Zero-Inflated Count Regression Models: A Review and Contemporary Perspective

Derek S. Young* and Eric S. Roemmele
Department of Statistics, University of Kentucky

July 28, 2017

Abstract

Count regression models provide a highly effective framework for relating discrete responses to covariates. Their utility in modern applied statistics is highlighted by important data problems in diverse areas like biometry, ecology, and insurance. However, a common practical problem with observed count data is the presence of excess zeros relative to the assumed count distribution. The seminal work of Lambert (1992) was one of the first papers to rigorously treat the problem of zero-inflated count data in the presence of covariates. Since then, numerous advancements have been made with zero-inflated count regression models, such as the development of score tests for inference in the presence of zero inflation, handling zero inflation in spatial data, diagnostic procedures for goodness-of-fit, flexible computational tools, and multivariate count regression models with different notions of inflation. The goal of this article is to provide a concise treatment of these advancements, with emphasis on recent methods for more complex data settings. This discussion will concurrently highlight important applied research problems where such models have provided significant insight.

Keywords: Data dispersion; Diagonal inflation; EM algorithm; Generalized linear models; Score tests; Two-component mixture model

*Contact: Derek S. Young (derek.young@uky.edu) Department of Statistics, University of Kentucky, Lexington, KY 40536.

1 Introduction

When modeling count data, the behavior of zero counts in the observed data often creates difficulties. For example, data may structurally exclude zeros or have different generating processes for the zero and non-zero counts. Models capable of characterizing such data are *zero-truncated models* (Cohen, Jr., 1960) and *hurdle models* (Mullahy, 1986), respectively. Observed data can also have excessive zeros with respect to the assumed underlying count distribution; i.e., *zero inflation*. Zero-inflated (ZI) models are used to characterize the zero counts as arising from a two-component mixture model, where one component is a point mass at zero and the other component is an assumed count distribution. One of the earliest works to develop ZI models in the presence of covariates is the seminal paper by Lambert (1992). In that paper, the ZI Poisson (ZIP) regression model is introduced, where the two components of the mixture structure are used to characterize the states of a manufacturing process. The two components are defined as, respectively, a *perfect state* where defects are extremely rare, and an *imperfect state* where defects are possible. In terms of the behavior of the zeros, these two states equate to *structural zeros* and *random zeros*, respectively.

Since the publication of Lambert (1992), numerous theoretical and computational advancements have been made to expand the utility of ZI count regression models. In the literature, estimation and inference are primarily conducted using likelihood methods. Since ZI count regression models are two-component mixture models, estimation via expectation-maximization (EM) algorithms (Dempster et al., 1977) is a natural choice. Lambert (1992) provided the details of an EM algorithm for estimating the ZIP regression model, as well as discussed practical issues relative to using a Newton-Raphson algorithm for maximizing the likelihood. In particular, the EM algorithm is straightforward to program, but it is usually slow to converge relative to the Newton-Raphson algorithm. However, this has become less of a practical concern with the advancement of computing power as well as the development of various strategies that can be employed to speed up convergence of EM algorithms; see

Chapter 4 of McLachlan and Krishnan (1997). A drawback of the Newton-Raphson algorithm is that it can fail to converge, as was noted in the numerical illustrations in Lambert (1992). Regardless, numerous optimization algorithms can be employed for estimating ZI regression models, but EM algorithms tend to be a pragmatic choice. Examples of EM algorithms for estimating more complex ZI count regression models include a ZI regression model with random effects (Hall, 2000), a class of Markov models for count time series with excess zeros (Wang, 2001), and a multilevel model that handles correlated counts (Lee et al., 2006).

Numerous inference considerations about ZI regression models have been investigated in the literature. Typically, one appeals to large-sample theory for estimating standard errors and conducting tests of parameter estimates of ZI regression models. This approach can also be used to construct normal-theory confidence intervals, but likelihood ratio confidence intervals are usually more trustworthy (Lambert, 1992; Hall, 2000). Tests for more model-specific hypotheses have also been developed in the literature. For example, testing whether a ZI count regression model is an improvement over its corresponding non-ZI count regression model can be accomplished with a score test (van den Broek, 1995; Janaskul and Hinde, 2002) or a boundary likelihood ratio test (see section 11.3.5 of Hilbe, 2011). Residual diagnostics for generalized linear models (GLMs) are typically employed when fitting ZI regression models. For example, one can assess Pearson, deviance, or Anscombe residuals. More recently, Sellers and Raim (2016) and Young et al. (2017) have highlighted the utility of using randomized quantile residuals (Dunn and Smyth, 1996) for assessing the fit of ZI regression models. Influence measures for ZI regression models have also been developed. For example, Garay et al. (2011) developed a global influence measure based on a generalized Cook's distance, and a local influence measure (see Cook, 1986) for ZI negative binomial (ZINB) regression models.

Beyond the significant theoretical and computational developments of ZI count regression models, the applied literature highlights the true utility of these models. ZI regression models are a commonly-used modeling strategy for numerous disciplines.

For example, ecological datasets tend to contain a large proportion of zero counts. Martin et al. (2005) discussed in detail the four ways zeros arise in ecological datasets. Specifically, the authors classify the zeros as *true zero counts* — which would equate to the degenerate component of the ZI regression model — or *false zero counts* — which would arise from the count regression component. The true zero counts arise because either the species "does not occur at a site because of the ecological process, or effect under study" or because it "does not saturate its entire suitable habitat by chance." The false zero counts arise because either the species "occurs at a site, but is not present during the survey period" or because it "occurs at a site and is present during the survey period, but the observer fails to detect it." Martin et al. (2005) further analyzed two ecological datasets — one on bird assemblages and one on woodland bird patch occupancy — using ZI regression models. Other ecological applications where ZI regression models have been successfully employed include the analysis of stream fish and abundance (Boone et al., 2012) and to assess the relationship between the abundance of a vulnerable plant species and the environment (Potts and Elith, 2006). The general importance of ZI regression models in ecology is also emphasized in Chapter 11 of Zuur et al. (2009).

Another discipline where ZI regression models are frequently used is insurance. Baetschmann and Winkelmann (2012) paralleled the notion of *strategic* and *incidental* zeros to the perfect and imperfect states introduced in Lambert (1992). For example, in determining the health policy of a (potential) policyholder, the number of physician visits can be an indication of overall health and, thus, affect the level of policy coverage. Baetschmann and Winkelmann (2012) noted that when modeling the number of physician visits, a person might have zeros during a particular time period (or *exposure*) because they follow alternative medicine and never visit a physician (strategic zero) or because they were healthy during the time period and had no reason to visit the physician (incidental zero). Yip and Yau (2005) noted similar states with how zeros occur in the no claim discount (NCD) system, which is widely used by automobile insurers. Policyholders could have zero claims either because

they typically will not file the claim if it is small (strategic zero) or because they simply did not have any issues regarding their automobile (incidental zero). Similar analyses involving ZI regression modeling for the risk classification of claim counts are performed in Boucher et al. (2007), Tang et al. (2014), and Sarul and Sahin (2015).

While ecology and insurance are two areas where ZI regression models are commonly used, the utility of such models has been demonstrated with numerous other diverse applications. Examples highlighting the broad range of interesting applications using ZI regression models include, the development of a functional relationship between truck accidents and the geometric design of road sections (Miaou, 1994), an analysis of economists seeking academic interviews after tenure denial (List, 2001), a study about the effects of cigarette price change on smoking behavior (Sheu et al., 2004), the assessment of dental cavities in low birth weight adolescents (Albert et al., 2014), and development of a set of models to characterize the on-going quality of census frames in preparation for the United States 2020 Census (Young et al., 2017). There was also a special section devoted to novel ZI models in *Biometrical Journal* (volume 58, issue 2).

Most research that uses ZI regression models rightfully cite Lambert (1992). Our main goal with this article is to provide a central location where the subsequent major developments in analyzing ZI data — with primary emphasis on the count regression setting — can be found. We concurrently highlight some of the important applied problems where ZI regression problems have provided considerable insight.

The rest of this article is organized as follows. Starting with the framework of Lambert (1992), Section 2 provides a formal discussion of the ZI count regression model with emphasis on the two most commonly assumed discrete distributions for the count component: the Poisson and the negative binomial. In this section, estimation and inference from a likelihood perspective are discussed in greater detail, as well as a review of available software. Section 3 concerns zero inflation in the context of modeling with non-standard discrete distributions when data dispersion is present. Section 4 discusses ZI count regression models in the presence of correlated observa-

tions. Section 5 provides a brief discussion of some Bayesian approaches used with ZI regression models, including how such approaches are used to model spatial data with zero inflation. Two different inflation mechanisms for multivariate count responses are discussed in Section 6. In Section 7, a brief overview of related models — such as zero-truncated and hurdle models — is presented. Finally, Section 8 provides some concluding remarks and future research ideas involving ZI count regression models.

2 Traditional ZI Models

Let the discrete random variable $Y \in \mathbb{N}$ be a count of interest; e.g., length of stay in a hospital (Yau et al., 2003), the counts of trees in a forest using grid-cell data (Nishii and Tanaka, 2013), or the number of added or deleted housing units in a census block (Young et al., 2017). Suppose that $\mathbf{X} \in \mathbb{R}^p$ and $\mathbf{W} \in \mathbb{R}^q$ are vectors of covariates measured with Y. Let $\mathbf{x}$, $\mathbf{w}$, and y be, respectively, the realizations of those variables. Suppose further that the conditional distribution of $Y|\mathbf{X} = \mathbf{x}$ is fully parametric with probability mass function (pmf) $p(\cdot; \mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\vartheta})$. Here, $\boldsymbol{\beta}$ appears through the conditional mean $\mathrm{E}[Y|\mathbf{X} = \mathbf{x}] = \eta$ specified via a known link function $g(\eta) = \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}$, and $\boldsymbol{\vartheta}$ pertains to any additional (possibly nuisance) parameters, such as the scale (or dispersion) parameter in negative binomial regression. $g(\cdot)$ is typically taken as the log link since it ensures that the value of η is positive. Also, the count of interest is sometimes measured in terms of its *exposure*, say, N. Assuming a log link implies

$$\log(\eta) = \log(N) + \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}, \tag{1}$$

where $\log(N)$ appears as an offset term in the right-hand side of the above expression.

For ZI data, we use a two-component mixture of a point mass at zero and a count distribution, where π (the mixing proportion) is the probability of observing a zero count such that $h(\pi) = \mathbf{w}^{\mathrm{T}}\boldsymbol{\alpha}$. $h(\cdot)$ is also a known link function, which is usually

taken as the logit link to ensure π is constrained to the unit interval; i.e.,

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \mathbf{w}^{\mathrm{T}}\boldsymbol{\alpha}. \tag{2}$$

Note that an offset term is not typically included in the above relationship. Letting $q(\cdot; \mathbf{w}, \boldsymbol{\alpha})$ be a zero-hurdle mass function — i.e., a mass function right-censored at 1 — we can write $\pi \equiv q(0; \mathbf{w}, \boldsymbol{\alpha})$. Thus, the pmf for a ZI count regression model is

$$f(y; \mathbf{x}, \mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\vartheta}) = q(0; \mathbf{w}, \boldsymbol{\alpha})\mathtt{I}\{y = 0\} + (1 - q(0; \mathbf{w}, \boldsymbol{\alpha}))p(y; \mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\vartheta}), \tag{3}$$

where $\mathtt{I}\{\cdot\}$ is the indicator function. Note that the predictors $\mathbf{w}$ may by uncoupled from those predictors in $\mathbf{x}$. Moreover, it is typical to assume that η and π are not functionally related, although this need not be the case.

2.1 Estimation and Inference

The most common count distributions used in ZI regression models (Lambert, 1992; Heilbron, 1994; Greene, 2007) are the Poisson, which has pmf

$$p(y; \lambda) = (y!)^{-1}\lambda^{y}e^{-\lambda}, \quad \lambda > 0, \tag{4}$$

and the negative binomial, which has one parameterization of its pmf as

$$p(y; \mu, \theta) = \frac{\Gamma(\theta + y)}{y!\Gamma(\theta)}\left(\frac{\mu}{\theta + \mu}\right)^{y}\left(\frac{\theta}{\theta + \mu}\right)^{\theta}, \quad \mu, \theta > 0. \tag{5}$$

The negative binomial distribution is often used to characterize overdispersion in data, even when accounting for excess zeros in a ZI model. In their respective pmfs, λ and μ are the means, which are related to a vector of covariates $\mathbf{x}$ through the log link function. In the negative binomial pmf, $\theta > 0$ is a dispersion parameter that is usually assumed constant, but could be modeled as a function of measured covariates if there is overdispersion or underdispersion relative to the negative binomial regression

model; see Chapter 7.5 of Hilbe (2011).

