Published in final edited form as:
Stat Med. 2013 June 30; 32(14): 2390–2405. doi:10.1002/sim.5691.

Distribution-free Models for Longitudinal Count Responses with Overdispersion and Structural zeros

Q. Yu[1], **R. Chen**[1], **W. Tang**[1], **H. He**[1,2], **R. Gallop**[3], **P. Crits-Christoph**[3], **J. Hu**[1], and **X.M. Tu**[1,2]

[1]Department of Biostatistics and Computational Biology, University of Rochester, 601 Elmwoord Ave, Rochester, NY 14642, USA

[2]Department of Psychiatry, University of Rochester, 601 Elmwoord Ave, Rochester, NY 14642, USA

[3]Department of Psychiatry, University of Pennsylvania, Philadelphia, PA 19104, USA

Summary

Overdispersion and structural zeros are two major manifestations of departure from the Poisson assumption when modeling count responses using Poisson loglinear regression. As noted in a large body of literature, ignoring such departures could yield bias and lead to wrong conclusions. Different approaches have been developed to tackle these two major problems. In this paper, we review available methods for dealing with overdispersion and structural zeros within a longitudinal data setting and propose a distribution-free modeling approach to address the limitations of these methods by utilizing a new class of functional response models (FRM). We illustrate our approach with both simulated and real study data.

Keywords

functional response models; monotone missing data pattern; negative binomial; zero-inflated Poisson; weighted generalized estimating equations

1 Introduction

Count (or frequency) responses such as number of heart attacks, days of hospitalization, suicide attempts or unprotected vaginal sex arise quite often in biomedical and psychosocial research. The Poisson distribution and more generally Poisson-based log-linear regression are widely used for modeling such data. However, heterogeneity in study populations such as data clustering often creates extra variability, which renders the Poisson distribution inappropriate for modeling count data in such instances. One approach for addressing this extra Poisson, or overdispersion, is the popular negative binomial (NB) distribution. This modeling strategy, however, is rendered ineffective when the extra variability is caused by an *excessive* number of zeros above and beyond the number of zeros expected by the Poisson law. For example, when modeling behavioral outcomes such as the number of unprotected vaginal sex over a period of time in HIV prevention research, the specific study population often contains a subgroup of individuals who are not at risk for such a behavior during the study period, in which case neither the Poisson nor NB is able to accommodate such cases of *structural zeros* in the study population. One popular approach for addressing such *inflated* zero counts is the zero-inflated Poisson (ZIP) model, which has been applied to a diverse range of studies(1-16). The inherent methodological problems with structural zeros have received a great deal of attention in the literature(4; 9; 10; 19; 17; 18).

When modelling count responses in the presence of overdispersion and structural zeros within a longitudinal data setting, one of the current strategies is to employ random effects within the context of the generalized linear mixed-effects model (GLMM) to account for correlated responses from repeated assessments over time(19). However, as it relies on parametric assumptions about random effects and response for inference, such an approach lacks robustness when real study data depart from the assumed distributional models. Further, the random effects induce overdispersion into the marginal model at each assessment, giving rise to quite different results and findings than those from the marginal models(20; 21; 22). In addition, such an approach computes estimates using the expectation/maximization (EM) algorithm, which can be problematic since EM is notorious for its slow convergence and may yield local rather than global maxima, making it difficult to apply such methods in routine analyses.

A popular alternative is to use the generalized estimating equations (GEE) to address correlated longitudinal responses. The GEE approach is widely used for modeling the mean response, or first-order moment. Unlike GLMM, model parameters have the same interpretations between the marginal and joint models across assessment times. In addition, as GEE models the marginal mean of the response variable at each assessment time, it ignores both layers of assumptions and thereby provides consistent estimates regardless of the complexity of the correlation structure and the distribution of the response. GEE estimates are also much easier to compute than those based on the GLMM approach.

As the key difference between the standard (Poisson) log-linear model and other models for count responses such as ZIP lies in the variance, or the second-order moment, GEE does not apply directly to extending such models to a longitudinal data setting(23; 24; 25). Also, since ZIP is a mixture of two distributions, we will not be able to identify the model parameters by simply modeling the mean response(24; 26). One approach is to model the zero and positive outcomes separately using a truncated Poisson for the positive response and a logistic regression for the zero outcome(27). However, as the structural and sampling zeros are mixed into a single category, this approach is unable to identify the parameters for modeling the structural zeros, which is often of great interest in practice. For example, in the hospitalization example, this approach will only model those who are hospitalized, since the at-risk subgroup for hospitalization is mixed with those who are healthy and are not at risk for hospitalization. In many studies, it is of great importance to model the at- and non-risk subgroups separately. An alternative to address the identifiability issue is to include a modeling component for the variance and apply GEE to both the specified mean and variance(24; 25; 28; 29). However, all these methods do not sufficiently address missing data, yielding biased inference if missing data does not follow the missing completely at random (MCAR) assumption(30; 31).

In this paper, we propose an approach to overcome the aforementioned difficulties by utilizing a class of functional response models (FRM) and the popular weighted generalized estimating equations (WGEE). In Section 2, we first give a brief overview of the problems with overdispersed and zero-inflated count data and popular models for addressing them. We then introduce FRM and discuss its application to the current setting. In Section 3, we discuss inference for the FRM-based models under both complete and missing data. In Section 4, we illustrate the proposed models with real study data and investigate their performance using simulated data. In Section 5, we give our concluding remarks.

2 Functional Response Models for Count Response

We start with a brief review of existing approaches for addressing overdispersion and structural zeros.

2.1 Models for Overdispersion and Structural Zeros

Consider first a cross-sectional study with n subjects, and let y_i denote a count response and $\mathbf{x}_i$ a vector of explanatory variables. The popular Poisson log-linear regression, a member of the generalized linear model (GLM) family, models the conditional mean of y_i given $\mathbf{x}_i$, $\mu_i = E(y_i \mid \mathbf{x}_i)$, by applying the log function to link μ_i to the linear predictor $\eta_i = \mathbf{x}_i^\top \beta(32)$:

$$y_i|\mathbf{x}_i \sim \text{i.d.} P(\mu_i), \log(\mu_i) = \mathbf{x}_i^\top \beta, 1 \le i \le n, \quad (1)$$

where i.d. denotes *independently distributed* and $P(\mu)$ the Poisson distribution with mean μ. Under (1), the conditional mean $E(y_i \mid \mathbf{x}_i)$ and variance $Var(y_i \mid \mathbf{x}_i)$ of y_i given $\mathbf{x}_i$ satisfy:

$$E(y_i|\mathbf{x}_i) = \text{Var}(y_i|\mathbf{x}_i) = \mu_i. \quad (2)$$

As mentioned, the conditional variance $Var(y_i \mid \mathbf{x}_i)$ often exceeds the conditional mean μ_i in real study applications, making (1) inappropriate for modeling such count data. When overdispersion occurs, the standard error of the parameter estimate of the Poisson model is artificially deflated, giving rise to artificially inflated effect size estimates and false significant findings.

Overdispersion can often be empirically detected by goodness of fit statistics or even formally tested(25; 32; 33). When deemed present, overdispersion may be corrected post hoc by using robust variance estimates(25). An alternative is to use models that explicitly address this issue. For example, the popular negative binomial (NB) model allows the variance to exceed the mean:

$$E(y_i|\mu_i,\alpha) = \mu_i, \text{Var}(y_i|\mu_i,\alpha) = \mu_i(1+\alpha\mu_i). \quad (3)$$

Unlike the Poisson, the NB has an extra parameter α to indicate the degree of overdispersion. As $\alpha \to 0$, $Var(y_i \mid \mu_i, \alpha) \to \mu_i$. Thus, unless $\alpha = 0$, the variance of NB is always larger than the mean, addressing overdispersion. Under NB, we can check overdispersion by testing the null: $H_0 : \alpha = 0$. Note, however that under H_0, $\alpha = 0$ is a boundary point of $\alpha \ge 0$ and the maximum likelihood estimate (MLE) $\hat{\alpha}$ of α cannot be used directly for testing H_0, and alternative score statistics must be used(33; 34; 35).

Count responses in many biomedical and psychosocial studies are dominated by a preponderance of zeros that exceeds the expected frequency of the Poisson. Such excess or *structural* zeros not only cause overdispersion, but also affect the conditional mean, leading to biased estimates of model parameters. The zero-inflated Poisson (ZIP) model is a popular approach to address the twin effects of structural zeros on both the mean and variance.

