MARGINALIZED ZERO-INFLATED POISSON REGRESSION

Dorothy Leann Long

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics.

Chapel Hill
2013

Approved by:

Dr. Amy H. Herring
Dr. John S. Preisser
Dr. Carol Golin
Dr. Brian Neelon
Dr. John W. Stamm
Dr. C. M. Suchindran

© 2013
Dorothy Leann Long
ALL RIGHTS RESERVED

Abstract

DOROTHY LEANN LONG: Marginalized Zero-inflated Poisson Regression
(Under the direction of Dr. Amy H. Herring and Dr. John S. Preisser)

The zero-inflated Poisson (ZIP) regression model is often employed in public health research to examine the relationships between exposures of interest and a count outcome exhibiting many zeros, in excess of the amount expected under sampling from a Poisson distribution. The regression coefficients of the ZIP model have latent class interpretations, which correspond to a susceptible subpopulation at risk for the condition, with counts generated from a Poisson distribution, and a non-susceptible subpopulation that provides the extra or excess zeros. The ZIP model parameters, however, are not well suited for inference targeted at overall exposure effects, specifically, in quantifying the effect of an explanatory variable in the overall mixture population. We develop a marginalized ZIP model for independent responses to model the population mean count directly, allowing straightforward inference for overall exposure effects and easy accommodation of offsets representing individuals' risk times, as well as empirical robust variance estimation for overall log incidence density ratios. Through simulation studies, the performance of maximum likelihood estimation of the marginalized ZIP model is assessed and compared to existing post-hoc methods for the estimation of overall effects in the traditional ZIP model framework. The marginalized ZIP model is applied to a recent study of a motivational interview-based safer sex counseling intervention, designed to reduce unprotected sexual act counts. Also, we develop a marginalized ZIP model with random effects to allow for more complicated data structures. SAS macros are developed for the marginalized ZIP model for independent data to assist applied

analysts in the direct modeling of the population mean in count data with excess zeros.

To my wonderful parents, you have always believed in me and encouraged me to pursue my dreams. To my fantastic husband Dustin, we both know this dissertation would have never happened without your loving patience, support, and encouragement. To Charlie, your smiles and unconditional love brighten every single day.

Acknowledgments

Most importantly I would like to thank my dissertation advisers. John Preisser has provided not only the ingenuity behind this work, but he has also given invaluable advice and guidance that have made me a better researcher. Through her outstanding mentoring and leadership, Amy Herring has been my strongest advocate, giving sage advice, offering much-needed encouragement, and challenging me to reach my full potential. I am extremely fortunate to have both Amy and John guide my work, both providing inspiration for the researcher I hope to be.

I would like to thank Dr. Suchindran, both for taking a chance on a pair of students from Tennessee and being so generous and supportive through my time at UNC. I would also like to thank my committee members, Carol Golin, Brian Neelon, and John Stamm, for their support and helpful comments which greatly improved this document. Also, I am very appreciative of the SafeTalk study team and participants for allowing me to work with and for them.

In addition to the support I have received from Dustin, I would like to thank several of my fellow classmates; Jennifer Clark, Annie Green Howard, Beth Horton, Andrea Byrnes, Elena Bordonali, and Matt Wheeler have all been sources of strength and comfort when coursework (and life) became overwhelming. Additionally, I could always depend on Melissa Hobgood for help, or words of wisdom, no matter the issue. You all will never know how truly vital you have been to my graduation.

This work was supported in part by NIH grants 2T32HD007237-25 (NICHD),

T32ES007018 (NIEHS), R01-MH069989, DK56350, and AI50410.

Table of Contents

List of Tables

List of Figures

Chapter 1

Literature Review

1.1 Motivation

As many fields of research involve count data, one of the most useful statistical models is Poisson regression; however the strong assumption of mean-variance equality of the Poisson distribution is rarely met because a vast proportion of count data exhibit variance in excess of the mean. While methods have been developed to incorporate the overdispersion of count data through the negative binomial distribution (McCullagh and Nelder, 1989), these methods usually do not adequately model data with excess zero counts. Data with excess zero counts, also called zero-inflated data, appear quite frequently in various fields of research, including health research, agricultural research, ecology and manufacturing (Ridout, Demétrio, and Hinde, 1998). In agriculture, zero-inflated data are often found when examining counts of roots produced from cuttings, where many cuttings produce no roots. When ecologists are studying the prevalence of a species within a geographical region, their data might include a count of the species within various subregions, of which many observations can be zero. Examining the numbers of defects from manufacturing machines, researchers might find that most machines create no defected items, and those machines that do create some number

substandard products have a separate distribution. Many such examples of zero-inflated count data stem from the presence of multiple underlying unique populations, where factors that distinguish these separate populations are often latent.

In public health, much research is performed with the intent of understanding the incidence of a health event and its relationship with some exposure(s) of interest. In dental research, suppose an investigator is interested in the incidence of dental caries among children and determining whether the incidence depends upon a number of covariates. Due to some underlying set of confounders such as household fluoride levels or genetic factors, many children have no dental caries present at screening and would perhaps represent a subpopulation of those not susceptible to the condition of interest. On the other hand, the children who are at risk for caries have counts, not necessarily strictly positive, of dental caries and represent a subpopulation of subjects that might have a different distribution than those children not at risk. However, often data that are zero-inflated can arise from one homogeneous population where the notion of latent subpopulation categorization is not meaningful. In the dental caries example, researchers might not believe that a subpopulation of insusceptible children is clinically meaningful and might desire inference on the entire sampled population. Despite the exact nature of the inference desired, investigators have several options for statistical analyses of zero-inflated data.

1.2 Zero-Inflated Methods

Many statistical methods designed to account for zero-inflated data are two-stage modeling processes. Those observations at zero, either all or some subset of them, are modeled through a distribution appropriate for binomial data and then the remaining realizations are modeled using a count distribution. When analyzing zero-inflated data, researchers have two distinct methodologies from which to choose, the zero-inflated

Poisson (ZIP) regression model and the hurdle model. The hurdle model employs a binomial process to model all the zero counts, and then positive realizations are modeled through a truncated-at-zero count distribution. In the dental caries example, the hurdle model would describe those children with no caries separately from all the children who had any caries. Instead of modeling all the zero observations in the first process, the ZIP model models the 'excess zeros' in binomial process and then a count distribution is fit using a full Poisson likelihood. The ZIP model allows for zeros to occur within the distribution of a second population. In terms of dental caries, the zero-inflated Poisson model describes the 'excess zero' children, perhaps those not at risk of caries, separately from all those children susceptible to caries, but not necessarily with observed dental caries. We will explore these two methods in the following sections.

1.2.1 Zero-Inflated Models

Lambert (1992) introduces the concept and some theory behind ZIP regression models, using a motivating example on solderability of boards (an experiment at AT&T). The ZIP model allows for modeling of the zeros in two ways, first the 'excess' zeros and then the zeros which occur in the Poisson distribution, which may occur for different reasons. The type of zero observed (excess or Poisson) is a latent variable. For Y_i, $i = 1, \ldots, n$, the ZIP data are structured

$$Y_i \sim \begin{cases} 0 & \text{with probability } \psi_i \\ \text{Poisson}(\mu_i) & \text{with probability } 1 - \psi_i \end{cases}$$

yielding

$$Y_i \;=\; \begin{cases} 0 & \text{with probability } \psi_i + (1-\psi_i)e^{-\mu_i} \\ k & \text{with probability } (1-\psi_i)e^{-\mu_i}\mu_i^k/k!, \quad k \in \mathcal{Z}^+. \end{cases}$$

For the i^{th} subject, ψ_i is the probability of being an excess zero and μ_i is the mean of the non-excess zero population. To model these parameters of ψ_i and μ_i, we define

$$\begin{aligned} \text{logit } (\psi_i) &= \boldsymbol{Z}_i\boldsymbol{\gamma} \\ \log\,(\mu_i) &= \boldsymbol{X}_i\boldsymbol{\beta} \end{aligned}$$

where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{p_1})'$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{p_2})'$ and $\boldsymbol{Z}_i$, $\boldsymbol{X}_i$ are $(1 \times p_1)$ and $(1 \times p_2)$ vectors of covariates for the i^{th} unit. The log-likelihood for the ZIP model can be

$$\begin{aligned} l(\boldsymbol{\beta}, \boldsymbol{\gamma}|\boldsymbol{y}) \;=\; & \sum_{y_i=0} \log\, \left[e^{\boldsymbol{Z}_i\gamma} + e^{-e^{\boldsymbol{X}_i\boldsymbol{\beta}}}\right] + \sum_{y_i>0}(y_i\boldsymbol{X}_i\boldsymbol{\beta} - e^{\boldsymbol{X}_i\beta}) \\ & - \sum_{i=1}^{n} \log\,(1 + e^{Z_i\gamma}) - \sum_{y_i>0} \log\,(y_i!). \end{aligned}$$

In the ZIP model, the parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ have latent class interpretations; that is, γ_j is the log-odds ratio of a one-unit increase in the j^{th} element of $\boldsymbol{Z}_i$ on the probability of being an *excess* zero and β_j is the log-incidence density ratio of a one-unit increase in the j^{th} element of $\boldsymbol{X}_i$ on the mean of the *suspectible* sub-population. No simple summary of the exposure effect on the overall mean of the outcome is directly available. Lambert admits that ZIP regression is difficult to interpret when the set of covariates affect ψ, μ and the mean number of defects $E(Y_i) = (1-\psi_i)\mu_i$ differently. That is, an explanatory variable's effect cannot be necessarily measured through its effect on ψ or μ alone; the overall effect on $E(Y_i) = (1-\psi_i)\mu_i$ is what needs to be examined.

Lambert identifies that there are two model situations: one in which $\boldsymbol{\psi}$ and $\boldsymbol{\mu}$ are unrelated and another in which $\boldsymbol{\psi}$ can be defined as a function of $\boldsymbol{\mu}$. Lambert examines the maximum likelihood estimation of the ZIP model parameters through the EM algorithm. For the special case where $\boldsymbol{\psi}$ is defined as a function of $\boldsymbol{\mu}$, Lambert suggests using the Newton-Raphson algorithm. Since the complete data log-likelihood can be separated into function of $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ alone, then one can maximize over these separately using the EM algorithm, where the complete data log-likelihood is

$$\begin{aligned} l_c(\boldsymbol{\beta}, \boldsymbol{\gamma} | \boldsymbol{y}, \boldsymbol{z}) &= \sum_{i=1}^{n} \{(z_i \boldsymbol{Z_i} \boldsymbol{\gamma} - \log\,(1 + e^{\boldsymbol{Z_i}\boldsymbol{\gamma}})) + \left[(1 - z_i)(y_i \boldsymbol{X_i} \boldsymbol{\beta} - e^{\boldsymbol{X_i}\boldsymbol{\beta}})\right] \\ &\qquad - (1 - z_i)\log\,(y_i!)\} \\ &= l_c(\boldsymbol{\gamma} | \boldsymbol{y}, \boldsymbol{z}) + l_c(\boldsymbol{\beta} | \boldsymbol{y}, \boldsymbol{z}) - \sum_{i=1}^{n} (1 - z_i)\log\,(y_i!). \end{aligned}$$

Here z_i is latent class indicator of whether the i^{th} observation originates from the zero process ($z_i = 1$) or the Poisson process ($z_i = 0$). Exploiting the mixture structure of the zero-inflated Poisson model, the EM algorithm iteratively fits weighted versions of simpler generalized linear models (Hall and Zhang, 2004). For large sample sizes, Lambert notes the MLE's for the ZIP model parameters are consistent and following a normal distribution with means $(\boldsymbol{\gamma}, \boldsymbol{\beta})$ and variances equal to the inverse of the observed information matrices.

Adapting Lambert's ZIP regression model, Hall (2000) formulates the zero-inflated binomial regression model to handle bounded count data and also expands the ZIP and binomial models to include cluster random effects. After briefly summarizing the zero-altered (hurdle) models, Hall discusses how the ZIP is preferable due to its interpretability and suitability for many types of data. For the example of pesticide efficacy on reducing the number of Whitefly, the parameters associated with the logistic model

quantify the effects of covariates on the probability that the pesticide is fully effective, and the parameters in the Poisson process explain the association between the covariates and the mean number of insects occurring when the pesticide is not fully effective. Hall argues that both sets of parameters are scientifically meaningful, either when tested jointly or separately. In the derivations, Hall discusses Lambert's EM algorithm for estimating the ZIP model parameters and notes that solving the M step via unweighted logistic regression is more straightforward than the weighted logistic regression on an augmented data set proposed by Lambert. With either method, both Lambert and Hall agree that the ZIP regression model is 'not hard to fit.'

To account for correlated observations, Hall proposes including a random effect in the count model portion of the zero-inflated Poisson and binomial models. Specifically, let $\boldsymbol{Y} = (\boldsymbol{Y}_1', \ldots, \boldsymbol{Y}_K')$ where K is the number of independent clusters and $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iT_i})'$ and T_i is the number of observations for the i^{th} cluster. Then

$$Y_{ij} \sim \begin{cases} 0 & \text{with probability } p_{ij} \\ \text{Poisson } (\lambda_{ij}) & \text{with probability } 1 - p_{ij}. \end{cases}$$

where ψ_{ij} is the probability of being an excess zero and μ_{ij} is the mean of the non-excess zero population for the i^{th} cluster and j^{th} observation. The log-linear and logistic regression models are

$$\begin{aligned} \text{logit}(\psi_{ij}) &= \boldsymbol{Z}_{ij}\boldsymbol{\gamma} \\ \log(\mu_{ij}) &= \boldsymbol{X}_{ij}\boldsymbol{\beta} + \sigma b_i \end{aligned}$$

where $b_1, \ldots, b_K \overset{i.i.d.}{\sim} N(0,1)$, $\boldsymbol{Z}_i$ and $\boldsymbol{X}_i$ are the design matrices for the logistic and Poisson processes, respectively. The log-likelihood for Hall's ZIP model with random

effects can be expressed

$$l(\boldsymbol{\theta}, \boldsymbol{y}) = \sum_{i=1}^{K} \log \int_{-\infty}^{\infty} \left[\prod_{j=1}^{T_i} \Pr(Y_{ij} = y_{ij} | b_i) \right] \phi(b_i)\ db_i$$

where $\boldsymbol{\theta} = (\boldsymbol{\gamma}', \boldsymbol{\beta}', \sigma)$, ϕ is the standard normal probability density and

$$\begin{aligned} \Pr(Y_{ij} = y_{ij} | b_i) &= \left[\psi_{ij} + (1 - \psi_{ij}) e^{-\mu_{ij}} \right]^{u_{ij}} \left[\frac{(1 - \psi_{ij}) e^{-\mu_{ij}} \mu_{ij}^{y_{ij}}}{y_{ij}!} \right]^{1 - u_{ij}} \\ &= (1 + e^{\boldsymbol{Z}_{ij}\boldsymbol{\gamma}})^{-1} \Bigg\{ u_{ij} \left[e^{\boldsymbol{Z}_{ij}\boldsymbol{\gamma}} + \exp(-e^{\boldsymbol{X}_{ij}\boldsymbol{\beta} + \sigma b_i}) \right] \\ &\qquad + (1 - u_{ij}) \frac{\exp[y_{ij}(\boldsymbol{X}_{ij}\boldsymbol{\beta} + \sigma b_i) - e^{\boldsymbol{X}_{ij}\boldsymbol{\beta} + \sigma b_i}]}{y_{ij}!} \Bigg\} \end{aligned}$$

where $u_{ij} = I(y_{ij} = 0)$. In order to handle the complexity of the estimation in this situation, Hall employs the EM algorithm with Gaussian quadrature with the complete data log-likelihood

$$\begin{aligned} l_c(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{b}) &= \log f(\boldsymbol{b}; \boldsymbol{\theta}) + \log f(\boldsymbol{y}, \boldsymbol{z} | \boldsymbol{b}; \boldsymbol{\theta}) \\ &= \sum_{i=1}^{K} \log \phi(b_i) + \sum_{i=1}^{K} \sum_{j=1}^{T_i} \{ [z_{ij} \boldsymbol{Z}_{ij}\boldsymbol{\gamma} - \log(1 + e^{\boldsymbol{Z}_{ij}\gamma})] \\ &\qquad + (1 - z_{ij}) [y_{ij}(\boldsymbol{X}_{ij}\boldsymbol{\beta} + \sigma b_i) - e^{\boldsymbol{X}_{ij}\boldsymbol{\beta} + \sigma b_i} - \log(y_{ij}!)]] \} \end{aligned}$$

with z_{ij} being the latent indicator of whether Y_{ij} comes from the zero state ($z_{ij} = 1$) or the Poisson(μ_{ij}) state ($z_{ij} = 0$). Hall applies the proposed ZIP with random effects to two data sets, one with pesticides and Whitefly larvae and also the Wiring Board data from Lambert (1992).

On the basis of marginal models and the general estimating equations (GEE) literature, Hall and Zhang (2004) propose an alternative expectation-maximization approach

to incorporate within-cluster correlation. Using a dependence working correlation matrix, Hall and Zhang alter the M step of the EM algorithm by replacing the weighted GLM score equation with the weighted GEE, accounting for the correlation among subjects. This work builds on the works by Rosen et al. (2000) and generalizes the EM algorithm to the ES, expectation-solution, algorithm, which gives both consistent and asymptotically normal parameter estimators under regularity conditions. Hall and Zhang recognize the need to address selection methods for the appropriate working correlation structure for the ES algorithm, but the authors note efficiency, not consistency, of the parameter estimators would be affected.

Beyond accommodating repeated measures data, Gilthorpe et al. (2009) outline extensions of the ZIP and binomial models to account for over-dispersion through zero-inflated negative binomial (ZINB) and beta-binomial models. In addition, these authors discuss the negative implications of excluding covariates from the zero process (ie. latent class membership prediction) without significant consideration. Except for randomization-based arguments, it is doubtful that balanced zero counts exist across the Poisson process covariates, since this would imply the proportion of excess zeros to be equal across all combinations of covariates. Gilthorpe et al. also discuss methods for choosing between zero-inflated and generic mixture models.

Focusing on zero-inflated model interpretations, Albert, Wang and Nelson (2011) present two estimators of overall exposure effects for zero-inflated count models. Even when the mixture distribution of the ZIP model is viewed as a single population, the authors note zero-inflated models are implemented, primarily for model fit (Mwalili, Lesaffre and Declerck, 2008). Although Lambert (1992) and Böhning et al. (1999) discuss estimators of the overall population mean, Albert et al. argue inference for an overall treatment effect while adjusting for covariates has been neglected. The authors then present two methods of achieving overall treatment effect estimates, the average

predicted value approach and the direct approach, particularly for the ZINB and zero-inflated beta-binomial distributions.

In the average predicted value (APV) approach, individual predicted response values under each binary exposure status x_i are calculated then averaged to obtain the estimated overall mean $E(y|x, w)$. The APV can provide either an average difference in predicted responses or average ratio of expected values. First, the model-predicted responses are calculated for each individual (such as $\hat{\mu}_i = \hat{\psi}_i\hat{\lambda}_i$ in ZINB), both as if the person was exposed ($x_i = 1$) and as if they were unexposed ($x_i = 0$), leaving all the other covariates, $\boldsymbol{w}_i$, fixed at that person's observed values. Then the average difference in predicted responses is

$$\theta_D \equiv \int [E(y_i|x_i = 1, \boldsymbol{w}_i) - E(y_i|x_i = 0, \boldsymbol{w}_i)]dF(\boldsymbol{w})$$

where $F(\boldsymbol{w})$ is the joint distribution function for the covariates $\boldsymbol{w}_i$ in reference population.

The next challenge is defining this joint distribution of the covariates. The authors suggest either assuming an appropriate distribution, based on the types of covariates used, or employing the empirical distribution function (EDF). For the parametric approach, Albert et al. suggest using a multinomial distribution for categorical variables and either normal or multivariate normal distributions for continuous covariates. Employing the observed data, one can also leave the covariate distributions unspecified and use the EDF; this method is averaging exposure effect over the EDF of the covariates. In ZINB for an individual i with covariates $\boldsymbol{w}_i$ and exposure status x, we have

$$E(y_i|x_i, \boldsymbol{w}_i) = \text{logit}^{-1}(\alpha_0 + \alpha_1 x_i + \boldsymbol{\alpha'}\boldsymbol{w}_i) \exp(\beta_0 + \beta_1 x_i + \boldsymbol{\beta'}\boldsymbol{w}_i)$$

and

$$\theta_D = \frac{1}{n_G}\sum_{i\in G}\{\text{logit}^{-1}(\hat{\alpha}_0+\hat{\alpha}_1+\hat{\boldsymbol{\alpha}}'\boldsymbol{w}_i)\exp(\hat{\beta}_0+\hat{\beta}_1+\hat{\boldsymbol{\beta}}'\boldsymbol{w}_i)-\text{logit}^{-1}(\hat{\alpha}_0+\hat{\boldsymbol{\alpha}}'\boldsymbol{w}_i)\exp(\hat{\beta}_0+\hat{\boldsymbol{\beta}}'\boldsymbol{w}_i)\}$$

for group G of size n_G. This estimator has the form of the 'standardization formula' from Hernán and Robins (2006). This method uses model-predicted values to perform a stratified analysis within a subpopulation looking for observed differences between exposed and unexposed groups. Depending on how the distributions of the covariates differ between the exposed and unexposed groups, the APV may require some extrapolation beyond the multivariate support of the data. If the ratio of expected values is desired instead of the average difference in responses, the authors provide

$$\theta_R \equiv \frac{\int E(y|x=1,\boldsymbol{w})dF(\boldsymbol{w})}{\int E(y|x=0,\boldsymbol{w})dF(\boldsymbol{w})}.$$

The variances for $\hat{\theta}_D$ and $\hat{\theta}_R$ can be estimated through the delta method or the bootstrap. The bootstrap has the added benefit of providing confidence intervals without requiring distributional assumptions. Also, depending on the form of the covariates, the delta method can be computationally intensive and tedious.