Suppose we have a sample of size n. The pmfs in (4) and (5) can be extended to the Poisson regression pmf and negative binomial pmf by modeling, respectively, $\lambda_i(\boldsymbol{\beta}) = \exp\{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\}$ and $\mu_i(\boldsymbol{\beta}) = \exp\{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\}$. Modeling the mixing proportions as $\pi_i(\boldsymbol{\alpha}) = \text{logit}^{-1}(\mathbf{w}_i^{\mathrm{T}}\boldsymbol{\alpha})$, and using the ZI count regression pmf in (3), it follows that the ZIP regression loglikelihood is

$$\begin{aligned}\ell_1(\boldsymbol{\lambda}, \boldsymbol{\pi}; \mathbf{y}) = &\sum_{y_i=0} \log\left(\pi_i(\boldsymbol{\alpha}) + (1-\pi_i(\boldsymbol{\alpha}))\exp\{-\lambda_i(\boldsymbol{\beta})\}\right) \\ &+ \sum_{y_i>0}\left[\log(1-\pi_i(\boldsymbol{\alpha})) - \lambda_i(\boldsymbol{\beta}) + y_i\log(\lambda_i(\boldsymbol{\beta})) - \log(y!)\right]\end{aligned} \tag{6}$$

and that the ZINB regression loglikelihood is

$$\begin{aligned}\ell_2(\boldsymbol{\mu}, \boldsymbol{\pi}, \theta; \mathbf{y}) = &\sum_{y_i=0} \log\left(\pi_i(\boldsymbol{\alpha}) + (1-\pi_i(\boldsymbol{\alpha}))\left(\frac{\theta}{\theta+\mu_i(\boldsymbol{\beta})}\right)^{\theta}\right) \\ &+ \sum_{y_i>0}[\log(1-\pi_i(\boldsymbol{\alpha})) + \log(\Gamma(\theta+y_i)) - \log(\Gamma(\theta)) - \log(y_i!) \\ &\qquad + y_i\log\left(\mu_i(\boldsymbol{\beta})\right) + \theta\log(\theta) - (\theta+y_i)\log(\theta+\mu_i(\boldsymbol{\beta}))],\end{aligned} \tag{7}$$

where $\boldsymbol{\lambda} = (\lambda_1(\boldsymbol{\beta}), \ldots, \lambda_n(\boldsymbol{\beta}))^{\mathrm{T}}$, $\boldsymbol{\mu} = (\mu_1(\boldsymbol{\beta}), \ldots, \mu_n(\boldsymbol{\beta}))^{\mathrm{T}}$, $\boldsymbol{\pi} = (\pi_1(\boldsymbol{\alpha}), \ldots, \pi_n(\boldsymbol{\alpha}))^{\mathrm{T}}$, and $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$. Note that the ZI geometric (ZIG) regression model is a special case of the ZINB regression model, which has been discussed in detail by Pandya et al. (2012). Also, if the response variable has an upper bounded count, then a ZI binomial (ZIB) regression model is easily obtained; see Hall (2000).

Maximum likelihood estimation for ZIP and ZINB regression models has been thoroughly treated in Lambert (1992) and Chapter 6 of Hilbe (2011), respectively. The asymptotic distribution of the maximum likelihood estimates (MLEs) is, of course, multivariate normal. While closed-form solutions for the MLEs do not exist, they are easily obtained using numerical methods. Moreover, Lambert (1992) and Garay et al. (2011) provide the formulas for the gradients and second derivatives, which can then be used for computing the observed information matrix for standard

errors.

Newton-Raphson and EM algorithms are commonly employed for calculating the ZIP regression and ZINB regression MLEs. The Newton-Raphson algorithm, when it converges, is typically faster than the EM algorithm. However, EM algorithms are quite easy to code and take advantage of the mixture structure of ZI regression models by iteratively fitting weighted versions of simpler GLMs (Hall and Zhang, 2004). An EM algorithm for estimating parameters in ZIP regression is given in Lambert (1992). ZINB regression also requires estimating a dispersion parameter. Instead of doing simultaneous maximization of the regression and dispersion parameters, we can break up the optimization into two conditional maximization steps via an expectation-conditional-maximization (ECM; Meng and Rubin, 1993). Thus, we perform iterative estimation of the dispersion parameter and the regression parameters, such that conditioning on the former allows the latter to be estimated via fitting a GLM. Details for an ECM algorithm for estimating parameters in a ZINB regression model are given in the Supplementary Material.

Novel semiparametric ZI count regression models have also been developed to provide greater flexibility relative to the fully parametric framework discussed thus far. Recalling that η is used to generically represent the conditional mean, most of the approaches in the literature have worked with some variant of the following:

$$\log(\eta) = k(\mathbf{X}) \quad \text{and} \quad \text{logit}^{-1}(\pi) = m(\mathbf{W}), \tag{8}$$

where $k(\cdot)$ and $m(\cdot)$ are unknown smooth functions. Lam et al. (2006) developed a semiparametric ZIP regression model by using a generalized partial linear model for the Poisson rate λ, namely, $\log(\lambda) = \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} + k(T)$, where a single predictor T is modeled nonparametrically. They made inference about k using the sieve method and established asymptotic properties of the sieve MLE. He et al. (2010) later extended this semiparametric ZIP regression model to also allow $\text{logit}^{-1}(\pi) = \mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha} + m(T)$, resulting in what they termed the *doubly semiparametric ZIP regression model*. The

authors also used the sieve method and established the corresponding asymptotic properties. Both Lam et al. (2006) and He et al. (2010) used their semiparametric ZIP regression models to analyze absenteeism data reported on a public health survey in Indonesia. Stasinopoulos and Rigby (2007) noted that generalized additive models (GAMs) for location, scale, and shape (GAMLSS; Rigby and Stasinopoulos, 2005) can be used when relating one or more of the ZI count regression parameters to additive covariates via suitable link functions. Liu and Chan (2011) constructed a partially constrained ZI count regression models using GAMs where some covariates affect η and π proportionally on the link scale. They highlighted the importance of such a model in ecological studies, including a real application involving jellyfish abundance data. Chiogna and Gaetan (2007) estimated the nonparametric components in (8) using penalized regression splines. Their semiparametric ZIP regression model was used to study the relationship between avian abundance and environmental variables. Similar to that model, Minami et al. (2007) took a penalized likelihood approach and estimated a semiparametric ZINB regression model for characterizing shark bycatch data.

Testing of regression coefficients for predictors in fully parametric ZI count regression models is typically based on the asymptotic normality of the MLEs. Such testing applies to predictors in either the count regression component or the zero-hurdle mass function that characterizes the proportion of zero inflation. Using the approximate standard errors, it is then straightforward to calculate Wald-based confidence intervals.

While tests of predictors are important, score tests on the zero-inflation structure in ZI count regression models have also been given considerable attention. In particular,

$$\begin{aligned} H_0 &: \pi = 0 \\ H_A &: 0 < \pi < 1 \end{aligned} \tag{9}$$

is used to test the null hypothesis of a count regression model against the alternative hypothesis of a ZI count regression model. Many score tests pertaining to the hy-

potheses above have been developed in the literature. van den Broek (1995) developed a score test for zero inflation in the Poisson setting, but where the zero-hurdle mass function was not modeled as a function of predictors. Janaskul and Hinde (2002) extended this score test to the setting where the zero-hurdle mass function could depend on predictors. Similar score tests were developed in the negative binomial setting by Janaskul and Hinde (2008).

Another test of interest is

$$\begin{aligned} H_0 &: \theta = 0 \\ H_A &: \theta > 0, \end{aligned} \tag{10}$$

which is used to test for the presence of overdispersion in the ZI count regression models. Specifically, the null distribution is the ZIP regression model and the alternative distribution is the ZINB regression model. Ridout et al. (2001) developed a score test for testing the hypotheses in (10). Deng and Paul (2005) provided a more comprehensive approach by developing score tests for each of the hypotheses in (9) and (10) as well as for testing both of them simultaneously.

Young et al. (2017) highlighted that since the null hypotheses for the tests in (9) and (10) are on a boundary of the parameter space, the standard asymptotic χ^2_1 distribution is conservative. An alternative is to employ a boundary likelihood ratio test (LRT) using a modified χ^2 distribution (Hilbe, 2011). The corresponding test statistic is characterized as having a limiting distribution that is a mixture of χ^2 distributions:

$$0.5\chi^2_0 + 0.5\chi^2_1, \tag{11}$$

where χ^2_0 is a degenerate distribution with all of its mass placed at 0. The distribution in (11) is also called the *chi-bar-squared distribution* when it pertains to the specific testing paradigm of unicomponent versus two-component mixture models with known component densities (Lindsay, 1995).

The Vuong non-nested test (Vuong, 1989) is also commonly used for testing the hypotheses in (9) and (10). However, Wilson (2015) pointed out that by Vuong's definition, nesting occurs on the boundary, so a model is not *strictly* nested in its

ZI counterpart. This does not imply that the models are non-nested and, hence, score tests or the boundary LRT should be used instead of the Vuong non-nested test. Finally, as noted in Hilbe (2011), model selection criteria, such as the Akaike information criterion (AIC) or Bayesian information criterion (BIC), can be used to select between all of the models discussed thus far; i.e., ZIP regression, ZINB regression, Poisson regression, and negative binomial regression.

Various pseudo-R^2 measures for assessing goodness-of-fit of ZI count regression models were developed by Martin and Hall (2016). One such measure the authors developed includes an adjustment to reward parsimony due to the chi-bar-squared limiting distribution of the boundary LRT. The adjusted-R^2 quantity for the ZIP regression model is given by

$$R^2_{\text{ZIP,adj}} = 1 - \frac{\ell_1(\mathbf{y}, \mathbf{z}; \mathbf{y}) - \ell_1(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\pi}}; \mathbf{y}) + p + q + 0.5}{\ell_1(\mathbf{y}, \mathbf{z}; \mathbf{y}) - \ell_1(\hat{\boldsymbol{\lambda}}_0, \mathbf{0}; \mathbf{y})}, \tag{12}$$

where $\hat{\boldsymbol{\lambda}} \equiv \boldsymbol{\lambda}(\hat{\boldsymbol{\beta}})$ and $\hat{\boldsymbol{\pi}} \equiv \boldsymbol{\pi}(\hat{\boldsymbol{\alpha}})$ are the MLEs under the full ZIP regression model, $\boldsymbol{\lambda}_0 = \bar{y}\mathbf{1}_n$, and $\mathbf{z} = (z_1, \ldots, z_n)^{\mathrm{T}}$ such that $z_i = \mathtt{I}\{y_i = 0\}$. Similarly, the adjusted-R^2 quantity for the ZINB regression model is given by

$$R^2_{\text{ZINB,adj}} = 1 - \frac{\ell_2(\mathbf{y}, \mathbf{z}, \hat{\theta}; \mathbf{y}) - \ell_2(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\pi}}, \hat{\theta}; \mathbf{y}) + p + q + 1.5}{\ell_2(\mathbf{y}, \mathbf{z}, \hat{\theta}; \mathbf{y}) - \ell_2(\hat{\boldsymbol{\mu}}_0, \mathbf{0}, \hat{\theta}; \mathbf{y})}, \tag{13}$$

where $\hat{\boldsymbol{\mu}} \equiv \boldsymbol{\mu}(\hat{\boldsymbol{\beta}})$, $\hat{\boldsymbol{\pi}} \equiv \boldsymbol{\pi}(\hat{\boldsymbol{\alpha}})$, and $\hat{\theta}$ are the MLEs under the full ZINB regression model, and $\boldsymbol{\mu}_0 = \bar{y}\mathbf{1}_n$. Note that in both expressions, the loglikelihoods evaluated at $\boldsymbol{\pi} = \mathbf{0}$ are simply the loglikelihoods for the corresponding non-ZI regression model. The above quantities are similar to the adjusted-R^2 formulas for non-ZI count regression models, as presented in Mittlböck and Waldhör (2000).