Let $\mathbf{u}_i$ and $\mathbf{v}_i$ be two subsets of $\mathbf{x}_i$, which may overlap one another or even identical, and thus may not be a partition of $\mathbf{x}_i$. The ZIP regression model is defined by:

$$y_i|\mathbf{x}_i \sim \text{i.d.ZIP}(\mu_i,\rho_i), \text{logit}(\rho_i) = u_i^\top \beta_{\mathbf{u}}, \log(\mu_i) = \mathbf{v}_i^\top \beta_v, 1 \le i \le n, \quad (4)$$

where ZIP(μ, ρ) denotes the ZIP distribution defined by:

$$f_{\mathrm{ZIP}}(y|\mu,\rho)=\begin{cases}\rho f_0(0)+(1-\rho)f_P(0|\mu) & \text{if} y=0\\ (1-\rho)f_P(y|\mu) & \text{if} y>0\end{cases}. \quad (5)$$

with $f_0(y)$ denoting a degenerate distribution centered at 0. In (5), the Poisson probability at 0, $f_P(0 \mid \mu)$, is modified by $\rho f_0(0) + (1 - \rho) f_P(0 \mid \mu)$ with $\rho f_0(0) = \rho$ to account for structural zeros.

Consider these models within a longitudinal setting with m assessments, with y_{it}, $\mathbf{x}_{it}$, $\mathbf{u}_{it}$ and $\mathbf{v}_{it}$ denoting the respective variables at time t ($1 \le t \le m$). We may model y_{it} as a function of $\mathbf{x}_{it}$ (or $\mathbf{u}_{it}$ and v_{it} for ZIP) by using either a parametric or distribution-free modeling approach. As mentioned, the former suffers interpretational and computational problems. A popular distribution-free alternative with inference based on the generalized estimating equations (GEE) is to specify the conditional mean of y_{it} given $\mathbf{x}_{it}$, which for count response has the following form,

$$E(y_{it}|\mathbf{x}_{it})=\exp\left(\mathbf{x}_{it}^{\top}\beta\right), 1 \le t \le m, 1 \le i \le n. \quad (6)$$

This *mean-based* specification, however, is not sufficient to distinguish the Poisson from the NB, as the two models only differ in the conditional variance $Var(y_{it} \mid \mathbf{x}_{it})$. The classic model specification also does not work for ZIP, since the conditional mean of y_{it} given $\mathbf{x}_{it}$ in this case is

$$E(y_{it}|\mathbf{x}_{it})=(1-\rho_{it})\mu_{it}, 1 \le t \le m, 1 \le i \le n, \quad (7)$$

which in general does not provide sufficient information to identify β_u and β_v.

To help distinguish among the three models, one can augment the GEE by including the distinct form of the conditional variance $Var(y_{it} \mid \mathbf{x}_{it})$ for each model and use the resulting GEE II for inference(23; 24; 28; 29). However, this approach is ad-hoc in the sense that GEE II is a method of inference primarily used for improving efficiency over GEE, rather than a formal model akin to (6), since the added response (or dependent variable) $Var(y_{it} \mid \mathbf{x}_{it})$ is a function of parameters(25). In addition, it does not effectively address missing data. Another approach is to model the zero and positive outcomes separately using a truncated Poisson for the positive response and a logistic regression for the zero outcome(27). However, this approach is unable to identify the parameters for modeling the structural zero, which is often of greater interest in practice. Next we utilize a new class of regression models to address the limitations of the aforementioned approaches.

2.2 Functional Response Models

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

$$E[\mathbf{f}(\mathbf{y}_{i_1},\ldots,\mathbf{y}_{i_q})|\mathbf{x}_{i_1},\ldots,\mathbf{x}_{i_q}]=\mathbf{h}(\mathbf{x}_{i_1},\ldots,\mathbf{x}_{i_q};\theta),(i_1,\ldots,i_q)\in C_q^n, 1 \le q, 1 \le i \le n, \quad (8)$$

where $\mathbf{y}_i = (y_{i1}, \ldots, y_{im})^{\top}$ denotes the vector of responses from the ith 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 of interest, q some positive integer, and C_q^n the set of $\binom{n}{q}$ combinations of q distinct elements ($i_1, \ldots, i_q$) from the integer set $\{1, \ldots, n\}$. The *functional response models* (FRM) (8) extend the single-subject response in the classic

GLM to a function of responses from multiple subjects. For example, by setting $q = 1$, we immediately obtain from (8) the class of distribution-free GLMs for longitudinal data with m assessments. With FRM, we can express a broader class of problems under a regression-like framework(25; 36; 37; 38; 39; 40). Below, we focus on the application of FRM within our setting for modeling count responses.

Consider first the simpler cross-sectional study setting. For the cross-sectional parametric ZIP in (4), let

$$\begin{array}{c}\mathbf{f}_i= \ \mathbf{f}(y_i)=(f_{1i},f_{2i})^\top, \quad \mathbf{h}_i=\mathbf{h}(\mathbf{u}_i,\mathbf{v}_i)=(h_{1i},h_{2i})^\top, \quad f_{1i}=y_i, \quad f_{2i}=y_i^2, \\ h_{1i}= \ (1-\rho_i)\mu_i, \quad h_{2i}=\mu_i(1-\rho_i)(1+\mu_i)^\top, \quad \mathrm{logit}(\rho_i)=\mathbf{u}_i^\top\beta_u, \quad \log(\mu_i)=\mathbf{v}_i^\top\beta_v,\end{array} \qquad (9)$$

where $\mathbf{u}_i(\mathbf{v}_i)$ denotes a subset of $\mathbf{x}_i$. Under (4), $E(\mathbf{f}_i \mid \mathbf{u}_i, \mathbf{v}_i) = \mathbf{h}_i(\mathbf{u}_i, \mathbf{v}_i)$. For NB, $\mathbf{f}$ (y$_i$) is defined the same as for ZIP in (9), but with $\mathbf{h_i} = (h_{1\mathbf{i}}, h_{2\mathbf{i}})^\top$ modified as follows:

$$h_{1i}=\mu_i, h_{2i}=\mu_i+\alpha\mu_i^2+\mu_i^2, \log(\mu_i)=\mathbf{x}_i^\top\beta. \qquad (10)$$

As a special case with $\alpha = 0$, the FRM for NB reduces to a distribution-free Poisson with $\mu_i=\exp\left(\mathbf{x}_i^\top\beta\right)$. Note that under the FRM-based NB, we can allow α to be negative and thus $\alpha = 0$ is no longer a boundary point. Thus, we can readily use the estimate of α to test the null $H_0 : \alpha = 0$ to determine whether the Poisson loglinear model is appropriate.

For longitudinal data, suppose that each subject is assessed m times, with y_{it} and $\mathbf{x}_{it}$ denoting the respective variables at time t ($1 \le t \le m$). Define the FRM-based ZIP model as follows:

$$\begin{aligned}\mathbf{f}_{it}=&\mathbf{f}(y_{it})\\ &=(f_{1it},f_{2it})^\top\\ &=\left(y_{it},y_{it}^2\right)^\top, h_{it}=(h_{1it},h_{2it})^\top, h_{1it}\\ &=(1-\rho_{it})\mu_{it}, h_{2it}\\ &=\mu_{it}(1\\ &\quad-\rho_{it})(1+\mu_{it})^\top, \mathrm{logit}(\rho_{it})\\ &=u_{it}^\top\beta_u, \log(\mu_{it})\\ &=v_{it}^\top\beta_v, 1\le t\le m.\end{aligned} \qquad (11)$$

Likewise, we obtain a longitudinal version of FRM-based NB by defining the same $\mathbf{f}_{it}$, but modifying $\mathbf{h}_{it}$ as follows:

$$h_{1it}=\mu_{it}, h_{2it}=\mu_{it}+\alpha\mu_{it}^2+\mu_{it}^2, \log(\mu_{it})=\mathbf{x}_{it}^\top\beta, 1 \le t \le m, 1 \le i \le n. \qquad (12)$$

Note that we have assumed a constant α for NB, though the model above readily accommodates a time-varying α. In many studies, clusters causing overdispersion such as those formed by the subjects sampled from a common habitat may not change over time during the study, and this assumption is reasonable.

Both the ZIP and NB models for longitudinal data in (11) and (12) yield the same first-and second-order moment as their respective cross-sectional versions in (9)-(10) at each time t (1

≤ t ≤ m). Thus, unlike their GLMM-based parametric counterparts, estimates from the FRM-based ZIP and NB models for longitudinal data can be readily compared to their corresponding cross-sectional versions. These distribution-free models are also called semiparametric or moment-based in the literature(41; 42). We refer to these as distribution-free models throughout the text unless otherwise stated.

3 Distribution-free Inference

We first discuss inference for cross-sectional data, and then extend the considerations to the longitudinal setting.

