \leq i \leq n, \quad \text{logit}(p_i) = \beta_{\mu 0}, \quad (13)$$

$$\log(\tilde{\mu}_i) = \beta_0 + x_i \beta_1 - \frac{1}{2} \sigma_b^2 + b_i, \quad x_i \sim N(\mu_x, \sigma_x^2), \quad b_i \sim N(0, \sigma_b^2).$$

We again set the parameters $(\mu_x, \sigma_x^2, \beta_0, \beta_1, \beta_{u0})$ to the same values as in the two examples above. But, for the variance σ_b^2 of the random effect b_i , we vary its values to investigate the robustness of the different methods.

Under (13), the conditional mean and variance of y_i given x_i for the Poisson component are given by [28]:

$$E(y_i|x_i) = \mu_i = \log^{-1}(\beta_0 + x_i \beta_1), \quad \text{Var}(y_i|x_i) > E(y_i|x_i).$$

The NRE-ZIP yields the same mean as in the above examples, but different overdispersed variances than the ZINB. Thus, data simulated in this setting is useful to assess the robustness of ZINB as well as the proposed approach.

Shown in Table 3 are the averaged estimates of θ and asymptotic standard errors over the M replicates, along with the empirical standard errors based on the M sets of parameter estimates. Note that Yu's method had some convergence problems and the results shown for this method were based on the converged runs (about 30% times). As before, Yu's method again showed severe bias in the parameter estimates, while the parameter estimates remained close to the true values for the other three models. As for standard and type I errors, the ZIP again underestimated the variability of the parameter estimates. Unlike the previous cases, however, the ZINB no longer provided valid inference, as it too underestimated the variability of its parameter estimates. In comparison, the proposed method continued to provide reliable standard and type I errors.

Shown in Table 4 are estimates of θ , standard errors (asymptotic and empirical) and type I errors for the same NRE-ZIP, but with $\sigma_b^2 = 1.5$. Yu's method in this case failed to converge and only estimates from the ZIP, ZINB and new approach are shown in the table. The results for the two parametric models were again biased, trending in the same direction as those in Table 3. With the increased overdispersion, the type I error was nearly 1 for the ZIP and almost 5 times of the nominal value for the ZINB. The proposed method remained robust.

5 Case Study

We also compared the proposed approach with the ZIP and ZINB using a multi-center study entitled "HIV/STD Safer Sex Skills Groups For Men In Methadone Maintenance Or Drugfree Outpatient Treatment Programs". This study was designed to examine the effectiveness of a 5-session motivational and skills training in HIV/AIDS group intervention developed to reduce sexual risk behaviors in male drug-users, as compared to an HIV

education only control condition. Unlike most community-based studies in which the HIV education provided was limited to information, this trial integrated a component to provide skill-training programs such as role plays to reduce sex risk behaviors. The primary outcome of the study is the number of unprotected vaginal and anal sexual intercourse occasions (USO) [2].

Out of 573 eligible subjects screened, 422 subjects completed assessment at baseline. The study has been analyzed in a number of publications using parametric ZIP and ZINB [2, 20], with ZINB showing a better fit than ZIP. In the current analysis, we applied ZIP and ZINB as well as the proposed approach to the 3-month outcomes from 381 (91.27%) subjects who came for the follow-up assessment.

Since earlier analyses showed that the intervention only had significant effect for the at-risk group, we included the intervention, a binary indicator with the value 1 (0) for the intervention (control) group, as the only predictor for the component of the count response of the FRM. Thus, the conditional mean of USO at 3-month y_i is modeled by:

$$\text{logit}(p_i) = \beta_{u0}, \quad \log(\mu_i) = (\beta_0) + x_i\beta_1, \quad (14)$$

where x_i is the binary indicator of treatment groups. We then fit each of the three models, with the log and logistic components of the models given in (14).

Note that this is a multi-site study, so the data are clustered by site. A random effect zero-inflated model may also be a viable alternative, if the site information is available. However, because this information is not available from the publicly available dataset, the random effect model is not included in the analysis.