2.2 Software and Numerical Demonstrations

Many statistical software programs have routines for estimating ZI count regression models, but the scope of such functions is usually limited to estimating ZIP and ZINB

n	Procedure	β_0 (= 3.000)	β_1 (= −1.500)	α_0 (= 0.500)	α_1 (= −0.500)
50	`zeroinfl`	2.998 (0.230)	-1.511 (0.240)	1.108 (10.197)	-1.387 (15.001)
	`vglm`	2.997 (0.229)	-1.506 (0.239)	0.844 (2.228)	-1.030 (3.874)
	`GENMOD`	2.998 (0.230)	-1.512 (0.240)	0.758 (1.203)	-0.852 (1.333)
	`NLMIXED`	2.998 (0.230)	-1.511 (0.240)	1.038 (5.681)	-1.307 (8.667)
	`COUNTREG`	2.998 (0.230)	-1.511 (0.240)	6.453 (262.742)	-11.971 (371.527)
250	`zeroinfl`	2.996 (0.084)	-1.500 (0.098)	0.523 (0.375)	-0.532 (0.356)
	`vglm`	2.996 (0.084)	-1.500 (0.098)	0.522 (0.375)	-0.532 (0.355)
	`GENMOD`	2.996 (0.084)	-1.500 (0.098)	0.523 (0.375)	-0.532 (0.355)
	`NLMIXED`	2.996 (0.084)	-1.500 (0.098)	0.523 (0.375)	-0.532 (0.356)
	`COUNTREG`	2.996 (0.084)	-1.500 (0.098)	0.523 (0.375)	-0.532 (0.356)

Table 1: Results for estimating simulated data from a ZIP regression model using the two `R` functions in the `pscl` and `VGAM` packages, and the three `SAS` procedures, `GENMOD`, `NLMIXED`, and `COUNTREG`. Notice the highly variable results across the five methods for $\boldsymbol{\alpha}$ when $n = 50$.

regression models. In `SAS` (SAS Institute Inc., 2013), three available procedures are `PROC GENMOD`, `PROC NLMIXED`, and `PROC COUNTREG`. In `R` (R Core Team, 2016), two of the major functions available are `zeroinfl` and `vglm`, which are within the `pscl` (Zeileis et al., 2008) and `VGAM` (Yee, 2015) packages, respectively. For estimation, all of the aforementioned functions employ gradient-based methods, such as Newton-Raphson or iteratively reweighted least squares (IRLS), by default.

We demonstrate the accuracy of the estimates obtained using the aforementioned five procedures through a brief simulation study. We generated $B = 1000$ datasets of sizes $n \in \{50, 250\}$ from a ZIP regression model and ZINB regression model. The parameters for these models are given in the headers of Tables 1 and 2, respectively. We report the mean and standard deviation of the ZI count regression estimates using the different procedures such that all arguments are set to their respective defaults. The ZIP regression estimates in Table 1 are nearly identical for the different procedures, but with fairly noticeable differences in the estimates of $\boldsymbol{\alpha}$ when $n = 50$, especially for `PROC COUNTREG`. The ZINB regression estimates in Table 2 are nearly identical for the different procedures and for both sample sizes. However, this time `PROC COUNTREG` demonstrates quite different results for both sample sizes. These numerical results are consistent with those obtained in Liu et al. (2017), who performed an extensive

n	Method	β_0 (= 3.000)	β_1 (= 1.200)	α_0 (= 0.500)	α_1 (= −0.500)	θ (= 4.482)
50	`zeroinfl`	2.986 (0.160)	1.168 (0.160)	-1.023 (0.563)	0.605 (0.536)	5.576 (1.221)
	`vglm`	2.986 (0.160)	1.168 (0.160)	-1.023 (0.563)	0.605 (0.536)	5.576 (1.221)
	`GENMOD`	2.986 (0.160)	1.168 (0.160)	-1.023 (0.563)	0.605 (0.536)	5.576 (1.221)
	`NLMIXED`	2.986 (0.160)	1.168 (0.160)	-1.023 (0.563)	0.605 (0.536)	5.576 (1.221)
	`COUNTREG`	2.871 (0.304)	1.011 (0.305)	-0.360 (63.363)	-106.078 (473.480)	2.208 (2.121)
250	`zeroinfl`	3.003 (0.076)	1.189 (0.077)	-0.573 (0.217)	0.486 (0.217)	4.617 (0.412)
	`vglm`	3.004 (0.076)	1.189 (0.077)	-0.573 (0.217)	0.486 (0.217)	4.617 (0.412)
	`GENMOD`	3.004 (0.076)	1.189 (0.077)	-0.573 (0.217)	0.486 (0.217)	4.617 (0.412)
	`NLMIXED`	3.004 (0.076)	1.189 (0.077)	-0.573 (0.217)	0.486 (0.217)	4.617 (0.412)
	`COUNTREG`	2.901 (0.216)	1.137 (0.134)	-1.844 (11.082)	-56.140 (533.695)	1.540 (2.398)

Table 2: Results for estimating simulated data from a ZINB regression model using the two `R` functions in the `pscl` and `VGAM` packages, and the three `SAS` procedures, `GENMOD`, `NLMIXED`, and `COUNTREG`. Notice the considerably different results for `COUNTREG`.

simulation study that addresses the performance of ZI estimation procedures in `SAS` and `R`.

We also performed a brief timing study for comparing the five procedures discussed above. We generated datasets of different sample sizes from a ZIP and ZINB regression model, and timed each of the five procedures using their default settings. `PROC GENMOD` was found to perform the quickest for nearly all of the settings considered. The `zeroinfl` function typically took the longest when estimating the ZIP regression model, while the `vglm` function typically took the longest when estimating the ZINB regression model. More details, including the actual timing results, are given in the Supplementary Material

We note that the `gamlss` package (Rigby and Stasinopoulos, 2005) can also estimate ZIP and ZINB regression models, as well as some related models discussed in Section 7, using the GAMLSS framework. Estimation is performed using a maximum penalized likelihood approach, which differs from the `pscl` and `VGAM` packages. Thus, we did not include a comparison with estimates obtained using the `gamlss` package.

Other statistical software have built-in routines to estimate ZIP and ZINB regression models. To estimate these models in Mplus (Muthén and Muthén, 1998–2012), place the `(i)` option after the count response variable in the `count` statement. In the Stata software (Stata Technical Support, 2015), the `zip` and `zinb` functions can be used, respectively, to estimate ZIP and ZINB regression models. Also, the NCSS

software (NCSS, LLC, 2016) has routines for estimating both of these models, which is found under the *Regression with Count Data* menu.

2.3 Example: Relationship Data

Cupach and Spitzberg (2004) presented data on $n = 387$ responses to a version of the Relational Pursuit-Pursuer Short Form (RP-PSF), which was used to study the unwanted pursuit behavior (UPB) of recently split couples. The form consisted of 28 questions about the pursuer's behavior — e.g., "Did the pursuer leave unwanted gifts?" — each measured on a five-point Likert scale (from 0 for *never* to 4 for *over five times*). The response UPB is a discrete summary index to these 28 questions, where higher scores indicate more perpetration. These data were analyzed using ZI count regression models by Loeys et al. (2012) and Sellers and Raim (2016), where the latter noted the clear presence of overdispersion since the mean UPB is 2.284 and the corresponding variance is 23.302. The predictors of interest are the anxious attachment level (continuous) between the previous couple, and a binary indicator for education level (0 for *lower than a bachelor's degree* and 1 for *at least a bachelor's degree*).

Figure 1 is a histogram of the frequency of UPB counts. The frequency was truncated at 15 in order to focus on the majority of the data. There are nine UPB counts greater than 15, and the maximum observed count is 34. We overlaid the fits based on the Poisson, negative binomial, ZIP, and ZINB distributions. Clearly the Poisson and ZIP fits are not appropriate as they noticeably deviate from the general shape of the data. The negative binomial and ZINB fits, however, provide a noticeable improvement. These fits were all obtained without accounting for covariates, but the observations we made with the histogram suggest that using the negative binomial or ZINB distribution for the count regression distribution should be a reasonable choice.

We next performed the boundary LRTs discussed in Section 2.1. The test of zero-inflation for the Poisson regression and negative binomial regression settings, as well as the test for using a ZIP versus a ZINB regression model, all have highly significant

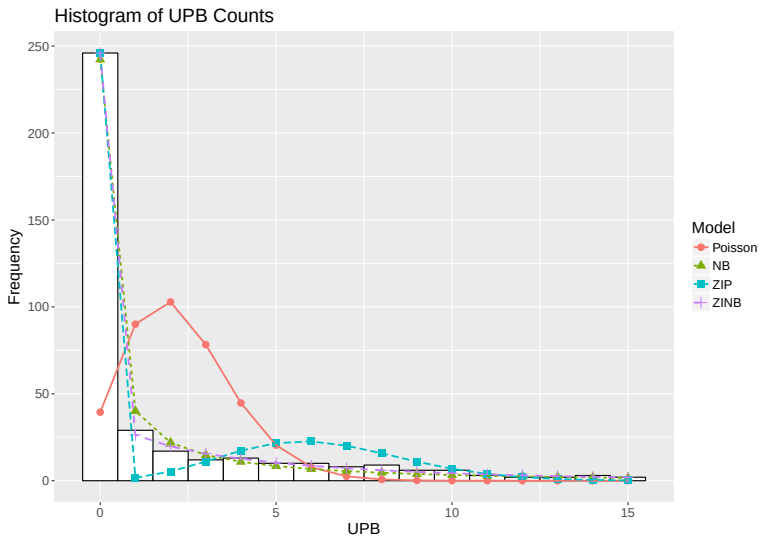

Figure 1: Histogram of the relationship data, truncated to show values of 15 or fewer for the UPB response. Fits for the four count distributions — Poisson, negative binomial, ZIP, and ZINB — are overlaid. A visually better fit can be seen with the estimated ZINB model.

results in favor of the alternative hypotheses; the largest p-value is 2.06×10^{-7}. Thus, these tests indicate the presence of zero-inflation and, more specifically, the use of the ZINB distribution. Table 3 gives the adjusted-R^2 values for the ZIP and ZINB regression models when including education level, anxiety attachment level, or both covariates in the respective model. These covariates were included in both the conditional mean model for the count distribution and the mixing proportion model for the zero inflation. For the ZIP regression models, the largest adjusted-R^2 is obtained for the model with only education level as a covariate. For the ZINB regression models, the largest adjusted-R^2 is obtained for the model with both covariates included. Since the boundary LRTs indicated the use of the ZINB distribution, we use the adjusted-R^2 results from this model and include both covariates in the model. As one final check of the fit, we calculated the randomized quantile residuals for the Poisson regression, negative binomial regression, ZIP regression, and ZINB regression models, where both covariates are included. The quantile-quantile (Q–Q) plots for these four estimated models are given in Figure 2. Clearly, better fits are obtained using neg-

Covariates	ZIP Regression	ZINB Regression
Education	0.523	0.016
Attachment	0.492	0.026
Education, Attachment	0.503	0.028

Table 3: Adjusted-R^2 results when including education level, anxiety attachment level, or both in the model. Results for both the ZIP regression and ZINB regression models are reported.

ative binomial regression or ZINB regression. In fact, the ZINB regression model provides a slightly better fit for those values in the right-hand tail of the distribution.

3 ZI Count Regression Models for Handling Data Dispersion

Relative to the Poisson distribution, many count datasets are heavily right-skewed and exhibit excess zero observations. As noted in Famoye and Singh (2006) and Sellers and Raim (2016), overdispersion has the tendency to increase the proportion of zeros such that other distributions, like the negative binomial, can improve the fit. However, better fits can be obtained through overdispersed models that simultaneously characterize excess zeros.