3.1 Distribution-free Inference for Cross-sectional Data

For the FRM-based ZIP model in (??), let $\theta=\left(\beta_u^\top,\beta_v^\top\right)^\top$ and

$$D_i=\frac{\partial}{\partial\theta}\mathbf{h}_i,\quad S_i=\mathbf{f}_i-\mathbf{h}_i,\quad V_i=\begin{pmatrix}\mathrm{Var}(f_{1i}|\mathbf{u}_i,\mathbf{v}_i) & \mathrm{Cov}((f_{1i},f_{2i})|\mathbf{u}_i,\mathbf{v}_i)\\ \mathrm{Cov}((f_{1i},f_{2i})|\mathbf{u}_i,\mathbf{v}_i) & \mathrm{Var}(f_{2i}|\mathbf{u}_i,\mathbf{v}_i)\end{pmatrix}. \quad (13)$$

We estimate θ by the following set of generalized estimating equations,

$$\mathbf{w}_n(\theta)=\sum_{i=1}^n D_i V_i^{-1} S_i=0. \quad (14)$$

Given the ZIP model in (4), the elements of V_i in (13) are functions of the conditional moments of y_i given $\mathbf{x}_i$ up to the 4th order, which can be expressed in closed form (see Appendix A). Thus, the quantities D_i, V_i and S_i in (13) are readily evaluated. Note that (14) bears a close resemblance to the generalized estimating equations II (GEE II) for generalized linear models(25; 28; 29; 43).

By defining D_i, V_i and S_i the same way as in (13), but with $\theta = (\beta^\top, \alpha)^\top$ and $\mathbf{h}_i$ defined in (10), the GEE in (14) can be used to obtain estimates of θ for NB as well.

Under (9), the GEE estimate $\hat{\theta}$ of θ obtained as the solution to (14) is consistent and asymptotically normal (see Theorem 1 below):

$$\sqrt{n}\left(\widehat{\theta}-\theta\right)\rightarrow_d N(0,\sum\nolimits_\theta),\, B=E\left(D_i^\top V_i^{-1} D_i\right),\sum\nolimits_\theta=B^{-1}E\left(D_i V_i^{-1} S_i S_i^\top V_i^{-1} D_i^\top\right)B^{-\top}, \quad (15)$$

where $\rightarrow_d$ denotes convergence in distribution(25). Unlike the MLE, the asymptotic results above do not require that y_i (given $\mathbf{u}_i$ and $\mathbf{v}_i$) follow the ZIP distribution in (4). If y_i does follow such a parametric model, Σ_θ in (15) simplifies to $\Sigma_\theta = B^{-1}$, which is the *model-based* asymptotic variance.

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

$$\widehat{\sum}_\theta=\widehat{B}^{-1}\left(\frac{1}{n-1}\sum_{i=1}^n \widehat{D}_i\widehat{V}_i^{-1}\widehat{S}_i\widehat{S}_i^\top\widehat{V}_i^{-1}\widehat{D}_i^\top\right)\widehat{B}^{-\top},\widehat{B}=\frac{1}{n-1}\sum_{i=1}^n\widehat{D}_i\widehat{V}_i^{-1}\widehat{D}_i^\top,$$

where $\hat{B}_i$, $\hat{D}$, $\hat{S}_i$ and $\hat{V}_i$ denote the corresponding quantities with θ replaced by $\hat{\theta}$. Our simulations indicate that the model-based asymptotic variance estimate $\hat{B}$ outperforms its sandwich alternative by yielding slightly more accurate type I error rates under the correct parametric model(44).

3.2 Distribution-free Inference for Longitudinal Data

We begin with inference under complete data and then extend the discussion to include missing data.

3.2.1 Inference under Complete Data—Let

$$\mathbf{f}_i=\left(\mathbf{f}_{i1}^{\top},\mathbf{f}_{i2}^{\top},\ldots,\mathbf{f}_{im}^{\top}\right)^{\top},h_i=\left(h_{i1}^{\top},h_{i2}^{\top},\ldots,h_{im}^{\top}\right)^{\top},D_i=\frac{\partial}{\partial\theta}h_i,S_i=f_i-h_i, \quad (16)$$

where $\mathbf{f}_{it}$ and $\mathbf{h}_{it}$ are defined by (11) for the ZIP or by (12) for the NB model. We again apply the GEE in (14), but with D_i and S_i revised to reflect the changed dimension, and V_i modified to reflect the correlation between the $\mathbf{f}_{it}$'s over time:

$$\begin{aligned} V_i&=A_i^{\frac{1}{2}}R(\alpha)A_i^{\frac{1}{2}},A_i=diag_t(A_{it}),1\le i\le n,1\le t\le m,\\ A_{it}&=diag_t\begin{pmatrix}\mathrm{Var}(f_{1it}|\mathbf{x}_{it}) & \mathrm{Cov}((f_{1it},f_{2it})|\mathbf{x}_{it})\\ \mathrm{Cov}((f_{1it},f_{2it})|\mathbf{x}_{it}) & \mathrm{Var}(f_{2it}|\mathbf{x}_{it})\end{pmatrix}, \end{aligned} \quad (17)$$

where $R(\tau)$ is a working correlation matrix among the components of $\mathbf{f}_i$ parameterized by τ. As in the cross-sectional data case, A_{it} is readily computed. For $R(\tau)$, the popular choices are the working independence model ($R(\tau) = \mathbf{I}_{2m}$) and the exchangeable correlation structure given by:

$$R(\rho)=\begin{pmatrix} I_2 & \rho J_2 & \cdots & \rho J_2\\ & I_2 & \cdots & \rho J_2\\ & & \ddots & \vdots\\ & & & I_2 \end{pmatrix},\quad I_2=\begin{pmatrix}1&0\\0&1\end{pmatrix},\quad J_2=\begin{pmatrix}1&1\\1&1\end{pmatrix},\quad 0<\rho<1.$$

Thus, τ is known for the working independence model, but unknown for the exchangeable correlation model with τ = ρ.

Note that since the GEE estimate may not be consistent under working correlation structures other than the independence model, especially in the presence of time-varying covariates(45), we focus on this model in what to follow unless otherwise stated. With this choice of $R(\tau)$, the GEE is readily solved for θ. However, when the working correlation model used involves an unknown τ, an estimate must be substituted before the GEE is solved to obtain estimates of θ.

As in the cross-sectional data case, the GEE estimate has nice asymptotic properties summarized in Theorem 1 below. Since this is a special case of Theorem 2, its proof is omitted. Since Theorem 1 is stated for general working correlation models, it includes the condition for the estimate of τ to ensure such nice properties.

Theorem 1: Let $\hat{\theta}$ denote the GEE estimate and let

$$B{=}E\left(D_iV_i^{-1}D_i^{\top}\right),\sum\nolimits_{\theta}{=}B^{-1}E\left(D_iV_i^{-1}S_iS_i^{\top}V_i^{-1}D_i^{\top}\right)B^{-\top}. \quad (18)$$

Under mild regularity conditions, $\hat{\theta}$ is consistent. Further, if $\hat{\tau}$ is $\sqrt{n}-\text{consistent}$, i.e., $\sqrt{n}(\widehat{\boldsymbol{\tau}}-\boldsymbol{\tau})$ is bounded in probability(25), then $\hat{\theta}$ is asymptotically normal with the asymptotic variance Σ_{θ}. A consistent estimate of Σ_{θ} is given by:

$$\widehat{\sum}_{\theta}{=}\widehat{B}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\widehat{D}_i\widehat{V}_i^{-1}\widehat{S}_i\widehat{S}_i^{\top}\widehat{V}_i^{-1}\widehat{D}_i^{\top}\right)\widehat{B}^{-\top},\widehat{B}{=}\frac{1}{n}\sum_{i=1}^{n}\widehat{D}_i\widehat{V}_i^{-1}\widehat{D}_i^{\top}, \quad (19)$$

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 given the limited choices for the working correlation matrix $R(\tau)$, $E\left(S_iS_i^{\top}|x_i\right){=}V_i$ generally is not true in practice. Thus, unlike the cross-sectional data case, there is no model-based asymptotic variance.

3.2.2 Inference under Missing Data—Missing data arise frequently in real studies. For mean-based distribution-free models such as the GLM, the weighted generalized estimating equations (WGEE) is the most popular for inference about model parameters. By integrating the inverse probability weighting (IPW) technique with the GEE, the WGEE ensures valid inference when the missing data follows the missing at random (MAR) model, a plausible and general missing data mechanism applicable to many studies in practice(25; 31; 41; 46; 47). We discuss below how to extend this IPW approach to the current FRM-based models for count responses.

Within the context of longitudinal data discussed in the preceding section, we define a missing (or rather observed) data indicator for each subject as follows:

$$r_{it}{=}\left\{\begin{array}{ll}1 & \text{if} y_{it} \text{is} \quad \text{observed}\\ 0 & \text{if} y_{it} \text{is} \quad \text{missing}\end{array}\right., \quad \mathbf{r}_i{=}(r_{i1},\ldots,r_{im})^{\top}, \quad 1\leq i\leq n.$$

We assume no missing data at baseline $t = 1$ such that $r_{i1} = 1$ for all $1 \le i \le n$. Let

$$\pi_{it}{=}\Pr(r_{it}{=}1|\mathbf{x}_i,y_i),\Delta_{it}{=}\frac{r_{it}}{\pi_{it}},\Delta_i{=}\text{diag}_t(\Delta_{it}). \quad (20)$$

In most applications, the weight function π_{it} is unknown and must be estimated. Under MCAR, $\mathbf{r}_i$ is independent of $\mathbf{x}_i$ and y_i and thus $\pi_{it} = \Pr(r_{it} = 1) = \pi_t$. In this case, π_t is a constant independent of $\mathbf{x}_i$ and y_i and is readily estimated by the sample moment: $\widehat{\pi}_t{=}\frac{1}{n}\sum\nolimits_{i=1}^{n}r_{it}(2\leq t\leq m)$.