In addition to the APV, Albert et al. also present the 'direct' option of using a log-linear model for the probability of an excess zero (ψ_i) instead of logistic regression. Thus the ZINB (log-log) models are

$$\begin{aligned}\log(\psi_i) &= \gamma_0 + \gamma_1 x_i + \boldsymbol{\gamma}'\boldsymbol{w}_i \\ \log(\mu_i) &= \beta_0 + \beta_1 x_i + \boldsymbol{\beta}'\boldsymbol{w}_i.\end{aligned}$$

Due to the common link function in each process, the ratio of the overall means (exposed

versus non-exposed) is

$$\theta_{RL} \equiv \frac{\mu_1}{\mu_0} = \frac{\exp(\gamma_0 + \gamma_1 + \boldsymbol{\gamma'w}_i)\exp(\beta_0 + \beta_1 + \boldsymbol{\beta'w}_i)}{\exp(\gamma_0 + \boldsymbol{\gamma'w}_i)\exp(\beta_0 + \boldsymbol{\beta'w}_i)} = e^{\gamma_1 + \beta_1}$$

While this approach is fairly simple and the variance of its estimator can be found using the delta method, it is limited by the appropriateness of the log-linear model for the first-step of the ZI model. In particular, the log-linear model is not limited to the range of 0-1 for predicted probabilities. The authors analyze dental caries data by the exposure of very low birth weight compared to normal birth weight, noting that the ZINB (logit-log), ZINB (log-log), and ZIBB (logit-logit) models produced the lowest AIC and BIC, with ZINB (log-log) model performing the best.

Through simulation studies, Albert et al. explore properties of their methods under correct and incorrect model selection, as well as for unbalanced covariate distribution across the exposure groups. An interesting fact is that even when the log-linear link was incorrect, the direct approach appeared to provide valid inference and was fairly robust; however, when the covariates were unbalanced, this method can be substantially biased, especially when the covariate has a large effect on the outcome.

While Albert et al. provide methods for producing estimates of overall exposure effects, the APV and 'direct' approaches are not straightforward and require either distributional assumptions on the extraneous covariates or the use of a log link for the excess zero process. Preisser et al. (2012) reviewed ZIP model usage in the dental caries literature and found that many health researchers have imprecise or misleading conclusions due to the complexity of the ZIP latent class structure.

1.2.2 Hurdle & Zero-Altered Models

Instead of modeling two latent class subpopulations, analysts can choose to use the hurdle model, which models all zeros separately from positive realizations, outlined by Mullahy (1986). In general, let $\phi_1(y, \theta_1)$ and $\phi_2(y, \theta_2)$ be two functions defined on $y \in \Gamma = \{0, 1, 2, ...\}$ satisfying $\phi_1,\ \phi_2\ > 0$ and

$$\phi_1(0, \theta_1) + \sum_{y \in \Gamma_+} (y, \theta_2) = 1$$

where $\Gamma_+ = \Gamma - \{0\}$. Note that a standard data model specifies $\phi_1(y, \theta_1) = \phi_2(y, \theta_2)$ for all $y \in \Gamma$ so that

$$\sum_{y \in \Gamma} \phi_1(y, \theta_1) = \sum_{y \in \Gamma} \phi_2(y, \theta_2) = 1$$

The hurdle model occurs when a binomial probability governs whether a count variable has zero or positive realization. If the realization is positive, then the 'hurdle' is crossed and the conditional distribution of the positives is governed by a truncated-at-zero count data model. The probability that the threshold is crossed is $\Phi_1(\theta_1) = \sum_{y=0} \phi(0, \theta_1)$, and the conditional distribution of the positives is $\phi_2(y, \theta_2)/\Phi_2(\theta_2)$, where Φ_2 is the summation of ϕ_2 on the support of the conditional density and the truncation normalization. Thus, the density function of y is

$$\begin{aligned} p(y) &= [\ \mathrm{P}(y = 0)]^{I(y=0)} * [\ \mathrm{P}(y|y > 0)\ \mathrm{P}(y > 0)]^{I(y>0)} \\ &= [1 - \Phi_i(\theta_1)]^{I(y=0)} * [(\phi_2(y, \theta_2)/\Phi_2(\theta_2)) * \Phi_1(\theta_1)]^{I(y>0)} \end{aligned}$$

The likelihood can be expressed

$$L(\theta_1, \theta_2) = \exp(\Lambda^H) = \prod_{t \in \Omega_0} [1 - \Phi_1(\theta_1)] \prod_{t \in \Omega_1} [(\phi_2(y, \theta_2)/\Phi_2(\theta_2)) * \Phi_1(\theta_1)]\,,$$

where $\Omega_0 = \{t|y_t = 0\}$, $\Omega_2 = \{t|y_t > 0\}$, $\Omega = \Omega_0 \cup \Omega_1$ and Λ^H represents the log-likelihood of the general form of the hurdle model. Mullahy notes that when $\Phi_1(\theta_1) = \Phi_2(\theta_2)$, the likelihood reduces to

$$L(\theta_1, \theta_2) = exp(\Lambda^H) = \prod_{t \in \Omega_0} [1 - \Phi_2(\theta_2)] \prod_{t \in \Omega_1} [\phi_2(y, \theta_2)]$$

Mullahy then states that the specifications where $\theta_1 = \theta_2$ are of primary interest and derives the likelihoods for hurdle models for both the Poisson and geometric distributions, and affirming that the hurdle model can be extended to any count data model.

Building upon the hurdle model approach to modeling zero-inflated counts, Heilbron (1994) recommends using the same distribution to describe both sub-populations by using a truncated-at-one count distribution to model the zero observations and then the truncated-at-zero count distribution for the positive values. This method, referred to as the zero-altered count model, differs from the hurdle model by requiring that both processes have the same distribution and link functions. Let $p_1(y|\lambda_1)$ represent the probability density function for the 'standard' sub-population and $p_2(y|\lambda_2)$ represent the probability density function for the 'zero' sub-population. Then by letting $\omega = p_2(0|\lambda_2)$, the distribution is

$$\begin{aligned} f(0) &= \omega + (1 - \omega) p_1(0|\lambda_1) \\ f(k) &= (1 - \omega) p_1(k|\lambda_1), \qquad k > 0. \end{aligned}$$

The zero-altered model is a hurdle model with additional assumptions: (1) p_1 and p_2 have identical distribution forms and overdispersion parameters, (2) the covariates $\{X_j\}$ and link functions η_k for modeling the means λ_k of the distributions are the same for the two parts (that is, $\eta_k = \beta_{0k} + \sum_{j=1}^{p} X_j \beta_{jk}$, $k = 1, 2$), and (3) λ_1 is a function of λ_2 and ancillary parameters. Heilbron proposes the relationship between λ_1 and λ_2 as

$\lambda_1 = \gamma_1 \lambda_2^{\gamma_2}$, where $\gamma_1 > 0,\ \gamma_2 \geq 0$, which implies $\beta_{2j} = \gamma_2 \beta_{j1},\quad j \geq 1$.

Heilbron affirms that interpretations of parameter estimates is simple for the zero-altered model. First, equality of all corresponding coefficients $\beta_{j1} = \beta_{j2}$ in the zero-altered Poisson model reduces to standard GLM based on p_1. Additionally, the difference $(\beta_{j2} - \beta_{j1})$ may be interpreted in terms of a difference in the mean λ_1 or in other features of p_1. The added-zero probability reduces to $p_1(0|\lambda_2) - p_1(0|\lambda_1)$ with overdispersion parameters being the same in both terms. Distributions p_1 where $p_1(0|\lambda)$ is decreasing in λ (such as Poisson and negative binomial) then the added-zero probability is positive if and only if $\lambda_2 < \lambda_1$.

It is the zero-altered Poisson model that Dobbie and Welsh (2001) modify to utilize general estimating equations to account for correlated observations, and Min and Agresti (2005) further extend it in the repeated measures setting through the use of random effects.

In the context of both the zero-inflated Poisson and hurdle models, Neelon, O'Malley and Normand (2010) detail the fitting of zero-inflated models for repeated measures using Bayesian techniques. By incorporating prior information, the Bayesian approach to fitting these models has the added benefit of straightforward estimation of functions of parameters.

1.3 Marginalized Models

Drawing from marginal approaches with likelihood-based inference, Heagerty (1999) offers an alternative parameterization of logistic-normal model with random effects where individual-level predictions are obtained through marginal regression parameters. By averaging over both measurement error and random individual heterogeneity, the marginalized model structures regression around the *marginal* mean rather than

the conditional mean. When comparing population-averaged and subject-specific coefficients, Heagerty argues that the conditional regression coeffecients have limited utility. For the marginalized logistic-normal model, the author adopts a pair of regression models, with the first focusing upon a population-averaged interpretation,

$$\text{logit } E(Y_{ij}|\boldsymbol{X}_{ij}) = \boldsymbol{X}_{ij}\boldsymbol{\beta},$$

and the second model incorporating the dependence among the longitudinal observations

$$\text{logit } E(Y_{ij}|\boldsymbol{b}_i, \boldsymbol{X}_i) = \Delta_{ij} + b_{ij}$$

where $\boldsymbol{b}_i = (b_{i1}, \ldots, b_{in_i})'$ and $\boldsymbol{b}_i|\boldsymbol{X}_i \sim N(0, \boldsymbol{D}_i)$. Here $\boldsymbol{D}_i$ is a covariance matrix that can be obtained as a function of observation $\boldsymbol{t}_i$ and the parameter vector $\boldsymbol{\alpha}$. Note the parameter Δ_{ij} is a function of $\eta_{ij} = \boldsymbol{X}_{ij}\boldsymbol{\beta}$, the marginal linear predictor, as well as $\sigma_{ij} = \sqrt{\text{var}(b_{ij})}$, the standard deviation of the random effects, where

$$h(\eta_{ij}) = \int h(\Delta_{ij} + \sigma_{ij}z)\phi(z)dz,$$

$h = \text{logit}^{-1}$ and ϕ is the standard normal probability density. Using both numerical integration and the Newton-Raphson iteration, Δ_{ij} is found, given (η_{ij}, σ_{ij}). Since Δ_{ij} is a function of both $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, Heagerty uses a 20-point Gauss-Hermite quadrature to solve the above convolution equation.

Merging some of the ideas of marginalized models and zero-inflated count methods, Lee et al. (2011) focus upon the hurdle model formulation of handling Poisson and negative binomial data with excess zeros while marginalizing over random effects for clustering. Let $Y_i = (Y_{i1}, \ldots, Y_{iN_i})'$ be the count response, where Y_{it} is the response for the i^{th} individual at time t, and let Y_{it} be conditionally independent given $b_i = (b_{i1}, b_{i2})'$. Also, let X_{it} be the covariates pertaining to Y_{it}. The marginalized Poisson hurdle model

is given by

$$P(Y_{it} = y_{it}|X_{it}) = \begin{cases} 1 - p_{it}^M & \text{if } y_{it} = 0 \\ p_{it}^M \frac{g(y_{it};\lambda_{it}^M)}{(1-e^{-\lambda_{it}^M})} & \text{if } y_{it} = 1, 2, \ldots, . \end{cases}$$

where $\text{logit}(p_{it}^M) = X_{it}'\gamma$, $g(y_{it};\lambda_{it}^M) = e^{-\lambda_{it}^M}(\lambda_{it}^M)^{y_{it}}/y_{it}!$, and $\lambda_{it}^M = \exp(X_{it}'\beta)$. To account for the clustered nature of the responses, Lee et al. (2011) draw from Heagerty (1999) to create the conditional hurdle model

$$P(Y_{it} = y_{it}; b_i) = \begin{cases} 1 - p_{it}^C(b_{i1}) & \text{if } y_{it} = 0 \\ p_{it}^C(b_{i1}) \frac{g(y_{it};\lambda_{it}^C(b_{i2}))}{(1-e^{-\lambda_{it}^C(b_{i2})})} & \text{if } y_{it} = 1, 2, \ldots, . \end{cases}$$

where $\text{logit}(p_{it}^C(b_{i1})) = \Delta_{it1} + z_{ij1}'b_{i1}$,

$$g(y_{it}; \lambda_{it}^C(b_{i2})) = e^{-\lambda_{it}^C(b_{i2})}(\lambda_{it}^C(b_{i2}))^{y_{it}}/y_{it}!,$$

$\log(\lambda_{it}^C(b_{i2})) = \Delta_{it2} + z_{ij2}'b_{i2}$ and $b_i = (b_{i1}, b_{i2}) \overset{i.i.d.}{\sim} N(0, \mathbf{\Sigma})$. Here, z_{it1} and z_{it2} are subsets of X_{it} and

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_1 & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{12} & \mathbf{\Sigma}_2 \end{bmatrix}$$

where $\mathbf{\Sigma}_1$, $\mathbf{\Sigma}_{12}$ and $\mathbf{\Sigma}_2$ are unknown positive-definite matrices. Here Δ_{it1}, Δ_{it2} are subject-specific intercepts and are functions of the marginal parameters ($\boldsymbol{\gamma}$, $\boldsymbol{\beta}$) and the dependence structure $\mathbf{\Sigma}$. However, these $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ are reported to have 'marginal' interpretations while handling the mixed model extension. Note that these 'marginal' interpretations are not 'overall' interpretations; the parameterization of the hurdle model implies that γ_k quantifies the effect of a k^{th} covariate on being an excess zero and β_k yields k^{th} covariate effect on the magnitude of some outcome given the outcome is not zero.

Chapter 2

A Marginalized Zero-inflated Poisson Regression Model

2.1 Introduction

Zero-inflated count data exist in many areas of medical and public health research. Because Poisson regression is often inadequate in describing count data with many zeros (Böhning et al., 1999), Lambert (1992) proposed the zero-inflated Poisson (ZIP) regression model, based on a mixture of a Poisson distribution and a degenerate distribution at zero. The ZIP model has two sets of regression parameters that have latent class interpretations, one for the Poisson mean and the other for the probability of being an excess zero. These latent classes are often thought to classify some *at-risk* and *not-at-risk* populations, indicating a difference in susceptibility between the two groups. Others have extended Lambert's model to cluster-specific random effects (Hall, 2000) and marginal models for clustered data (Hall and Zhang, 2004). A separate but related branch of methodological research has focused upon hurdle models, where all zeros are modeled separately from positive counts (Mullahy, 1986; Heilbron, 1994; Dobbie and Welsh, 2001; Min and Agresti, 2005).

Despite the increasing popularity of the ZIP model in health-related fields, the idea

of latent class effects can be troublesome for many investigators to communicate, often yielding misleading or incorrect statements. For example, Preisser et al. (2012) found that many dental researchers interpreted the ZIP Poisson parameters with respect to the overall caries incidence, rather than the caries incidence within the *at-risk* population. In many such situations, the ZIP model is simply a convenient modeling tool for handling data with excess zeros, often used without interest in the latent classes constructed in the analysis (Mwalili, Lesaffre and Declerck, 2008). While the ZIP model parameters have latent class interpretations on these two subpopulations, researchers sometimes seek to make inference on the entire population sampled. Albert, Wang and Nelson (2011) argue that insufficient emphasis has been given to the effects of risk factors on the overall population from which the study sample was drawn and propose estimators of overall exposure effects using the causal inference literature under the zero-inflated modeling framework. Although such marginal effects of predictors are commonly sought, estimating them can be difficult in the traditional ZIP model framework. While transformation techniques, such as those employing the delta method for variance estimation, may be employed to estimate marginal effects of an exposure of interest, these can prove tedious, and the treatment of covariates is not straightforward.

The search for easily implementable overall exposure effect estimation in the ZIP model leads to the consideration of the marginalized models literature. Heagerty (1999) proposed marginalized multilevel models, which directly model the marginal means by linking marginal and conditional models with a function of covariates, marginal parameters and random effects specification. Lee et al. (2011) explore hurdle models in the context of marginalized models to analyze clustered data with excess zeros, marginalizing over the random effects. Combining overdispersion, random effects and marginalized models methods, Iddi and Molenberghs (2012) obtain population-averaged interpretations. These methods for regression of correlated outcomes combine the desire

for population average interpretations with the convenience of estimation with a likelihood function constructed with random effects. In a comparatively simple adaptation of these methods, the marginalized models approach can be extended in the ZIP model for independent data in order to achieve population-wide parameter interpretations for independent count responses with many zeros. Instead of integrating (averaging) over mixtures of distributions defined by random effects, our approach marginalizes over the Poisson and degenerate components of the two-part ZIP model to obtain overall effects.

In studies of risky sexual behavior among HIV-positive individuals, one zero-inflated count variable often studied is the Unprotected Anal and Vaginal Intercourse count (UAVI), the number of unprotected anal or vaginal intercourse acts with any partner over a specified time period. Golin et al. (2010) developed the SafeTalk program, a multicomponent, motivational interviewing-based, safer sex intervention for this at-risk population to reduce the number of unprotected sexual acts. In several populations, sexual behavior count data have displayed a distribution with excess zeros (Heilbron, 1994; Ghosh and Tu, 2009), and population-averaged effects of covariates on sexual behavior are often desired.

To obtain inference across the marginalized means of the ZIP model, this manuscript proposes a new method for zero-inflated counts in which overall exposure effect estimates are easily obtained. Section 2 introduces the marginalized ZIP model that includes parameters with log-incidence density ratio (IDR) interpretations which are estimated by a maximum likelihood procedure. Section 3 presents a simulation study, which examines the properties of the marginalized ZIP and compares it to existing post-hoc methods for estimating overall effects. Section 4 presents analysis of the SafeTalk sexual behavior data, using the marginalized ZIP model. A discussion follows in Section 5.

2.2 Traditional ZIP Model

Originally proposed by Lambert (1992) with application to counts of manufacturing defects, the ZIP regression model allows the count variable of interest, say Y_i, $i = 1, \ldots, n$ to take the value of zero from a Bernoulli distribution, with probability ψ_i, or be drawn from a Poisson distribution, with mean μ_i, with probability $1 - \psi_i$. That is,

$$Y_i \sim \begin{cases} 0 & \text{with probability } \psi_i \\ \text{Poisson}(\mu_i) & \text{with probability } 1 - \psi_i \end{cases}$$

Thus,

$$Y_i = \begin{cases} 0 & \text{with probability } \psi_i + (1 - \psi_i)e^{-\mu_i} \\ k & \text{with probability } (1 - \psi_i)e^{-\mu_i}\mu_i^k/k!, \quad k \in \mathcal{Z}^+. \end{cases}$$

The likelihood for this ZIP model is

$$L(\boldsymbol{\psi}, \boldsymbol{\mu}|\boldsymbol{y}) = \prod_{y_i=0}[(\frac{\psi_i}{1-\psi_i} + e^{-\mu_i})(1 - \psi_i)] \prod_{y_i>0}[(1 - \psi_i)e^{-\mu_i}\mu_i^{y_i}/(y_i!)]. \tag{2.1}$$

Lambert proposed models for the parameters μ_i and ψ_i

$$\begin{aligned} \text{logit } (\psi_i) &= \boldsymbol{Z}_i'\boldsymbol{\gamma} \\ \log (\mu_i) &= \boldsymbol{X}_i'\boldsymbol{\beta} \end{aligned}$$

where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{p_1})'$ is a $(p_1 \times 1)$ column vector of parameters associated with the excess zeros, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{p_2})'$ is a $(p_2 \times 1)$ vector of parameters associated with the Poisson process, and $\boldsymbol{Z}'_{i_{(1\times p_1)}}$ and $\boldsymbol{X}'_{i_{(1\times p_2)}}$ are the vectors of covariates for the i^{th} individual for excess zero and Poisson processes, respectively.

Importantly, the parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ have latent class interpretations; that is, γ_j is the log-odds ratio of a one-unit increase in the j^{th} element of $\boldsymbol{Z}$ on the probability of being an *excess* zero and β_j is the log-incidence density ratio of a one-unit increase in the j^{th} element of $\boldsymbol{X}$ on the mean of the *susceptible* sub-population. In general, no simple summary of the exposure effect on the overall mean of the outcome is directly available. Specifically, consider the overall mean of Y_i, say $\nu_i \equiv E[Y_i]$, often the primary interest of investigators. The relationship between ν_i and the parameters from the ZIP model is

$$\nu_i = (1 - \psi_i)\mu_i = \frac{e^{X_i'\beta}}{1 + e^{Z_i'\gamma}}. \tag{2.2}$$

In (2.2), the population mean is a function of *all* covariates and parameters from both model parts. For the j^{th} covariate in a ZIP model where $Z_i = X_i$ as is commonly specified, the ratio of means for a one-unit increase in x_{ij} is

$$\frac{E(Y_i|x_{ij} = j + 1, \tilde{x}_i = \tilde{x}_i)}{E(Y_i|x_{ij} = j, \tilde{x}_i = \tilde{x}_i)} = \exp(\beta_j)\frac{1 + \exp(j\gamma_j + \tilde{x}_i'\tilde{\gamma})}{1 + \exp[(j + 1)\gamma_j + \tilde{x}_i'\tilde{\gamma}]}$$

where $\tilde{x}_i$ indicates all covariates except x_{ij} and $\tilde{\gamma}$ is created by removing γ_j from $\boldsymbol{\gamma}$. Thus, unless $\gamma_j = 0$, the incidence density ratio (IDR) is not constant across various levels of the extraneous covariates included in the logistic portion of the ZIP model. Additionally, in order to make statements regarding the variability of any IDR estimates at fixed levels of the non-exposure covariates, formal statistical techniques, such as the delta method or bootstrap resampling methods, are required (Albert, et al., 2011). The computational tools needed for these transformations are typically not readily available in standard software packages, meaning that these calculations can be arduous for many applied analysts.