When the negative binomial still fails to provide a good fit to data, the generalized Poisson distribution (Consul and Jain, 1973; Consul, 1989) can often provide an improved fit. For two given parameters, $\mu > 0$ and $\max\{-1, -\mu/4\} \leq \alpha < 1$, one parameterization of the generalized Poisson pmf (Famoye, 1993) is

$$p(y; \mu, \alpha) = \begin{cases} \mu(\mu + \alpha y)^{y-1} \exp\{-(\mu + \alpha y)\}/y!, & \text{for } y \in \mathbb{N}; \\ 0, & \text{if } y > m \text{ when } \alpha < 0, \end{cases} \tag{14}$$

where m is the largest positive integer for which $\alpha + m\mu > 0$ when the disper-

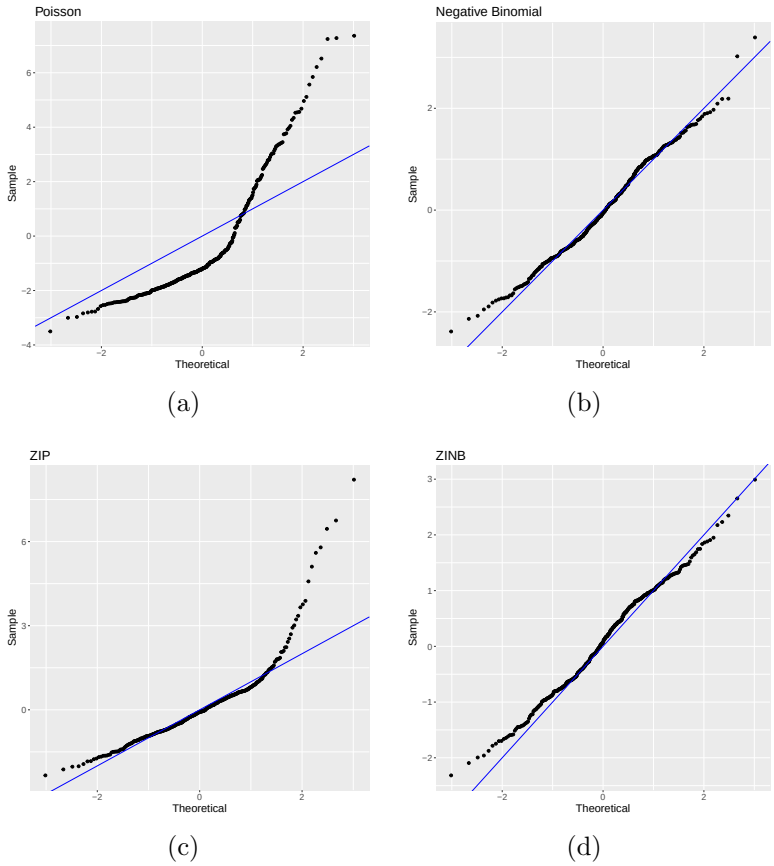

Figure 2: Q–Q plots of the randomized residuals for the fitted models: (a) Poisson regression, (b) negative binomial regression, (c) ZIP regression, and (d) ZINB regression. Better fits are indicated by the negative binomial regression and ZINB regression Q–Q plots.

sion parameter α is negative. When $\alpha = 0$, the above reduces to the Poisson pmf (equidispersion), while $\alpha > 0$ and $\alpha < 0$ represent count data with overdispersion and underdispersion, respectively.

Czado and Min (2005) and Famoye and Singh (2006) were the earliest works to study the ZI generalized Poisson (ZIGP) model, where the mean μ is related to the vector of covariates $\mathbf{x}$ through the log link function. The former also established the consistency and asymptotic normality of the MLEs for the parameters in the ZIGP regression model. Gupta et al. (2005) developed a score test to determine whether the ZIGP regression model is necessary over the ZIP or ZINB regression models. Czado

et al. (2007) provided an extension of the ZIGP regression model that allows the dispersion parameter to be related to a vector of covariates. Computational routines for this model were made available in the `R` package `ZIGP`, which is archived as of July 2017. Applications where the ZIGP regression model has been demonstrated to provide a better fit compared to the ZIP and ZINB regression models are data on domestic violence occurrences (Famoye and Singh, 2006), outsourcing of patent filing processes (Czado et al., 2007), and mapping quantitative trait loci (Cui and Yang, 2009).

The Conway-Maxwell-Poisson (CMP) distribution of Conway and Maxwell (1962) is another flexible distribution for count data expressing overdispersion or underdispersion (Sellers and Shmueli, 2013). This two-parameter distribution has pmf given by

$$p(y;\lambda,\nu) = \frac{\lambda^y}{(y!)^\nu} Z(\lambda,\nu), \quad \lambda > 0, \nu \geq 0, \tag{15}$$

where ν is a dispersion parameter and $Z(\lambda,\nu) = \sum_{j=0}^{\infty} \frac{\lambda^j}{(j!)^\nu}$ normalizes the distribution. Similar to the generalized Poisson pmf, when the dispersion parameter $\nu = 1$, (15) reduces to the Poisson pmf, while $\nu > 0$ and $\nu < 0$ characterize overdispersion and underdispersion, respectively. The flexibility with the CMP distribution is that it can capture two other classic discrete distributions, namely the geometric with success probability $(1-\lambda)$ when $\nu = 0$ and $\lambda < 1$, and the Bernoulli distribution with success probability $\lambda/(1+\lambda)$ when $\nu \to \infty$. Imoto (2014) later proposed a generalized CMP distribution that generalizes both the CMP distribution and the negative binomial distribution. Sellers and Shmueli (2010) first proposed a CMP regression model, where λ is related to a vector of covariates $\mathbf{x}$ using the log link function. Just like the CMP distribution generalizes several different discrete distributions, the CMP regression model generalizes both Poisson and logistic regression models.

Sellers and Raim (2016) introduced a ZICMP regression model when excess zeros are present in a CMP regression setting. The authors further allowed the dispersion parameter to be modeled as a function of covariates via a log link. The probability of observing a zero count, π, is again allowed to be modeled as a function of covariates

via a logit link. Just like the discussion of ZI count regression models in Section 2, the covariates used when modeling the parameters λ, ν, and π need not all be the same. Sellers and Raim (2016) also developed the LRT for the presence of significant data dispersion, derived the Fisher information matrix for computing the estimated parameter standard errors, and conducted a broad simulation study comparing the ZICMP regression model fit to other standard ZI count regression fits. The model was demonstrated to provide a nearly similar fit (in terms of its loglikelihood) relative to the ZINB and ZIG regression fits, thus indicating the ZICMP regression's ability to characterize data dispersion. The authors have also made available their functions related to this work in the `R` package `COMPoissonReg` (Sellers et al., 2017).

4 ZI Models for Longitudinal Data and Count Time Series

Longitudinal or panel study designs can also result in longitudinal or clustered ZI count data. As noted in Feng and Zhu (2011), ignoring the within-cluster correlation of longitudinal data will lead to loss of efficiency and incorrect inference of the regression coefficients. Most research in handling longitudinal ZI count data has been restricted to the ZIP regression setting. In particular, a marginal model and a conditional model for ZIP regression are two approaches commonly taken in the literature.

Hall and Zhang (2004) framed the approach for finding the MLEs in marginal ZIP regression models by using generalized estimating equations (GEEs). Following their discussion, let $\mathbf{y}_i \in \mathbb{R}^{n_i}$ be a vector of responses for the i^{th} cluster, $i = 1, \ldots, M$. In a marginal ZI count regression model, the random variable Y_{ij} associated with the observation y_{ij}, $j = 1, \ldots, n_i$, follows a ZI distribution as defined in Section 2, but where the count distribution must belong to the exponential dispersion family (Jørgensen, 1987). Let Z_{ij} be the indicator variable that Y_{ij} came from the degenerate distribution at 0. Under independence, the complete data loglikelihood based on

$\mathbf{y} = (\mathbf{y}_1^{\mathrm{T}}, \ldots, \mathbf{y}_M^{\mathrm{T}})^{\mathrm{T}}$ and $\mathbf{Z} = (z_{11}, \ldots, z_{Mn_M})^{\mathrm{T}}$ separates as follows:

$$\ell_c(\boldsymbol{\eta}(\boldsymbol{\beta}), \boldsymbol{\pi}(\boldsymbol{\alpha}), \phi; \mathbf{y}, \mathbf{x}, \mathbf{w}, \mathbf{Z}) = \ell_c(\boldsymbol{\pi}(\boldsymbol{\alpha}); \mathbf{y}, \mathbf{w}, \mathbf{Z}) + \ell_c(\boldsymbol{\eta}(\boldsymbol{\beta}), \phi; \mathbf{y}, \mathbf{x}, \mathbf{Z}),$$

where $\boldsymbol{\eta}(\boldsymbol{\beta})$ has been used to generically represent the conditional mean and we have replaced $\boldsymbol{\vartheta}$ in (3) with the univariate scale parameter ϕ as defined in the exponential dispersion family. Using an EM algorithm, at the $(t+1)^{\text{th}}$ iteration we maximize

$$\begin{aligned} \mathcal{Q}\left(\boldsymbol{\eta}(\boldsymbol{\beta}), \boldsymbol{\pi}(\boldsymbol{\alpha}), \phi | \boldsymbol{\eta}(\boldsymbol{\beta}^{(t)}), \boldsymbol{\pi}(\boldsymbol{\alpha}^{(t)}), \phi^{(t)}\right) = \ell_c\left(\boldsymbol{\pi}(\boldsymbol{\alpha}); \mathbf{y}, \mathbf{w}, \hat{\mathbf{Z}}^{(t)}\right) \\ + \ell_c\left(\boldsymbol{\eta}(\boldsymbol{\beta}), \phi; \mathbf{y}, \mathbf{x}, \hat{\mathbf{Z}}^{(t)}\right), \end{aligned} \tag{16}$$

where $\hat{\mathbf{Z}}^{(t)}$ is an estimate of the posterior membership probabilities calculated in the E-step. The M-step requires maximizing $\mathcal{Q}$ with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, and solving the following respective equations:

$$\sum_{i=1}^{M} \left\{ \frac{\partial \boldsymbol{\pi}_i(\boldsymbol{\alpha})^{\mathrm{T}}}{\partial \boldsymbol{\alpha}} \right\} \left[\left(\mathbf{A}_i(\boldsymbol{\pi}_i(\boldsymbol{\alpha}))\right)^{1/2} \mathbf{I}_{n_i} \left(\mathbf{A}_i(\boldsymbol{\pi}_i(\boldsymbol{\alpha}))\right)^{1/2} \right]^{-1} (\hat{\mathbf{Z}}_i^{(t)} - \boldsymbol{\pi}_i(\boldsymbol{\alpha})) = \mathbf{0} \tag{17}$$

$$\sum_{i=1}^{M} \left\{ \frac{\partial \boldsymbol{\eta}_i(\boldsymbol{\beta})^{\mathrm{T}}}{\partial \boldsymbol{\beta}} \right\} \left[\left(\mathbf{B}_i(\boldsymbol{\eta}_i(\boldsymbol{\beta}))\right)^{1/2} \mathbf{I}_{n_i} \left(\mathbf{B}_i(\boldsymbol{\eta}_i(\boldsymbol{\beta}))\right)^{1/2} \right]^{-1} \mathbf{W}_i^{(t)} (\mathbf{y}_i - \boldsymbol{\eta}_i(\boldsymbol{\beta})) = \mathbf{0}. \tag{18}$$

In the above, $\mathbf{I}_{n_i}$ is the $(n_i \times n_i)$ identity matrix, $\mathbf{A}_i(\boldsymbol{\pi}_i(\boldsymbol{\alpha})) = \mathrm{diag}(\pi_{i1}(\boldsymbol{\alpha})(1 - \pi_{i1}(\boldsymbol{\alpha})), \ldots, \pi_{in_i}(\boldsymbol{\alpha})(1 - \pi_{in_i}(\boldsymbol{\alpha})))$, $\mathbf{W}_i^{(t)} = \mathrm{diag}(1 - \hat{Z}_{i1}^{(t)}, \ldots, 1 - \hat{Z}_{in_i}^{(t)})$, and $\mathbf{B}_i(\boldsymbol{\eta}_i(\boldsymbol{\beta}))$ is an $(n_i \times n_i)$ diagonal matrix with entries composed of the conditional variance; see Hall and Zhang (2004) for how this last quantity is explicitly defined. In the above, the conditional mean η and mixing proportion π from Section 2 have been vectorized and written explicitly as functions of the parameters to be estimated; i.e., $\boldsymbol{\eta}_i(\boldsymbol{\beta})$ and $\boldsymbol{\pi}_i(\boldsymbol{\alpha})$. Then, the scale parameter ϕ needs to be estimated.