Under MAR, π_{it} becomes dependent on the observed $\mathbf{x}_i$ and y_i, making it difficult to model and estimate π_{it} without imposing the monotone missing data pattern (MMDP) assumption because of the large number of missing data patterns(25; 37; 41). Under MMDP, y_{it} ($\mathbf{x}_{it}$) is observed only if all y_{is} ($\mathbf{x}_{is}$) prior to time t are observed ($1 \le s \le t \le m$).

Let

$$H_{it}=\{\tilde{\mathbf{x}}_{it},\tilde{y}_{it};2\le t\le m\},\tilde{\mathbf{x}}_{it}=\left(\mathbf{x}_{i1}^{\top},\ldots,\mathbf{x}_{i(t-1)}^{\top}\right)^{\top},\tilde{y}_{it}=\left(y_{i1},\ldots,y_{i(t-1)}\right)^{\top},$$

where $\tilde{\mathrm{X}}_{it}$ and $\tilde{\mathbf{y}}_{it}$ contain the explanatory and response variables prior to time t, respectively. Under MAR we have:

$$\pi_{it}=\Pr(r_{it}=1|\mathbf{x}_i,y_i)=\Pr(r_{it}=1|H_{it}),2\le t\le m.$$

Let p_{it} = Pr (r_{it} = 1 | $r_{i(t-1)}$ = 1, H_{it}), the one-step transition probability for observing the response from time t − 1 to t. We can model p_{it} using logistic regression:

$$\mathrm{logit}(p_{it}(\gamma_t))=\gamma_{0t}+\gamma_{\mathbf{x}t}^{\top}\tilde{\mathbf{x}}_{it}+\gamma_{yt}^{\top}\tilde{y}_{it},2\le t\le m,\quad (21)$$

where $\gamma_t=\left(\gamma_{0t},\gamma_{\mathbf{x}t}^{\top},\gamma_{yt}^{\top}\right)^{\top}$. Let $\gamma=\left(\gamma_2^{\top},\ldots,\gamma_m^{\top}\right)^{\top}$. Then, under MMDP,

$$\pi_{it}(\gamma)=p_{it}\Pr(r_{i(t-1)}=1|H_{i(t-1)})=\prod_{s=2}^{t}p_{is}(\gamma_s),2\le t\le m,1\le i\le n.$$

The above provides a relationship to estimate π_{it} from the model for p_{it} in (21).

We may estimate γ using the following estimating equations:

$$\begin{array}{c}\mathbf{Q}_n(\gamma)=\sum_{i=1}^{n}\left(\mathbf{Q}_{i2}^{\top},\ldots,\mathbf{Q}_{im}^{\top}\right)^{\top}=0,\quad 2\le t\le m,\quad 1\le i\le n,\\ \mathbf{Q}_{it}=\frac{\partial}{\partial\gamma_t}\{r_{i(t-1)}[r_{it}\log(p_{it})-(1-r_{it})\log(1-p_{it})]\},\end{array}\quad (22)$$

With estimated π_{it}, we can estimate θ by generalizing the WGEE for mean-based response models to a WGEE for the current context as follows:

$$\mathbf{w}_n(\theta)=\sum_{i=1}^{n}\mathbf{w}_{ni}=\sum_{i=1}^{n}D_iV_i^{-1}\widehat{\Delta}_iS_i=\sum_{i=1}^{n}D_iV_i^{-1}\widehat{\Delta}_i(\mathrm{y}_i-h_i)=0,\quad (23)$$

where D_i, V_i and S_i are defined the same as in the GEE in the complete data case, and $\hat{\Delta}_i$ denotes Δ_i in (21) with estimated π_{it}. Also, as in the complete data case, V_i may be a function of τ if working dependence correlation models are used, which must replaced with an estimate before (23) is used for inference about θ.

The WGEE estimate $\hat{\theta}$ has nice asymptotic properties, as summarized by the theorem below (see Appendix B for a proof).

Theorem 2: Let $\hat{\theta}$ denote the WGEE II estimate. Under mild regularity conditions,

1. $\hat{\theta}$ is consistent.

2. If $\hat{\tau}$ is $\sqrt{n}-\text{consistent}$ $\hat{\theta}$ is asymptotically normal with asymptotic variance given by:

$$\begin{gathered}\sum\nolimits_{\theta}= B^{-1}(\sum\nolimits_{U}+\Phi)B^{-\top}, \quad \sum\nolimits_{U}=E\left(D_iV_i^{-1}\Delta_iS_iS_i^{\top}\Delta_iV_i^{-1}D_i^{\top}\right), \\ B= E\left(D_iV_i^{-1}\Delta_iD_i^{\top}\right), \quad G=E\left(D_iV_i^{-1}\Delta_iS_i\mathbf{Q}_{ni}^{\top}H^{-\top}C^{\top}\right), \quad C=E\left[\tfrac{\partial}{\partial\gamma}\left(D_iV_i^{-1}\Delta_iS_i\right)\right]^{\top}, \\ H= E\left(\tfrac{\partial}{\partial\gamma}\mathbf{Q}_{ni}\right)^{\top}, \quad \Phi=CH^{-\top}C^{\top}-G-G^{\top}.\end{gathered} \qquad (24)$$

A consistent estimate of Σ_θ is given by:

$$\widehat{\sum}_{\theta}=\widehat{B}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\widehat{D}_i\widehat{V}_i^{-1}\widehat{\Delta}_i\widehat{S}_i\widehat{S}_i^{\top}\widehat{\Delta}_i\widehat{V}_i^{-1}\widehat{D}_i^{\top}\right)\widehat{B}^{-\top}, \widehat{B}=\frac{1}{n}\sum_{i=1}^{n}\widehat{D}_i\widehat{V}_i^{-1}\widehat{\Delta}_i\widehat{V}_i^{-1}\widehat{D}_i^{\top}, \widehat{G}=\left(\frac{1}{n}\sum_{i=1}^{n}\widehat{D}_i\widehat{V}_i^{-1}\widehat{\Delta}_i\widehat{S}_i\widehat{\mathbf{Q}}_{ni}^{\top}\right)\widehat{H}^{-\top}\widehat{C}^{\top},$$

Note that the asymptotic variance in (24) contains a correction term $B^{-1}\Phi B^{-\top}$ to account for the sampling variability in the estimated $\hat{\gamma}$.

3.2.3 Score Test

As Wald-type tests are typically anti-conservative(21; 48; 49), score statistics are often used as an alternative to reduce bias, especially in type I error rates for small to moderate samples. Within the current context, let $\theta=\left(\theta_{(1)}^{\top}, \theta_{(2)}^{\top}\right)^{\top}$, with p and q denoting the dimension of $\theta_{(1)}$ and $\theta_{(2)}$, respectively. Consider testing the null $H_0 : \theta_{(2)} = \theta_{(20)}$, with $\theta_{(20)}$ denoting a vector of known constants.

Under $H_0 : \theta_{(2)} = \theta_{(20)}$,

$$\begin{gathered}D_i=\left(\begin{array}{c}\frac{\partial\mathbf{h}(\theta)}{\partial\theta_{(1)}}\\ \frac{\partial\mathbf{h}(\theta)}{\partial\theta_{(2)}}\end{array}\right)=\left(\begin{array}{c}D_{i(1)}\\ D_{i(2)}\end{array}\right), \quad D_iV_i^{-1}\widehat{\Delta}_iS_i=\left(\begin{array}{c}D_{i(1)}V_i^{-1}\widehat{\Delta}_iS_i\\ D_{i(2)}V_i^{-1}\widehat{\Delta}_iS_i\end{array}\right), \\ \mathbf{w}_n(\theta)=\left(\begin{array}{c}\mathbf{w}_{n(1)}(\theta)\\ \mathbf{w}_{n(2)}(\theta)\end{array}\right)=\tfrac{1}{n}\sum_{i=1}^{n}D_iV_i^{-1}\widehat{\Delta}_iS_i=\tfrac{1}{n}\left(\begin{array}{c}\sum_{i=1}^{n}D_{i(1)}V_i^{-1}\widehat{\Delta}_iS_i\\ \sum_{i=1}^{n}D_{i(2)}V_i^{-1}\widehat{\Delta}_iS_i\end{array}\right).\end{gathered} \qquad (26)$$

Let $\tilde{\theta}_{(1)}$ denote the estimate from solving the reduced WGEE:

$$\mathbf{w}_{n(1)}\left(\theta_{(1)}, \theta_{(20)}\right)=\frac{1}{n}\sum_{i=1}^{n}D_{i(1)}V_i^{-1}\widehat{\Delta}_iS_i=0. \qquad (27)$$