Shown in Table 5 are the estimated parameters, asymptotic errors and associated p-values from the three different models. For the fitted ZINB, the estimated dispersion parameter was $\hat{\tau}=1.59$. Although some differences existed, the parameter estimates were in general agreement across the different models. As expected, the ZIP underestimated the standard errors of the estimates of β_0 and β_1 for the Poisson component, causing a highly false significant intervention effect on reducing USO for the at-risk group. Both the ZINB and proposed method corrected the underestimated standard errors, indicating no significant effect of the intervention for this at-risk group. However, the big difference in the point estimates of β_{u0} between the ZINB and proposed method suggests that the NB may not be a correct distribution to address the overdispersion in the count response from the at-risk group.

6 Discussion

Population mixtures defined by zero-inflated count outcomes arise quite often in biomedical and psychosocial research and practice. Since overdispersion is quite common in practice, ZINB generally provides a better fit than ZIP. However, ZINB only addresses a special type of overdispersion and too often it fails to fit study data. If sources of overdispersion are known, other parametric models such as the normal-random-effect ZIP discussed in Section 4.3 may be applied. However, as aptly indicated by the multi-site study in Section 5,

parametric models may not be applicable, even if sources of overdispersion are known, because of a lack of available data such as site information as in this study. In contrast, the proposed FRM-based approach requires no such elaborate assumptions and provides more robust inference than these parametric alternatives. Further, the proposed approach also seems quite efficient, as evidenced by results from the simulation studies.

Note that hurdle models may be applied, if the subgroups of the mixture are observed, such as in the case of a mixture of zeros and zero-truncated Poisson (NB) or a mixture of observed structural zeros and Poisson (NB). When the subgroups are unobservable, then the ZIP (ZINB) are generally used to model such two-group mixed populations.

Although likelihood ratio and score tests are widely used for assessing goodness of fit for parametric models, their use within the current context is quite limited, because of a lack of consensus on whether the Poisson (NB) is nested within the ZIP (ZINB) [29, 30]. Vuong's statistic is arguably the most popular test for choosing between different parametric models such as ZIP and ZINB [19, 31]. If a count response of interest in a study is overdispersed, one may use the ZINB instead of the ZIP. However, if there is a lack of evidence that the overdispersion follows ZINB, it is safer to go with distribution-free methods such as the proposed approach.

In this paper, we have focused on the robustness of the different approaches for crosssectional data. For longitudinal data, the weighted generalized estimating equations (WGEE) developed by Yu et al. [12] within the context of modeling zero-inflated responses using the FRM may also be applied to our approach to address missing follow-up data under the missing at random (MAR) mechanism. Performance of this approach as applied to the current model requires future investigations.

Parametric hurdle and ZIP (NB) models can be fit using popular software packages such as R and SAS. For example, both hurdle and zero-inflated models can be fit using the SAS experimental procedure FMM. For the proposed distribution-free approach, we have developed both SAS and R codes, which are available from the authors upon request.

Acknowledgment

This research was supported in part by grants DA027521 and GM108337 from the National Institutes of Health.

Appendix A: Variance Matrix of f_i

Under the ZIP model in (1), it is readily checked that the elements of V_i above are given by the following:

$$\begin{aligned} Var(f_{1i}|\mathbf{u}_i, \mathbf{v}_i) &= Pr(f_{1i}|\mathbf{u}_i, \mathbf{v}_i) [1 - Pr(f_{1i}|\mathbf{u}_i, \mathbf{v}_i)] \\ &= \text{logit}^{-1} \left(\mathbf{u}_i^T \beta_u \right) + \frac{\exp(-\exp(\mathbf{v}_i^T \beta_v))}{1 + \exp(\mathbf{u}_i^T \beta_u)} \\ &\quad - \left[\text{logit}^{-1} \left(\mathbf{u}_i^T \beta_u \right) + \frac{\exp(-\exp(\mathbf{v}_i^T \beta_v))}{1 + \exp(\mathbf{u}_i^T \beta_u)} \right]^2, \end{aligned}$$