2.3 Marginalized ZIP Model

Because population-wide parameter interpretations are desired, the overall mean ν_i can be modeled directly to give overall exposure effect estimates. The marginalized ZIP model specifies

$$\begin{aligned} \text{logit}(\psi_i) &= \boldsymbol{Z_i'\gamma} \\ \log\,(\nu_i) &= \boldsymbol{X_i'\alpha} + \log(N_i) \end{aligned} \tag{2.3}$$

where an offset term N_i is included to allow more flexibility in the modeling process. Then,

$$\nu_i = N_i \exp(\boldsymbol{X_i'\alpha}) \tag{2.4}$$

allows log-IDR interpretations of the elements of $\boldsymbol{\alpha}$. Thus, $\exp(\alpha_j)$ is the amount by which the mean ν_i, or in the case of offsets the incidence density ν_i/N_i, is multiplied per unit change in x_j, providing the same interpretation as in Poisson regression. In order to utilize the ZIP model likelihood framework, we redefine $\mu_i = \exp(\delta_i)$, where δ_i is not necessarily a linear function of model parameters. Rather, solving $\nu_i = (1-\psi_i)\mu_i$, with substitution for (2.3), provides

$$\delta_i = \boldsymbol{X_i'\alpha} + \log[1 + \exp(\boldsymbol{Z_i'\gamma})] + \log(N_i).$$

Substituting $\psi_i = \exp(\boldsymbol{Z_i'\gamma})/(1 + \exp(\boldsymbol{Z_i'\gamma}))$ and $\mu_i = \exp(\delta_i)$ into (2.1), the likelihood of the marginalized ZIP model for ($\boldsymbol{\gamma}$,$\boldsymbol{\alpha}$) is

$$\begin{aligned} L(\boldsymbol{\gamma}, \boldsymbol{\alpha}|\boldsymbol{y}) = &\prod_{y_i}(1 + e^{\boldsymbol{Z_i'\gamma}})^{-1} \prod_{y_i=0}(e^{\boldsymbol{Z_i'\gamma}} + e^{-N_i(1+\exp(\boldsymbol{Z_i'\gamma}))\exp(\boldsymbol{X_i'\alpha})}) \\ &\prod_{y_i>0}[e^{-N_i(1+\exp(\boldsymbol{Z_i'\gamma}))\exp(\boldsymbol{X_i'\alpha})}(1 + e^{\boldsymbol{Z_i'\gamma}})^{y_i} e^{\boldsymbol{X_i'\alpha} y_i} N_i^{y_i}/(y_i!)] \end{aligned} \tag{2.5}$$

with score equations $\boldsymbol{U}_i = \begin{bmatrix} \frac{\partial l(\gamma,\alpha)}{\partial \boldsymbol{\gamma}} & \frac{\partial l(\gamma,\alpha)}{\partial \boldsymbol{\alpha}} \end{bmatrix}'$ where

$$\begin{aligned}
\frac{\partial l(\boldsymbol{\gamma},\boldsymbol{\alpha})}{\partial \boldsymbol{\gamma}} &= \sum_i \left[\frac{I(y_i=0)\psi_i(1-\psi_i)^{-1}(e^{\nu_i(1-\psi_i)^{-1}}-\nu_i)}{\psi_i(1-\psi_i)^{-1}e^{\nu_i(1-\psi_i)^{-1}}+1} \right. \\
&\qquad \left. + \psi_i(y_i-1) - I(y_i>0)\psi_i(1-\psi_i)^{-1}\nu_i \right] \boldsymbol{Z_i'} \\
\frac{\partial l(\boldsymbol{\gamma},\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} &= \sum_i \left[(y_i - \nu_i(1-\psi_i)^{-1})I(y_i>0) - \frac{\nu_i(1-\psi_i)^{-1}I(y_i=0)}{\psi_i(1-\psi_i)^{-1}e^{\nu_i(1-\psi_i)^{-1}}+1} \right] \boldsymbol{X_i'}
\end{aligned}$$

and $\nu_i = \nu_i(\boldsymbol{\alpha})$ and $\psi_i = \psi_i(\boldsymbol{\gamma})$. Thus the Fisher information is

$$I(\boldsymbol{\gamma},\boldsymbol{\alpha}) = \begin{bmatrix} -E[\frac{\partial^2 l}{\partial \boldsymbol{\gamma}\partial \boldsymbol{\gamma}'}] & -E[\frac{\partial^2 l}{\partial \boldsymbol{\gamma}\partial \boldsymbol{\alpha}'}] \\ -E[\frac{\partial^2 l}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\gamma}'}] & -E[\frac{\partial^2 l}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha}'}] \end{bmatrix}$$

and the model-based standard errors of the parameter estimates are

$$\mathrm{se}_M(\hat{\boldsymbol{\gamma}},\hat{\boldsymbol{\alpha}}) = \sqrt{\mathrm{diag}(I(\boldsymbol{\gamma},\boldsymbol{\alpha})^{-1})}.$$

To address possibly overdispersed counts relative to the ZIP model, the robust (empirical) estimates of the standard errors are

$$\mathrm{se}_R(\hat{\boldsymbol{\gamma}},\hat{\boldsymbol{\alpha}}) = \{\mathrm{diag}[I(\boldsymbol{\gamma},\boldsymbol{\alpha})^{-1}\sum_{i=1}^{n}\boldsymbol{U}_i\boldsymbol{U}_i'I(\boldsymbol{\gamma},\boldsymbol{\alpha})]^{-1}\}^{1/2},$$

with substitution of the MLE's $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\alpha}}$ for $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$, respectively.

While parameter estimation can be implemented using various techniques, such as MCMC methods or the EM algorithm, all results herein are obtained through nonlinear optimization by the quasi-Newton method, implemented in SAS 9.3 IML (SAS Institute, Cary, NC). The likelihood derivations, as well as those used to obtain the Fisher information, are provided in the Appendix.

2.4 Simulation Study

Simulation studies were performed to examine the properties of the new marginalized ZIP model under different scenarios, implemented in SAS 9.3 IML. Let Y_i be the zero-inflated Poisson outcome of interest for the i^{th} participant. Also, let x_{i1} be the exposure variable of interest and let x_{i2} be an additional covariate desired in a regression model. In the SafeTalk example, Y_i is the UAVI count, x_{i1} is an indicator of randomization to SafeTalk intervention, and the additional covariate x_{i2} is a site indicator, necessary due to the randomization scheme. Thus the simulated marginalized ZIP regression model is

$$\begin{aligned} \text{logit}(\psi_i) &= \gamma_0 + \gamma_1 x_{i1} + \gamma_2 x_{i2} \\ \log(\nu_i) &= \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} \end{aligned}$$

To examine the finite sample performance of the marginalized ZIP in estimating specific parameter estimates, we simulated data using the above model. Specifically, $x_{i1} \sim$ Bernoulli(0.25) and $x_{i2} \sim$ Bernoulli(0.4), where x_{i1}, x_{i2} are generated separately for a fixed sample size. Together with fixed vectors of $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$, these x_{i1} and x_{i2} were used to define ψ_i and μ_i, which were employed to randomly generate excess zeros and Poisson counts, the latter through $\mu_i = \nu_i/(1-\psi_i)$. Then the marginalized ZIP model was fit to these simulated data and all parameter estimates retained for examination; the simulation was performed 10,000 times and summary measures were calculated. Specifically, for sample sizes of 100, 200 and 1000, Table 2.1 presents the percent relative median bias, simulation standard deviation, median model-based and robust standard errors and their corresponding coverage probabilities for each parameter in the model; 95% Wald-type confidence intervals are used. In Table 2.1, the true parameter values are $\{\gamma_0 = 0.60, \alpha_0 = -0.25, \gamma_1 = -1, \alpha_1 = \log(1.5), \gamma_2 = \alpha_2 = 0.25\}$.

From Table 2.1, we note that the marginalized ZIP has low bias for $\boldsymbol{\alpha}$ and the bias generally decreases with increasing sample size. For most parameters, the model-based standard errors are similar to the standard deviation of the simulated parameter estimates, implying adequate estimation of the standard error of the parameter estimates. For all sample sizes, Wald-type confidence intervals of the marginalized ZIP parameters have model-based coverage probabilities near the expected 0.95, and coverage probabilities created using the robust standard error have fractionally less coverage.

Additionally, a simulation study was performed to compare the new marginalized ZIP model to several existing methods, namely the traditional ZIP model employing a delta method transformation for post-hoc estimation of overall effects due to x_{i1}. We also used Poisson regression to model each simulated data set and obtain estimates of IDR, both with and without deviance scaling for overdispersion. Since Preisser et al. (2012) found that many researchers interpreted the ZIP latent class parameters as population-level parameters, we examined the properties of these naïve ZIP interpretations as well.

For the traditional ZIP, the relationship between the parameter estimates and the IDR is

$$\frac{E(Y_i|x_{i1}=1,x_2)}{E(Y_i|x_{i1}=0,x_2)} = e^{\beta_1}\frac{1+\exp(\gamma_0+\gamma_2 x_{i2})}{1+\exp(\gamma_0+\gamma_1+\gamma_2 x_{i2})}. \tag{2.6}$$

Note that this relationship produces multiple IDR's, one for each value of the extraneous covariate x_{i2}. For these simulations, $\bar{x}_2 = \frac{1}{n}\sum_{i=1}^{n} x_{i2}$ was used to calculate the log-IDR and its standard error for the traditional ZIP model and delta method. However, this IDR estimate represents the IDR for an 'average' individual, potentially unobservable in the sample population. The use of the naïve ZIP parameter interpretations would present e^{β_1} as the IDR, failing to recognize the relationship between IDR and the zero-inflated parameters $\boldsymbol{\gamma}$. Using data generation as described above, the marginalized ZIP

and traditional ZIP were both performed, then the delta method transformation was used to obtain the latter's estimates of the log-IDR and standard error of log-IDR. Additionally, 95% Wald-type confidence intervals for the log-IDR were created using the point estimate and respective standard error. For each of the methods described, Table 2.2 presents the relative median bias in estimating the IDR and log-IDR, Table 2.3 presents coverage probabilities and Table 2.4 displays power. For the marginalized ZIP and Poisson regression models, robust estimators of the covariance matrix were also employed to calculate the 95% Wald-type confidence intervals, as well as their corresponding coverage probabilities and power. Results are presented for varying levels of the true incidence density ratio e^{α_1}, where $\{\gamma_0 = 0.60,\ \alpha_0 = -0.25,\ \gamma_1 = -1,$ $\alpha_1 = \{\log(1.25), \log(1.5), \log(2)\},\ \gamma_2 = \alpha_2 = 0.25\}$.

With regards to bias, Table 2.2 shows that the marginalized ZIP, ZIP with delta method transformation and Poisson regression models all have low relative bias in estimating both the log-IDR and IDR. However, the naïve ZIP parameter interpretation yields very biased estimates for both log-IDR and IDR. For the fixed parameter values and (2.6), we can determine the expected relative bias in IDR under the naïve ZIP model to be

$$\begin{aligned} \text{Percent Relative Bias} &= \frac{e^{\beta_1} - e^{\alpha_1}}{e^{\alpha_1}} \times 100 \\ &= \left(1 - \frac{1 + \exp(\gamma_0 + \gamma_1 + \gamma_2 \bar{x}_{i2})}{1 + \exp(\gamma_0 + \gamma_2 \bar{x}_{i2})}\right) \times 100 \\ &= 42.24 \end{aligned}$$

regardless of true IDR. This quantity is driven by the magnitude of the exposure parameter in the zero-inflated process.

Table 2.3 displays the coverage probabilities for the 95% confidence intervals for

each method described. Again, since the naïve ZIP parameter interpretation is estimating the wrong quantity, the coverage of the true IDR goes to zero as the sample size increases. The marginalized ZIP, ZIP with delta method transformation, Poisson with robust variance estimator and overdispersed Poisson models all have appropriate coverage, with the Poisson with model-based variance estimator having less coverage than desirable.

Examining the power from each method, the marginalized ZIP has slightly more power than the ZIP with delta method transformation, Poisson with robust variance estimate and overdispersed Poisson under nearly every scenario. For a given sample size, we observed a non-monotone relationship between power and true IDR for the naïve ZIP interpretation. This phenomenon is a result of rejections of the null hypothesis with estimated IDR below 1 when the true IDR is 1.25. For large sample sizes, this naïve interpretation has high power to detect an *incorrect* IDR, emphasizing the need for direct methods to achieve marginal interpretations.

2.5 Motivational Interviewing Intervention Example

Reducing risky sexual behavior among people living with HIV/AIDS is one area of focus among infectious disease researchers, and one measure of risky behavior is the UAVI count, the number of Unprotected Anal or Vaginal sexual Intercourse acts within a given time period. The SafeTalk program was developed as a motivational interviewing-based intervention to reduce sexual behavior, particularly UAVI (Golin et al., 2010). To assess SafeTalk's efficacy at reducing unprotected sex acts in this population, a randomized clinical trial was performed with subjects recruited at three sites being randomized to receive either SafeTalk or a nutritional intervention as control. The participants were then surveyed every four months for one year to measure their self-reported sexual acts in the previous three-month period. Since the primary research

question for this study is whether those in the SafeTalk intervention have lower UAVI than those in the control at the eight-month follow-up visit, the marginalized ZIP is employed to quantify the effect of treatment on UAVI.

For this analysis, there are 357 participants with non-missing UAVI counts at the 8-month visit, excluding eight participants with UAVI counts greater than 18. Figure 2.1 shows the distribution of UAVI counts, which contains 300 (84%) zeros and 8 '10+' counts (2.2%). Since the randomization scheme stratified by site, the marginalized ZIP model to be fit is

$$\begin{aligned} \text{logit}(\psi_i) &= \gamma_0 + \gamma_1 x_{i1} + \gamma_2 x_{i2} + \gamma_3 x_{i3} + \gamma_4 x_{i4} \\ \log(\nu_i) &= \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \alpha_3 x_{i3} + \alpha_4 x_{i4} \end{aligned}$$

where x_{i1} is an indicator of whether the i^{th} participant received the SafeTalk intervention and x_{i2} and x_{i3} are indicators of whether the i^{th} participant was randomized at the second and third study sites, respectively. Additionally, the analysis controls for baseline UAVI count x_{i4}.

In order to compare the traditional ZIP model fit to the marginalized ZIP model, the standardized Pearson residuals of each method were computed and plotted in Figure (2.2). We investigated potential outliers in this manner, finally electing to remove all observations with UAVI greater than 18.

In order to calculate the 'overall' effect of the SafeTalk treatment for the traditional ZIP model, the proportions observed at Site 2 (0.3221) and Site 3 (0.0588) and mean baseline UAVI count (0.9748) are used for the delta method calculations. For the traditional ZIP with delta method, the log-IDR for the intervention is 0.2133 (0.2872), which yields an IDR estimate of 1.2378 and 95% confidence interval (0.705, 2.173). For

Sites 1, 2 and 3, the IDR (and corresponding 95% confidence intervals) from the transformed ZIP with fixed mean baseline UAVI count are 1.2360 (0.706, 2.165), 1.2399 (0.704, 2.184), and 1.2429 (0.701, 2.203), respectively. Examining the range of IDR across baseline UAVI counts, the IDR and corresponding 95% confidence intervals for zero and 18 baseline UAVI counts are 1.2487 (0.702, 2.222) and 0.9598 (0.727, 1.267). For this particular example, there does not appear to be much difference in the IDR of treatment across sites, but note the moderate change in IDR estimates for the different baseline UAVI counts. Although none of these estimates are statistically significant, the estimates for different combinations of covariates demonstrate the lack of a single IDR measure when using traditional ZIP with the delta method. In fact, particular transformed ZIP analyses may yield very different IDR estimates for various combinations of covariate values. Also, notice the transformed ZIP methods require significantly more effort and expertise in deriving and programming than the direct estimation of the log-IDR through the marginalized ZIP model.

Table 2.5 presents the results of the marginalized ZIP analysis on the SafeTalk example. By exponentiating α_1, the estimate of the IDR for treatment is $\exp(-0.0666) = 0.9355$; thus, the marginalized ZIP model reveals those on SafeTalk intervention have 6% fewer unprotected sexual acts at the eight-month followup visit than those participants randomized to control. The 95% model-based Wald-type confidence interval for the treatment IDR is $(0.559, 1.567)$, implying there is no significant difference between the two treatment groups. However, this illustrative analysis is not considered definitive due to the deletion of large UAVI counts. Because the traditional ZIP with delta method is limited by the substitution of specific levels of the extraneous covariates, the overall effect of SafeTalk is difficult to summarize briefly. However, the marginalized ZIP model gives one IDR of SafeTalk intervention, adjusted for all the other covariates. In terms of model fit, the full likelihood values for the marginalized ZIP and traditional

ZIP models are -291.28 and -288.51, indicating that the two models have similar fit to the SafeTalk data.

2.6 Conclusion

In this manuscript, we develop a new marginalized ZIP model to achieve population-average estimates rather than the traditional ZIP latent class estimates, whose interpretations are difficult to convey. The primary advantage of the marginalized ZIP model is the direct estimation of the population mean, offering meaningful statements about an exposure effect on an entire sampled population rather than an unobservable latent class. Thus, arguably the marginalized ZIP model yields more interpretable results than the traditional ZIP model, for which additional calculations, involving more time and statistical expertise, are required to achieve population-level inference. Particularly, exposure effect estimation in the presence of covariates requires no additional assumptions or estimation. Not only does the marginalized ZIP have relatively low bias, but it also outperforms traditional ZIP in estimating overall exposure effect estimates in power in a simulation study.

Noting that the traditional ZIP and marginalized ZIP models are non-nested, we expect that goodness-of-fit measures will often have similar values for models of comparable complexity and structure. Thus the decision on whether to utilize a marginalized ZIP or traditional ZIP should generally be made with the desired interpretations in mind. When interest lies in describing the two latent subpopulations, the traditional ZIP model is preferable, whereas the marginalized ZIP model should be strongly considered by researchers in various fields wishing to make population statements. Future research is needed to extend the marginalized ZIP model to include data with clustering, as well as data containing overdispersion in addition to zero-inflation.

Figure 2.1: Histogram of UAVI Counts

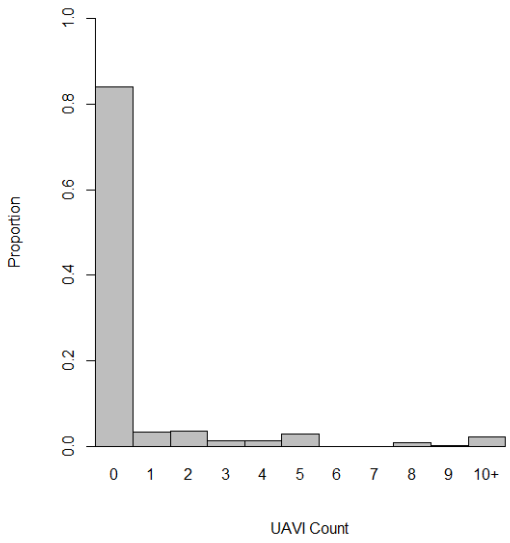

Figure 2.2: Standardized Pearson Residuals of SafeTalk Marginalized ZIP and Traditional ZIP Models

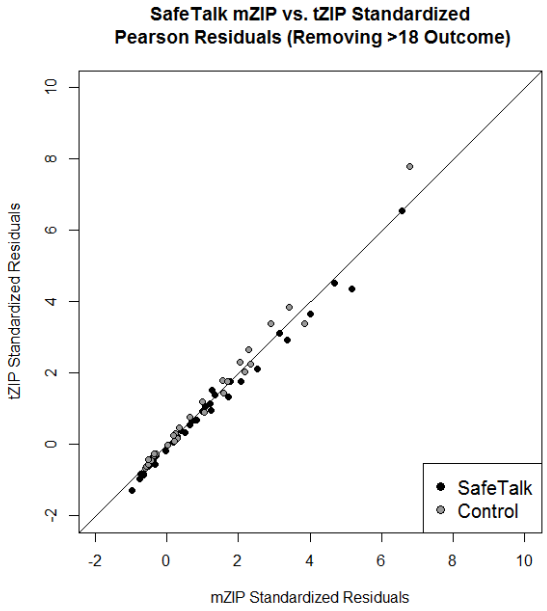

Table 2.1: Marginalized ZIP Performance with 10,000 Simulations and Varying Sample Size

Sample Size	Parameter	Relative Median Bias (%)	Simulation Std Dev	Median Model-Based Std Error	Median Robust Std Error	Model-Based Coverage Probability	Robust Coverage Probability
100	γ_0	-3.44	0.3487	0.3341	0.3256	0.9541	0.9457
	γ_1	-6.54	0.8312	0.5558	0.5362	0.9626	0.9567
	γ_2	9.38	0.4980	0.4627	0.4525	0.9468	0.9409
	α_0	5.22	0.2530	0.2371	0.2283	0.9400	0.9287
	α_1	-1.99	0.3405	0.3176	0.3038	0.9420	0.9281
	α_2	-1.70	0.3471	0.3232	0.3114	0.9440	0.9336
200	γ_0	-1.64	0.2405	0.2330	0.2279	0.9438	0.9387
	γ_1	-2.28	0.4108	0.3822	0.3719	0.9513	0.9435
	γ_2	3.43	0.3327	0.3219	0.3164	0.9453	0.9416
	α_0	2.03	0.1760	0.1674	0.1631	0.9404	0.9333
	α_1	-0.36	0.2363	0.2241	0.2168	0.9418	0.9321
	α_2	-0.63	0.2373	0.2275	0.2224	0.9414	0.9343
1000	γ_0	-0.06	0.1054	0.1031	0.1013	0.9471	0.9426
	γ_1	-0.80	0.1726	0.1678	0.1639	0.9463	0.9397
	γ_2	0.70	0.1450	0.1423	0.1404	0.9484	0.9437
	α_0	0.90	0.0768	0.0748	0.0735	0.9468	0.9405
	α_1	-0.51	0.1028	0.0997	0.0974	0.9433	0.9363
	α_2	0.07	0.1031	0.1013	0.1000	0.9490	0.9456

Table 2.2: Comparison of Relative Median Biases for Estimation of Overall Exposure Effects with Marginalized ZIP, ZIP with Delta Transformation, ZIP with Naïve Interpretations, & Poisson

True	Sample	Marginalized ZIP		ZIP w/ Delta Method		Naïve ZIP		Poisson	
IDR	Size	Log-IDR	IDR	Log-IDR	IDR	Log-IDR	IDR	Log-IDR	IDR
1.25	100	3.42	0.77	5.27	1.18	-248.56	-42.57	2.43	0.54
	200	-2.56	-0.57	-2.25	-0.50	-250.44	-42.81	-1.62	-0.36
	1000	-0.37	-0.08	-1.17	-0.26	-249.56	-42.70	-0.43	-0.10
1.5	100	-1.99	-0.80	-0.95	-0.39	-139.17	-43.12	-1.58	-0.64
	200	-0.36	-0.15	-0.07	-0.03	-137.49	-42.73	-0.75	-0.30
	1000	-0.51	-0.21	-0.87	-0.35	-137.07	-42.64	-0.49	-0.20
2	100	-1.03	-0.71	-0.13	-0.09	-80.79	-42.88	-0.62	-0.43
	200	0.01	0.01	0.26	0.18	-80.26	-42.67	-0.02	-0.02
	1000	0.17	0.12	-0.05	-0.03	-79.99	-42.56	0.35	0.24