Set

$$\tilde{\theta}=\left(\begin{array}{c}\tilde{\theta}_{(1)}\\ \theta_{(2)}\end{array}\right), \quad B=E(D_iV_i^{-1}\Delta_iD_i)=\left(\begin{array}{cc}B_{11} & B_{12}\\ B_{12}^{\top} & B_{22}\end{array}\right), \quad G=(-B_{21}B_{11}^{-1}, I_v), \qquad (28)$$

where q is the dimension of $w_{n(2)}$, B_{11} denotes the $p \times p$ submatrix, B_{12} the $p \times q$ submatrix, and B_{22} the $q \times q$ submatrix from the partitioned $(p + q) \times (p + q)$ matrix B. Then, under $H_0 : \theta_{(2)} = \theta_{(20)}$, the following score statistic has as an asymptotic (central) χ_q^2 distribution with q degrees of freedom (see Appendix C for a proof):

$$T_s\left(\left(\tilde{\theta}_{(1)},\theta_{(2)}\right)\right)=n\tilde{w}_{n(2)}^{\top}\left(\left(\tilde{\theta}_{(1)},\theta_{(2)}\right)\right)\tilde{\sum}_{(2)}^{-1}\left(\left(\tilde{\theta}_{(1)},\theta_{(2)}\right)\right)\tilde{w}_{n(2)}\left(\left(\tilde{\theta}_{(1)},\theta_{(2)}\right)\right)\to_d\chi_q^2, \quad (29)$$

where $\tilde{\Sigma}_{(2)} = \tilde{G}\tilde{\Sigma}_\theta \tilde{G}^\top$ with $\tilde{G}$ and $\tilde{\Sigma}_\theta$ denoting the corresponding quantities with θ replaced by $\tilde{\theta}$.

4 Applications

We first investigate the performance of the approach with small to moderate sample sizes by simulation and then present a real data application. In all the examples, we set the statistical significance level at $\alpha = 0.05$.

4.1 Simulation Study

For space considerations, we only report results from the ZIP model for longitudinal data with sample size $n = 50$, 100 and 200. All simulations were performed with a Monte Carlo sample of 1,000. We start with data simulations under complete data.

4.1.1 Complete Data Case—For notational brevity, we considered a relatively simple pre-post longitudinal study design, with only one explanatory variable x_i following a normal distribution $N(1,1)$, and simulated the bivariate count response, $\mathbf{y}_i = (y_{i1}, y_{i2})^\top$, to satisfy the following marginal ZIP model:

$$y_{it}|x_i\sim\text{ZIP}(\rho_i,\mu_i),\text{logit}(\rho_i)=\beta_{u0},\log(\mu_i)=\beta_0+x_i\beta_1,1\le t\le 2,1\le i\le n. \quad (30)$$

We set $\beta_{u0} = -1$, $\beta_0 = \beta_1 = 1$. We first simulated x_i from $N(1, 1)$, and then conditional on x_i, generated y_{it} by using a copula approach(50; 51; 52). The copula method can generate correlated multivariate responses for any specified marginal distribution and correlation structure. For our simulation study, we set $Corr(y_{i1}, y_{i2} \mid x_i) = 0.5$.

To examine type I error rates, we considered the null, $H_0 : \beta_1 = 1$, and computed the Wald statistic, $Q_n=n\left(\hat{\sigma}_{\beta 1}^2\right)^{-1}\left(\hat{\beta}_1-1\right)^2$, where $\hat{\sigma}_{\beta 1}^2$ denotes the element of the estimated asymptotic variance $\tilde{\Sigma}_\theta$ corresponding to $\tilde{\beta}_1$. Let $Q_n^{(k)}$ denote this statistic at the kth MC simulation ($1 \le k \le 1000$). The type I error rate for testing H_0 was estimated by: $\hat{\alpha}=\frac{1}{1000}\sum_{k=1}^{1000}I_{\{Q_n^{(k)}\ge q_{1,0.95}\}}$, with $q_{1,0.95}$ denoting the 95th percentile of a central χ_1^2 with one degree of freedom.

Since Wald statistics are often anti-conservative, we also applied the score test in Section 3.2. Let $\theta=\left(\theta_{(1)}^\top,\theta_{(2)}\right)^\top$, where $\theta_{(1)} = (\beta_{u0}, \beta_0)^\top$ and $\theta_{(2)} = \beta_1$. Under H_0, $\theta_{(2)} = 1$, the score statistic $T_s(\tilde{\theta}_{(1)}, 1)$ in (29) has an χ_1^2 distribution. The type I error rate for testing H_0 was again estimated by: $\hat{\alpha}=\frac{1}{1000}\sum_{k=1}^{1000}I_{\{T_s^{(k)}\ge q_{1,0.95}\}}$, where $T_s^{(k)}$ denotes this statistic at the kth MC simulation ($1 \le m \le 1000$).

Shown in Table 1 are the estimates of θ, standard errors, and type I errors for the ZIP model in (30). For comparison purposes, we also included "Empirical" variance estimates and type I error rates based on such a variance estimate. The "Empirical" type I error rates were computed based on substituting Σ_θ with the Empirical variance estimate in the Wald test statistic. It is seen that type I error rates were a bit inflated for sample sizes 50 and 100 under the Wald test, but were closer to the nominal 0.05 under the "Score" and "Empirical" tests even for samples as small as $n = 50$.

To compare our approach with GEE II, we also estimated the parameters using a program developed for such an alternative by Hall and Zhang (2004)(24). As noted earlier, their method modeled the conditional variance, rather than the second moment. In addition, they assumed working independence between the mean and variance. We obtain quite similar results (not shown), which may not be surprising, as such differences are likely to have minor impact on inference given the marginal ZIP model in (30).

4.1.2 Missing Data Case—

Assuming no missing data at baseline $t = 1$, we simulated missing responses under MCAR and MAR with about 20% missing data at $t = 2$. By applying the discussion in Section 3.2 to the context of the pre-post design, we modeled the missingness at time $t = 2$ under MAR by:

$$\text{logit}(\pi_{i2}(\gamma)) = \gamma_0 + \gamma_x^\top x_{i1} + \gamma_y^\top y_{i1}, \gamma_x = \gamma_y = \frac{1}{2}.$$

We again considered the null H_0: $\beta_1 = 1$, and computed the Wald and score statistics and the associated type I error rates. The Wald statistic Q_n is computed the same way as in the complete data case except that the estimate of θ is obtained from the WGEE in (23).

Shown in Table 2(3) are the estimates of θ, standard errors, and type I errors for the ZIP model under MCAR (MAR). As in the complete data case, the score test again performed a marvelous job in correcting the upward bias in type I error rates by the Wald statistic in testing the null H_0: $\beta_1 = 1$, especially for the sample size $n = 50, 100$. For inference under MAR, the Wald statistic again yielded inflated type I error rates for testing the null, but the score test corrected the upward bias and maintained a type I error rate consistently near 0.05 across all sample sizes.

4.2 Real Study Data

To illustrate the approach to real study data, we applied it to a multi-center, NIDA-sponsored study entitled "HIV/STD Safer Sex Skills Groups For Men In Methadone Maintenance Or Drug-free Outpatient Treatment Programs," known as CTN0018 within the Clinical Trials Network (CTN) studies. This study was designed to examine the effectiveness of 5 session motivational and skills training in HIV/AIDS group interventions developed to reduce sexual risk behaviors in men, as compared to an HIV education only control condition. Unlike most community-based studies in which the HIV education provided is limited to information, this trial integrated a component to provide skill-training programs such as role plays to reducing sex risk behaviors. The primary outcome of the study is the number of unprotected vaginal and anal sexual intercourse occasions (USO) which was assessed at baseline, 2 weeks, 3- and 6-months(53; 54).

Out of 573 eligible subjects screened, 422 subjects completed assessment at baseline. Among these, 381 (91.27%) and 345 (60.2 %) came for assessment at 3- and 6-months. Since 2 weeks was too short to observe a reasonably large USO, we limited our analysis to the period from baseline to 3- and 6-months follow-up visits.

Shown in Table 4 are the mean USOs and percent of zero USO at baseline, 3- and 6-months for the two treatment groups. It is evident that there was a preponderance of zeros in the distribution of this outcome at each assessment time. Accordingly, we modeled the USO at

3-month (y_{i1}) and 6-month (y_{i2}) as a function of treatment condition, time and time by treatment interaction, controlling for baseline USO, y_{i0}, using the FRM-based ZIP model in (11) with

$$\text{logit}(\rho_{it})=\beta_{u0}+\beta_{u1}x_i+\beta_{u2}y_{i0}+\beta_{u3}t+\beta_{u4}t\cdot x_i, \log(\mu_{it})=\beta_0+x_i\beta_1=\beta_0+\beta_1x_i+\beta_2y_{i0}+\beta_3t+\beta_4t\cdot x_i, 1\le t\le 2, 1\le i\le n, \quad (31)$$

where x_i was an indicator with $x_i = 1$ for the intervention and 0 otherwise.