$$\begin{aligned} Cov((f_{1i}, f_{2i}) | \mathbf{u}_i, \mathbf{v}_i) &= E(f_{1i}f_{2i} | \mathbf{u}_i, \mathbf{v}_i) - E(f_{1i} | \mathbf{u}_i, \mathbf{v}_i) E(f_{2i} | \mathbf{u}_i, \mathbf{v}_i) \\ &= - \left[\logit^{-1}(\mathbf{u}_i^T \beta_u) + \frac{\exp(-\exp(\mathbf{v}_i^T \beta_v))}{1 + \exp(\mathbf{u}_i^T \beta_u)} \right] \frac{\exp(\mathbf{v}_i^T \beta_v)}{1 + \exp(\mathbf{u}_i^T \beta_u)}, \end{aligned}$$

$$\begin{aligned} Var(f_{2i} | \mathbf{u}_i, \mathbf{v}_i) &= \frac{\exp(\mathbf{v}_i^T \beta_v) [1 + \exp(\mathbf{v}_i^T \beta_v)]}{1 + \exp(\mathbf{u}_i^T \beta_u)} - \left[\frac{\exp(\mathbf{v}_i^T \beta_v)}{1 + \exp(\mathbf{u}_i^T \beta_u)} \right]^2 \\ &= \frac{\exp(\mathbf{v}_i^T \beta_v)}{1 + \exp(\mathbf{u}_i^T \beta_u)} - \frac{\exp(\mathbf{u}_i^T \beta_u + 2\mathbf{v}_i^T \beta_v)}{1 + \exp(\mathbf{u}_i^T \beta_u)}. \end{aligned}$$

Appendix B: Proof of Distribution-free Inference

Consider the normalized $\frac{1}{n} \sum_{i=1}^n D_i V_i^{-1} S_i$ and for notational brevity we continue to denote the normalized estimating equations by $\mathbf{w}_n$. It follows from the iterated conditional expectation that $E(D_i V_i^{-1} S_i) = E[D_i V_i^{-1} E(S_i | \mathbf{x}_i)] = \mathbf{0}$. Thus, the GEE is unbiased and the estimate $\hat{\theta}$ obtained as the solution to the equations is consistent.

By applying a Taylor series expansion to the GEE in (9), we have:

$$\sqrt{n} \mathbf{w}_n = - \left(\frac{\partial}{\partial \theta} \mathbf{w}_n \right)^T \sqrt{n} (\hat{\theta} - \theta) + \mathbf{o}_p(1). \quad (15)$$

where $\mathbf{o}_p(1)$ denotes the stochastic $\mathbf{o}(1)$ [18]. Solving the above for $\sqrt{n}(\hat{\theta} - \theta)$ yields:

$$\sqrt{n}(\hat{\theta} - \theta) = \left(- \frac{\partial}{\partial \theta} \mathbf{w}_n \right)^{-T} \sqrt{n} \mathbf{w}_n + \mathbf{o}_p(1). \quad (16)$$

Since

$$\frac{\partial}{\partial \theta} \mathbf{w}_n = \frac{1}{n} \sum_{i=1}^n \left(\frac{\partial}{\partial \theta} S_i \right) D_i V_i^{-1} + \mathbf{o}_p(1) = - \frac{1}{n} \sum_{i=1}^n D_i V_i^{-1} D_i^T + \mathbf{o}_p(1) \rightarrow_p -B. \quad (17)$$

where $\rightarrow_p$ denotes convergence in probability, it follows from (16) and (17) that

$$\sqrt{n}(\hat{\theta} - \theta) = B^{-1} \frac{\sqrt{n}}{n} \sum_{i=1}^n D_i V_i^{-1} D_i^T + \mathbf{o}_p(1). \quad (18)$$

By applying the central limit and Slutsky's theorems to (18), $\hat{\theta}$ is asymptotically normal with the asymptotic variance given by Σ_{θ} in (10).