The formulas in (17) and (18) have the form of (weighted) GEEs with working correlation matrix equal to the identity matrix. Hall and Zhang (2004) and Dobbie and Welsh (2001) explore substituting the working correlation structures in the marginal model approach with something other than the identity matrix, such as

an exchangeable or AR(1) structure. To guard against correlation misspecification, Hall and Zhang (2004) advocate using the GEE-1 approach of Liang et al. (1992), which treats the first and second moment parameters orthogonally. Finally, Iddi and Molenberghs (2013) extended the framework of Hall and Zhang (2004) and presented a marginalized, ZI, overdispersed model for correlated data.

The basic framework of the conditional model approach is to use mixed effects models for $g(\eta)$ and $h(\pi)$. This approach was first considered in Hall (2000) for ZIP and ZIB regression with random intercepts, where the parameters were estimated using an EM algorithm. Wang et al. (2002) obtained the penalized likelihood function by treating the random effects as unknown parameters, and then using residual maximum likelihood (REML) for estimation. Min and Agresti (2005) and Lam et al. (2006) have also proposed random effects models to accommodate within-subject and between-subject heterogeneity in the presence of zero inflation. Alfò and Maruotti (2010) take a semiparametric approach to the model in Min and Agresti (2005) by relaxing the normality assumption of the random effects and leaving the corresponding distribution unspecified.

Count time series with extra zeros have also been explored in the literature. Let $(Y_t, \mathbf{X}_t, \mathbf{W}_t, N_t)$, $t = 1, \ldots, n$, be a sequence of random variables, where Y_t is the count response variable associated with exposure N_t, and $\mathbf{X}_t$ and $\mathbf{W}_t$ are vectors of measured covariates. The index t is used to indicate a time ordering. Wang (2003) was one of the first papers to consider this general setup, and developed a Markov ZIP regression model that allows for the frequency distribution to change according to the states of a two-state discrete time Markov chain with the transition probabilities associated by the covariates through a logit link function. This model was then used for analyzing the daily number of phone calls on a fault report. A similar Markov ZIP regression model was developed in Yang et al. (2013), who employed a partial likelihood for conducting statistical inference.

Yau et al. (2004) used the ZIP mixed regression model for correlated count data of Wang et al. (2002) to model the serial dependency between successive responses

Y_t and Y_{t+1}. The model was used to analyze workplace injuries from a participatory ergonomics intervention. Specifically, the serial dependency can be modeled explicitly via random effects attached to each time point as follows:

$$\log(\eta_t) = \log(N_t) + \mathbf{X}_t^{\mathrm{T}}\boldsymbol{\beta} + u_t.$$

For simplicity, the random component u_t is assumed to follow an AR(1) process with $\mathrm{Var}(u) = \sigma^2\mathbf{V}$, where

$$\mathbf{V} = \frac{1}{1-\rho^2}\begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-2} & \rho^{n-3} & \rho^{n-4} & \cdots & \rho \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{pmatrix}$$

such that ρ is the autocorrelation parameter. Zhao et al. (2009) developed a score test to assess if zero inflation is significant to warrant use of the above ZIP autoregression model.

Other frameworks for count time series with excess zeros have also been explored. For example, ZI integer-valued generalized autoregressive conditional heteroskedasticity (GARCH) models for the Poisson, negative binomial, and compound Poisson distributions have been developed in Zhu (2012) and Gonçalves et al. (2016). Yang et al. (2015) developed a state-space model that accommodates overdispersion, zero-inflation, and temporal correlation, similar to the setup of Iddi and Molenberghs (2013). For parameter estimation, Yang et al. (2015) proposed a Monte Carlo EM algorithm where particle filtering and particle smoothing methods are employed to approximate the high-dimensional integrals in the E-step. Computational routines for the work of Yang et al. (2013, 2015) are available in the `R` package `ZIM` (Yang et al., 2017).

5 Bayesian Approaches and Spatial Models

Bayesian approaches for analyzing ZI count regression models have received increasing attention in the literature. Dagne (2004) is one of the earliest papers where such a Bayesian analysis is performed. The paper presented a Bayesian hierarchical ZIP regression model that simultaneously models covariates and correlated count data. The approach was applied to count data on the efficacy of pesticides in controlling the reproduction of whiteflies. Ghosh et al. (2006) presented the ZI power series (ZIPS) regression model, which provides a generalized setting for the ZIP and ZINB regression models. To define the power series distribution, let $b_0, b_1, b_2, \ldots$ be a sequence of nonnegative real numbers. The partial sum of order $n \in \mathbb{N}$ is given by $g_n(\eta) = \sum_{k=0}^{n} b_k \eta^k,\ \eta \in \mathbb{R}$. The power series g is defined by $g(\eta) = \lim_{n\to\infty} g_n(\eta)$ and is denoted by $g(\eta) = \sum_{n=0}^{\infty} b_n \eta^n$. Letting $r \geq 0$ denote the radius of convergence of this series, the pmf of the power series distribution is given by

$$p(y;\eta) = \frac{b_y \eta^y}{g(\eta)}, \quad y \in \mathbb{N},\ 0 \leq \eta < r. \tag{19}$$

For Bayesian fitting of the ZIPS regression model, Ghosh et al. (2006) assume that the regression parameters $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are *a priori* independent and specify multivariate normal priors with the identity matrix as the variance-covariance matrix. The authors present their MCMC algorithm for generating samples from the respective full conditional distributions. They also provide their code, which is written in `WinBUGS` (Lunn et al., 2000). Jang et al. (2010) had a similar setup as Ghosh et al. (2006), but focused strictly on performing a Bayesian analysis of ZIP and ZINB regression models using a power prior as an informative prior. Their Bayesian approach was used to analyze data on road safety countermeasures. Klein et al. (2015) proposed a class of Bayesian GAMs for ZI count responses in the GAMLSS framework. Their approach was used to develop ZI count regression models for patent citations and claim frequencies from an automobile insurance company.

Bayesian approaches to test and construct influence diagnostics have also been

proposed in the literature. Bayarri et al. (2008) proposed a Bayes factor based on a suitable objective prior for testing a Poisson regression model versus a ZIP regression model. Bayesian inference involving specific ZI count regression models has also been treated in the literature, including approaches for ZINB regression (Garay et al., 2015), ZIGP regression (Xie et al., 2014), and ZICMP regression (Barriga and Louzada, 2014). In each paper, the authors present an MCMC sampler for the particular ZI model under consideration, followed by a discussion of Bayesian case influence diagnostics and relevant model selection criteria. For Bayesian influence diagnostics, the primary approach is based on *case-deletion*, where the impact of deleting an observation on the estimates is directly assessed by measures such as the likelihood distance and Cook's distance. Some of the model criteria discussed include the deviance information criterion (DIC; Spiegelhalter et al., 2002), the expected BIC (EBIC; Carlin and Louis, 2008), and the log pseudo-marginal likelihood (LPML) statistic. The LPML statistic requires calculation of the conditional predictive ordinate (CPO) statistic of Gelfand et al. (1992). A CPO can be calculated for each observation, which provides a measure of the height of the marginal density at the time of the event y_i. Since no closed-form solution for this quantity exists, a single MCMC sample from the posterior distribution can be used to provide a Monte Carlo approximation, $\widehat{\text{CPO}}_i$. Then, LPML $= \sum_{i=1}^{n} \log(\widehat{\text{CPO}}_i)$, such that larger values of LPML indicate a better fit.

Zero inflation in spatial count data has also been investigated in the literature, but such problems are primarily addressed using Bayesian hierarchical models. One of the earliest works taking this approach is due to Agarwal et al. (2002). The authors incorporated spatial association into the ZIP regression model through random effects in the GLM models in (1) and (2), thus yielding a hierarchical model that they estimate using a Bayesian framework. The approach is used to understand the living location choice of terrestrial isopods. Rathbun and Fei (2006) used a ZI framework to characterize the excess zeros that often occur outside the range of the distribution of a given species. They determined the species range using a spatial probit model in which

they include physical variables as covariates. The elicitation of priors had not been studied for such an application, so the authors adopted the g-prior of Zellner (1986) to obtain a separable prior structure on the entire parameter vector of interest. The approach was used to understand regeneration of oak trees in central Pennsylvania. Finally, Bayesian approaches that include both spatial and temporal random effects (i.e., spatial-temporal effects) have also been explored in the literature; cf. Musenge et al. (2013), Wang et al. (2015), and Arab (2015).

6 Zero-Inflation and Diagonal-Inflation in Multivariate Count Responses

Multivariate ZI models have been treated far less in the literature compared to their univariate counterparts, especially in the presence of covariates. The practical implication of multivariate ZI count regression models is that they foster descriptions of how a vector of correlated ZI count variables respond simultaneously to changes in measured covariates. Some applications of multivariate ZI regression models include development of a bivariate ZIP regression model for analyzing two types of occupational injuries (musculoskeletal and non-musculoskeletal) at a teaching hospital during different intervention trial time periods (Wang et al., 2003), a semiparametric bivariate ZIP regression model for analyzing two populations of fish (common carp and channel catfish) as a function of various environmental variables (Arab et al., 2012), and a bivariate ZINB regression model (Wang, 2003) and a bivariate ZIGP regression model (Faroughi and Ismail, 2017) for analyzing healthcare utilization (doctor and non-doctor health professional visits) as a function of various socio-economic variables.

It is illustrative to first define the bivariate Poisson distribution, which simplifies the necessary notation. The bivariate Poisson distribution utilizes a general multivariate reduction scheme. Let R_j be independent Poisson random variables with rates λ_j, $j = 1, 2, 3$. Then $\mathbf{Y} = (Y_1, Y_2)^{\mathrm{T}} = (R_1 + R_3, R_2 + R_3)^{\mathrm{T}}$ is bivariate Poisson

distributed with rates λ_1, λ_2, and λ_3 and has pmf

$$p(\mathbf{y};\lambda_1,\lambda_2,\lambda_3) = e^{-\sum_{j=1}^{3}\lambda_j}\frac{\lambda_1^{y_1}\lambda_2^{y_2}}{y_1!y_2!}\sum_{t=0}^{y_1\wedge y_2} t!\binom{y_1}{t}\binom{y_2}{t}\left(\frac{\lambda_3}{\lambda_1\lambda_2}\right)^t, \tag{20}$$

where the mean vector and covariance matrix are, respectively,

$$\mathrm{E}[\mathbf{Y}] = \begin{pmatrix} \lambda_1+\lambda_3 \\ \lambda_2+\lambda_3 \end{pmatrix} \quad \text{and} \quad \mathrm{Var}[\mathbf{Y}] = \begin{pmatrix} \lambda_1+\lambda_3 & \lambda_3 \\ \lambda_3 & \lambda_2+\lambda_3 \end{pmatrix}. \tag{21}$$

For dimensions greater than 2, calculating the pmf becomes challenging, but can be accomplished using a recursive scheme (Karlis and Meligkotsidou, 2005).

For bivariate Poisson regression, we incorporate the predictors $\mathbf{x}_1, \mathbf{x}_2$, and $\mathbf{x}_3$ through the relationships

$$\log(\lambda_j) = \mathbf{x}_j^{\mathrm{T}}\boldsymbol{\beta}_j + \log(N_j), \quad j = 1,2,3, \tag{22}$$

where N_j is an (optional) exposure term. Note that the predictor variables used to model each λ_j need not be the same, however, models with a constant λ_3 are typically used because they are easier to interpret (Karlis and Ntzoufras, 2005). We then write the version of the pmf in (20) with predictors as $p(\mathbf{y};\mathbf{x},\boldsymbol{\beta})$, where $\mathbf{x}$ and $\boldsymbol{\beta}$ are used to represent all observed predictor variables and regression coefficients that appear through the linear predictors in (22).