To account for potential response-dependent MAR missingness, we modeled the missingness under MMDP using logistic regression:

$$\text{logit}(p_{it}(\gamma_t))=\gamma_{0t}+\gamma_{xt}x_i+\gamma_{yt}y_{i(t-1)}, 1\le t\le 2, 1\le i\le n. \quad (32)$$

We assumed a Markov condition in (32) so that the missingness only depended on the most recent observed response.

Shown in Table 5 are the estimates of parameters from the logistic regression, their standard errors and corresponding p-values. The results show that the missingness at time $t = 1$ depended on the treatment assignment, while at time $t = 2$ depended on the observed response at time $t = 1$. In other words, the subjects in the intervention group were more likely to drop out than those in the control at time $t = 2$, while those with smaller values of USO at $t = 1$ were also more likely to drop out at $t = 2$. Based on these results, we proceeded with inference under MAR.

Shown in Table 6 are the estimates of parameters of the ZIP model, their standard errors and associated p-values. As the interaction term involving time and intervention was neither significant in the logistic (ρ_{it}) nor in the Poisson (μ_{it}) component of the model, we refit the model without this term, with the results from the revised model shown in Table 7.

For treatment effectiveness based on the results from the additive model, the logistic part of the model indicates that the intervention increased the likelihood of no risk for USO during the study, while the log-linear component shows that the intervention also significantly reduced the mean frequency of USO for the at-risk subgroup. The ratio of the mean USO of the treated to that of the control condition is exp(-0.09) = 0.9, suggesting a 10% decrease in USO for the treated subjects.

Baseline USO also played a significant role. The logistic component indicates that lower baseline USO would significantly increase the likelihood of being at no risk for USO during the study period. The log-linear part of the model shows that higher baseline USO was significantly associated with higher USO during the study. The findings suggest that substance abuse treatment programs should consider offering motivational exercises and skills training to achieve greater reductions in risky sexual activities.

5 Discussion

Count responses are a common type of outcome in biomedical, psychosocial and related services research. We discussed two major manifestations of departure from the Poisson assumption, overdispersion and structural zeros, and reviewed existing methods for addressing these two important issues. In particular, we focused on the limitations of available approaches with respect to longitudinal data analysis and proposed an approach to systematically tackle these problems under a unified modeling framework.

We applied the proposed approach to a real study in HIV prevention, allowing us to address important methodological issues in a timely application. In addition, the results from the simulation study show that the proposed approach works well for longitudinal study data under both complete and missing data settings. Although inference is derived based on large samples, the approach seems to provide valid inference for samples with sample size as small as 50.

Acknowledgments

This research was supported in part by NIH grant R21 DA027521-01. We want to thank two anonymous reviewers for very careful reviews of the manuscript, with constructive comments and edits that led to a significantly improved manuscript.

Appendix

Appendix A

The variance $Var(\mathbf{f}_i \mid \mathbf{x}_i)$ for the cross-sectional data case is readily computed using the moments up to the 4th order under either ZIP or NB distribution. The first two order moments for ZIP and NB are given in (9) and (10), while the 3rd and 4th order moments for the two models are given by:

$$\begin{gathered}\text{ZIP:}E(y^4)=\mu+7(\alpha+1)\mu^2+6(\alpha+1)(2\alpha+1)\mu^3+(\alpha+1)(2\alpha+1)(3\alpha+1)\mu^4,:E(y^3)=\mu+3(\alpha+1)\mu^2+(\alpha+1)(2\alpha+1)\mu^3,\\ \text{NB: } E(y^4|\mathbf{x})=(1-\rho)\mu(1+7\mu+6\mu^2+\mu^3),\quad E(y^3|\mathbf{x})=(1-\rho)\mu(1+3\mu+\mu^2),\end{gathered} \qquad (33)$$

Appendix B. Proof of Theorem 2

Let $G_i=D_iV_i^{-1}$ and $\pi_i = (\pi_{i1}, \dots, \pi_{im1})^\top$. Then, $w_n=\frac{1}{n}\sum_{i=1}^n G_i\Delta_iS_i$, with $G_i\Delta_iS_i = G_i(\mathbf{x}_i, \theta, \alpha)\Delta_i(\mathbf{r}_i, \pi_i, \gamma)S_i(\mathbf{y}_i, \mathbf{x}_i, \theta)$. It follows from the iterated conditional expectation that $E(G_i\Delta_iS_i) = E\,[G_iS_iE(\Delta_i \mid \mathbf{r}_i, \mathbf{y}_i, \mathbf{x}_i)]$. By definition, Δ_i is a $m \times m$ block diagonal matrix with the tth block diagonal matrix given by $\frac{r_{it}}{\pi_{it}}\mathbf{I}_2(1\le t\le m)$, with $\mathbf{I}_m$ denoting the $m \times m$ identify matrix. Since $E\left(\frac{\mathbf{r}_{it}}{\pi_{it}}\mathbf{I}_2|r_i,\mathbf{y}_i,\mathbf{x}_i\right)=\mathbf{I}_2$, it follows that $E(G_i\Delta_\text{i}S_i) = E(G_iS_i) = 0$. Thus, the WGEE II is unbiased and the estimate $\hat{\theta}$ obtained as the solution to the equations is consistent.

Let $\hat{\gamma}$ be the solution to the (22). By a Taylor expansion of the estimating equations in (22) and solving for $\hat{\gamma}-\gamma$, we obtain

$$\sqrt{n}\,(\widehat{\gamma}-\gamma)=-H^{-1}\frac{\sqrt{n}}{n}\sum_{i=1}^{n}\mathbf{Q}_{ni}+\mathbf{o}_p(1), \qquad (34)$$

where $\mathbf{o}_p(1)$ denotes the stochastic $\mathbf{o}(1)$(25). Also, by applying a Taylor series expansion to the WGEE II in (23), we have

$$\sqrt{n}\mathbf{w}_n=-\left(\frac{\partial}{\partial\theta}\mathbf{w}_n\right)^\top\sqrt{n}(\widehat{\theta}-\theta)-\left(\frac{\partial}{\partial\alpha}\mathbf{w}_n\right)^\top\sqrt{n}(\widehat{\alpha}-\alpha)--\left(\frac{\partial}{\partial\gamma}\mathbf{w}_n\right)^\top\sqrt{n}(\widehat{\gamma}-\gamma)+\mathbf{o}_p(1). \qquad (35)$$

If $\hat{\alpha}$ is $\sqrt{n}-\text{consistent}$, it follows that

$$\left(\frac{\partial}{\partial\alpha}\mathbf{w}_n(\theta,\alpha)\right)^{\top}\sqrt{n}(\widehat{\alpha}-\alpha)=\mathbf{o}_p(1)\ \sqrt{n}(\widehat{\alpha}-\alpha)=\mathbf{o}_p(1).$$

By substituting $\mathbf{o}_p(1)$ for $\left(\frac{\partial}{\partial\alpha}\mathbf{w}_n(\theta,\alpha)\right)^{\top}\sqrt{n}(\widehat{\alpha}-\alpha)$ in (35) and solving for $\sqrt{n}\left(\widehat{\theta}-\theta\right)$ $(\hat{\theta}-\theta)$, we obtain

$$\sqrt{n}\left(\widehat{\theta}-\theta\right)=\left(-\frac{\partial}{\partial\theta}\mathbf{w}_n\right)^{-\top}\sqrt{n}[\mathbf{w}_n+C(\widehat{\gamma}-\gamma)]+\mathbf{o}_p(1). \quad (36)$$

It follows from (34) and (36) that

$$\sqrt{n}\left(\widehat{\theta}-\theta\right)=\left(-\frac{\partial}{\partial\theta}\mathbf{w}_n\right)^{-\top}\frac{\sqrt{n}}{n}\sum_{i=1}^{n}(\mathbf{w}_{ni}-CH^{-1}\mathbf{Q}_{ni})+\mathbf{o}_p(1). \quad (37)$$

Since

$$\frac{\partial}{\partial\theta}\mathbf{w}_n=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial}{\partial\theta}\Delta_i S_i\right)G_i^{\top}+\mathbf{o}_p(1)=\frac{1}{n}\sum_{i=1}^{n}D_i\Delta_i G_i^{\top}+\mathbf{o}_p(1)\rightarrow_p -B^{\top}. \quad (38)$$

where $\rightarrow_p$ denotes convergence in probability, it follows from (37) and (38) that

$$\sqrt{n}\left(\widehat{\theta}-\theta\right)=-B^{-\top}\frac{\sqrt{n}}{n}\sum_{i=1}^{n}(w_{ni}-CH^{-1}Q_{ni})+o_p(1). \quad (39)$$

By applying the central limit theorem and Slutsky's theorem to (39)(25), $\hat{\theta}$ is asymptotically normal with the asymptotic variance given by Σ_{θ} in (24).