Table 2.3: Comparison of Coverage Probabilities for Estimation of Overall Exposure Effects with Marginalized ZIP, ZIP with Delta Transformation, ZIP with Naïve Interpretations, & Poisson

				ZIP with				Overdispersed
		Marginalized ZIP		Delta Method	Naïve ZIP	Poisson		Poisson
True IDR	Sample Size	Model	Robust	Model	Model	Model	Robust	Model
1.25	100	0.9422	0.9301	0.9480	0.5022	0.8194	0.9427	0.9681
	200	0.9390	0.9296	0.9493	0.1687	0.8163	0.9465	0.9666
	1000	0.9393	0.9318	0.9469	0.0000	0.8160	0.9467	0.9670
1.5	100	0.9420	0.9281	0.9490	0.4042	0.8083	0.9441	0.9642
	200	0.9418	0.9321	0.9505	0.1078	0.8049	0.9479	0.9639
	1000	0.9433	0.9363	0.9479	0.0000	0.8044	0.9486	0.9642
2	100	0.9474	0.9364	0.9529	0.3067	0.7885	0.9479	0.9589
	200	0.9450	0.9388	0.9483	0.0576	0.7751	0.9465	0.9543
	1000	0.9494	0.9463	0.9530	0.0000	0.7866	0.9523	0.9571

Table 2.4: Comparison of Power for Estimation of Overall Exposure Effects with Marginalized ZIP, ZIP with Delta Transformation, ZIP with Naïve Interpretations, & Poisson

				ZIP with				Overdispersed
		Marginalized ZIP		Delta Method	Naïve ZIP	Poisson		Poisson
True IDR	Sample Size	Model	Robust	Model	Model	Model	Robust	Model
1.25	100	0.1247	0.1449	0.1159	0.1941	0.2854	0.1213	0.0883
	200	0.1858	0.2021	0.1686	0.3932	0.3588	0.1701	0.1307
	1000	0.5817	0.6017	0.5557	0.9766	0.7657	0.5527	0.4833
1.5	100	0.2629	0.2907	0.2490	0.0799	0.4725	0.2541	0.2107
	200	0.4445	0.4653	0.4267	0.1298	0.6662	0.4238	0.3747
	1000	0.9762	0.9789	0.9722	0.4861	0.9937	0.9701	0.9597
2	100	0.5980	0.6235	0.5864	0.0924	0.8088	0.5933	0.5578
	200	0.8640	0.8727	0.8558	0.1496	0.9556	0.8532	0.8378
	1000	1.0000	1.0000	1.0000	0.5249	1.0000	1.0000	1.0000

Table 2.5: Marginalized ZIP Model Results: SafeTalk Example

	Parameter	Parameter Estimate	Model-Based Std Error	Robust Std Error
Zero-Inflation Model				
Intercept	γ_0	1.8485	0.2373	0.2444
Treatment	γ_1	-0.0242	0.2905	0.3488
Site 2	γ_2	0.1055	0.3141	0.3396
Site 3	γ_3	-0.1856	0.5824	0.6183
Baseline UAVI	γ_4	-0.1679	0.0421	0.0476
Marginalized Mean Model				
Intercept	α_0	-0.7338	0.2189	0.2335
Treatment	$\boldsymbol{\alpha_1}$	$\mathbf{-0.0666}$	0.2630	0.3837
Site 2	α_2	0.3146	0.2863	0.3648
Site 3	α_3	4.4169	0.4974	0.5487
Baseline UAVI	α_4	0.1169	0.0266	0.0378

Chapter 3

Marginalized ZIP Regression Model with Random Effects

3.1 Introduction

Infectious disease researchers are often concerned with reducing risky sexual behavior among HIV-positive individuals. One measure of risky sexual behavior is the Unprotected Anal and Vaginal Intercourse (UAVI) count, the number of unprotected anal or vaginal intercourse acts with any partner over a specified period of time. The SafeTalk program was developed by Golin et al. (2010) to reduce the number of unprotected sexual acts through a multicomponent, motivational interviewing-based, safer sex intervention. Sexual behavior count data can display a distribution with excess zeros (Heilbron, 1994; Ghosh and Tu, 2009). To examine the efficacy of the SafeTalk program over time, a randomized controlled clinical trial collected risky sexual behavior data at baseline and up to three follow-up visits.

Methods have been developed for modeling correlated count data with excess zeros, both under the zero-inflated and hurdle model frameworks. Building upon the zero-inflated Poisson (ZIP) regression model established by Lambert (1992), Hall (2000) extends the ZIP regression model to include random effects in the Poisson process.

In order to account for overdispersion beyond the excess zeros, Yau, Wang and Lee (2003) modify the zero-inflated negative-binomial (ZINB) regression model to include random effects. Instead of using random effects to handle correlated data, Hall and Zhang (2004) employ GEE methodology for zero-inflated models in order to achieve population-averaged interpretations. For each of these zero-inflated methods, two sets of parameter estimates are produced, those associated with the excess zero process and those associated with the count process. Although many health-related fields are implementing zero-inflated techniques, these two sets of parameter estimates can be difficult to interpret, in many cases leading to incorrect statements (Preisser, et al., 2012). Often health researchers wish to make inference upon an entire sampled population rather than the latent classes modeled by ZIP methodology. Transformation methods, with variance estimation by the delta method or resampling methods, may be used to make inference on overall estimates of exposure effect for ZIP and ZINB models (Albert, et al., 2011). However, such transformations can be tedious for many analysts, and the treatment of covariates is not necessarily apparent.

While closely related to the zero-inflated methodology, hurdle models (including zero-altered models) account for excess zeros by modeling all zeros separately from positive counts (Mullahy, 1986; Heilbron, 1994). One set of parameters measures effects on the probability of being a zero and one set measures effects on the mean conditional on the observation being positive. Dobbie and Welsh (2001) use the zero-altered Poisson model, modified to utilize GEE, to account for correlated observations. Min and Agresti (2005) extend the zero-altered model to include random effects. Like ZIP models, hurdle models do not produce a direct overall estimate of exposure effect for the marginal mean count.

The choice between the hurdle and zero-inflated model classes has been approached from various angles. Much of the literature pertaining to the analysis of count data

with excess zeros focuses on model fit, using fit statistics to provide justification of model class choice. Gilthorpe, et al. (2009) argue that *a priori* knowledge of the data-generating mechanism could be used to identify the class of models from which to choose, supported by statements in Neelon et al. (2010) and Buu et al. (2012). Applications in which all zeros can be considered as arising from an identical process indicate a hurdle model, rather than a zero-inflated model, where zeros can occur from the two different processes. Albert et al. (2009) contend that model interpretations have been generally overlooked in the zero-inflated literature; they propose methods for the assessment of overall exposure effects. In many applications containing count data with many zeros, the two latent class interpretations are not clinically supported, and the zero-inflated methodology is just a modeling technique to account for excess zeros in a single population (Mwalili, et al., 2008).

Proposing the marginalized model for longitudinal binary data, Heagerty (1999) employs joint models by directly modeling the marginal mean and simultaneously using a linked random effects model to account for correlated responses. Through this joint model, marginalization over random effects achieves population-averaged parameters, while accounting for correlated measures. Extending the marginalized model approach, Lee et al. (2011) focus on the hurdle model formulation for Poisson and negative binomial data with excess zeros while marginalizing over random effects for clustering. Since Lee et al. focus on marginalizing over the random effects, the two sets of parameters from their marginalized hurdle models have the same interpretations as hurdle models for independent responses.

Where marginalized models often average over random effects to obtain population-average effect estimates, this manuscript proposes marginalizing over the two ZIP model processes to achieve overall effect estimates for expected counts. Section 2 briefly reviews the ZIP model with random effects from Hall (2000). In Section 3, we propose

the marginalized ZIP model with random effects, which has subject-specific parameters and discuss the situation where those parameters have equivalent population-averaged interpretations. Section 4 presents simulation study results examining the finite sample performance of the new model. In Section 5, we consider data from the SafeTalk randomized controlled clinical trial. A discussion is provided in Section 6.

3.2 ZIP Model with Random Effects

Extending Lambert's ZIP model to incorporate correlated zero-inflated count data, Hall (2000) developed the ZIP model with random effects. Let $\boldsymbol{Y} = (\boldsymbol{Y_1'}, \ldots, \boldsymbol{Y_K'})$ where K is the number of independent clusters and $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iT_i})'$, where T_i is the number of observations for the i^{th} cluster. Let $s_{ij} = 1$ if Y_{ij} is from the first process (i.e. Y_{ij} is an excess zero) and $s_{ij} = 2$ if Y_{ij} is from the second (Poisson) process. Then

$$Y_{ij} \sim \begin{cases} 0 & \text{with probability } P(s_{ij} = 1) = \psi_{ij} \\ \text{Poisson } (\mu_{ij}^C) & \text{with probability } P(s_{ij} = 2) = 1 - P(s_{ij} = 1) = 1 - \psi_{ij} \end{cases} \tag{3.1}$$

where $\mu_{ij}^C = E(Y_{ij}|s_{ij} = 2, b_i)$. The notation μ_{ij}^C indicates that the Poisson mean is conditional on the random effect b_i. The log-linear and logistic regression models are

$$\begin{aligned} \text{logit}(\psi_{ij}) &= \boldsymbol{Z}_{ij}'\boldsymbol{\gamma} \\ \log(\mu_{ij}^C) &= \boldsymbol{X}_{ij}'\boldsymbol{\beta} + \sigma b_i, \end{aligned}$$

where $b_1, \ldots, b_K \overset{i.i.d.}{\sim} N(0, 1)$, and $\boldsymbol{Z}_{ij}$ and $\boldsymbol{X}_{ij}$ are the covariate vectors for the logistic and Poisson processes, respectively. The log-likelihood can be expressed

$$l(\boldsymbol{\theta}, \boldsymbol{y}) = \sum_{i=1}^{K} \log \int_{-\infty}^{\infty} \left[\prod_{j=1}^{T_i} \Pr(Y_{ij} = y_{ij}|b_i) \right] \phi(b_i)\ db_i$$

where $\boldsymbol{\theta} = (\boldsymbol{\gamma}', \boldsymbol{\beta}', \sigma)$, ϕ is the standard normal probability density and

$$\begin{aligned}
\Pr(Y_{ij} = y_{ij}|b_i, \boldsymbol{\theta}) &= \left[\psi_{ij} + (1-\psi_{ij})e^{-\mu_{ij}^C}\right]^{u_{ij}} \left[\frac{(1-\psi_{ij})e^{-\mu_{ij}^C}(\mu_{ij}^C)^{y_{ij}}}{y_{ij}!}\right]^{1-u_{ij}} \qquad (3.2)\\
&= (1+e^{\boldsymbol{Z}_{ij}'\boldsymbol{\gamma}})^{-1}\Bigg\{u_{ij}\left[e^{\boldsymbol{Z}_{ij}'\boldsymbol{\gamma}} + \exp(-e^{\boldsymbol{X}_{ij}'\boldsymbol{\beta}+\sigma b_i})\right]\\
&\qquad + (1-u_{ij})\frac{\exp[y_{ij}(\boldsymbol{X}_{ij}'\boldsymbol{\beta}+\sigma b_i) - e^{\boldsymbol{X}_{ij}'\boldsymbol{\beta}+\sigma b_i}]}{y_{ij}!}\Bigg\},
\end{aligned}$$

where $u_{ij} = I(y_{ij} = 0)$. Using the EM algorithm framework that Lambert (1992) proposed, Hall fits this ZIP model with random effects with the EM algorithm with Gaussian quadrature. Generally, the overall conditional mean $E(Y_{ij}|b_i) = (1-\psi_{ij})\mu_{ij}^C$ will depend on $\boldsymbol{\gamma}$, $\boldsymbol{\beta}$ and b_i through a complicated function that does not permit easy and direct inference for overall effects, here defined as ratios of such means when a single covariate is allowed to vary.

3.3 Marginalized ZIP Model with Random Effects

3.3.1 Subject-specific marginalized ZIP model

Using a marginalized model approach, we now present a marginalized adaptation of the ZIP model with random effects for repeated measures data. The marginalized ZIP model for clustered data directly models the overall subject-specific mean $\nu_{ij}^C = E(Y_{ij}|\boldsymbol{d}_i)$ through

$$\begin{aligned}
\text{logit}(\psi_{ij}^C) &= \boldsymbol{Z}_{ij}'\boldsymbol{\gamma} + \boldsymbol{w}_{1ij}'\boldsymbol{c}_i\\
\log\,(\nu_{ij}^C) &= \boldsymbol{X}_{ij}'\boldsymbol{\alpha} + \log(N_i) + \boldsymbol{w}_{2ij}'\boldsymbol{d}_i, \qquad (3.3)
\end{aligned}$$

where $\psi_{ij}^C = P(s_{ij} = 1|\boldsymbol{c}_i)$ and $\boldsymbol{b}_i = (\boldsymbol{c}_i, \boldsymbol{d}_i)'$ follows the multivariate normal distribution with mean zero and covariance matrix $\boldsymbol{\Sigma} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$. Because ν_{ij}^C is modeled directly in this marginalized ZIP with random effects model, α_k is interpreted as the subject-specific log-incidence density ratio (IDR) for the k^{th} covariate; that is, for a one-unit increase in corresponding covariate x_k, $\exp(\alpha_k)$ is the amount by which the mean ν_{ij}^C for a particular subject is multiplied, which is the same interpretation as in a Poisson random effects model.

For $\boldsymbol{\theta} = (\boldsymbol{\gamma}', \boldsymbol{\alpha}', \boldsymbol{\Sigma})'$, the log-likelihood for this marginalized ZIP model with random effects can be written

$$l(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{i=1}^{K} \log \int_{-\infty}^{+\infty} \left[\prod_{j=1}^{T_i} P(Y_{ij} = y_{ij}|\boldsymbol{b}_i, \boldsymbol{\theta}) \right] \Phi(\boldsymbol{b}_i) d\boldsymbol{b}_i, \tag{3.4}$$

where Φ is the multivariate normal density $(\boldsymbol{0}, \boldsymbol{\Sigma})$. In order to use the ZIP likelihood presented in (3.2), we redefine $\mu_{ij}^C = \exp(\delta_{ij}^C)$, where δ_{ij}^C is not necessarily a linear function of covariates. Then

$$P(Y_{ij} = y_{ij}|\boldsymbol{b}_i, \boldsymbol{\theta}) = \left[\psi_{ij}^C + (1 - \psi_{ij}^C) e^{-\exp(\delta_{ij}^C)} \right]^{I(y_{ij}=0)} \left[\frac{(1 - \psi_{ij}^C) e^{-\exp(\delta_{ij}^C)} e^{\delta_{ij}^C y_{ij}}}{y_{ij}!} \right]^{I(y_{ij}>0)} \tag{3.5}$$

Substitution of (3.3) into $\nu_{ij}^C = (1 - \psi_{ij}^C)\mu_{ij}^C$ and solving for $\delta_{ij}^C = \log(\mu_{ij}^C)$ gives

$$\delta_{ij}^C = \log(N_i) + \log[1 + \exp(\boldsymbol{Z}_{ij}'\boldsymbol{\gamma} + \boldsymbol{w}_{1ij}'\boldsymbol{c}_i)] + \boldsymbol{X}_{ij}'\boldsymbol{\alpha} + \boldsymbol{w}_{2ij}'\boldsymbol{d}_i. \tag{3.6}$$

Through substitution of (3.6) into (3.5), this subject-specific marginalized ZIP model with random effects may be fit using SAS NLMIXED (SAS Institute Inc, 2013), which employs an adaptive Gauss-Hermite quadrature to approximate the integral of the likelihood (3.4) over the random effects. Also, SAS NLMIXED can provide robust

(empirical) standard error estimates of the parameters, through the likelihood-based 'sandwich' estimator, to address model misspecification (White, 1982).

3.3.2 Population-averaged marginalized ZIP model for clustered data

The primary objective in the marginalized models literature (e.g. Heagerty, 1999) is to obtain marginalized (or population-averaged) parameters rather than subject-specific parameters. In Section 3.3.1, we described the marginalized ZIP model with random effects, where the 'marginalization' is over the two latent classes of the ZIP model to achieve overall exposure effect estimates. However, because the marginalized ZIP with random effects models $\nu_{ij}^C = E(Y_{ij}|\boldsymbol{d}_i)$, it yields parameters with subject-specific interpretations.

For data with repeated measures, statistical analysts usually choose between methods employing subject-specific (SS) parameters (mixed models) and methods having population-average (PA) parameters (GEE). However, Ritz and Spiegelman (2004) and Young et al. (2007) investigate the exact nature of the relationship between SS and PA parameters for Poisson count data, using methods established in McCulloch and Searle (2001). For models with log links and normally distributed random effects, the mathematical relationships between SS and PA parameters can be quite straightforward.

To explore the connection between SS and PA parameters for the marginalized ZIP model with random effects, we restate the model as

$$\begin{aligned} \text{logit}(\psi_{ij}^C) &= \boldsymbol{Z}_{ij}'\boldsymbol{\gamma}^{\boldsymbol{SS}} + \boldsymbol{w}_{1ij}'\boldsymbol{c}_i \\ \log\,(\nu_{ij}^C) &= \boldsymbol{X}_{ij}'\boldsymbol{\alpha}^{\boldsymbol{SS}} + \log(N_i) + \boldsymbol{w}_{2ij}'\boldsymbol{d}_i, \end{aligned} \tag{3.7}$$

where the SS superscript indicates that subject-specific interpretations are appropriate

for these parameters. Then

$$E(Y_{ij}|\boldsymbol{d}_i) = \exp[\boldsymbol{X_i'}\boldsymbol{\alpha^{SS}} + \log(N_i) + \boldsymbol{w}'_{2ij}\boldsymbol{d}_i]$$

and

$$\begin{aligned} E(Y_{ij}) &= E[E(Y_{ij}|\boldsymbol{d}_i)] \\ &= N_i \exp(\boldsymbol{X_i'}\boldsymbol{\alpha^{SS}}) E(\exp(\boldsymbol{w}'_{2ij}\boldsymbol{d}_i)) \\ &= N_i \exp(\boldsymbol{X_i'}\boldsymbol{\alpha^{SS}}) \exp(0.5\boldsymbol{w}'_{2ij}\Sigma_{22}\boldsymbol{w}_{2ij}) \end{aligned} \tag{3.8}$$

where $\boldsymbol{d}_i \sim N(\boldsymbol{0}, \Sigma_{22})$. From (3.8), defining $\nu_{ij}^M = E(Y_{ij})$,

$$\log(\nu_{ij}^M) = \boldsymbol{X}'_{ij}\boldsymbol{\alpha^{SS}} + \log(N_i) + 0.5\boldsymbol{w}'_{2ij}\Sigma_{22}\boldsymbol{w}_{2ij}. \tag{3.9}$$

Now consider the fully marginal model (3.10), where PA denotes population-averaged parameters

$$\log(\nu_{ij}^M) = \boldsymbol{X}'_{ij}\boldsymbol{\alpha^{PA}} + \log(N_i). \tag{3.10}$$

The PA parameters in (3.10) are multiplicatively offset from the SS parameters by the function $\exp(0.5\boldsymbol{w}'_{2ij}\Sigma_{22}\boldsymbol{w}_{2ij})$ of the Poisson random effects and respective covariance matrix. Thus, for all covariates that do not have corresponding random effects in $\boldsymbol{w}'_{2ij}$, the corresponding parameters $\boldsymbol{\alpha^{SS}}$ are equivalent to $\boldsymbol{\alpha^{PA}}$. Consider the model with only a random intercept ($\boldsymbol{w}'_{2ij} = 1$) and $\Sigma_{22} = \sigma_b^2$; then

$$\log(\nu_{ij}^M) = [\alpha_0^{SS} + (\sigma_b^2/2)] + \tilde{\boldsymbol{X}}'_{\boldsymbol{i}}\tilde{\boldsymbol{\alpha}}^{\boldsymbol{SS}} + \log(N_i),$$

where $\tilde{\boldsymbol{X}}'_{\boldsymbol{i}}$ and $\tilde{\boldsymbol{\alpha}}^{\boldsymbol{SS}}$ contain all the covariates and corresponding parameters excluding

the intercept. In this situation, $\tilde{\boldsymbol{\alpha}}^{\boldsymbol{SS}}$ also have population-averaged interpretations. While analysts may choose to include further normal random effects, such as a random slope over time, all parameters without a corresponding random effect have population-averaged as well as subject-specific interpretations because of the log link and normal random effects.

3.4 Simulation Study

To examine the properties of the marginalized ZIP model with random effects, a simulation study was performed using SAS 9.3 NLMIXED. Let Y_{ij} be a zero-inflated Poisson outcome for the i^{th} participant at time j, and let g_i be a time-constant exposure variable of interest for each subject. The simulation scenario is motivated by the constant treatment assignment in the SafeTalk clinical trial. In the SafeTalk motivating example, Y_{ij} is the UAVI count outcome and g_i is an indicator of randomization to the SafeTalk intervention group. For this simulation study, three time points were used with $I(j = 2)$ and $I(j = 3)$ being the indicators of whether an observation occurs at follow-up time 2 or 3. Data were simulated using the marginalized ZIP model with random effects given by

$$\begin{aligned} \text{logit}(\psi_{ij}^C) &= \gamma_0 + \gamma_1 I(j=2) + \gamma_2 I(j=2) g_i + \gamma_3 I(j=3) + \gamma_4 I(j=3) g_i + c_i \\ \log(\nu_{ij}^C) &= \alpha_0 + \alpha_1 I(j=2) + \alpha_2 I(j=2) g_i + \alpha_3 I(j=3) + \alpha_4 I(j=3) g_i + d_i \end{aligned} \quad (3.11)$$

where c_i, d_i are independent normal random intercepts with variances σ_1^2 and σ_2^2 used to account for correlated outcomes for the i^{th} participant. Although we designed the simulation study using independent random intercepts, correlated effects across the two portions of this model could be fit specified. For a fixed sample, g_i was generated from

a Bernoulli(0.5) and (c_i, d_i) were independently generated as $N(0,1)$. The parameters ψ_{ij}^C and ν_{ij}^C are calculated with the specified values of $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$. Using the first model part in equation (3.11) and $\mu_{ij}^C = \nu_{ij}^C/(1-\psi_{ij}^C))$, excess zeros and Poisson counts were randomly generated. These simulations were performed for 300, 500 and 1000 participants, respectively, with $\boldsymbol{\gamma}$, $\boldsymbol{\alpha}$ vectors chosen such that $\psi_i^C = \{0.45, 0.50, 0.50\}$, $\nu_i^C = \{1.75, 1.70, 1.70\}$ for $g_i = 0$ and $\psi_i^C = \{0.45, 0.65, 0.65\}$, $\nu_i^C = \{1.75, 1.275, 1.11\}$ for $g_i = 1$. For each cluster size, 1,000 simulations were attempted, but the SAS NLMIXED procedure failed to converge for 74 ($K = 300$), 60 ($K = 500$), and 76 ($K = 1000$) iterations.