References

- [1]. Cheung YB. Zero-inflated models for regression analysis of count study of growth and development. *Statistics in Medicine*. 2002; 21:1461–1469. [PubMed: 12185896]

- [2]. Calsyn DA, Hatch-Maillette M, Tross S, et al. Motivational and skills training HIV/sexually transmitted infection sexual risk reduction groups for men. *Journal of Substance Abuse Treatment*. 2009; 37(2):138–150. [PubMed: 19150206]
- [3]. Cameron AC, Trivedi PK. Econometric models based on count data: Comparisons and applications of some estimators and tests. *Journal of Applied Econometrics*. 1986; 1:29–53.
- [4]. Crepon B, Duguet E. Research and development, competition and innovation — pseudo-maximum likelihood and simulated maximum likelihood methods applied to count data models with heterogeneity. *Journal of Econometrics*. 1997; 79:355–378.
- [5]. Gurmu S, Trivedi P. Excess zeros in count models for recreational trips. *Journal of Business & Economic Statistics*. 1996; 14:469–477.
- [6]. Hall DB. Zero-Inflated Poisson and binomial regression with random effects: A case study. *Biometrics*. 2000; 56:1030–1039. [PubMed: 11129458]
- [7]. Hur K, Hedeker D, Henderson W, Khuri S, Daley J. Modeling clustered count data with excess zeros in health care outcomes research. *Health Services and Outcomes Research Methodology*. 2002; 3:5–2.
- [8]. Lachenbruch PA. Analysis of data with excess zeros. *Statistical Methods in Medical Research*. 2002; 11:297–302. [PubMed: 12197297]
- [9]. Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*. 1992; 34:1–14.
- [10]. Miaou SP. The relationship between truck accidents and geometric design of road sections — Poisson versus negative binomial regressions. *Accident Analysis & Prevention*. 1994; 26:471–482. [PubMed: 7916855]
- [11]. Welsh A, Cunningham RB, Donnelly CF, Lindenmayer DB. Modeling the abundance of rare species: statistical-models for counts with extra zeros. *Ecological Modelling*. 1996; 88:297–308.
- [12]. Yu Q, Chen R, Tang W, He H, Gallop R, Crits-Christoph P, Hu J, Tu XM. Distribution-free models for longitudinal count responses with over-dispersion and structural zeros. *Statistics in Medicine*. 2013; 32:2390–2405. [PubMed: 23239019]
- [13]. Heilbro D. Zero-altered and other regression models for count data with added zeros. *Biometrical Journal*. 1994; 36:531–547.
- [14]. Ghosh S, Kim H. Semiparametric inference based on a class of zero-altered distributions. *Statistical Methodology*. 2007; 4:371–383.
- [15]. Cameron, AC.; Trivedi, PK. *Regression analysis of count data*. Cambridge university press; 2013.
- [16]. Tang, W.; He, H.; Tu, XM. *Applied categorical and count data analysis*. CRC Press; Boca Raton: 2012.
- [17]. Dean CB, Lawless JF. Tests for detecting overdispersion in Poisson regression models. *J. Amer. Statist. Assoc.* 1989; 84:467–472.
- [18]. Kowalski, J.; Tu, XM. *Modern Applied U Statistics*. Wiley; New York: 2007.
- [19]. Xia Y, Morrison-Beedy D, Ma J, Feng C, Cross W, Tu XM. Modeling count outcomes from HIV risk reduction interventions: A comparison of competing statistical models for count responses. *AIDS Research and Treatment*. 2012 Article ID 593569.
- [20]. Crits-Christoph P, Gallop R, Sadicario JS, Markell HM, Calsyn DA, Tang W, He H, Tu XM, Woody G. Predictors and moderators of outcomes of HIV/STD safer sex skills groups in substance abuse treatment programs, a pooled analysis of two randomized controlled trials. *Substance Abuse Treatment, Prevention, and Policy*. 2014; 9:3. DOI: 10.1186/1747-597X-9-3.
- [21]. McCullagh, P.; Nelder, JA. *Generalized Linear Models*. 2nd. Chapman and Hall; London: 1989.
- [22]. Zhang H, Xia Y, Chen R, Lu N, Tang W, Tu X. On Modeling Longitudinal Binomial Responses — Implications from Two Dueling Paradigms. *Journal of Applied Statistics*. 2011; 38:2373–2390.
- [23]. Liang KY, Zeger SL, Qaqish B. Multivariate regression analyses for categorical data. *J. R. Statist. Soc.,B*. 1992; 54:3–40.