The support of $\mathbf{Y}$ is $\mathbb{N}^2$. Suppose that the maximum observed outcomes for these variables are m_1 and m_2, respectively, where m_1 need not equal m_2. Consider the following matrix of observed counts for Y_1 and Y_2:

$$
\begin{array}{cc} & Y_2 \\ \end{array}
$$

$$
\begin{array}{cc}
 & \begin{array}{ccccc} 0 & 1 & 2 & \cdots & m_2 \end{array} \\
Y_1 & \begin{pmatrix}
0 & c_{00} & c_{01} & c_{02} & \cdots & c_{0m_2} \\
1 & c_{10} & c_{11} & c_{12} & \cdots & c_{1m_2} \\
2 & c_{20} & c_{21} & c_{22} & \cdots & c_{2m_2} \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
m_1 & c_{m_1 0} & c_{m_1 1} & c_{m_1 2} & \cdots & c_{m_1 m_2}
\end{pmatrix}
\end{array} \tag{23}
$$

Inflation in multivariate count data is handled in two primary ways. First, the most common setting is straightforward zero inflation. This amounts to an inflated count of the $(0, 0)$ cell, c_{00}. Under this approach, a ZI bivariate Poisson regression model is given by

$$
f(\mathbf{y}; \mathbf{x}, \boldsymbol{\beta}, \mathbf{w}, \boldsymbol{\alpha}) = q(0; \mathbf{w}, \boldsymbol{\alpha})\mathtt{I}\{y_1 = y_0 = 0\} + (1 - q(0; \mathbf{w}, \boldsymbol{\alpha}))p(\mathbf{y}; \mathbf{x}, \boldsymbol{\beta}), \tag{24}
$$

where the probability of observing $(0, 0)$ is, again, modeled by a binomial GLM using the linear predictor $\mathbf{w}^{\mathrm{T}}\boldsymbol{\alpha}$.

The second setting is *diagonal-inflation*, where all of the counts on the diagonal, $\{c_{00}, c_{11}, c_{22}, \ldots\}$, are inflated. The diagonally-inflated Poisson (DIP) regression model is given by

$$
\begin{aligned}
f(\mathbf{y}; \mathbf{x}, \boldsymbol{\beta}, \mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\vartheta}) = {} & q(y_1; \mathbf{w}, \boldsymbol{\alpha}) p_D(y_1; \boldsymbol{\vartheta})\mathtt{I}\{y_1 = y_2\} \\
& + (1 - q(y_1; \mathbf{w}, \boldsymbol{\alpha}))p(\mathbf{y}; \mathbf{x}, \boldsymbol{\beta}),
\end{aligned} \tag{25}
$$

where $p_D(y_1; \boldsymbol{\vartheta})$ is the pmf of a discrete distribution defined on $\mathbb{N}$. Note that p_D is only a function of y_1 since it is only computed when $y_1 = y_2$. Similarly, this impacts how the probability of observing (j, j), $j \in \mathbb{N}$, is modeled, which results in $q(0; \mathbf{w}, \boldsymbol{\alpha})$ being replaced by $q(y_1; \mathbf{w}, \boldsymbol{\alpha})$. DIP regression models have been used to model pre- and post-treatment studies, where the treatment may not have an effect on some patients for an unknown reason, and the number of draws that result in various sports games; see Karlis and Ntzoufras (2003, 2005).

We are only aware of one computational package relevant to the multivariate models discussed in this section. The `R` package `bivpois` (Karlis and Ntzoufras, 2005) was available for estimating bivariate ZIP and DIP regression models, however, this package is archived as of July 2017. Young et al. (2017) noted that they attempted to use the `bivpois` package for their application, but that there were various numerical issues that arose (e.g., only a small sample size could be used to coax the EM algorithm to converge).

7 Related Models

There are various models available for handling other issues with zero counts in an observed dataset. Such models are often discussed along with ZI count regression models. We briefly highlight some of these models in this section.

In contrast to ZI count regression models, which are two-component mixture models, *hurdle regression models* are two-part models where it is assumed that the positive counts are generated from a different process than the zero counts. Using the same notation as in (3), a hurdle regression pmf combines a count data model left-truncated at 1 with a zero hurdle model right-censored at 1 as follows:

$$\begin{aligned} f_{\mathrm{H}}(y; \mathbf{x}, \mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\vartheta}) = q(0; \mathbf{w}, \boldsymbol{\alpha})\mathtt{I}\{y = 0\} \\ + (1 - q(0; \mathbf{w}, \boldsymbol{\alpha}))\frac{p(y; \mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\vartheta})}{1 - p(y; \mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\vartheta})}\mathtt{I}\{y > 0\}. \end{aligned} \tag{26}$$

Hurdle models were first proposed by Mullahy (1986) to analyze survey data on beverage consumption. The `pscl` package and `PROC NLMIXED` can both be used to estimate hurdle regression models.

A large number of zeros can also be present in continuous data, but the probability of yielding a zero under continuous distributions is 0. This setting can often be characterized with a *semi-continuous variable*, which has a portion of responses equal to a single value (typically 0) and a continuous, often right-skewed distribution, for the

remaining values (Olsen and Schafer, 2001). As noted in Hall and Zhang (2004), "for independent semi-continuous data, there is little motivation for a model such as a ZI normal, because all observed zeros are unambiguous; they necessarily come from the degenerate distribution, rather than from the nondegenerate continuous distribution." Thus, the likelihood for such a model factors into terms for the zero and non-zero data, similar to a hurdle regression model. See Mills (2013) for a discussion of ZI gamma regression and ZI lognormal regression models, who also used those models to analyze data involving Parkinson's disease and driving capabilities.

Related to the ZI regression models for semi-continuous data is the zero-one-inflated beta (ZOIB) regression model, which was introduced by Ospina and Ferrari (2012). ZOIB regression can be used to model proportions with a high amount of observed zero and one proportions. For example, Wieczorek and Hawala (2011) took a Bayesian approach to estimate the parameters of a ZOIB regression model for modeling US county poverty rates, which yielded comparable results to the US Census Bureau's current small-area model for county poverty estimation. Nishii and Tanaka (2013) developed a ZOIB regression model to analyze grid-cell data of a forest coverage ratio as a function of two covariates. The `zoib` package (Liu and Kong, 2015) in `R` can perform Bayesian estimation and inference for ZOIB regression models.

The term *zero-altered models* is found in the literature, and may refer to either hurdle models or, more generally, any model that reflects some secondary behavior of zero counts. Besides the models we discussed here, one could also have *zero-truncated models*, where the value of zero cannot occur, as well as *zero-deflated models*, where the mixing proportion in the ZI model is allowed to be negative and, hence, is no longer a true mixture distribution. More discussion on the differences between these different types of zero-altered models can be found in Chapter 11 of Hilbe (2011) and Chapter 17 of Yee (2015).

8 Concluding Remarks and Future Research

The seminal paper of Lambert (1992) introduced the ZIP regression model and provided details about likelihood estimation of the parameters. Many advancements with ZI count regression models have been made in the 25 years since the publication of that paper. In our paper, we have provided a thorough review of the major advancements, while simultaneously highlighting important applied data problems addressed by such models as well as available software. We further provided a contemporary perspective on ZI count regression models by addressing strategies and methodology employed for modern complex data problems, such as Bayesian approaches to ZI count modeling, spatial analysis in the presence of ZI data, and notions of inflation in multivariate count responses. We conclude by listing some interesting directions for research about ZI count regression models.

- Zhang et al. (2011) presented a novel measurement error (ME) model that accounts for both MEs in the covariates and zero inflation for estimating the distributions of dietary intakes. The authors performed extensive empirical work and demonstrated the efficacy of this model in relating multiple dietary components and patterns with health outcomes. However, research needs to be performed in the context of ZI count regression models under different ME structures. For example, suppose that we are interested in relating our ZI count Y to a vector of covariates, $\mathbf{X}$; however, $\mathbf{X}$ cannot be observed in practice. Instead, we observe $\mathbf{V}$. *Classical ME* is where $\mathbf{V} = \mathbf{X} + \mathbf{U}$ such that each variable in the vector $\mathbf{U}$ is normally distributed with mean 0 and constant variance. *Berkson ME* is where $\mathbf{X} = \mathbf{V} + \mathbf{U}$ (additive) or $\mathbf{X} = \mathbf{V}\mathbf{U}$ (multiplicative), such that $\mathbf{V}$ and $\mathbf{U}$ are independent and, respectively, $\mathrm{E}(\mathbf{U}) = \mathbf{0}$ or $\mathrm{E}(\mathbf{U}) = \mathbf{1}$. Thus, $\mathrm{E}(\mathbf{X}|\mathbf{V}) = \mathbf{V}$. Estimators and their properties for ZI count regression models need to be studied, as well as when ME occurs in the covariates for modeling the mixing proportion π. Some of the work of Guo and Li (2002), who discussed ME in the context of non-ZI Poisson regression, could be leveraged for

this research.

- Big data problems are of broad and current interest to researchers and data analysts. Zero inflation can also occur in such big data problems, as highlighted with the census application in Young et al. (2017). One issue highlighted by the authors is the need for efficient computing routines when estimating ZI models applied to big data. In particular, routines are necessary to handle ultrahigh dimensional variable selection in ZI count regression models. Such routines could be developed in the spirit of the *iteratively sure independent screening* approach of Fan et al. (2009). Perhaps even more beneficial will be including these computational routines in a statistical package devoted to modeling and inference tools for ZI count regression models. In Section 2.2 and throughout the subsequent sections, we highlighted major routines available in statistical software packages. However, most of these simply estimate ZIP and ZINB regression models, with options for obtaining simple residual summaries. A package that encompasses many of the modern methods that we discussed, including routines for big data problems, will make an invaluable contribution.

- In Section 5, we noted some novel Bayesian hierarchical models that have been developed for ZI counts in spatial data. One specific type of spatial data is *areal data*, which is aggregated quantities for each measured (areal) unit within some meaningful partition of a given region, such as counties within a state. A growing research topic is developing efficacious spatial regression models that capture not only zero inflation, but more generally characterize data dispersion for areal count data. Such models could better address problems related to the spread of diseases (Gschlößl and Czado, 2008), trends in emergency department visits (Neelon et al., 2013, 2016), and changes in the status of housing units for conducting censuses (Musgrove et al., 2017). One alternative to the models proposed for these applied problems is development of a spatial CMP regression model, which could provide a flexible framework for capturing the data

dispersion.

- In Section 6, we discussed the notion of zero inflation and diagonal inflation in multivariate count regression models, with an emphasis on multivariate Poisson regression models. We noted some applied work where zero inflation has been investigated for other multivariate count regression models. However, there is a need for a more rigorous development and treatment of ZI and DI count regression models beyond the multivariate Poisson regression setting. More generally, it would be beneficial to develop a unified framework about zero inflation and diagonal inflation in multivariate count regression models, regardless of the assumed count distribution. Such work could further inform more complex data structures, such as ZI counts in tensor regression. Zhou et al. (2013) developed an effective framework for tensor regression models that allows for discrete responses. However, the notion of zero inflation has, to our knowledge, not been investigated.

Finally, we hope that our review has contributed an insightful perspective on ZI count regression model, and that it will stimulate additional interest in this very important class of statistical models.

SUPPLEMENTARY MATERIAL

Supplemental File: A file with the details of the ECM algorithm for estimating a ZINB regression model and full results for the timing study discussed in Section 2.2. (.pdf file)

Code: All `R` and `SAS` code used for the numerical portions in Section 2. (.txt file)

References

D. K. Agarwal, A. E. Gelfand, and S. Citron-Pousty. Zero-Inflated Models with Application to Spatial Count Data. *Environmental and Ecological Statistics*, 9(4): 409–426, 2002.

J. M. 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*, 23(3):257–278, 2014.

M. Alfò and A. Maruotti. Two-Part Regression Models for Longitudinal Zero-Inflated Count Data. *The Canadian Journal of Statistics*, 38(2):197–216, 2010.

A. Arab. Spatial and Spatio-Temporal Models for Modeling Epidemiological Data with Excess Zeros. *International Journal of Environmental Research and Public Health*, 12(9):10536–10548, 2015.