Table 3.1 presents the percent relative median bias, simulation standard deviation and median standard errors (model-based and robust) of each estimate from the marginalized ZIP model. The vectors of parameters to simulate the above values of ψ_{ij} and ν_{ij} are $\boldsymbol{\gamma} = \{-0.2007,\ 0.2007,\ 0.8197,\ 0.2007,\ 0.8197\}$ and $\boldsymbol{\alpha} = \{0.5596,\ -0.0290,\ -0.2877,\ -0.0290,\ -0.4263\}$.

In Table 3.1, the percent relative median bias is small for each cluster size K, and both the model-based and robust standard errors are close to the standard deviation of the parameter estimates, indicating adequate estimation of the variability in parameter estimates. The largest percent relative bias in estimating $\boldsymbol{\alpha}$ occur for α_1 and α_3, which have true values very close to 0, inflating the relative bias. For $K = 500$, the true α_3 is -0.0290 and the median bias is 0.00688, yielding a percent relative median bias of -23.73%.

In addition to the marginalized ZIP model with random effects, both a Poisson population-average model with GEE estimation and a Poisson random intercept model were fit in SAS 9.3 GENMOD and NLMIXED, respectively, for comparison in estimating the population-average IDR. The model for the Poisson population-average model

is

$$\log(\nu_{ij}^{M}) = \alpha_0^* + \alpha_1 I(j=2) + \alpha_2 I(j=2) g_i + \alpha_3 I(j=3) + \alpha_4 I(j=3) g_i, \qquad (3.12)$$

with unstructured covariance and model-based standard errors scaled with Pearson's chi-square for potential overdispersion, as well as empirical (robust) standard errors; (3.11) expresses the model for the Poisson random intercept model with ν_{ij}^{C} representing the Poisson mean $E(Y_{ij}|d_i)$. As discussed in Section 3.3.2, the parameters (α_1, α_2, α_3, α_4) from (3.11) have population-average interpretations (since intercept is the only random effect), so the parameters from the Poisson population-average model with GEE estimation in (3.12) are estimating the same quantities. For time 2, Table 3.2 presents the relative median bias in estimating both the log-IDR (α_2) and IDR ($\exp(\alpha_2)$) for all three models, as well as the 95% Wald-type coverage probabilities and power. Similar results were obtained for time 3 (data not shown).

In Table 3.2, note that the marginalized ZIP model with random effects has lower percent relative median bias for each K (number of participants), as well as appropriate coverage. With the model-based standard errors in the Poisson random intercept model, the coverage probabilities are much less than the expected 0.95, indicating these standard errors are underestimating the extra-Poisson variability in the ZIP data due to the excess zeros. The robust standard errors for both Poisson models provide appropriate coverage of the IDR, but the marginalized ZIP model has increased power to detect significance in IDR over both Poisson methods. Using the Pearson scaled model-based standard errors, the Poisson PA models have only slightly less bias and coverage than the marginalized ZIP model, but there is a marked difference in power with the Poisson PA model having significantly less ability to detect differences in IDR.

3.5 Motivating Example

In safer sex counseling for people living with HIV/AIDS, an outcome of interest is Unprotected Anal or Vaginal Intercourse acts (UAVI), defined as the number of unprotected sexual acts with any partner. Researchers developed the motivational interview-based intervention SafeTalk to reduce the number of unprotected sexual acts (Golin et al., 2007; Golin et al., 2010). For the clinical trial examining SafeTalk efficacy, participants were randomized to receive either SafeTalk intervention counseling or a control nutritional counseling. These participants completed questionnaires about both nutritional and sexual behavior at baseline as well as at three follow-up visits. After data cleaning, the sample sizes at each time point are 476, 399, 363 and 301. In these data, MAR is assumed; the assumption of MCAR is not valid because those participants with any risky baseline behavior have 54.1% retention at the final visit versus 65.6% retention in those with non-risky baseline behavior. In order to evaluate the efficacy of the SafeTalk intervention over time, the marginalized ZIP with random effects is fit to the UAVI counts at all four time points. The model of interest is

$$\begin{aligned}
\text{logit}(\psi_{ij}^C) &= \gamma_0 + \gamma_1 x_{i1} + \gamma_2 x_{i2} + \gamma_3 I(j=2) + \gamma_4 I(j=2) g_i \\
&\quad + \gamma_5 I(j=3) + \gamma_6 I(j=3) g_i + \gamma_7 I(j=4) + \gamma_8 I(j=4) g_i + c_i \\
\log(\nu_{ij}^C) &= \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \alpha_3 I(j=2) + \alpha_4 I(j=2) g_i \\
&\quad + \alpha_5 I(j=3) + \alpha_6 I(j=3) g_i + \alpha_7 I(j=4) + \alpha_8 I(j=4) g_i + d_i,
\end{aligned}$$

where $c_i,\ d_i$ are bivariate normal random intercepts with covariance $\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}$, j is the visit number, g_i is an indicator of randomization to SafeTalk intervention group, and x_{i1} and x_{i2} are fixed effects for study site.

Using SAS NLMIXED (for which the code is presented in the Appendix), the SafeTalk analysis results are presented in Table 3.3. The contrast testing treatment effect over time $H_0 : (\alpha_4, \alpha_6, \alpha_8)' = (0, 0, 0)'$ is highly significant ($p = 0.0003$), indicating that the SafeTalk intervention affects UAVI count. At the second follow-up visit, for which the IDR (and 95% Wald-type model-based confidence interval) is 0.542 (0.360, 0.815), a participant randomized to SafeTalk has 46% fewer unprotected sexual acts with any partner than he or she would have if randomized to the nutritional intervention. Because the only random effect for the above model is a random intercept, the parameters associated with treatment effect from this analysis additionally have population-averaged interpretations. Thus, at the second follow-up visit, those participants randomized to SafeTalk had on average 46% fewer unprotected sexual acts with any partner than the participants randomized to the nutritional intervention. The SafeTalk intervention appears to have the largest effect on UAVI count at the first follow-up survey, where the estimated IDR (and 95% Wald-type model-based confidence interval) of treatment effect is 0.280 (0.182, 0.431). By the third follow-up survey, we observe less reduction in UAVI count due to SafeTalk, with an IDR of 0.769 (0.461, 1.282). Figure 3.1 displays the predicted mean UAVI over time, as well as the IDR of treatment at each time point. The SafeTalk intervention appears to have a significant effect in reducing UAVI counts at the first follow-up visit, but the difference between the two treatment groups is reduced at each subsequent follow-up visit. From Figure 3.1 and Table 3.3, note that the nutritional control arm has a significant reduction in predicted UAVI count at the final visit, numerically represented through α_7.

When the SafeTalk data is examined using a Poisson population-average model with GEE estimation, the contrast testing treatment effect is non-significant (p=0.8259). At the second follow-up, the GEE model estimates the IDR to be 0.768 with 95% Wald-type model-based and empirical confidence intervals (0.391, 1.508) and (0.403,

1.466), respectively. Using the Poisson random intercept model, the treatment efficacy contrast is significant when using the model-based standard errors (p=0.0303) but non-significant when robust standard errors are used (p=0.8443). At the second follow-up, the random intercept model estimates the IDR to be 0.711 with model-based and robust 95% Wald-type confidence intervals of (0.556, 0.908) and (0.336, 1.502). Since the simulations in Section 3.4 suggest that the model-based standard errors in the Poisson random intercept model underestimate the variability due to the excess zero process, the conclusions of the robust methods are preferred.

3.6 Conclusion

Motivated by the difficulty in estimating overall exposure effects from the traditional ZIP model parameters, we have proposed a marginalized ZIP model with random effects to account for correlated observations. Since the overall subject-specific mean is modeled directly, the parameters from this new model allow subject-specific inference rather than on the mixture model components of the subject-specific ZIP model. Additionally, when the log link is used for the marginal mean and normal random effects are used, those parameters without random effects have both subject-specific and population-average interpretations.

In the simulation study, we experienced convergence issues similar to ZIP model instability usually associated with those effects in the excess zero portion of the ZIP analysis (Min and Agresti, 2005). In marginalized ZIP regression, the excess zero model parameters are considered nuisance parameters, as the primary hypotheses concern the marginal mean. Thus, the relatively small number of simulation iterations with failed NLMIXED convergence is not excessively worrisome.

In contrast to reliance on fit statistics and conjectures about data-generating mechanisms as a basis for selecting the type of count regression model for handling data with

many zeros, we affirm that the choice between marginalized ZIP, ZIP and hurdle model classes should be motivated by the interpretations desired. When inference upon the overall marginal mean is desired, the marginalized ZIP model is preferred. The *a priori* choice of model class for zero-inflation is analogous to the *a priori* choice between PA and SS models for longitudinal data where the interpretations of regressions parameters differ in models with non-identity link functions.

Rather than marginalizing over the two processes of the ZIP model, the ZIP model with random effects could be marginalized over the random effects, similar to the marginalized hurdle model in Lee et al. (2011). Additionally, one could marginalize over both the random effects and two ZIP processes to achieve a 'fully' marginalized ZIP model. We have argued that the marginalized ZIP model proposed in this article can be used not only for subject-specific inference on overall conditional effects but also for population-average inference for overall effects in many problems.

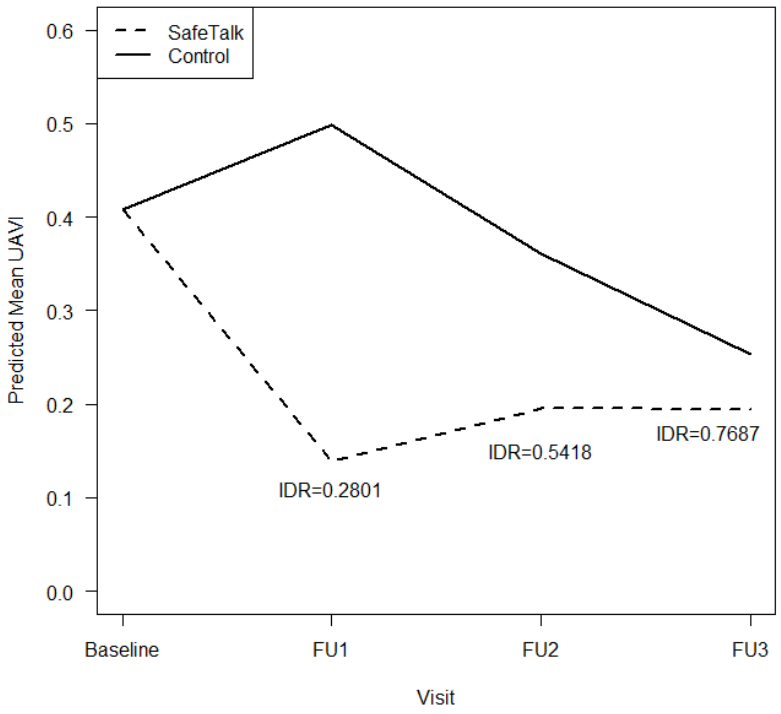

Figure 3.1: Predicted UAVI Means Over Time

Table 3.1: Marginalized ZIP w/ RE Performance with 1,000 Simulations and Varying Number of Subjects

K	Parameter	Percent Relative Median Bias	Simulation Std Dev	Median Std Error	Median Robust Std Error
300	γ_0	-12.1130	0.1664	0.1632	0.1624
	γ_1	0.7608	0.2471	0.2441	0.2408
	γ_2	0.6558	0.3005	0.3035	0.3023
	γ_3	-4.2015	0.2511	0.2439	0.2419
	γ_4	2.0586	0.3126	0.3060	0.3046
	α_0	0.6814	0.1020	0.1028	0.1016
	α_1	19.1520	0.1232	0.1266	0.1248
	α_2	-0.3477	0.1899	0.1926	0.1913
	α_3	-38.9690	0.1284	0.1260	0.1255
	α_4	0.8086	0.1956	0.1944	0.1927
	σ_1^2	0.6927	0.3455	0.3526	0.3469
	σ_2^2	-3.4939	0.1764	0.1701	0.1653
500	γ_0	-7.6883	0.1353	0.1272	0.1261
	γ_1	-0.7127	0.1886	0.1884	0.1870
	γ_2	-0.1939	0.2256	0.2336	0.2323
	γ_3	-6.9450	0.1927	0.1891	0.1872
	γ_4	2.1543	0.2315	0.2363	0.2348
	α_0	1.1634	0.0805	0.0804	0.0796
	α_1	9.4651	0.0987	0.0977	0.0969
	α_2	-2.3118	0.1463	0.1481	0.1474
	α_3	-23.7300	0.0987	0.0974	0.0969
	α_4	0.4757	0.1494	0.1492	0.1492
	σ_1^2	-0.7635	0.2728	0.2735	0.2691
	σ_2^2	-2.4187	0.1382	0.1337	0.1300
1000	γ_0	-4.8208	0.0873	0.0895	0.0892
	γ_1	-1.6741	0.1304	0.1334	0.1327
	γ_2	-0.6487	0.1703	0.1647	0.1640
	γ_3	-1.5368	0.1269	0.1332	0.1328
	γ_4	-0.0560	0.1569	0.1659	0.1652
	α_0	0.4007	0.0590	0.0574	0.0569
	α_1	0.4916	0.0675	0.0689	0.0688
	α_2	0.6070	0.1122	0.1046	0.1044
	α_3	13.1970	0.0688	0.0689	0.0685
	α_4	1.7074	0.1039	0.1058	0.1058
	σ_1^2	-2.8465	0.1904	0.1951	0.1913
	σ_2^2	-1.3215	0.1044	0.0961	0.0940

Table 3.2: Percent Relative Median Bias, Coverage & Power for Estimating Time 2 IDR ($\exp(\alpha_2)$) and log-IDR (α_2)

K	Model*	Percent Relative Median Bias (IDR)	Percent Relative Median Bias (Log-IDR)	Model-Based Coverage	Model-Based Power	Robust Coverage	Robust Power
300	mZIP	0.100	-0.3480	0.9676	0.3089	0.9644	0.3099
	Poisson PA	-0.177	0.5951	0.9331	0.1793	0.9266	0.1955
	Poisson RI	-1.206	4.2191	0.4244	0.7138	0.9266	0.1847
500	mZIP	0.667	-2.3120	0.9543	0.4734	0.9479	0.4809
	Poisson PA	-2.965	10.4620	0.9287	0.3021	0.9277	0.2957
	Poisson RI	-4.004	14.2040	0.3957	0.7702	0.9266	0.2915
1000	mZIP	-0.174	0.6070	0.9362	0.7760	0.9372	0.7749
	Poisson PA	-0.765	2.6708	0.9210	0.4156	0.9242	0.4275
	Poisson RI	-2.970	10.4790	0.3907	0.8669	0.9383	0.3853

* mZIP: Marginalized ZIP model with random effects;
Poisson RA: Poisson population average model with GEE estimation;
Poisson RI: Poisson random intercept model

Table 3.3: Marginalized ZIP Model with Random Effects Results: SafeTalk Example

	Parameter	Parameter Estimate	Model-Based Std Error	Robust Std Error
Zero-Inflation Model				
Intercept	γ_0	2.1187	0.3581	0.3665
Site 2	γ_1	0.1026	0.4311	0.4184
Site 3	γ_2	0.2445	0.8782	0.9548
Follow-up 1	γ_3	1.2709	0.3287	0.3468
Follow-up 1*Treatment	γ_4	0.8849	0.4144	0.4627
Follow-up 2	γ_5	1.7071	0.3611	0.7011
Follow-up 2*Treatment	γ_6	-0.6021	0.5022	0.9185
Follow-up 3	γ_7	1.0214	0.4577	0.6881
Follow-up 3*Treatment	γ_8	-0.3331	0.6034	1.0968
Marginalized Mean Model				
Intercept	α_0	-0.8966	0.2803	0.2965
Site 2	α_1	0.0362	0.2941	0.2893
Site 3	α_2	-0.0220	0.6191	0.6442
Follow-up 1	α_3	0.2011	0.1471	0.1969
Follow-up 1*Treatment	α_4	-1.2725	0.2197	0.3365
Follow-up 2	α_5	-0.1217	0.1632	0.2264
Follow-up 2*Treatment	α_6	-0.6128	0.2082	0.3742
Follow-up 3	α_7	-0.4762	0.2203	0.3521
Follow-up 3*Treatment	α_8	-0.2630	0.2611	0.4691
Variance Parameters				
	σ_{11}	9.7487	2.1328	2.4313
	σ_{12}	-4.5957	0.8270	0.7345
	σ_{22}	3.4461	0.6929	0.6599

Chapter 4

A SAS/IML Macro for Marginalized ZIP Regression

4.1 Introduction

When analyzing count data with excess zeros, there are several methods from which to choose. One of the more popular methods in health research is the zero-inflated Poisson (ZIP) regression model (Lambert, 1992) based on a mixture of a Poisson distribution and a degenerate distribution at zero. The ZIP model has two sets of regression parameters with latent class interpretations, one for the Poisson mean and the other for the probability of being an excess zero. Indicating a potential difference in susceptibility between two subpopulations, these latent classes are often thought to classify some *at-risk* and *not-at-risk* populations. Due to the latent class formulation, the ZIP model parameters can prove difficult for many investigators to interpret, leading to misleading statements about an exposure's effect on prevalence or incidence (Mwalili, et al., 2008; Preisser, et al., 2012). In order to estimate overall exposure effects in the ZIP model, parameter transformations are sometimes undertaken, requiring computationally tedious methods such as the delta method or resampling method to estimate variances (Albert, et al., 2011). Also, the treatment of additional covariates is not

straightforward.

Zero-inflated count regression models have found popularity and utility in highway research, for example, where some road sections are considered safe, while others are potentially unsafe, even if the unsafe sections do not have any observed accidents (Shankar, Milton and Mannering, 1997). In manufacturing, machines may be considered not at risk of failure if all parts are properly aligned, but at-risk for producing faulty parts under misalignment (Lambert, 1992). ZIP models have also been applied to health care utilization (Moon and Shin, 2006), medicine (Bulsara, et al., 2004), political science (Zorn, 1996) and occupational safety (Carrivick, Lee and Yau, 2003).

Alternatively, analysts can fit hurdle models (Mullahy, 1986), which model all zero observations separately from the positive realizations. Including the zero-altered model (Heilbron, 1994), the hurdle model class produces two sets of parameters, one set estimating the effects of predictors on the probability of being a zero and one set estimating the effect on the mean conditional on a positive observation. As in the ZIP case, no direct estimation of overall exposure effects are produced by the class of hurdle models.

For these reasons, Long et al. (Chapter 2 of dissertation) develop the marginalized zero-inflated Poisson regression model (MZIP), which marginalizes over the two ZIP processes to achieve population-average parameter interpretations. Because the marginalized ZIP methodology has a different framework from traditional ZIP models for connecting the observed data to the set of parameters in question, the standard software used to perform ZIP regression (such as SAS/STAT©, GENMOD procedure) can not be used to fit marginalized ZIP models (SAS Institute Inc. 2013a). Although the marginalized ZIP model for independent data can be fit in SAS NLMIXED, this

manuscript presents a SAS/IML (SAS Institute Inc. 2013b) macro to fit the marginalized ZIP likelihood in a straightforward manner, providing additional output and generally requiring less computation time. In addition, the empirical standard error estimates are not available in NLMIXED, giving our macro advantage over the existing procedure. By making marginalized ZIP methods and software widely available to statistical analysts and epidemiologists, we hope to initiate significant improvement in interpretations in the many disciplines employing count regression models for data with many zeros.

In Section 2, the marginalized ZIP methodology from Long et al. is briefly presented. Section 3 outlines the m_ZIP SAS/IML macro fields. Section 4 utilizes the SafeTalk motivating example from Long et al. to display the macro usage; Section 5 presents a discussion.

4.2 Marginalized ZIP Methodology

Let Y_i, $i = 1, \ldots, n$ be a zero-inflated Poisson outcome of interest, with ψ_i being the probability of being an excess zero and μ_i the mean of the Poisson latent class. Then,

$$Y_i \sim \begin{cases} 0 & \text{with probability } \psi_i \\ \text{Poisson}(\mu_i) & \text{with probability } 1 - \psi_i. \end{cases}$$

The population mean $\nu_i = E(Y_i)$ is a function of both ψ_i and μ_i (Lambert, 1992; Böhning et al., 1999) through

$$\nu_i = (1 - \psi_i)\mu_i.$$

For the traditional ZIP model, ψ_i and μ_i are modeled through functions of covariates for which transformations are necessary to make inference on ν_i. Often of interest to researchers is the incidence density ratio (IDR), the ratio of two ν_i's with different

covariate values, often indicating an exposure of interest. As discussed in Chapter 2, any IDR calculated for a particular exposure of interest is not only a function of the parameter associated with ν_i, but also is a function of all the parameters associated with ψ_i, producing an IDR for each combination of extraneous covariates used in modeling ψ_i.