- [24]. Chen R, Wu P, Ma F, Han Y, Chen T, Tu XM, Kowalski J. Extending the Mann-Whitney-Wilcoxon Rank Sum Test for multiple treatment groups and longitudinal study data. *Clinical Research in HIV/AIDS*. 2014; 1:1005.
- [25]. Gunzler D, Tang W, Lu N, Wu P, Tu XM. A class of distribution-free models for longitudinal mediation analysis. *Psychometrika*. 2014; 79:543–568. [PubMed: 24271505]
- [26]. Wu P, Han Y, Chen T, Tu XM. Causal inference for Mann-Whitney-Wilcoxon rank sum and other nonparametric statistics. *Statistics in Medicine*. 2014; 38:1261–1271. [PubMed: 24132928]
- [27]. Prentice RL, Zhao LP. Estimating Equations for Parameters in Means and Co-variances of Multivariate Discrete and Continuous Responses. *Biometrics*. 1991; 47:825–839. [PubMed: 1742441]
- [28]. Zhang H, Tang W, Yu Q, Feng C, Gunzler D, Tu XM. A new look at the difference between GEE and GLMM when modeling longitudinal count responses. *Journal of Applied Statistics*. 2012; 39:2067–2079.
- [29]. Van den Broek J. A score test for zero inflation a Poisson distribution. *Biometrics*. 1995; 51:738–743. [PubMed: 7662855]
- [30]. Sheu ML, Hu TW, Keeler TE, Ong M, Sung HY. The effect of a major cigarette price change on smoking behavior in California: a zero-inflated negative binomial model. *Health Economics*. 2004; 13:781–791. [PubMed: 15322990]
- [31]. Vuong QH. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*. 1989; 57:307–333.

Table 1

MLE and GEE estimates of parameters, asymptotic and empirical standard errors, and type I error rates based on the asymptotic standard errors for the hypothesis considered with data simulated from ZIP.

Parameter estimates, standard errors and type I errors ($H_0: \beta_1 = 0.5$)						
under ZIP: $\beta_{u0} = -1, \beta_0 = 6, \beta_1 = 0.5, n = 500$						
Parameter	Mean	Standard error		Mean	Standard error	
		Asymptotic	Empirical		Asymptotic	Empirical
	ZIP			ZINB		
β_{u0}	-1.00047	0.1010	0.0988	-1.00048	0.1010	0.0987
β_0	6.00016	0.0034	0.0035	6.00015	0.0035	0.0035
β_1	0.49993	0.0019	0.0019	0.49993	0.0019	0.0019
Type I error	0.047			0.045		
	Yu's method			New method		
β_{u0}	-1.00049	0.1010	0.0988	-1.00049	0.1010	0.0989
β_0	6.00016	0.0034	0.0035	6.00015	0.0034	0.0035
β_1	0.49992	0.0019	0.0019	0.49993	0.0019	0.0019
Type I error	0.052			0.051		

Table 2

MLE and GEE estimates of parameters, asymptotic and empirical standard errors, and type I error rates based on the asymptotic standard errors for the hypothesis considered with data simulated from ZINB.