Appendix C. Asymptotic Normality of Score Statistic

First, assume no missing data. Then, $B=E(D_iV_i^{-1}D_i)$ By applying the law of large numbers,

$$\frac{\partial}{\partial\theta}\mathbf{w}_n(\theta)=\begin{pmatrix}\frac{\partial}{\partial\theta_{(1)}}\mathbf{w}_{n(1)}(\theta) & \frac{\partial}{\partial\theta_{(1)}}\mathbf{w}_{n(2)}(\theta)\\ \frac{\partial}{\partial\theta_{(2)}}\mathbf{w}_{n(2)}(\theta) & \frac{\partial}{\partial\theta_{(2)}}\mathbf{w}_{n(2)}(\theta)\end{pmatrix}\rightarrow_p B=E(D_iV_i^{-1}D_i)=\begin{pmatrix}B_{11} & B_{12}\\ B_{12}^{\top} & B_{22}\end{pmatrix}. \quad (40)$$

It follows from a Taylor's series expansion and (40) that

$$0=\mathbf{w}_{n(1)}\left(\tilde{\theta}_{(1)},\theta_{(20)}\right)\mathbf{w}_{n(1)}(\theta)-B_{11}^{-\top}\left(\tilde{\theta}_{(1)}-\theta_{(1)}\right)+\mathbf{o}_p\left(n^{-\frac{1}{2}}\right).$$

Thus,

$$\tilde{\theta}_{(1)}-\theta_{(1)}=B_{11}^{-1}\mathbf{w}_{n(1)}(\theta)+\mathbf{o}_p\left(n^{-\frac{1}{2}}\right). \quad (41)$$

Similarly, since $B_{12}^{\top}=B_{21}$, we have:

$$\begin{aligned}\mathbf{w}_{n(2)}\left(\tilde{\theta}_{(1)},\theta_{(20)}\right)&=\mathbf{w}_{n(2)}(\theta)-\left(\frac{\partial^{\top}}{\partial\theta_{(1)}}\mathbf{w}_{n(2)}\right)\left(\tilde{\theta}_{(1)}-\theta_{(1)}\right)+\mathbf{o}_p\left(n^{-\frac{1}{2}}\right)\\&=\mathbf{w}_{n(2)}(\theta)-B_{21}\left(\tilde{\theta}_{(1)}-\theta_{(1)}\right)+\mathbf{o}_p\left(n^{-\frac{1}{2}}\right).\end{aligned} \quad (42)$$

It follows from (41) and (42) that

$$\begin{aligned}\mathbf{w}_{n(2)}\left(\tilde{\theta}_{(1)},\theta_{(20)}\right)&=\mathbf{w}_{n(2)}(\theta)-B_{21}\left[B_{11}^{-1}\mathbf{w}_{n(1)}(\theta)+\mathbf{o}_p\left(n^{-\frac{1}{2}}\right)\right]+\mathbf{o}_p\left(n^{-\frac{1}{2}}\right)\\&=G\mathbf{w}_n(\theta)+\mathbf{o}_p\left(n^{-\frac{1}{2}}\right).\end{aligned}$$

By the central limit theorem,

$$\sqrt{n}\mathbf{w}_{n(2)}\left(\tilde{\theta}_{(1)},\theta_{(20)}\right)=\sqrt{n}G\mathbf{w}_n(\theta)+\mathbf{o}_p(1)\rightarrow_d N\left(0,\sum_{(2)}=G\sum_{\theta}G^{\top}\right). \quad (43)$$

where G is defined in (28) and Σ_{θ} in (24).

In the presence of missing data, $B=E(D_iV_i^{-1}\Delta_iD_i)$ as defined in (28). By a similar argument, $\mathbf{w}_{n(2)}$ $(\tilde{\theta}_{(1)}, \theta_{(20)})$ has an asymptotic normal distribution, which implies that the score statistic $T_s((\tilde{\theta}_{(1)}, \theta_{(2)}))$ has the asymptotic χ_q^2 distribution.

References

1. Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics. 1992; 34:1–14.
2. 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.
3. 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]
4. 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.
5. Faddy, M. Stochastic models for analysis of species abundance data. In: Fletcher, DJ.; Kavalieris, L.; Manly, BF., editors. Statistics in Ecology and Environmental Monitoring 2: Decision Making and Risk Assessment in Biology. University of Otago Press; 1998. p. 33-40.
6. Gurmu S, Trivedi P. Excess zeros in count models for recreational trips. Journal of Business & Economic Statistics. 1996; 14:469–477.
7. Gurmu S. Semi-parametric estimation of hurdle regression models with an application to Medicaid utilization. Journal of Applied Econometrics. 1997; 12:225–242.
8. Shonkwiler J, Shaw W. Hurdle count-data models in recreation demand analysis. Journal of Agricultural and Resource Economics. 1996; 21:210–219.
9. Hall DB. Zero-Inflated Poisson and binomial regression with random effects: A case study. Biometrics. 2000; 56:1030–1039. [PubMed: 11129458]

10. Yau KW, Lee AH. Zero-inflated Poisson regression with random effects to evaluate an occupational injury prevention programme. Statistics in Medicine. 2001; 20:2907–2920. [PubMed: 11568948]
11. World Health Organization. Optimal duration of exclusive breastfeeding. Geneva: WHO; 2001.
12. Donath S, Amir LH. Rates of breastfeeding in Australia by State and socio-economic status: Evidence from the 1995 National Health Survey. Journal of Pediatrics and Child Health. 2000; 36(2):164–168.
13. Cheung YB. Zero-infated models for regression analysis of count study of growth and development. Statistics in Medicine. 2002; 21:1461–1469. [PubMed: 12185896]
14. Wyman PA, Cross W, Brown HC, Yu Q, Tu XM. Intervention to strengthen emotional self-regulation in children with emerging mental health problems: Proximal impact on school behavior. Journal of Abnormal Child Psychology. in press.
15. Abma JC, Martinez GM, Mosher WD, Dawson BS. Teenagers in the United States: Sexual activity, contraceptive use, and child bearing. Vital Health Statistics. 2002; 23(24)
16. Abe T, Martin I, Roche L. Clusters of Census Tracts with High Proportions of Men with Distant-Stage Prostate Cancer Incidence in New Jersey, 1995 to 1999. American Journal of Preventive Medicine. 2006; 30(2):S60–S66. [PubMed: 16458791]
17. 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; 3:5–20.
18. Lachenbruch PA. Analysis of data with excess zeros. Statistical Methods in Medical Research. 2002; 11:297–302. [PubMed: 12197297]
19. Lee AH, Wang K, Scott JA, Yau KKW, McLachlan GJ. Multi-level zero-inflated Poisson regression modelling of correlated count data with excess zeros. Statistical Methods in Medical Research. 2006; 15:47–61.
20. Ritz J, Spiegelman D. Equivalence of conditional and marginal regression models for clustered and longitudinal data. Statistical Methods in Medical Research. 2004; 13:309–323.
21. 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.
22. Zhang H, Tang W, Yu Q, Feng C, Gunzler D, Tu X. A New Look at the Differerence between GEE and GLMM When Modeling Longitudinal Count Responses. Journal of Applied Statistics.
23. Estimating Equations. Oxford University Press; New York: 1991. Estimating equations for mixed Poisson models; p. 35-46.
24. Hall DB, Zhang ZG. Marginal models for zero inflated clustered data. Statistical Modeling. 2004; 4:161–180.
25. Kowalski, J.; Tu, XM. Modern Applied U Statistics. Wiley; New York: 2007.
26. Crowder M. On linear and quadratic estimating functions. Biometrika. 1987; 74:591–97.
27. Dobbie MJ, Welsh AH. Modeling correlated zero-inflated count data. Australian & New Zealand Journal of Statistics. 2001; 43:431–444.
28. Prentice RL, Zhao LP. Estimating Equations for Parameters in Means and Covariances of Multivariate Discrete and Continuous Responses. Biometrics. 1991; 47:825–839. [PubMed: 1742441]
29. Liang KY, Zeger SL, Qaqish B. Multivariate regression analyses for categorical data. J R Statist Soc B. 1992; 54:3–40. Rubeussin and Liang, 1998.
30. Rubin DB. Inference and Missing Data. Biometrika. 1976; 63:581–592.
31. Little, RJA.; Rubin, DB. Statistical Analysis with Missing Data. New York: Wiley; 1987.
32. McCullagh, P.; Nelder, JA. Generalized Linear Models. 2nd. Chapman and Hall; London: 1989.
33. Dean CB, Lawless JF. Tests for detecting overdispersion in Poisson regression models. J Amer Statist Assoc. 1989; 84:467–472.
34. 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.
35. Lee LF. Specification test for Poisson regression models. International Economic Review. 1986; 27:689–706.

36. Tu XM, Feng C, Kowalski J, Tang W, Wang H, Wan C, Ma Y. Correlation analysis for longitudinal data: Applications to HIV and psychosocial research. Statistics in Medicine. 2007; 26:4116–4138. [PubMed: 17342700]

37. Ma Y, Tang W, Feng C, Tu XM. Inference for kappas for longitudinal study data: applications to sexual health research. Biometrics. 2008; 64:781–789. [PubMed: 18047535]

38. Ma Y, Tang W, Yu Q, Tu XM. Modeling concordance correlation coefficient for longitudinal study data. Psychometrika. 2010; 75:99–119.