A. Arab, S. H. Holan, C. K. Wikle, and M. L. Wildhaber. Semiparametric Bivariate Zero-Inflated Poisson Models with Application to Studies of Abundance for Multiple Species. *Environmetrics*, 23(2):183–196, 2012.

G. Baetschmann and R. Winkelmann. Modelling Zero-Inflated Count Data when Exposure Varies: With an Application to Sick Leave. Technical Report 61, Department of Economics, University of Zurich, 2012.

G. D. C. Barriga and F. Louzada. The Zero-Inflated Conway-Maxwell-Poisson Distribution: Bayesian Inference, Regression Modeling and Influence Diagnostic. *Statistical Methodology*, 21:23–34, 2014.

M. J. Bayarri, J. O. Berger, and G. S. Datta. Objective Bayes Testing of Poisson Versus Inflated Poisson Models. In *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh*, volume 3 of *B. Clarke and S. Ghosal (eds.) IMS Collections*, pages 105–121. Institute of Mathematical Statistics, Beachwood, OH, 2008.

E. L. Boone, B. Stewart-Koster, and M. J. Kennard. A Hierarchical Zero-Inflated Poisson Regression Model for Stream Fish Distribution and Abundance. *Environmetrics*, 23(3):207–218, 2012.

J.-P. Boucher, M. Guillén, and M. Denuit. Risk Classification for Claim Counts: A Comparative Analysis of Various Zero-Inflated Mixed Poisson and Hurdle Models. *North American Actuarial Journal*, 11(4):110–131, 2007.

B. P. Carlin and T. A. Louis. *Bayesian Methods for Data Analysis*. Chapman & Hall/CRC Press, Boca Raton, FL, 2nd edition, 2008.

M. Chiogna and C. Gaetan. Semiparametric Zero-Inflated Poisson Models with Application to Animal Abundance Studies. *Environmetrics*, 18(3):303–314, 2007.

A. C. Cohen, Jr. Estimating the Parameter in a Conditional Poisson Distribution. *Biometrics*, 16(2):203–211, 1960.

P. C. Consul. *Generalized Poisson Distributions: Properties and Applications*. Marcel Dekker, New York, NY, 1989.

P. C. Consul and G. C. Jain. A Generalization of the Poisson Distribution. *Technometrics*, 15(4):791–799, 1973.

R. W. Conway and W. L. Maxwell. A Queuing Model with State Dependent Service Rates. *Journal of Industrial Engineering*, 12(2):132–136, 1962.

R. D. Cook. Assessment of Local Influence. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 48(2):133–169, 1986.

Y. Cui and W. Yang. Zero-Inflated Generalized Poisson Regression Mixture Model for Mapping Quantitative Trait Loci Underyling Count Trait with Many Zeros. *Journal of Theoretical Biology*, 256(2):276–285, 2009.

W. R. Cupach and B. H. Spitzberg. *The Dark Side of Relationship Pursuit. From Attraction to Obsession and Stalking*. Lawrence Erlbaum Associates, Mahwah, NJ, 2004.

C. Czado and A. Min. Consistency and Asymptotic Normality of the Maximum Likelihood Estimator in a Zero-Inflated Generalized Poisson Regression. Technical report, Discussion Paper No. 423, Sonderforschungsbereich 386 der Ludwig-Maximilians-Universität München, 2005. http://nbn-resolving.de/urn:nbn:de:bvb:19-epub-1792-8.

C. Czado, V. Erhardt, A. Min, and S. Wagner. Zero-Inflated Generalized Poisson Models with Regression Effects on the Mean, Dispersion and Zero-Inflation Level Applied to Patent Outsourcing Rates. *Statistical Modelling*, 7(2):125–153, 2007.

G. A. Dagne. Hierarchical Bayesian Analysis of Correlated Zero-Inflated Count Data. *Biometrical Journal*, 46(6):653–663, 2004.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 39(1):1–38, 1977.

D. Deng and S. R. Paul. Score Tests for Zero-Inflation and Over-Dispersion in Generalized Linear Models. *Statistica Sinica*, 15(1):257–276, 2005.

M. J. Dobbie and A. H. Welsh. Modelling Correlated Zero-Inflated Count Data. *Australian and New Zealand Journal of Statistics*, 43(4):431–444, 2001.

P. K. Dunn and G. K. Smyth. Randomized Quantile Residuals. *Journal of Computational and Graphical Statistics*, 5(3):236–244, 1996.

F. Famoye. Restricted Generalized Poisson Regression Model. *Communications in Statistics - Theory and Methods*, 22(5):1335–1354, 1993.

F. Famoye and K. P. Singh. Zero-Inflated Generalized Poisson Regression Model with an Application to Domestic Violence Data. *Journal of Data Science*, 4:117–130, 2006.

J. Fan, R. Samworth, and Y. Wu. Ultrahigh Dimensional Feature Selection: Beyond The Linear Model. *Journal of Machine Learning Research*, 10:2013–2038, 2009.

P. Faroughi and N. Ismail. Bivariate Zero-Inflated Generalized Poisson Regression Model with Flexible Covariance. *Communications in Statistics - Theory and Methods*, 46(15):7769–7785, 2017.

J. Feng and Z. Zhu. Semiparametric Analysis of Longitudinal Zero-Inflated Count Data. *Journal of Multivariate Analysis*, 102(1):61–72, 2011.

A. M. Garay, E. M. Hashimoto, E. M. M. Ortega, and V. H. Lachos. On Estimation and Influence Diagnostics for Zero-Inflated Negative Binomial Regression Models. *Computational Statistics and Data Analysis*, 55(3):1304–1318, 2011.

A. M. Garay, V. H. Lachos, and H. Bolfarine. Bayesian Estimation and Case Influence Diagnostics for the Zero-Inflated Negative Binomial Regression Model. *Journal of Applied Statistics*, 42(6):1148–1165, 2015.

A. E. Gelfand, D. Dey, and H. Chang. Model Determination Using Predictive Distributions with Implementation via Sampling-Based Methods. In *Bayesian Statistics 4: Proceedings of the Fourth Valencia International Meeting*, J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (eds.), pages 147–167. Oxford University Press, Oxford, UK, 1992.

S. K. Ghosh, P. Mukhopadhyay, and J.-C. Lu. Bayesian Analysis of Zero-Inflated Regression Models. *Journal of Statistical Planning and Inference*, 136(4):1360–1375, 2006.

E. Gonçalves, N. Mendes-Lopes, and F. Silva. Zero-Inflated Compound Poisson Distributions in Integer-Valued GARCH Models. *Statistics: A Journal of Theoretical and Applied Statistics*, 50(3):558–578, 2016.

W. H. Greene. Fixed and Random Effects Models for Count Data. *SSRN eLibrary*, 2007.

S. Gschlößl and C. Czado. Modelling Count Data with Overdispersion and Spatial Effects. *Statistical Papers*, 49(3):531–552, 2008.

J. Q. Guo and T. Li. Poisson Regression Models with Errors-in-Variables: Implications and Treatment. *Journal of Statistical Planning and Inference*, 104(2):391–401, 2002.

P. L. Gupta, R. C. Gupta, and R. C. Tripathi. Score Test for Zero Inflated Generalized Poisson Regression Model. *Communications in Statistics - Theory and Methods*, 33(1):47–64, 2005.

D. B. Hall. Zero-Inflated Poisson and Binomial Regression with Random Effects: A Case Study. *Biometrics*, 56(4):1030–1039, 2000.

D. B. Hall and Z. Zhang. Marginal Models for Zero Inflated Clustered Data. *Statistical Modelling*, 4(3):161–180, 2004.

X. He, H. Xue, and N.-Z. Shi. Sieve Maximum Likelihood Estimation for Doubly Semiparametric Zero-Infalted Poisson Models. *Journal of Multivariate Analysis*, 101(9):2026–2038, 2010.

D. C. Heilbron. Zero-Altered and Other Regression Models for Count Data with Added Zeros. *Biometrical Journal*, 36(5):531–547, 1994.

J. M. Hilbe. *Negative Binomial Regression*. Cambridge University Press, Cambridge, UK, 2nd edition, 2011.

S. Iddi and G. Molenberghs. A Marginalized Model for Zero-Inflated, Overdispersed and Correlated Count Data. *Electronic Journal of Applied Statistical Analysis*, 6 (2):149–165, 2013.

T. Imoto. A Generalized ConwayMaxwellPoisson Distribution which Includes the Negative Binomial Distribution. *Applied Mathematics and Computation*, 247:824–834, 2014.

N. Janaskul and J. P. Hinde. Score Tests for Zero-Inflated Poisson Models. *Computational Statistics and Data Analysis*, 40(1):75–96, 2002.

N. Janaskul and J. P. Hinde. Score Tests for Extra-Zero Models in Zero-Inflated Negative Binomial Models. *Communications in Statistics - Simulation and Computation*, 38(1):92–108, 2008.

H. Jang, S. Lee, and S. W. Kim. Bayesian Analysis for Zero-Inflated Regression Models with the Power Prior: Applications to Road Safety Countermeasures. *Accident Analysis and Prevention*, 42(2):540–547, 2010.

B. Jørgensen. Exponential Dispersion Models (with Discussion). *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 49(2):127–162, 1987.

D. Karlis and L. Meligkotsidou. Multivariate Poisson Regression with Covariance Structure. *Statistics and Computing*, 15(4):255–265, 2005.

D. Karlis and I. Ntzoufras. Analysis of Sports Data by Using Bivariate Poisson Models. *Journal of the Royal Statistical Society, Series D (The Statistician)*, 52 (3):381–393, 2003.

D. Karlis and I. Ntzoufras. Bivariate Poisson and Diagonal Inflated Bivariate Poisson Regression Models in R. *Journal of Statistical Software*, 14(10):1–36, 2005.

N. Klein, T. Kneib, and S. Lang. Bayesian Generalized Additive Models for Location, Scale, and Shape for Zero-Inflated and Overdispersed Count Data. *Journal of the American Statistical Association*, 110(509):405–419, 2015.

K. F. Lam, H. Xue, and Y. B. Cheung. Semiparametric Analysis of Zero-Inflated Count Data. *Biometrics*, 62(4):996–1003, 2006.

D. Lambert. Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. *Technometrics*, 34(1):1–14, 1992.

A. H. Lee, K. Wang, J. A. Scott, K. K. W. Yau, and G. J. McLachlan. Multi-Level Zero-Inflated Poisson Regression Modelling of Correlated Count Data with Excess Zeros. *Statistical Methods in Medical Research*, 15(1):47–61, 2006.

K.-Y. Liang, S. L. Zeger, and B. Qaqish. Multivariate Regression Analyses for Categorical Data. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 54(1):3–40, 1992.

B. G. Lindsay. *Mixture Models: Theory, Geometry and Applications*, volume 5 of *NSF-CBMS Regional Conference Series in Probability and Statistics*. Institute of Mathematical Statistics and the American Statistical Association, 1995.

J. A. List. Determinants of Securing Academic Interviews After Tenure Denial: Evidence from a Zero-Inflated Poisson Model. *Applied Economics*, 33(11):1423–1431, 2001.

F. Liu and Y. Kong. zoib: An R Package for Bayesian Inference for Beta Regression and Zero/One Inflated Beta Regression. *The R Journal*, 7(2):34–51, 2015.

H. Liu and K.-S. Chan. Generalized Additive Models for Zero-Inflated Data with Partial Constraints. *Scandinavian Journal of Statistics*, 38(4):650–665, 2011.

X. Liu, B. Winter, L. Tang, B. Zhang, Z. Zhang, and H. Zhang. Simulating Comparisons of Different Computing Algorithms Fitting Zero-Inflated Poisson Models for Zero Abundant Counts. *Journal of Statistical Computation and Simulation (in press)*, 2017.

T. Loeys, B. Moerkerke, O. De Smet, and A. Buysse. The Analysis of Zero-Inflated Count Data: Beyond Zero-Inflated Poisson Regression. *British Journal of Mathematical and Statistical Psychology*, 65(1):163–180, 2012.

D. J. Lunn, A. Thomas, N. Best, and D. Spiegelhalter. WinBUGS - A Bayesian Modelling Framework: Concepts, Structure, and Extensibility. *Statistics and Computing*, 10(4):325–337, 2000.