Parameter estimates, standard errors and type I errors ($H_0: \beta_1 = 0.5$)						
under ZINB: $\beta_{u0} = -1, \beta_0 = 6, \beta_1 = 0.5, \tau = 1.5, n = 500$						
Parameter	Mean	Standard error		Mean	Standard error	
		Asymptotic	Empirical		Asymptotic	Empirical
	ZIP			ZINB		
β_{u0}	-1.00056	0.1010	0.0993	-1.00142	0.1011	0.0994
β_0	5.99946	0.0029	0.0459	5.99898	0.0469	0.0456
β_1	0.50078	0.0054	0.0974	0.50219	0.0961	0.0940
Type I error	0.926			0.051		
	Yu's method			New method		
β_{u0}	0.23892	0.0801	0.0822	-1.00059	0.1010	0.0993
β_0	6.50803	0.0598	0.0602	5.99946	0.0473	0.0459
β_1	0.49092	0.1138	0.1281	0.50079	0.0983	0.0974
Type I error	0.104			0.052		

Table 3

MLE and GEE estimates of parameters, asymptotic and empirical standard errors, and type I error rates based on the asymptotic standard errors for the hypothesis considered with data simulated from NRE-ZIP for $\sigma_b^2 = 0.2$.

Parameter estimates, standard errors and type I errors ($H_0: \beta_1 = 0.5$) for $\sigma_b^2 = 0.2$						
under NRE-ZIP: $\beta_{u0} = -1, \beta_0 = 6, \beta_1 = 0.5, n = 1,000$						
Parameter	Mean	Standard error		Mean	Standard error	
		Asymptotic	Empirical		Asymptotic	Empirical
	ZIP			ZINB		
β_{u0}	-1.004	0.071	0.070	-1.008	0.072	0.071
β_0	5.998	0.002	0.041	5.997	0.037	0.041
β_1	0.500	0.004	0.091	0.501	0.076	0.087
Type I error	0.941			0.083		
	Yu's method			New method		
β_{u0}	0.484	0.069	0.061	-1.004	0.071	0.070
β_0	6.616	0.061	0.043	5.998	0.042	0.041
β_1	0.489	0.116	0.142	0.500	0.087	0.091
Type I error	0.105			0.047		

Table 4

MLE and GEE estimates of parameters, asymptotic and empirical standard errors, and type I error rates based on the asymptotic standard errors for the hypothesis considered with data simulated from NRE-ZIP for $\sigma_b^2 = 1.5$.

Parameter estimates, standard errors and type I errors ($H_0: \beta_1 = 0.5$) for $\sigma_b^2 = 1.5$						
under NRE-ZIP: $\beta_{u0} = -1, \beta_0 = 6, \beta_1 = 0.5, n = 1,000$						
Parameter	Mean	Standard error		Mean	Standard error	
		Asymptotic	Empirical		Asymptotic	Empirical
	ZIP			ZINB		
β_{u0}	-0.989	0.071	0.071	-1.155	0.086	0.088
β_0	5.997	0.002	0.101	5.952	0.059	0.094
β_1	0.512	0.003	0.218	0.527	0.121	0.205
Type I error	0.966			0.245		
	Yu's method			New method		
β_{u0}	-	-	-	-0.989	0.071	0.070
β_0	-	-	-	5.997	0.098	0.101
β_1	-	-	-	0.512	0.193	0.218
Type I error	-			0.055		

Table 5

MLE (ZIP, ZINB) and GEE (New Method) estimates (Est.) of parameters, asymptotic standard errors (S.E.) and p-values (p-value) from parametric and distribution-free models for real study.

	β_{u0}			β_0			β_1		
Method	Est.	S.E.	p-value	Est.	S.E.	p-value	Est.	S.E.	p-value
ZIP	-0.701	0.106	<0.001	3.392	0.037	<0.001	-0.070	0.024	0.0029
ZINB	-1.068	0.175	<0.001	3.292	0.243	<0.001	-1.075	0.153	0.6210
New	-0.701	0.106	<0.001	3.392	0.226	<0.001	-0.071	0.150	0.6396