39. Ma Y, Gonzalez Della Valle A, Zhang H, Tu XM. A U-statistics based approach for modeling Cronbach Coefficient Alpha within a longitudinal data setting. Statistics in Medicine. 2011; 29(6): 659–670.

40. Yu Q, Tang W, Kowalski J, Tu XM. Multivariate U-Statistics: A Tutorial with applications. Wiley Interdisciplinary Reviews – Computational Statistics. 2011; 3:457–471.

41. Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. JASA. 1995; 90:106–121.

42. Cameron, AC.; Trivedi, PK. Regression analysis of counter data. Cambridge Univ. Press; London: 1998.

43. Reboussin BA, Liang KY. An estimating equations approach for the LISCOMP Model. Psychometrika. 1998; 63:165–182.

44. Yu, Q. Department of Biostatistics and Computational Biology School of Medicine and Dentistry. University of Rochester; Rochester, New York: 2009. Distribution-free models for longitudinal count data. Ph.D. Thesis.

45. Pepe MS, Anderson GL. A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. Communications in Statistics: Simulation and Computation. 1994; 23:939–951.

46. Scharfstein, DO.; Rotnitzky, A.; Robins, JM. Adjusting for nonignorable drop-out using semi-parametric nonresponse models. Vol. 94. Journal of the American Statistical Association; 1999. p. 1096-1146.

47. Tsiatis, AA. Semiparametric Theory and Missing Data. New York: Spring; 2006.

48. Rotnitzky A, Jewell NP. Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. Biometrika. 1990; 77:485–497.

49. Pan W. On the robust variance estimator in generalized estimating equations. Biometrika. 2001; 88:901–906.

50. Freesm EW, Valdez EA. Understanding relationships using copulas. North American Actuarial Journal. 1998; 2:1–25.

51. Nelsen, RB. An introduction to Copulas. Springer; New York: 2006.

52. Yan, JR. Package copula on CRAN, multivariate dependence with copula. 2009. http://cran.r-project.org/web/packages/copula/index.html

53. Calsyn, DA.; Wells, EA.; Saxon, AJ.; Jackson, R.; Heiman, JR. Sexual activity under the influence of drugs is common among methadone clients. In: Harris, L., editor. Problems of Drug Dependence 1999. Vol. 315. National Institute on Drug Abuse; 2000. NIH Pub. No. 00-4773

54. 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]

Table 1

GEE estimates of parameters, standard errors, and type I error rates based on Wald and score tests, along with empirical standard errors and type I error rates for ZIP under complete data from 1,000 MC simulations.

Simulation summary for ZIP under complete data						
$\beta_{u0} = -1, \beta_0 = 1, \beta_1 = 1$						
Parameter	**Mean**	**Standard errors**		**Type I error for $H_0 : \beta_1 = 1$**		
		WGEE	**Empirical**	**Wald**		**Score**
				WGEE	**Empirical**	
Sample size of 50						
β_{u0}	−1.052	0.363	0.385			
β_0	1.000	0.090	0.100			
β_1	0.998	0.039	0.048			
				0.095	0.061	0.045
Sample size of 100						
β_{u0}	−1.021	0.252	0.256			
β_0	1.000	0.063	0.067			
β_1	0.999	0.027	0.031			
				0.076	0.052	0.054
Sample size of 200						
β_{u0}	−1.012	0.177	0.176			
β_0	0.999	0.044	0.046			
β_1	1.000	0.019	0.021			
				0.065	0.042	0.042

Table 2

WGEE estimates of parameters, standard errors, and type I error rates based on Wald and Score tests, along with empirical standard errors and type I error rates for ZIP under MCAR from 1,000 MC simulations.

Simulation summary for ZIP under missing data following MCAR						
$\beta_{u0} = -1, \beta_0 = 1, \beta_1 = 1$						
Parameter	**Mean**	**Standard errors**		**Type I error for $H_0 : \beta_1 = 1$**		
		GEE	**Empirical**	**Wald**		**Score**
				GEE	**Empirical**	
Sample size of 50						
β_{u0}	−1.077	0.378	0.402			
β_0	0.991	0.112	0.120			
β_1	0.997	0.115	0.135			
				0.108	0.061	0.046
Sample size of 100						
β_{u0}	−1.026	0.257	0.258			
β_0	0.997	0.080	0.082			
β_1	0.998	0.082	0.088			
				0.075	0.057	0.044
Sample size of 200						
β_{u0}	−1.016	0.180	0.183			
β_0	0.998	0.057	0.055			
β_1	1.000	0.059	0.060			
				0.055	0.049	0.045

Table 3

WGEE estimates of parameters, standard errors, and type I error rates based on Wald and Score tests, along with empirical standard errors and type I error rates for ZIP under MAR from 1,000 MC simulations.

Simulation summary for ZIP under missing data following MAR						
$\beta_{u0} = -1, \beta_0 = 1, \beta_1 = 1$						
Parameter	**Mean**	**Standard errors**		**Type I error for $H_0 : \beta_1 = 1$**		
		WGEE	**Empirical**	**Wald**		**Score**
				WGEE	**Empirical**	
Sample size of 50						
β_{u0}	−1.05	0.402	0.400			
β_0	0.995	0.128	0.105			
β_1	1.000	0.168	0.151			
				0.094	0.062	0.052
Sample size of 100						
β_{u0}	−1.02	0.253	0.261			
β_0	1.001	0.064	0.066			
β_1	0.998	0.088	0.080			
				0.087	0.058	0.043
Sample size of 200						
β_{u0}	−1.01	0.176	0.177			
β_0	0.999	0.044	0.044			
β_1	1.000	0.066	0.060			
				0.055	0.056	0.051

Table 4

Comparison of mean USO and percent of zero USO between the two treatment groups at baseline, 3- and 6-month follow-up for the CTN0018 Study.

Mean USO and number of zeros at each assessment time for CTN0018 study			
	Intervention (S.D.)	**Without intervention (S.D.)**	**zeros (%)**
Baseline	21.46(26.66)	22.34(27.77)	65(15.40)
USO at 3 months	15.71(25.43)	18.14(27.21)	125(32.80)
USO at 6 months	15.05(23.35)	17.19(25.89)	132(38.26)

Table 5

Estimates of logistic regression for modeling missingness under MAR and MMDP for CTN0018 Study.

Estimates of logistic regression for modeling missingness for CTN0018 study			
Assessment time *t* = 1			
Predictors	**Estimates**	**Standard errors**	**P-values**
Intercept	2.777	0.319	< 0.001
y_{i1}	−0.002	0.006	0.752
intervention	−0.869	0.351	0.013
Assessment time *t* = 2			
Intercept	1.443	0.206	< 0.001
y_{i2}	0.019	0.007	0.007
intervention	−0.325	0.257	0.206

Table 6

WGEE estimates of parameters, standard errors, and p-values from FRM-based ZIP model with treatment by time interaction based on Wald and score tests under MAR and MMDP for the CTN0018 Study.

Results of FRM-based ZIP model for CTN0018 study				
			P-value for $H_0 : \beta = 0$	
Parameter	**Estimate**	**Standard errors**	**Wald**	**Score**
Log-linear part (μ_{it})				
β_0	2.69	0.196	< 0.001	< 0.001
β_1 (intervention)	−0.08	0.028	< 0.001	< 0.001
β_2 (baseline USO)	0.012	0.001	< 0.001	< 0.001
β_3 (time)	−0.017	0.118	0.885	0.883
β_4(intervention*time)	−0.062	0.187	0.742	0.741
Logistic part (ρ_{it})				
β_{u0}	−0.52	0.354	0.142	0.140
β_{u1} (intervention)	0.301	0.499	0.564	0.562
β_{u2} (baseline USO)	−0.017	0.004	< 0.001	< 0.001
β_{u3}(time)	0.126	0.221	0.568	0.566
β_{u4}(intervention*time)	−0.121	0.314	0.701	0.700

Table 7

WGEE estimates of parameters, standard errors, and p-values from revised additive ZIP model based on Wald and score tests under MAR and MMDP for the CTN0018 Study.

Results from revised additive ZIP model for CTN0018 study				
Parameter	**Estimate**	**Standard errors**	**P-value for $H_0 : \beta = 0$**	
			Wald	**Score**
Log-linear part (μ_{it})				
β_0	2.90	0.021	< 0.001	< 0.001
β_1 (intervention)	−0.09	0.025	< 0.001	< 0.001
β_2 (baseline USO)	0.012	0.0004	< 0.001	< 0.001
Logistic part (ρ_{it})				
β_{u0}	−0.68	0.144	< 0.001	< 0.001
β_{u1} (intervention)	0.371	0.200	0.065	0.068
β_{u2} (baseline USO)	−0.015	0.004	< 0.001	< 0.001