J. Martin and D. B. Hall. R^2 Measures for Zero-Inflated Regression Models for Count Data with Excess Zeros. *Journal of Statistical Computation and Simulation*, 86(18): 3777–3790, 2016.

T. G. Martin, B. A. Wintle, J. R. Rhodes, P. M. Kunhert, S. A. Field, S. J. Low-Choy, A. J. Tyre, and H. P. Possingham. Zero Tolerance Ecology: Improving Ecological Inference by Modelling the Source of Zero Observations. *Ecology Letters*, 8(11): 1235–1246, 2005.

G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley, New York, 1997.

X.-L. Meng and D. B. Rubin. Maximum Likelihood Estimation Via the ECM Algorithm: A General Framework. *Biometrika*, 80(2):267–278, 1993.

S.-P. Miaou. The Relationship Between Truck Accidents and Geometric Design of Road Sections: Poisson Versus Negative Binomial Regressions. *Accident Analysis and Prevention*, 26(4):471–482, 1994.

E. D. Mills. *Adjusting for Covariates in Zero-Inflated Gamma and Zero-Inflated Log-Normal Models for Semicontinuous Data*. PhD thesis, University of Iowa, 2013.

Y. Min and A. Agresti. Random Effect Models for Repeated Measures of Zero-Inflated Count Data. *Statistical Modelling*, 5(1):1–19, 2005.

M. Minami, C. E. Lennert-Cody, W. Gao, and M. Román-Verdesoto. Modeling Shark Bycatch: The Zero-Inflated Negative Binomial Regression Model with Smoothing. *Fisheries Research*, 84(2):210–221, 2007.

M. Mittlböck and T. Waldhör. Adjustments for R^2-Measures for Poisson Regression Models. *Computational Statistics and Data Analysis*, 34(4):461–472, 2000.

J. Mullahy. Specification and Testing of Some Modified Count Data Models. *Journal of Econometrics*, 33(3):341–365, 1986.

E. Musenge, T. F. Chirwa, K. Kahn, and P. Vounatsou. Bayesian Analysis of Zero Inflated Spatiotemporal HIV/TB Child Mortality Data Through the INLA and SPDE Approaches: Applied to Data Observed Between 1992 and 2010 in Rural North East South Africa. *International Journal of Applied Earth Observation and Geoinformation*, 22:86–98, 2013.

D. Musgrove, D. S. Young, J. Hughes, and L. E. Eberly. A Sparse Areal Mixed Model for Multivariate Outcomes, with an Application to Zero-Inflated Census Data. *Submitted*, 2017.

L. K. Muthén and B. O. Muthén. *Mplus Users Guide, 7^{th} Edition*. Muthén and Muthén, Los Angeles, CA, 1998–2012.

NCSS, LLC. *NCSS 11 Statistical Software*. Kaysville, UT, 2016.

B. Neelon, P. Ghosh, and P. F. Loebs. A Spatial Poisson Hurdle Model for Exploring Geographic Variation in Emergency Department Visits. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, 176(2):389–413, 2013.

B. Neelon, H. H. Chang, Q. Liang, and N. S. Hastings. Spatiotemporal Hurdle Models for Zero-Inflated Count Data: Exploring Trends in Emergency Department Visits. *Statistical Methods in Medical Research*, 25(6):2558–2576, 2016.

R. Nishii and S. Tanaka. Modeling and Inference of Forest Coverage Ratio Using Zero-One Inflated Distributions with Spatial Dependence. *Enviornmental and Ecological Statistics*, 20(2):315–336, 2013.

M. K. Olsen and J. L. Schafer. A Two-Part Random-Effects Model for Semicontinuous Longitudinal Data. *Journal of the American Statistical Association*, 96(454):730–745, 2001.

R. Ospina and S. L. P. Ferrari. A General Class of Zero-or-One Inflated Beta Regression Models. *Computational Statistics and Data Analysis*, 56(6):1609–1623, 2012.

M. Pandya, H. Pandya, and S. Pandya. Bayesian Inference on Mixture of Geometric with Degenerate Distribution: Zero Inflated Geometric Distribution. *International Journal of Research and Reviews in Applied Sciences*, 13(1):53–66, 2012.

J. M. Potts and J. Elith. Comparing Species Abundance Models. *Ecological Modelling*, 199:153–163, 2006.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL `https://www.R-project.org/`.

S. L. Rathbun and S. Fei. A Spatial Zero-Inflated Poisson Regression Model for Oak Regeneration. *Environmental and Ecological Statistics*, 13(4):409–426, 2006.

M. Ridout, J. Hinde, and C. G. B. Demétrio. A Score Test for Testing a Zero-Inflated Poisson Regression Model Against Zero-Inflated Negative Binomial Alternatives. *Biometrics*, 57(1):219–223, 2001.

R. A. Rigby and D. M. Stasinopoulos. Generalized Additive Models for Location, Scale and Shape (with Discussion). *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 54(3):507–554, 2005.

L. S. Sarul and S. Sahin. An Application of Claim Frequency Data Using Zero Inflated and Hurdle Models in General Insurance. *Journal of Business, Economics and Finance*, 4(4):732–743, 2015.

SAS Institute Inc. *SAS/STAT® 9.4 User's Guide*. SAS Institute Inc., Cary, NC, 2013.

K. F. Sellers and A. Raim. A Flexible Zero-Inflated Model to Address Data Dispersion. *Computational Statistics and Data Analysis*, 99:68–80, 2016.

K. F. Sellers and G. Shmueli. A Flexible Regression Model for Count Data. *The Annals of Applied Statistics*, 4(2):943–961, 2010.

K. F. Sellers, T. Lotze, and A. Raim. *COMPoissonReg: Conway-Maxwell Poisson (COM-Poisson) Regression*, 2017. URL `https://CRAN.R-project.org/package=COMPoissonReg`. R package version 0.4.0.

K.F. Sellers and G. Shmueli. Data Dispersion: Now You See It... Now You Don't. *Communications in Statistics - Theory and Methods*, 42(17):3134–3147, 2013.

M-L. Sheu, T.-W. Hu, T. E. Keeler, M. Ong, and H.-Y. Sung. The Effect of a Major Cigarette Price Change on Smoking Behavior in California: A Zero-Inflated Negative Binomial Model. *Health Economics*, 13(8):781–791, 2004.

D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. van der Linde. Bayesian Measures of Model Complexity and Fit. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 64(4):583–639, 2002.

D. M. Stasinopoulos and R. A. Rigby. Generalized Additive Models for Location Scale and Shape (GAMLSS) in R. *Journal of Statistical Software*, 23(7):1–46, 2007. http://www.jstatsoft.org/v23/i07/.

Stata Technical Support. *Stata Statistical Software: Release 14*. StataCorp LP, College Station, TX, 2015.

Y. Tang, L. Xiang, and Z. Zhu. Risk Factor Selection in Rate Making: EM Adaptive LASSO for Zero-Inflated Poisson Regression Models. *Risk Analysis*, 34(6):1112–1127, 2014.

J. van den Broek. A Score Test for Zero Inflation in a Poisson Distribution. *Biometrics*, 51(2):738–743, 1995.

Q. H. Vuong. Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica*, 57(2):307–333, 1989.

K. Wang, K. K. W. Yau, and A. H. Lee. A Zero-Inflated Poisson Mixed Model to Analyze Diagnosis Related Groups with Majority of Same-Day Hospital Stays. *Computer Methods and Programs in Biomedicine*, 68(3):195–203, 2002.

K. Wang, A. H. Lee, K. K. W. Yau, and P. J. W. Carrivick. A Bivariate Zero-Inflated Poisson Regression Model for Analyzing Occupational Injuries. *Accident Analysis and Prevention*, 35(4):625–629, 2003.

P. Wang. Markov Zero-Inflated Poisson Regression Models for a Time Series of Counts with Excess Zeros. *Journal of Applied Statistics*, 28(5):623–632, 2001.

P. Wang. A Bivariate Zero-Inflated Negative Binomial Regression Model for Count Data with Excess Zeros. *Economics Letters*, 78(3):373–378, 2003.

X. Wang, M.-H. Chen, R. C. Kuo, and D. K. Dey. Bayesian Spatial-Temporal Modeling of Ecological Zero-Inflated Count Data. *Statistica Sinica*, 25(1):189–204, 2015.

J. Wieczorek and S. Hawala. A Bayesian Zero-One Inflated Beta Model for Estimating Poverty in U.S. Counties. In *JSM Proceedings, Section on Survey Research Methods*, pages 2812–2822, Alexandria, VA, 2011. American Statistical Association.

P. Wilson. The Misuse of the Vuong Test for Non-Nested Models to Test for Zero-Inflation. *Economics Letters*, 127:51–53, 2015.

F.-C. Xie, J.-G. Lin, and B.-C. Wei. Bayesian Zero-Inflated Generalized Poisson Regression Model: Estimation and Case Influence Diagnotics. *Journal of Applied Statistics*, 41(6):1383–1392, 2014.

M. Yang, G. K. D. Zamba, and J. E. Cavanaugh. Markov Regression Models for Count Time Series with Excess Zeros: A Partial Likelihood Approach. *Statistical Methodology*, 14:26–38, 2013.

M. Yang, J. E. Cavanaugh, and G. K. D. Zamba. State-Space Models for Count Time Series with Excess Zeros. *Statistical Modelling*, 15(1):70–90, 2015.

M. Yang, G. K. D. Zamba, and J. E. Cavanaugh. *ZIM: Zero-Inflated Models for Count Time Series with Excess Zeros*, 2017. URL `https://CRAN.R-project.org/package=ZIM`. R package version 1.0.3.

K. K. W. Yau, K. Wang, and A. H. Lee. Zero-Inflated Negative Binomial Mixed Regression Modeling of Over-Dispersed Count Data with Extra Zeros. *Biometrical Journal*, 45(4):437–452, 2003.

K. K. W. Yau, A. H. Lee, and P. J. W. Carrivick. Modeling Zero-Inflated Count Series with Application to Occupational Health. *Computer Methods and Programs in Biomedicine*, 74(1):47–52, 2004.

T. W. Yee. *Vector Generalized Linear and Additive Models, With an Implementation in R*. Springer, New York, NY, 2015.

K. C. H. Yip and K. K. W. Yau. On Modeling Claim Frequency Data in General Insurance with Extra Zeros. *Insurance: Mathematics and Economics*, 36(2):153–163, 2005.

D. S. Young, A. M. Raim, and N. R. Johnson. Zero-Inflated Modelling for Characterizing Coverage Errors of Extracts from the US Census Bureau's Master Address File. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180 (1):73–97, 2017.

A. Zeileis, C. Kleiber, and S. Jackman. Regression Models for Count Data in R. *Journal of Statistical Software*, 27(8):1–25, 2008. http://www.jstatsoft.org/v27/i08/.

A. Zellner. On Assessing Prior Distributions and Bayesian Regression Analysis with g-Prior Distributions. In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, P. Goel and A. Zellner (eds.), pages 233–243. North Holland Publishing Company, New York, NY, 1986.

S. Zhang, D. Midthune, P. M. Guenther, S. M. Krebs-Smith, V. Kipnis, K. W. Dodd, D. W. Buckman, J. A. Tooze, L. Freedman, and R. J. Carroll. A New Multivariate Measurement Error Model with Zero-Inflated Dietary Data, and Its Application to Dietary Assessment. *The Annals of Applied Statistics*, 5(2B):1456–1487, 2011.

Y. Zhao, A. H. Lee, V. Burke, and K. K. W. Yau. Testing for Zero-Inflation in Count Series: Application to Occupational Health. *Journal of Applied Statistics*, 36(12): 1353–1359, 2009.

H. Zhou, L. Li, and H. Zhu. Tensor Regression with Applications to Neuroimaging Data Analysis. *Journal of the American Statistical Association*, 108(502):540–552, 2013.

F. Zhu. Zero-Inflated Poisson and Negative Binomial Integer-Valued GARCH Models. *Journal of Statistical Planning and Inference*, 142(4):826–839, 2012.

A. F. Zuur, E. N. Ieno, N. Walker, A. A. Saveliev, and G. M. Smith. *Mixed Effects Models and Extensions in Ecology with R.* Springer, New York, NY, 2009.