Rather than modeling the latent class mean μ_i, the marginalized ZIP model specifies

$$\begin{aligned} \text{logit}(\psi_i) &= \boldsymbol{Z_i'\gamma} \\ \log(\nu_i) &= \boldsymbol{X_i'\alpha} + \log(N_i) \end{aligned} \tag{4.1}$$

where an offset term N_i is included to allow the modeling of rates (ν_i/N_i) based on varying exposure times. Here, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{p_1})'$ is a $(p_1 \times 1)$ column vector of parameters associated with the excess zeros, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{p_2})'$ is a $(p_2 \times 1)$ vector of parameters associated with the marginal mean, and $\boldsymbol{Z'}_{i_{(1\times p_1)}}$ and $\boldsymbol{X'}_{i_{(1\times p_2)}}$ are the vectors of covariates for the i^{th} individual for the excess zero process and marginal means, respectively. Because the marginal mean ν_i is modeled directly, the elements of $\boldsymbol{\alpha}$ have log-IDR interpretations, providing the same interpretation as in Poisson regression. That is, $\exp(\alpha_j)$ is the amount by which the mean ν_i, or in the case of offsets the incidence density ν_i/N_i, is multiplied per unit change in x_j. Long et al. present the likelihood of the marginalized ZIP model for $(\boldsymbol{\gamma},\boldsymbol{\alpha})$ to be

$$\begin{aligned} L(\boldsymbol{\gamma}, \boldsymbol{\alpha}|\boldsymbol{y}) = &\prod_{y_i}(1 + e^{\boldsymbol{Z_i'\gamma}})^{-1} \prod_{y_i=0}(e^{\boldsymbol{Z_i'\gamma}} + e^{-N_i(1+\exp(\boldsymbol{Z_i'\gamma}))\exp(\boldsymbol{X_i'\alpha})}) \\ &\times \prod_{y_i>0}[e^{-N_i(1+\exp(\boldsymbol{Z_i'\gamma}))\exp(\boldsymbol{X_i'\alpha})}(1 + e^{\boldsymbol{Z_i'\gamma}})^{y_i} e^{\boldsymbol{X_i'\alpha} y_i} N_i^{y_i}/(y_i!)]. \end{aligned} \tag{4.2}$$

A SAS macro for fitting the marginalized ZIP model is described in the next section.

4.3 Marginalized ZIP Model Macro

For the marginalized ZIP model with independent data, the SAS macro employs SAS 9.3 IML to maximize the likelihood (4.2) through the SAS/IML NLPNRA algorithm, which performs nonlinear optimization by the Newton-Raphson method. Not only is this SAS programming straightforward to implement, but the maximization occurs relatively quickly. The macro yields the log-likelihood values, parameter estimates and model-based standard errors, as well as robust (empirical) standard errors for possibly overdispersed counts relative to the ZIP model. In addition, the macro produces and outputs both the model-based and robust covariance matrices, as well as Wald tests of significance for each parameter estimate.

The following code calls the marginalized ZIP model for independent data macro.

```
%macro m_ZIP( DATA=,OUTCOME=,ZI_PRED=,M_PRED=,OUTPUT_MB_COV=,SAVE_MB_COV=,
              MB_COV_DATA=,OUTPUT_R_COV=,SAVE_R_COV=,R_COV_DATA=);
```

where the required fields for the SAS/IML macro are

$$\begin{aligned} \text{DATA} &= \text{analysis data set} \\ \text{OUTCOME} &= \text{outcome of interest } Y_i \\ \text{ZI\_PRED} &= \text{vector of predictors for } \psi_i \\ \text{M\_PRED} &= \text{vector of predictors for } \nu_i \end{aligned}$$

The user is required to input a vector of ones in both ZI_PRED and M_PRED to include an intercept. The following optional fields provide and save the model-based and robust

(empirical) covariance matrices:

OUTPUT_MB_COV	=	binary indicator of whether to print the model-based covariance matrix, default 0
SAVE_MB_COV	=	binary indicator of whether to save the model-based covariance matrix, default 0
MB_COV_DATA	=	location to save model-based covariance matrix as SAS data set, if indicated by SAVE_MB_COV, default work.mb_cov
OUTPUT_R_COV	=	binary indicator of whether to print the robust covariance matrix, default 0
SAVE_R_COV	=	binary indicator of whether to save the robust covariance matrix, default 0
R_COV_DATA	=	location to save robust covariance matrix as SAS data set, if indicated by SAVE_R_COV, default work.r_cov

4.4 Motivating Example

The following is the m_ZIP macro usage based upon the SafeTalk motivating example from Long et al. (2013). In order to make the data available to users, we have simulated a new data set based on the SafeTalk parameter estimates in Long et al. (2013). Researchers designed SafeTalk, a multicomponent, motivational interviewing-based, safer sex intervention to reduce risky sexual behavior among people living with HIV/AIDS. One measure of risky sexual behavior is the count of unprotected anal or vaginal sexual intercourse acts (UAVI), which is known to be zero-inflated. Examining

the data from a clinical trial of the SafeTalk intervention, the research question addressed here is whether the intervention is efficacious at the second follow-up visit. As in Chapter 2, Y_i is the UAVI count for a three-month period before the primary endpoint, x_{i1} is the primary exposure of interest, the randomization to receive the SafeTalk intervention, and x_{i2}, x_{i3} are necessary randomization effects for study site. Also, this analysis accounts for x_{i4}, the baseline UAVI count. Thus the model fit is

$$\begin{aligned} \text{logit}(\psi_i) &= \gamma_0 + \gamma_1 x_{i1} + \gamma_2 x_{i2} + \gamma_3 x_{i3} + \gamma_4 x_{i4} \\ \log(\nu_i) &= \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \alpha_3 x_{i3} + \alpha_4 x_{i4}. \end{aligned} \tag{4.3}$$

The m_ZIP macro call is

```
%macro m_ZIP( DATA          = work.safetalk,
              OUTCOME       = uavi,
              ZI_PRED       = one site2 site3 b_uavi,
              M_PRED        = one site2 site3 b_uavi,
              OUTPUT_MB_COV = 1,
              OUTPUT_R_COV  = 1);
```

which produces the output for this model in Tables 1 through 4.

Also produced by the m_ZIP macro call above are the model-based and robust covariance matrices in Tables 5 and 6 because the OUTPUT_MB_COV and OUTPUT_R_COV options are employed.

As in Tables 4.1 and 4.2, the SAS macro outputs the parameters for the MZIP model, as well as both model-based and robust standard error estimates. In addition, Wald chi-square statistics and p-values are given for each parameter. In all macro output, the ZI_ prefix denotes a $\boldsymbol{\gamma}$ parameter used to model the probability of being an excess zero ψ_i, as in (4.1). Similarly, the M_ prefix in the output designates an $\boldsymbol{\alpha}$

parameter modeling the marginal mean ν_i. For example, ZI_B_UAVI is x_{i4} in (4.4), and its associated parameter γ_4 is estimated to be -0.1679 ($p = 0.0001$), indicating that baseline UAVI count is highly predictive of being an excess zero at the second follow-up visit. The primary research question of the efficacy of the SafeTalk intervention can be addressed with the parameter α_1 in (4.4), which corresponds to the x_{i1} variable M_ARM in Tables 4.1 and 4.2. The parameter estimate for M_ARM is -0.0666, which is not statistically significant using either the model-based or robust standard errors (p=0.80; p=0.86).

Table 4.3 displays the odds ratio estimates and 95% confidence intervals (both model-based and robust) for each variable used to model ψ_i. For the ZI_B_UAVI, the odds ratio for being an excess zero is 0.8455, meaning that the odds of being an excess zero decreases by 15% for each one-unit increase in baseline UAVI count. These estimates are the exponentiated zero-inflated parameter estimates and Wald-type confidence intervals from Tables 4.1 and 4.2.

Table 4.4 provides the incidence density ratios (IDR) and corresponding model-based and robust 95% confidence intervals for each predictor of ν_i. Using the SafeTalk intervention variable M_ARM, the IDR for the SafeTalk treatment is 0.9355, indicating that those participants randomized to receive the SafeTalk intervention had 6.5% fewer unprotected sexual acts at the second followup visit than those participants randomized to control.

Because the particular options were identified in the SAS macro call statement, Tables 4.5 and 4.6 present the model-based and robust variance-covariance matrices, from which the standard errors in Tables 4.1 and 4.2 are derived.

4.5 Conclusion

Using the methods developed in Chapter 2, this manuscript provides a SAS/IML macro which performs MZIP regression. Using this macro, analysts can readily fit the MZIP model and directly make inference on the sampled population, rather than interpreting effects on the latent class subpopulations modeled in the traditional ZIP framework. The macro provides parameter estimates, model-based and robust standard errors, odds ratios and incidence density ratios, and their respective 95% Wald-type confidence intervals. By making the computational methods of Chapter 2 widely available, researchers with zero-inflated data have more analytic options, particularly when seeking population-level inference.

Table 4.1: m_ZIP Macro Output: Model-based Results

Parameters	Estimates	Model-based Std Errors	Model-based Wald Chi-Square	Model-based p-value
ZI_ONE	1.8485	0.2373	60.6544	< 0.0001
ZI_ARM	-0.0242	0.2905	0.0069	0.9337
ZI_SITE2	0.1055	0.3141	0.1128	0.7370
ZI_SITE3	-0.1856	0.5824	0.1016	0.7499
ZI_B_UAVI	-0.1679	0.0421	15.9264	0.0001
M_ONE	-0.7338	0.2189	11.2346	0.0008
M_ARM	-0.0666	0.263	0.0642	0.8000
M_SITE2	0.3146	0.2863	1.2071	0.2719
M_SITE3	1.4169	0.4974	8.1161	0.0044
M_B_UAVI	0.1169	0.0266	19.2561	< 0.0001
Log-likelihood	-291.2787			

Table 4.2: m_ZIP Macro Output: Robust (Empirical) Results

Parameters	Estimates	Robust Std Errors	Robust Wald Chi-Square	Robust p-value
ZI_ONE	1.8485	0.2444	57.1940	<0.0001
ZI_ARM	-0.0242	0.3488	0.0048	0.9448
ZI_SITE2	0.1055	0.3396	0.0965	0.7561
ZI_SITE3	-0.1856	0.6183	0.0901	0.7640
ZI_B_UAVI	-0.1679	0.0476	12.4281	0.0004
M_ONE	-0.7338	0.2335	9.8762	0.0017
M_ARM	-0.0666	0.3837	0.0302	0.8621
M_SITE2	0.3146	0.3648	0.7436	0.3885
M_SITE3	1.4169	0.5487	6.6686	0.0098
M_B_UAVI	0.1169	0.0378	9.5615	0.0020
Log-likelihood	-291.2787			

Table 4.3: m_ZIP Macro Output: Odds Ratios for Zero-inflated Parameters $\boldsymbol{\gamma}$

Zero-Inflation Parameters	Odds Ratio	Model-based Lower 95% CI	Model-based Upper 95% CI
ZI_ONE	6.3501	3.9879	10.1115
ZI_ARM	0.9761	0.5523	1.7251
ZI_SITE2	1.1112	0.6004	2.0566
ZI_SITE3	0.8306	0.2652	2.6009
ZI_B_UAVI	0.8455	0.7785	0.9181

		Robust Lower 95% CI	Robust Upper 95% CI
ZI_ONE	6.3501	3.9330	10.2527
ZI_ARM	0.9761	0.4928	1.9336
ZI_SITE2	1.1112	0.5711	2.1621
ZI_SITE3	0.8306	0.2472	2.7908
ZI_B_UAVI	0.8455	0.7701	0.9282

Table 4.4: m_ZIP Macro Output: Incidence Density Ratios (IDR) for Marginal Mean Parameters $\boldsymbol{\alpha}$

Maringal Mean Parameters	IDR	Model-based Lower 95% CI	Model-based Upper 95% CI
M_ONE	0.4801	0.3126	0.7373
M_ARM	0.9355	0.5587	1.5666
M_SITE2	1.3697	0.7814	2.4008
M_SITE3	4.1244	1.556	10.9324
M_B_UAVI	1.124	1.0668	1.1842

		Robust Lower 95% CI	Robust Upper 95% CI
M_ONE	0.4801	0.3038	0.7587
M_ARM	0.9355	0.4410	1.9845
M_SITE2	1.3697	0.6700	2.8001
M_SITE3	4.1244	1.4070	12.0896
M_B_UAVI	1.1240	1.0437	1.2104

Table 4.5: m_ZIP Macro Output: Model-based Covariance Matrix

	ZI_ONE	ZI_ARM	ZI_SITE2	ZI_SITE3	ZI_B_UAVI	M_ONE	M_ARM	M_SITE2	M_SITE3	M_B_UAVI
ZI_ONE	0.05633	-0.03907	-0.02937	-0.03417	-0.00172	-0.04375	0.03047	0.02133	0.02714	0.00095
ZI_ARM	-0.03907	0.08442	-0.00329	-0.00353	-0.00146	0.03070	-0.06629	0.00271	0.00141	0.00128
ZI_SITE2	-0.02937	-0.00329	0.09864	0.03422	-0.00145	0.02058	0.00289	-0.07738	-0.02549	0.00172
ZI_SITE3	-0.03417	-0.00353	0.03422	0.33917	0.00123	0.02696	-0.00040	-0.02607	-0.27035	-0.00036
ZI_B_UAVI	-0.00172	-0.00146	-0.00145	0.00123	0.00177	0.00086	0.00153	0.00150	-0.00106	-0.00086
M_ONE	-0.04375	0.03070	0.02058	0.02696	0.00086	0.04793	-0.03168	-0.02339	-0.03070	-0.00092
M_ARM	0.03047	-0.06629	0.00289	-0.00040	0.00153	-0.03168	0.06919	-0.00300	0.00043	-0.00142
M_SITE2	0.02133	0.00271	-0.07738	-0.02607	0.00150	-0.02339	-0.00300	0.08199	0.02866	-0.00186
M_SITE3	0.02714	0.00141	-0.02549	-0.27035	-0.00106	-0.03070	0.00043	0.02866	0.24736	0.00049
M_B_UAVI	0.00095	0.00128	0.00172	-0.00036	-0.00086	-0.00092	-0.00142	-0.00186	0.00049	0.00071

Table 4.6: m_ZIP Macro Output: Robust (Empirical) Covariance Matrix

	ZI_ONE	ZI_ARM	ZI_SITE2	ZI_SITE3	ZI_B_UAVI	M_ONE	M_ARM	M_SITE2	M_SITE3	M_B_UAVI
ZI_ONE	0.05974	-0.05066	-0.02072	-0.03230	-0.00321	-0.04060	0.04071	0.00450	0.01794	0.00034
ZI_ARM	-0.05066	0.12163	-0.02916	-0.01264	0.00052	0.03565	-0.11340	0.04219	0.02551	0.00456
ZI_SITE2	-0.02072	-0.02916	0.11532	0.02247	-0.00118	0.01007	0.03535	-0.09528	-0.01294	-0.00138
ZI_SITE3	-0.03230	-0.01264	0.02247	0.38235	0.00248	0.01556	0.02927	-0.01074	-0.31773	-0.00289
ZI_B_UAVI	-0.00321	0.00052	-0.00118	0.00248	0.00227	0.00189	0.00136	0.00165	-0.00330	-0.00100
M_ONE	-0.04060	0.03565	0.01007	0.01556	0.00189	0.05452	-0.04217	-0.02022	-0.02594	-0.00069
M_ARM	0.04071	-0.11340	0.03535	0.02927	0.00136	-0.04217	0.14720	-0.04735	-0.04360	-0.00703
M_SITE2	0.00450	0.04219	-0.09528	-0.01074	0.00165	-0.02022	-0.04735	0.13310	0.02574	0.00215
M_SITE3	0.01794	0.02551	-0.01294	-0.31773	-0.00330	-0.02594	-0.04360	0.02574	0.30106	0.00429
M_B_UAVI	0.00034	0.00456	-0.00138	-0.00289	-0.00100	-0.00069	-0.00703	0.00215	0.00429	0.00143

Chapter 5

Conclusion

The statistical literature for count data with excess zeros has largely overlooked potential difficulties of parameter interpretation, focusing mostly on model fit. Proposed by Lambert (1992), the zero-inflated Poisson (ZIP) model produces two sets of parameter estimates, one set quantifying the effect of covariates upon the probability of being an excess zero and one set quantifying the effect of covariates on the mean of the non-excess zero population. While these two sets of latent class parameters might be meaningful in certain applications, often investigators seek to make inference on the sampled population rather than two latent subpopulations. Overall effects of an exposure of interest are often functions of many ZIP model parameters, requiring transformation techniques to estimate variability for inference. In a distinct but related field, the hurdle methodology (Mullahy, 1986) models all zeros separately from positive counts; however, this analytic method also yields two sets of parameters, one set modeling the probability of being a zero and one set modeling the conditional mean of the non-zero observations. Like ZIP models, hurdle models do not provide direct estimates of overall exposure effects (Albert, et al., 2011).

Building upon the ZIP methodology of Lambert (1992), we propose marginalized ZIP regression, a new model type that directly estimates the marginal mean rather than

the ZIP latent class mean. This new method yields overall estimates of exposure effect and is of particular interest to those analysts wishing to make inference on the sampled population. In particular, a primary zero-inflated outcome in a study of SafeTalk, a multicomponent, motivational interviewing-based, safer sex intervention, is used as a motivating example because inference is desired on the population of all participants randomized.

In order to account for correlated count data with excess zeros, we introduce the marginalized ZIP model with random effects, extending the ZIP with random effects of Hall (2000). Using established relationships between the log link and normal random effects, we outline situations in which the parameters from the marginalized ZIP have population-average interpretations in addition to subject-specific interpretations.

Finally a SAS/IML macro is created to implement the marginalized ZIP model for independent data. This program yields marginalized ZIP parameter estimates, model-based and robust standard errors, Wald chi-square tests, and p-values. In addition, odds ratios and incidence density ratios are presented with 95% Wald-type confidence intervals. Also, both model-based and robust covariance matrices are available for the analyst.

When analyzing count data with excess zeros, the statistical analyst should carefully consider methods which directly answer the research question of interest. With the methods contained in this work, a new type of model has been made available to those seeking inference on the sampled population with zero-inflated count data. Future work includes extending these methods to account for overdispersion beyond that of the excess zeros by marginalizing over the two processes of the zero-inflated negative binomial distribution. Also, marginalized ZIP methods for complex survey data are needed, as well as methods and software programs for sample size calculation in marginalized ZIP scenarios.

Appendix I

Likelihood Derivations for Chapter 2

First, we focus on the derivation of the MLE of $(\boldsymbol{\gamma}, \boldsymbol{\alpha})$ by constructing the likelihood. From Equation (2.1), we can derive the MLE's of $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$ using Newton-Raphson algorithm, as well as derive the analytic variance of these MLE's.

$$
\begin{aligned}
L(\boldsymbol{\gamma}, \boldsymbol{\delta}|\boldsymbol{y}) &= \prod_{y_i=0} [(e^{\boldsymbol{Z_i'}\boldsymbol{\gamma}} + e^{-\exp(\delta_i)})(1 + e^{\boldsymbol{Z_i'}\boldsymbol{\gamma}})^{-1}] \prod_{y_i>0} [(1 + e^{\boldsymbol{Z_i'}\boldsymbol{\gamma}})^{-1} e^{-\exp(\delta_i)} e^{\delta_i y_i}/(y_i!)] \\
&= \prod_{y_i} (1 + e^{\boldsymbol{Z_i'}\boldsymbol{\gamma}})^{-1} \prod_{y_i=0} (e^{\boldsymbol{Z_i'}\boldsymbol{\gamma}} + e^{-\exp(\delta_i)}) \prod_{y_i>0} [e^{-\exp(\delta_i)} e^{\delta_i y_i}/(y_i!)] \\
L(\boldsymbol{\gamma}, \boldsymbol{\alpha}|\boldsymbol{y}) &= \prod_{y_i} (1 + e^{\boldsymbol{Z_i'}\boldsymbol{\gamma}})^{-1} \prod_{y_i=0} (e^{\boldsymbol{Z_i'}\boldsymbol{\gamma}} + e^{-N_i(1+\exp(\boldsymbol{Z_i'}\boldsymbol{\gamma}))\exp(\boldsymbol{X_i'}\boldsymbol{\alpha})}) \\
&\quad \prod_{y_i>0} [e^{-N_i(1+\exp(\boldsymbol{Z_i'}\boldsymbol{\gamma}))\exp(\boldsymbol{X_i'}\boldsymbol{\alpha})} (1 + e^{\boldsymbol{Z_i'}\boldsymbol{\gamma}})^{y_i} e^{\boldsymbol{X_i'}\boldsymbol{\alpha} y_i} N_i^{y_i}/(y_i!)] \\
l(\boldsymbol{\gamma}, \boldsymbol{\alpha}|\boldsymbol{y}) &= -\sum_i \log(1 + e^{\boldsymbol{Z_i'}\boldsymbol{\gamma}}) + \sum_{y_i=0} \log(e^{\boldsymbol{Z_i'}\boldsymbol{\gamma}} + e^{-N_i \exp(\boldsymbol{X_i'}\boldsymbol{\alpha})(1+\exp(\boldsymbol{Z_i'}\boldsymbol{\gamma}))}) \\
&\quad + \sum_{y_i>0} (-N_i(1 + e^{\boldsymbol{Z_i'}\boldsymbol{\gamma}}) e^{\boldsymbol{X_i'}\boldsymbol{\alpha}} + y_i \log(1 + e^{\boldsymbol{Z_i'}\boldsymbol{\gamma}}) + \boldsymbol{X_i'}\boldsymbol{\alpha} y_i \\
&\quad + y_i \log(N_i) - \log(y_i!))
\end{aligned}
$$

Using this log-likelihood, the score equations are

$$
\begin{aligned}
\frac{\partial l(\gamma, \alpha)}{\partial \boldsymbol{\gamma}} &= \sum_{y_i=0} \frac{e^{\boldsymbol{Z_i'}\boldsymbol{\gamma}} \boldsymbol{Z_i'} + e^{-N_i \exp(\boldsymbol{X_i'}\boldsymbol{\alpha})(1+e^{\boldsymbol{Z_i'}\boldsymbol{\gamma}})} (-N_i e^{\boldsymbol{X_i'}\boldsymbol{\alpha} + \boldsymbol{Z_i'}\boldsymbol{\gamma}}) \boldsymbol{Z_i'}}{e^{\boldsymbol{Z_i'}\boldsymbol{\gamma}} + e^{-N_i \exp(\boldsymbol{X_i'}\boldsymbol{\alpha})(1+\exp(\boldsymbol{Z_i'}\boldsymbol{\gamma}))}} \\
&\quad + \sum_{y_i>0} \frac{y_i e^{\boldsymbol{Z_i'}\boldsymbol{\gamma}}}{1 + e^{\boldsymbol{Z_i'}\boldsymbol{\gamma}}} \boldsymbol{Z_i'} - N_i e^{\boldsymbol{X_i'}\boldsymbol{\alpha} + \boldsymbol{Z_i'}\boldsymbol{\gamma}} \boldsymbol{Z_i'} - \sum_i \frac{e^{\boldsymbol{Z_i'}\boldsymbol{\gamma}} \boldsymbol{Z_i'}}{1 + e^{\boldsymbol{Z_i'}\boldsymbol{\gamma}}} \\
&= \sum_i \left[\frac{I(y_i = 0) e^{\boldsymbol{Z_i'}\boldsymbol{\gamma}} (e^{N_i \exp(\boldsymbol{X_i'}\boldsymbol{\alpha})(1+\exp(\boldsymbol{Z_i'}\boldsymbol{\gamma}))} - N_i e^{\boldsymbol{X_i'}\boldsymbol{\alpha}})}{e^{\boldsymbol{Z_i'}\boldsymbol{\gamma}} e^{N_i \exp(\boldsymbol{X_i'}\boldsymbol{\alpha})(1+\exp(\boldsymbol{Z_i'}\boldsymbol{\gamma}))} + 1} + \frac{e^{\boldsymbol{Z_i'}\boldsymbol{\gamma}}(y_i - 1)}{1 + e^{\boldsymbol{Z_i'}\boldsymbol{\gamma}}} \right. \\
&\quad \left. - I(y_i > 0) N_i e^{\boldsymbol{X_i'}\boldsymbol{\alpha} + \boldsymbol{Z_i'}\boldsymbol{\gamma}} \right] \boldsymbol{Z_i'} \\
\frac{\partial l(\gamma, \alpha)}{\partial \boldsymbol{\alpha}} &= \frac{\partial \log L}{\partial \boldsymbol{\alpha}} = -\sum_{y_i=0} \frac{N_i(1 + e^{\boldsymbol{Z_i'}\boldsymbol{\gamma}}) e^{\boldsymbol{X_i'}\boldsymbol{\alpha}} e^{-N_i \exp(\boldsymbol{X_i'}\boldsymbol{\alpha})(1+\exp(\boldsymbol{Z_i'}\boldsymbol{\gamma}))} \boldsymbol{X_i'}}{e^{\boldsymbol{Z_i'}\boldsymbol{\gamma}} + e^{-N_i \exp(\boldsymbol{X_i'}\boldsymbol{\alpha})(1+\exp(\boldsymbol{Z_i'}\boldsymbol{\gamma}))}} \\
&\quad + \sum_{y_i>0} (y_i - N_i e^{\boldsymbol{X_i'}\boldsymbol{\alpha}}(1 + e^{\boldsymbol{Z_i'}\boldsymbol{\gamma}})) \boldsymbol{X_i'} \\
&= \sum_i \left[(y_i - N_i e^{\boldsymbol{X_i'}\boldsymbol{\alpha}}(1 + e^{\boldsymbol{Z_i'}\boldsymbol{\gamma}})) I(y_i > 0) - \frac{N_i(1 + e^{\boldsymbol{Z_i'}\boldsymbol{\gamma}}) e^{\boldsymbol{X_i'}\boldsymbol{\alpha}} I(y_i = 0)}{e^{\boldsymbol{Z_i'}\boldsymbol{\gamma}} e^{N_i \exp(\boldsymbol{X_i'}\boldsymbol{\alpha})(1+\exp(\boldsymbol{Z_i'}\boldsymbol{\gamma}))} + 1} \right] \boldsymbol{X_i'}
\end{aligned}
$$

Substituting the link functions $\text{logit}(\psi_i) = \boldsymbol{Z_i'\gamma}$ and $\log(\nu_i) = \boldsymbol{X_i'\alpha} + \log(N_i)$, these expressions of the score equations are equivalent to those presented in Section 2.2. The matrix of second derivatives of the log-likelihood has the form

$$\begin{bmatrix} \frac{\partial^2 l}{\partial \boldsymbol{\gamma}\partial \boldsymbol{\gamma'}} & \frac{\partial^2 l}{\partial \boldsymbol{\gamma}\partial \boldsymbol{\alpha'}} \\ \frac{\partial^2 l}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\gamma'}} & \frac{\partial^2 l}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha'}} \end{bmatrix}$$

where

$$\begin{aligned}
\frac{\partial^2 l}{\partial \boldsymbol{\gamma}\partial \boldsymbol{\gamma'}} &= \frac{\partial}{\partial \boldsymbol{\gamma}}\left[\frac{\partial l}{\partial \boldsymbol{\gamma}}\right]' \\
&= \frac{\partial}{\partial \boldsymbol{\gamma}}\Bigg\{\sum_i \boldsymbol{Z_i}\Bigg[\frac{I(y_i=0)e^{\boldsymbol{Z_i'\gamma}}(e^{N_i\exp(\boldsymbol{X_i'\alpha})(1+\exp(\boldsymbol{Z_i'\gamma}))} - N_i e^{\boldsymbol{X_i'\alpha}})}{e^{\boldsymbol{Z_i'\gamma}}e^{N_i\exp(\boldsymbol{X_i'\alpha})(1+\exp(\boldsymbol{Z_i'\gamma}))}+1} \\
&\qquad + \frac{e^{\boldsymbol{Z_i'\gamma}}(y_i-1)}{1+e^{\boldsymbol{Z_i'\gamma}}} - I(y_i>0)N_i e^{\boldsymbol{X_i'\alpha}+\boldsymbol{Z_i'\gamma}}\Bigg]\Bigg\} \\
&= -\sum_i \boldsymbol{Z_i}\Big[\frac{I(y_i=0)e^{\boldsymbol{Z_i'\gamma}}[N_i e^{\boldsymbol{X_i'\alpha}} - e^{N_i\exp(\boldsymbol{X_i'\alpha})(1+\exp(\boldsymbol{Z_i'\gamma}))}(1+N_i e^{\boldsymbol{Z_i'\gamma}+\boldsymbol{X_i'\alpha}} + N_i^2 e^{2\boldsymbol{Z_i'\gamma}+2\boldsymbol{X_i'\alpha}})]}{(e^{\boldsymbol{Z_i'\gamma}}e^{N_i\exp(\boldsymbol{X_i'\alpha})(1+\exp(\boldsymbol{Z_i'\gamma}))}+1)^2} \\
&\qquad - \frac{e^{\boldsymbol{Z_i'\gamma}}(y_i-1)}{(1+e^{\boldsymbol{Z_i'\gamma}})^2} + I(y_i>0)N_i e^{\boldsymbol{X_i'\alpha}+\boldsymbol{Z_i'\gamma}}\Big]\boldsymbol{Z_i'} \\
&= -\sum_i \boldsymbol{Z_i}\Big[\frac{I(y_i=0)\frac{\psi_i}{1-\psi_i}[\nu_i - e^{\nu_i(1-\psi_i)^{-1}}(1+\frac{\psi_i}{1-\psi_i}\nu_i + (\frac{\psi_i}{1-\psi_i}\nu_i)^2)]}{(\frac{\psi_i}{1-\psi_i}e^{\nu_i(1-\psi_i)^{-1}}+1)^2} - \psi_i(1-\psi_i)(y_i-1) \\
&\qquad + I(y_i>0)\frac{\psi_i}{1-\psi_i}\nu_i\Big]\boldsymbol{Z_i'}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 l}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha'}} &= \frac{\partial}{\partial \boldsymbol{\alpha}}\left[\frac{\partial l}{\partial \boldsymbol{\alpha}}\right]' \\
&= \frac{\partial}{\partial \boldsymbol{\alpha}}\Bigg\{\sum_i \boldsymbol{X_i}\Bigg[(y_i - N_i e^{\boldsymbol{X_i'\alpha}}(1+e^{\boldsymbol{Z_i'\gamma}}))I(y_i>0) - \frac{N_i(1+e^{\boldsymbol{Z_i'\gamma}})e^{\boldsymbol{X_i'\alpha}}I(y_i=0)}{e^{\boldsymbol{Z_i'\gamma}}e^{N_i\exp(\boldsymbol{X_i'\alpha})(1+\exp(\boldsymbol{Z_i'\gamma}))}+1}\Bigg]\Bigg\} \\
&= \sum_i \boldsymbol{X_i}\Big[(-N_i e^{\boldsymbol{X_i'\alpha}}(1+e^{\boldsymbol{Z_i'\gamma}}))I(y_i>0) \\
&\qquad - \frac{I(y_i=0)N_i(1+e^{\boldsymbol{Z_i'\gamma}})e^{\boldsymbol{X_i'\alpha}}[e^{\boldsymbol{Z_i'\gamma}}e^{N_i\exp(\boldsymbol{X_i'\alpha})(1+\exp(\boldsymbol{Z_i'\gamma}))}(1-N_i e^{\boldsymbol{X_i'\alpha}}(1+e^{\boldsymbol{Z_i'\gamma}}))+1]}{(e^{\boldsymbol{Z_i'\gamma}}e^{N_i\exp(\boldsymbol{X_i'\alpha})(1+\exp(\boldsymbol{Z_i'\gamma}))}+1)^2}\Big]\boldsymbol{X_i'} \\
&= -\sum_i \boldsymbol{X_i}\Big[\frac{\nu_i}{1-\psi_i}I(y_i>0) + \frac{I(y_i=0)\frac{\nu_i}{1-\psi_i}[\frac{\psi_i}{1-\psi_i}e^{\nu_i(1-\psi_i)^{-1}}(1-\frac{\nu_i}{1-\psi_i})+1]}{(\frac{\psi_i}{1-\psi_i}e^{\nu_i(1-\psi_i)^{-1}}+1)^2}\Big]\boldsymbol{X_i'}
\end{aligned}$$

$$
\begin{aligned}
\frac{\partial^2 l}{\partial \boldsymbol{\gamma}\partial \boldsymbol{\alpha}'} &= \frac{\partial}{\partial \boldsymbol{\gamma}}\left[\frac{\partial l}{\partial \boldsymbol{\alpha}}\right]' \\
&= \sum_i \boldsymbol{X_i}\frac{\partial}{\partial \boldsymbol{\gamma}}\left[(y_i - N_i e^{\boldsymbol{X_i'\alpha}}(1+e^{\boldsymbol{Z_i'\gamma}}))I(y_i>0) - \frac{N_i(1+e^{\boldsymbol{Z_i'\gamma}})e^{\boldsymbol{X_i'\alpha}}I(y_i=0)}{e^{\boldsymbol{Z_i'\gamma}}e^{N_i\exp(\boldsymbol{X_i'\alpha})(1+\exp(\boldsymbol{Z_i'\gamma}))}+1}\right] \\
&= -\sum_i \boldsymbol{X_i}\Big[N_i e^{\boldsymbol{X_i'\alpha}+\boldsymbol{Z_i'\gamma}}I(y_i>0) \\
&\quad + \frac{I(y_i=0)N_i e^{\boldsymbol{Z_i'\gamma}+\boldsymbol{X_i'\alpha}}[1-e^{N_i\exp(\boldsymbol{X_i'\alpha})(1+\exp(\boldsymbol{Z_i'\gamma}))}(1+N_i e^{\boldsymbol{Z_i'\gamma}+\boldsymbol{X_i'\alpha}}+N_i e^{2\boldsymbol{Z_i'\gamma}+\boldsymbol{X_i'\alpha}})]}{(e^{\boldsymbol{Z_i'\gamma}}e^{N_i\exp(\boldsymbol{X_i'\alpha})(1+\exp(\boldsymbol{Z_i'\gamma}))}+1)^2}\Big]\boldsymbol{Z_i'} \\
&= -\sum_i \boldsymbol{X_i}\left[\frac{\psi_i}{1-\psi_i}\nu_i I(y_i>0) + \frac{I(y_i=0)\frac{\psi_i}{1-\psi_i}\nu_i[1-e^{\nu_i(1-\psi_i)^{-1}}(1+\frac{\psi_i}{1-\psi_i}\nu_i+(\frac{\psi_i}{1-\psi_i})^2\nu_i)]}{(\frac{\psi_i}{1-\psi_i}e^{\nu_i(1-\psi_i)^{-1}}+1)^2}\right]\boldsymbol{Z_i'} \\
\frac{\partial^2 l}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\gamma}'} &= \left[\frac{\partial^2 l}{\partial \boldsymbol{\gamma}\partial \boldsymbol{\alpha}'}\right]'
\end{aligned}
$$

In order to obtain the Fisher information matrix, we calculate the negative expectations of the above second derivatives. First, we note that

$$
\begin{aligned}
P(Y_i=0) &= \psi_i + (1-\psi_i)e^{-\nu_i(1-\psi_i)^{-1}} = \frac{1-\psi_i}{e^{\nu_i(1-\psi_i)^{-1}}}\left(\frac{\psi_i}{1-\psi_i}e^{\nu_i(1-\psi_i)^{-1}}+1\right) \\
P(Y_i>0) &= (1-\psi_i)(1-e^{-\nu_i(1-\psi_i)^{-1}}) \\
E(Y_i) &= \nu_i
\end{aligned}
$$

Then

$$
\begin{aligned}
-E\left[\frac{\partial^2 l}{\partial \boldsymbol{\gamma}\partial \boldsymbol{\gamma}'}\right] &= \sum_i \boldsymbol{Z_i}\Big[\frac{P(y_i=0)\frac{\psi_i}{1-\psi_i}[\nu_i - e^{\nu_i(1-\psi_i)^{-1}}(1+\frac{\psi_i}{1-\psi_i}\nu_i+(\frac{\psi_i}{1-\psi_i}\nu_i)^2)]}{(\frac{\psi_i}{1-\psi_i}e^{\nu_i(1-\psi_i)^{-1}}+1)^2} \\
&\quad -\psi_i(1-\psi_i)(E[y_i]-1) + P(y_i>0)\frac{\psi_i}{1-\psi_i}\nu_i\Big]\boldsymbol{Z_i'} \\
&= \sum_i \boldsymbol{Z_i}\Big[\frac{(1-\psi_i)(\frac{\psi_i}{1-\psi_i}e^{\nu_i(1-\psi_i)}+1)}{e^{\nu_i(1-\psi_i)}}\frac{\frac{\psi_i}{1-\psi_i}[\nu_i - e^{\nu_i(1-\psi_i)^{-1}}(1+\frac{\psi_i}{1-\psi_i}\nu_i+(\frac{\psi_i}{1-\psi_i}\nu_i)^2)]}{(\frac{\psi_i}{1-\psi_i}e^{\nu_i(1-\psi_i)^{-1}}+1)^2} \\
&\quad -\psi_i(1-\psi_i)(\nu_i-1) + (1-\psi_i)(1-e^{-\nu_i(1-\psi_i)^{-1}})\frac{\psi_i}{1-\psi_i}\nu_i\Big]\boldsymbol{Z_i'} \\
&= \sum_i \boldsymbol{Z_i}\Big[\frac{\psi_i^2(1-\psi_i)(\frac{\psi_i}{1-\psi_i}\nu_i+1)(e^{\nu_i(1-\psi_i)^{-1}}-\frac{\nu_i}{1-\psi_i}-1)}{\psi_i e^{\nu_i(1-\psi_i)^{-1}}+(1-\psi_i)}\Big]\boldsymbol{Z_i'}
\end{aligned}
$$

$$
\begin{aligned}
-E\left[\frac{\partial^2 l}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha'}}\right] &= \sum_i \boldsymbol{X_i}\Big[\frac{\nu_i}{1-\psi_i}P(y_i>0) + \frac{P(y_i=0)\frac{\nu_i}{1-\psi_i}[\frac{\psi_i}{1-\psi_i}e^{\nu_i(1-\psi_i)^{-1}}(1-\frac{\nu_i}{1-\psi_i})+1]}{(\frac{\psi_i}{1-\psi_i}e^{\nu_i(1-\psi_i)^{-1}}+1)^2}\Big]\boldsymbol{X_i'} \\
&= \sum_i \boldsymbol{X_i}\Big[\nu_i(1-e^{-\nu_i(1-\psi_i)^{-1}}) + \\
&\qquad \frac{(1-\psi_i)(\frac{\psi_i}{1-\psi_i}e^{\nu_i(1-\psi_i)^{-1}}+1)}{e^{\nu_i(1-\psi_i)^{-1}}}\frac{\frac{\nu_i}{1-\psi_i}[\frac{\psi_i}{1-\psi_i}e^{\nu_i(1-\psi_i)^{-1}}(1-\frac{\nu_i}{1-\psi_i})+1]}{(\frac{\psi_i}{1-\psi_i}e^{\nu_i(1-\psi_i)^{-1}}+1)^2}\Big]\boldsymbol{X_i'} \\
&= \sum_i \boldsymbol{X_i}\Big[\frac{\nu_i[\psi_i(e^{\nu_i(1-\psi_i)^{-1}}-\frac{\nu_i}{1-\psi_i}-1)+1]}{\psi_i e^{\nu_i(1-\psi_i)^{-1}}+(1-\psi_i)}\Big]\boldsymbol{X_i'}
\end{aligned}
$$

$$
\begin{aligned}
-E\left[\frac{\partial^2 l}{\partial \boldsymbol{\gamma}\partial \boldsymbol{\alpha'}}\right] &= \sum_i \boldsymbol{X_i}\left[\frac{\psi_i\nu_i}{1-\psi_i}P(y_i>0) + \frac{P(y_i=0)\frac{\psi_i\nu_i}{1-\psi_i}[1-e^{\nu_i(1-\psi_i)^{-1}}(1+\frac{\psi_i}{1-\psi_i}\nu_i+(\frac{\psi_i}{1-\psi_i})^2\nu_i)]}{(\frac{\psi_i}{1-\psi_i}e^{\nu_i(1-\psi_i)^{-1}}+1)^2}\right]\boldsymbol{Z_i'} \\
&= \sum_i \boldsymbol{X_i}\Bigg[\psi_i\nu_i(1-e^{-\mu_i}) + \\
&\qquad \frac{(1-\psi_i)(\frac{\psi_i}{1-\psi_i}e^{\nu_i(1-\psi_i)^{-1}}+1)}{e^{\nu_i(1-\psi_i)^{-1}}}\frac{\frac{\psi_i\nu_i}{1-\psi_i}[1-e^{\nu_i(1-\psi_i)^{-1}}(1+\frac{\psi_i}{1-\psi_i}\nu_i+(\frac{\psi_i}{1-\psi_i})^2\nu_i)]}{(\frac{\psi_i}{1-\psi_i}e^{\nu_i(1-\psi_i)^{-1}}+1)^2}\Bigg]\boldsymbol{Z_i'} \\
&= \sum_i \boldsymbol{X_i}\left[\nu_i\psi_i\left(1-e^{-\nu_i(1-\psi_i)^{-1}}+\frac{e^{-\nu_i(1-\psi_i)^{-1}}-(1+\frac{\psi_i}{1-\psi_i}\nu_i+\nu_i(\frac{\psi_i}{1-\psi_i})^2)}{\frac{\psi_i}{1-\psi_i}e^{\nu_i(1-\psi_i)^{-1}}+1}\right)\right]\boldsymbol{Z_i'}
\end{aligned}
$$

Appendix II

SAS Code from Chapter 3

The following SAS NLMIXED code was used for the SafeTalk motivating example.

```
proc nlmixed data=safetalk seed=31415;

  parms b0 0 b1 0 b2 0 b3 0 b4 0 b5 0 b6 0 b7 0 b8 0
    a0 0 a1 0 a2 0 a3 0 a4 0 a5 0 a6 0 a7 0 a8 0
    sigma1 1 sigma12 0 sigma2 1;

   /* linear predictor for the zero-inflation probability      */

   logit_psi = a0 + a1*site2 + a2*site3 + a3*v2 + a4*v2*st + a5*v3 + a6*v3*st
                  + a7*v4 + a8*v4*st + c1;
   *logit(\psi)=Z\gamma + c;

   /* useful functions of \psi */
   psi1  = exp(logit_psi)/(1+exp(logit_psi));
   *\psi = exp(Z\gamma+c)/(1+exp(Z\gamma+c));
   psi2  = 1/(1+exp(logit_psi));
   *1-\psi = (1+exp(Z\gamma+c))^-1;

   /* Overall mean \nu */
   log_nu   = b0 + b1*site2 + b2*site3 + b3*v2 + b4*v2*st + b5*v3 + b6*v3*st
                 + b7*v4 + b8*v4*st + d1;
```

```
   delta = log(psi2**(-1)) + log_nu;

   /* Build the mZIP + RE log likelihood */
   if outcome=0 then
        ll = log(psi1 + psi2*(exp(-exp(delta))));
   else ll = log(psi2) - exp(delta)  + outcome*(delta) - lgamma(outcome + 1);
   model outcome ~ general(ll);
   random c1 d1~normal([0,0],[sigma1,sigma12,sigma2]) SUBJECT=urn;
   contrast "TX" b4, b6, b8;
run;
```

Appendix III

SAS/IML Macro from Chapter 4

```
/* file: mZIP 101.sas

MZIP for Independent Data - Version 1.00

SAS/IML MACRO FOR MARGINALIZED ZERO-INFLATED POISSON REGRESSION
FOR INDEPENDENT DATA
BY D. LEANN LONG - THE UNIVERSITY OF NORTH CAROLINA-CHAPEL HILL

VERSION NOTES:
v 1.01
-All of the following are suggestions from Preisser 02/2013:
-Adding p-values for all parameters; GENMOD performs Wald chi-
square 1 df for each, so add these p-values for both model-
based and robust SE.
-Adding output of IDR's and 95% confidence intervals for each
predictor in the marginal mean model. Also adding OR and
95% confidence intervals for excess zeros model.
-Adding option to output both model-based and robust covariances
matrices to either default work directory or user-defined
location
```

```
PURPOSE
Fit the marginalized zero-inflated Poisson (mZIP) model, presenting both
model-based and robust standard error estimates.
As in Long, Preisser, Herring & Golin (Biometrics, under review).

PARAMETERS SPECIFIED {Default value}

%m_ZIP  (
   DATA           = SAS dataset,                              { &syslast }
   OUTCOME        = dependent variable,                       { required }
   ZI_PRED        = variables in excess zero model,           { required }
   M_PRED         = variables in marginal mean model,         { required }
   OUTPUT_MB_COV  = binary indicator of whether to output     { 0 }
                    model-based covariance matrix (0 | 1)
                    (Inverse of Fisher information)
   SAVE_MB_COV    = binary indicator of whether to output     { 0 }
                    model-based covariance matrix (0 | 1)
   MB_COV_DATA    = Location to save model-based covariance   {work.mb_cov}
                    matrix
   OUTPUT_R_COV   = binary indicator of whether to output     { 0 }
                    robust covariance matrix (0 | 1)
   SAVE_R_COV     = binary indicator of whether to output     { 0 }
                    robust covariance matrix (0 | 1)
   R_COV_DATA     = Location to save robust covariance        {work.r_cov}
                    matrix
)
```

```
NOTE:
In both ZI_PRED and M_PRED, you must specify a vector of ones in order to
have an intercept in each portion of the model.

REQUIRED MACRO SPECIFICATIONS
To run the macro, the user is required to provide a SAS dataset
(DATA) with response variable (OUTCOME), a list of independent variables
for the excess zero model (ZI_PRED), and a list of independent variables
for the marginal mean model (M_PRED).
**********************************************************************/
%macro m_ZIP (
     DATA=&syslast,    /* Data set to use; Syslast if not specified    */
     OUTCOME= ,        /* Outcome of interest                          */
     ZI_PRED= ,        /* Predictors for Zero-inflated portion of model*/
     M_PRED= ,        /* Predictors for Poisson latent class          */
     OUTPUT_MB_COV= 0, /* Indicates whether to output model-based      */
                       /*   covariance into SAS data set               */
     SAVE_MB_COV= 0,   /* Indicates whether to save model-based        */
                       /*   covariance into SAS data set               */
     MB_COV_DATA=work.mb_cov,
                       /* Where to save the model-based covariance     */
     OUTPUT_R_COV= 0,  /* Indicates whether to output model-based      */
                       /*   covariance into SAS data set               */
     SAVE_R_COV= 0,    /* Indicates whether to save model-based        */
                       /*   covariance into SAS data set               */
```

```
     R_COV_DATA=work.r_cov
                        /* Where to save the model-based covariance     */
);

proc iml worksize = 30000 symsize = 1000;

reset noprint noname fuzz;
use &DATA;
label_zipred = {&ZI_PRED};
label_ppred  = {&M_PRED};

p_zi = ncol(label_zipred);/* Number of predictors in Zero-inflated Portion*/
p_p  = ncol(label_ppred); /* Number of predictors in Marginal Mean Model  */

read all var {&OUTCOME} INTO yvar;
read all var {&ZI_PRED} INTO zivar;
read all var {&M_PRED}  INTO pvar;
read all var {&ZI_PRED &M_PRED &OUTCOME} INTO dat;

/* write the log-likelihood function for mZIP */
start LogLik(param) global(dat, zivar, pvar, yvar, p_zi, p_p);

gamma = J(p_zi,1,0);
alpha = J(p_p,1,0);

do ii = 1 to p_zi; *assigning parameters to be estimated;
```

```
gamma[ii,] = param[ii];
end;
do jj = 1 to p_p;
alpha[jj,] = param[p_zi+jj];
end;

outcome = yvar;

z_gamma =  zivar * gamma;
x_alpha =  pvar  * alpha;

  f   = -sum(log(1+exp(z_gamma)))
        +sum(log(exp(z_gamma)+exp(-exp(x_alpha)#
        (1+exp(z_gamma))))#(outcome=0))
        -sum(exp(x_alpha)#(1+exp(z_gamma))#(outcome>0))
        +sum(outcome#log(1+exp(z_gamma))#(outcome>0))
        +sum(((x_alpha)#outcome)#(outcome>0))
        -sum((lfact(outcome))#(outcome>0));
  return ( f );
finish;

/*    parameter constraint matrix */
con = J(2,(p_zi+p_p),.);

/* Call an optimization routine*/
```

```
/*You can now call an optimization routine to find the MLE estimate
for the data. You need to provide an initial guess to the
optimization routine, and you need to tell it whether you are
finding a maximum or a minimum. There are other options that you
can specify, such as how much printed output you want. */

p=J(1,(p_zi+p_p),0); /* initial values for solution */

opt = {1,    /* find maximum of function   */
       0};   /* prints nothing      */

call NLPNRA(rc, result, "LogLik", p, opt, con);
/*Nonlinear optimization by Newton-Raphson method*/

log_like = LogLik(result`);
result2 = result || log_like;

columns = J (1,(p_zi+p_p+1)," ");
col_varM= J (1,(p_zi+p_p)," ");
col_varR= J (1,(p_zi+p_p)," ");

/*Column Naming*/
do kk = 1 to p_zi;
columns[,kk] = concat("ZI_",label_zipred[,kk]);
col_varM[,kk]= concat("ZI_M_SE_",label_zipred[,kk]);
col_varR[,kk]= concat("ZI_R_SE_",label_zipred[,kk]);
```

```
end;
do ll = 1 to p_p;
columns[,(p_zi + ll)] = concat("M_",label_ppred[,ll]);
col_varM[,(p_zi + ll)]= concat("M_M_SE_",label_ppred[,ll]);
col_varR[,(p_zi + ll)]= concat("M_R_SE_",label_ppred[,ll]);
end;
columns[,p_zi+p_p+1] = "Log-likelihood";
zi_columns=strip(columns[,1:p_zi]);
m_columns =strip(columns[,(p_zi+1):(p_zi+p_p)]);

/* Using estimates from Newton-Raphson to calculate
model-based and robust SE */

outcome = yvar;
k = nrow(yvar);
z = zivar`;
x = pvar`;
gamma_hat   = result2[,1:p_zi];
alpha_hat = result2[,(p_zi+1):(p_zi+p_p)];
z_gamma_hat  = zivar * gamma_hat`;  /* n x 1 vector*/
x_alpha_hat  = pvar  * alpha_hat`;   /* n x 1 vector*/

psi_hat  = exp(z_gamma_hat)/(1+exp(z_gamma_hat));/* n x 1 vector*/
nu_hat  = exp(x_alpha_hat);  /* n x 1 vector*/
psi_hat2 = 1/(1-psi_hat);
```

```
/* d/dg [d/dg log L] */
diag_gg  = (psi_hat##2#(1-psi_hat)#(psi_hat#psi_hat2#nu_hat+1)#
             (exp(nu_hat#psi_hat2)-nu_hat#psi_hat2-1))/
             (psi_hat#exp(nu_hat#psi_hat2)+(1-psi_hat));

/* d/da [d/da log L] */
diag_aa   = (nu_hat#(psi_hat#(exp(nu_hat#psi_hat2)-nu_hat#psi_hat2-1)+1))/
             (psi_hat#exp(nu_hat#psi_hat2)+(1-psi_hat));

/* d/dg [d/da log L] */
diag_ga  = nu_hat#psi_hat#(1-exp(-nu_hat#psi_hat2)+(-(1+nu_hat#psi_hat#
             psi_hat2+nu_hat#(psi_hat#psi_hat2)##2+exp(-nu_hat#psi_hat2))/
             ((psi_hat#psi_hat2)#exp(nu_hat#psi_hat2)+1)));

I_gg = z * DIAG(diag_gg) * z‘;
I_aa = x * DIAG(diag_aa) * x‘;
I_ga = x * DIAG(diag_ga) * z‘;
I_ag = I_ga‘;

Inform = (I_gg || I_ag) // (I_ga || I_aa);
Inv_inform = GINV(Inform);
M1 = J(p_zi+p_p,p_zi+p_p,0);
score_g = J(p_zi,1,0);
score_a = J(p_p,1,0);
do qq = 1 to k;
```

```
y = outcome[qq,];
ph= psi_hat[qq,];
ph2=psi_hat2[qq,];
nu=nu_hat[qq,];

score_g = ((y=0)#(ph#ph2#(exp(nu#ph2)-nu))/(ph#ph2#exp(nu#ph2)+1)
                + ph#(y - 1)-(y>0)#ph#ph2#nu) * (z[,qq])‘;
score_a = ((y-nu#ph2)#(y>0)-(y=0)#(nu#ph2)/(ph#ph2#exp(nu#ph2)+1))*(x[,qq])‘;

score = score_g || score_a;
M1 = M1 + score‘ * score;

if (qq=k) then do;
robust = inv_inform * M1 * inv_inform;
m_se = sqrt(vecdiag(inv_inform)‘);
r_se = sqrt(vecdiag(robust)‘);

if &save_mb_cov = 1 then do;
    create &mb_cov_data FROM Inv_inform[colname=columns];
    append from Inv_inform;
    close &mb_cov_data;
end;

if &save_r_cov = 1 then do;
    create &r_cov_data FROM robust[colname=columns];
    append from robust;
```

```
    close &r_cov_data;
end;


end;
end;

/* Calculating Confidence Intervals & Wald Chi-Square P-Values */

wald_m=J(1,(p_zi+p_p),0);
wald_r=J(1,(p_zi+p_p),0);
p_m=J(1,(p_zi+p_p),0);
p_r=J(1,(p_zi+p_p),0);
p_m2=J(1,(p_zi+p_p),"              ");
p_r2=J(1,(p_zi+p_p),"              ");

or_zi=J(1,p_zi,0); /* Creating OR and CI vectors for ZI process */
or_ci_l_m=J(1,p_zi,0);
or_ci_u_m=J(1,p_zi,0);
or_ci_l_r=J(1,p_zi,0);
or_ci_u_r=J(1,p_zi,0);

idr_m=J(1,p_p,0); /* Creating IDR and CI vectors for Marginal Mean */
idr_ci_l_m=J(1,p_p,0);
idr_ci_u_m=J(1,p_p,0);
idr_ci_l_r=J(1,p_p,0);
idr_ci_u_r=J(1,p_p,0);
```

```
/* Wald Statistics and P-values */
do rr = 1 to (p_zi+p_p);
wald_m[,rr] = (result2[,rr]/m_se[,rr])**2; *Model-based;
wald_r[,rr] = (result2[,rr]/r_se[,rr])**2; *Robust;
p_m[,rr] = round(1-cdf('CHISQ',wald_m[,rr],1),0.0001);
if (p_m[,rr] < 0.0001) then p_m2[,rr] = "<0.0001";
else p_m2[,rr] = char(p_m[,rr]);
p_r[,rr] = round(1-cdf('CHISQ',wald_r[,rr],1),0.0001);
if (p_r[,rr] < 0.0001) then p_r2[,rr] = "<0.0001";
else p_r2[,rr] = char(p_r[,rr]);
end;

/* IDR/OR & Confidence Intervals */
do ss = 1 to p_zi;
or_zi[,ss] = round(exp(result2[,ss]),0.0001);
or_ci_l_m[,ss] = round(exp(result2[,ss] - 1.96*m_se[,ss]),0.0001);
or_ci_u_m[,ss] = round(exp(result2[,ss] + 1.96*m_se[,ss]),0.0001);
or_ci_l_r[,ss] = round(exp(result2[,ss] - 1.96*r_se[,ss]),0.0001);
or_ci_u_r[,ss] = round(exp(result2[,ss] + 1.96*r_se[,ss]),0.0001);
end;
do w = 1 to (p_p);
idr_m[,w] = round(exp(result2[,p_zi+w]),0.0001);
idr_ci_l_m[,w] = round(exp(result2[,p_zi+w] - 1.96*m_se[,p_zi+w]),0.0001);
idr_ci_u_m[,w] = round(exp(result2[,p_zi+w] + 1.96*m_se[,p_zi+w]),0.0001);
idr_ci_l_r[,w] = round(exp(result2[,p_zi+w] - 1.96*r_se[,p_zi+w]),0.0001);
```

```
idr_ci_u_r[,w] = round(exp(result2[,p_zi+w] + 1.96*r_se[,p_zi+w]),0.0001);
end;

t_col = columns';
t_res = round(result2,0.0001)';
t_mse = round(m_se,0.0001)';
t_w_m = round(wald_m,0.0001)';
t_p_m = p_m2';
t_rse = round(r_se,0.0001)';
t_w_r = round(wald_r,0.0001)';
t_p_r = p_r2';

t_or_zi  = or_zi';
t_or_ci_l_m = or_ci_l_m';
t_or_ci_u_m = or_ci_u_m';
t_or_ci_l_r = or_ci_l_r';
t_or_ci_u_r = or_ci_u_r';
t_idr_m  = idr_m';
t_idr_ci_l_m= idr_ci_l_m';
t_idr_ci_u_m= idr_ci_u_m';
t_idr_ci_l_r= idr_ci_l_r';
t_idr_ci_u_r= idr_ci_u_r';
t_zi_columns= zi_columns';
t_m_columns = m_columns';

C1 = {"Parameters"};
```

```
C2 = {"Estimates"} ;

C3 = {"Model-Based SE"};

C4 = {"Model-Based Wald"};

C5 = {"Model-Based p-value"};

C6 = {"Robust SE"} ;

C7 = {"Robust Wald"};

C8 = {"Robust p-value"};


C_ZI_COL = {"Zero-Inflation Parameters"};

C_OR = {"Odds Ratio"};

C_OR_LCI= {"Lower 95% CI: Model-Based"};

C_OR_UCI= {"Upper 95% CI: Model-Based"};

C_OR_LCI_R= {"Lower 95% Robust CI"};

C_OR_UCI_R= {"Upper 95% Robust CI"};


C_M_COL = {"Marginal Mean Parameters"};

C_IDR = {"IDR"};

C_IDR_LCI= {"Lower 95% CI: Model-Based"};

C_IDR_UCI= {"Upper 95% CI: Model-Based"};

C_IDR_LCI_R= {"Lower 95% Robust CI"};

C_IDR_UCI_R= {"Upper 95% Robust CI"};


print "Marginalized ZIP Model Output:",,"Model-Based Results",,

      t_col[colname = C1]

      t_res[colname=C2]

      t_mse[colname=C3]
```

```
      t_w_m[colname=C4]
      t_p_m[colname=C5],,
"Robust (Empirical) Results",,
t_col[colname = C1]
t_res[colname=C2]
t_rse[colname=C6]
t_w_r[colname=C7]
t_p_r[colname=C8],,
"Odds Ratios for Zero-inflated Parameters",,
t_zi_columns[colname=C_ZI_COL]
t_or_zi[colname=C_OR]
t_or_ci_l_m[colname=C_OR_LCI]
t_or_ci_u_m[colname=C_OR_UCI],,
t_zi_columns[colname=C_ZI_COL]
t_or_zi[colname=C_OR]
t_or_ci_l_r[colname=C_OR_LCI_R]
t_or_ci_u_r[colname=C_OR_UCI_R],,
"Incidence Density Ratios for Marginal Mean Parameters",,
t_m_columns[colname=C_M_COL]
t_idr_m[colname=C_IDR]
t_idr_ci_l_m[colname=C_IDR_LCI]
t_idr_ci_u_m[colname=C_IDR_UCI],,
t_m_columns[colname=C_M_COL]
t_idr_m[colname=C_IDR]
t_idr_ci_l_r[colname=C_IDR_LCI_R]
t_idr_ci_u_r[colname=C_IDR_UCI_R],,;
```

```
if &output_mb_cov = 1 then
print "Model-based Covariance Matrix:",,
Inv_inform[colname=columns],,;

if &output_r_cov = 1 then
print "Robust Covariance Matrix:",,
robust[colname=columns],,;

close &DATA;
quit;

%mend;
```

Bibliography

[1] J. Albert, W. Wang, and S. Nelson. Estimating overall exposure effects for zero-inflated regression models with application to dental caries. *Statistical Methods in Medical Research*, doi:10.1177/0962280211407800, 2011.

[2] D. Böhning, E. Dietz, P. Schlattmann, L. Mendonça, and U. Kirchner. The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 162:195–209, 1999.

[3] M. Bulsara, C. Holman, E. Davis, and T. Jones. Evaluating risk factors associated with severe hypoglycaemia in epidemiology studieswhat method should we use? *Diabetic Medicine*, 21(8):914–919, 2004.

[4] A. Buu, R. Li, X. Tan, and R. A. Zucker. Statistical models for longitudinal zero-inflated count data with applications to the substance abuse field. *Statistics in Medicine*, 31(29):4074–4086, 2012.

[5] P. J. Carrivick, A. H. Lee, and K. K. Yau. Zero-inflated Poisson modeling to evaluate occupational safety interventions. *Safety Science*, 41(1):53–63, 2003.

[6] M. Dobbie and A. Welsh. Theory & Methods: Modelling correlated zero-inflated count data. *Australian & New Zealand Journal of Statistics*, 43(4):431–444, 2001.

[7] P. Ghosh and W. Tu. Assessing sexual attitudes and behaviors of young women: a joint model with nonlinear time effects, time varying covariates, and dropouts. *Journal of the American Statistical Association*, 104(486):474–485, 2009.

[8] M. Gilthorpe, M. Frydenberg, Y. Cheng, and V. Baelum. Modelling count data with excessive zeros: The need for class prediction in zero-inflated models and the issue of data generation in choosing between zero-inflated and generic mixture models for dental caries data. *Statistics in Medicine*, 28(28):3539–3553, 2009.

[9] C. Golin, R. Davis, S. Przybyla, B. Fowler, S. Parker, J. Earp, E. Quinlivan, S. Kalichman, S. Patel, and C. Grodensky. Safetalk, a multicomponent, motivational interviewing-based, safer sex counseling program for people living with HIV/AIDS: A qualitative assessment of patients' views. *AIDS Patient Care and STDs*, 24(4):237–245, 2010.

[10] C. Golin, S. Patel, K. Tiller, E. Quinlivan, C. Grodensky, and M. Boland. Start talking about risks: development of a motivational interviewing-based safer sex program for people living with HIV. *AIDS and Behavior*, 11:72–83, 2007.

[11] D. Hall and Z. Zhang. Marginal models for zero inflated clustered data. *Statistical Modelling*, 4(3):161–180, 2004.

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

[13] P. Heagerty. Marginally specified logistic-normal models for longitudinal binary data. *Biometrics*, 55(3):688–698, SEP 1999.

[14] D. Heilbron. Zero-altered and other regression models for count data with added zeros. *Biometrical Journal*, 36:531–547, 1994.

[15] M. Hernán and J. Robins. Estimating causal effects from epidemiological data. *Journal of Epidemiology and Community Health*, 60(7):578–586, 2006.

[16] S. Iddi and G. Molenberghs. A combined overdispersed and marginalized multilevel model. *Computational Statistics & Data Analysis*, 56:1944–1951, 2012.

[17] D. Lambert. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34:1–14, 1992.

[18] K. Lee, Y. Joo, J. Song, and D. Harper. Analysis of zero-inflated clustered count data: A marginalized model approach. *Computational Statistics & Data Analysis*, 55(1):824–837, 2011.

[19] P. McCullagh and J. Nelder. *Generalized Linear Models*, volume 37. Chapman & Hall/CRC, 1989.

[20] C. McCulloch and S. Searle. *Generalized, Linear, and Mixed Models*. Wiley, 2001.

[21] Y. Min and A. Agresti. Random effect models for repeated measures of zero-inflated count data. *Statistical Modelling*, 5:1–19, 2005.

[22] S. Moon and J. Shin. Health care utilization among Medicare-Medicaid dual eligibles: a count data analysis. *BMC Public Health*, 6(1):88, 2006.

[23] J. Mullahy. Specification and testing of some modified count data models. *Journal of Econometrics*, 33:341–365, 1986.

[24] S. M. Mwalili, E. Lesaffre, and D. Declerck. The zero-inflated negative binomial regression model with correction for misclassification: an example in caries research. *Statistical Methods in Medical Research*, 17(2):123–139, 2008.

[25] B. Neelon, A. O'Malley, and S. Normand. A Bayesian model for repeated measures zero-inflated count data with application to outpatient psychiatric service use. *Statistical Modelling*, 10(4):421–439, 2010.

[26] J. S. Preisser, J. W. Stamm, D. L. Long, and M. E. Kincade. Review and recommendations for zero-inflated count regression modeling of dental caries indices in epidemiological studies. *Caries Research*, 46(4):413–423, 2012.

[27] M. Ridout, C. Demétrio, and J. Hinde. Models for count data with many zeros. In *Proceedings of the XIXth International Biometric Conference*, volume 19, pages 179–192, 1998.

[28] J. Ritz and D. Spiegelman. Equivalence of conditional and marginal regression models for clustered and longitudinal data. *Statistical Methods in Medical Research*, 13(4):309–323, 2004.

[29] O. Rosen, W. Jiang, and M. Tanner. Mixtures of marginal models. *Biometrika*, 87(2):391–404, 2000.

[30] SAS Institute Inc. *The SAS System* Cary, NC. Version 9.3. http://support.sas.com/documentation/93/index.html, 2013a.

[31] SAS Institute Inc. *SAS/IML Software* Cary, NC. Version 9.3. http://support.sas.com/documentation/cdl/en/imlug/65547/HTML/default/viewer.htm#imlug_imlstart_toc.htm, 2013b.

[32] SAS Institute Inc. *SAS/STAT Software, The NLMIXED Procedure* Cary, NC. Version 9.3. http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#nlmixed_toc.htm, 2013b.

[33] V. Shankar, J. Milton, and F. Mannering. Modeling accident frequencies as zero-altered probability processes: an empirical inquiry. *Accident Analysis & Prevention*, 29(6):829–837, 1997.

[34] H. White. Maximum likelihood estimation of misspecified models. *Econometrica*, pages 1–25, 1982.

[35] K. Yau, K. Wang, and A. Lee. Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros. *Biometrical Journal*, 45(4):437–452, 2003.

[36] M. Young, J. Preisser, B. Qaqish, and M. Wolfson. Comparison of subject-specific and population averaged models for count data from cluster-unit intervention trials. *Statistical Methods in Medical Research*, 16(2):167–184, 2007.

[37] C. J. Zorn. Evaluating zero-inflated and hurdle Poisson specifications. *Midwest Political Science Association*, 18(20):1–16, 1996.