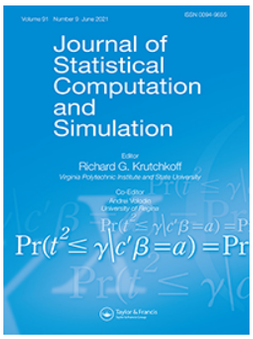


Taylor & Francis
Taylor & Francis Group

Journal of Statistical Computation and Simulation

ISSN: 0094-9655 (Print) 1563-5163 (Online) Journal homepage: www.tandfonline.com/journals/gscs20

A flexible regression model for zero- and *k*-inflated count data

Monika Arora, N. Rao Chaganty & Kimberly F. Sellers

To cite this article: Monika Arora, N. Rao Chaganty & Kimberly F. Sellers (2021) A flexible regression model for zero- and *k*-inflated count data, Journal of Statistical Computation and Simulation, 91:9, 1815-1845, DOI: 10.1080/00949655.2021.1872077

To link to this article: https://doi.org/10.1080/00949655.2021.1872077

JOURNAL OF STATISTICAL COMPUTATION AND SIMULATION
2021, VOL. 91, NO. 9, 1815–1845
https://doi.org/10.1080/00949655.2021.1872077

A flexible regression model for zero- and *k*-inflated count data

Monika Arora[a], N. Rao Chaganty[b] and Kimberly F. Sellers [c]

[a]Department of Mathematics, IIIT Delhi, Delhi, India; [b]Department of Mathematics and Statistics, Old Dominion University, Norfolk, VA, USA; [c]Department of Mathematics and Statistics, Georgetown University, Washington, DC, USA

ABSTRACT

Count data with inflated zeros commonly occur in numerous research studies. Accordingly, there is substantive literature regarding zero-inflated Poisson and analogous generalizable count regression models that account for data dispersion via excess zeros. Scenarios exist, however, where another count $k > 0$ tends to be inflated, thus there remains the need to develop a flexible regression model that can accommodate both inflated frequencies and any inherent data dispersion. This work achieves this goal by employing the Conway–Maxwell–Poisson (CMP) distribution. We develop a zero- and *k*-inflated Conway–Maxwell–Poisson (ZkICMP) distribution and corresponding regression that addresses over- and under-dispersed count data. We further discuss parameter estimation and other diagnostics by analytical and numerical methods, and illustrate superior performance of the ZkICMP regression via real data examples.

ARTICLE HISTORY
Received 25 June 2020
Accepted 2 January 2021

KEYWORDS
Zero-inflated regression models; Conway–Maxwell–Poisson distribution; over-dispersion; under-dispersion

1. Introduction

Count data arise in numerous applications ranging from medical and scientific investigations to social science research. The most commonly used models to analyse such data are developed using the Poisson probability distribution which possesses the equi-dispersion property (that is, its variance and mean are equal). In real-life examples, however, the data are usually over- (under-)dispersed, i.e. the variation is larger (smaller) than the mean. There are several potential causes of data over-dispersion in count data. A well-known factor associated with over-dispersion is an inflated number of zeros in excess of the number expected under the Poisson distribution. In such cases, an appropriate model is the zero-inflated Poisson (ZIP) which has been well-studied in the literature. The earliest paper on the ZIP model is [1] that considers a random variable Y with a degenerate distribution at 0 with probability π and Poisson(λ_*) with probability $1-\pi$. The ZIP(λ_*) probability mass function (pmf) for Y is

$$P(Y=y) = \left(\pi + (1-\pi)e^{-\lambda_*}\right)^u \left((1-\pi)\frac{e^{-\lambda_*}\lambda_*^y}{y!}\right)^{1-u}, \quad y = 0, 1, 2, \ldots, \qquad (1)$$

CONTACT N. Rao Chaganty rchagant@odu.edu Department of Mathematics and Statistics, Old Dominion University, Norfolk, VA 23529, USA

where $u = I_{\{y=0\}}$, is the indicator function. Lambert [2] utilized this form to describe a ZIP regression model, using the log link function to model the rate parameter vector λ_* and the logit function to associate the success probability vector π with predictors, and studies the ZIP regression model using the Expectation–Maximization (EM) approach [3]. The ZIP model with random effects was studied by Yau and Lee [4] and Min and Agresti [5]. Ghosh et al. [6] explore the Bayesian approach for ZIP regression for small and moderate sample sizes. A Bayesian approach for ZIP modelling of spatial data was studied by Agarwal et al. [7]. Furthermore, ZIP models for censored data were studied in [8,9].

A common solution for handling over-dispersion is to replace the Poisson distribution with a negative binomial distribution. These count models have likewise been generalized to consider zero-inflation in order to address underlying over-dispersion above and beyond that captured via the negative binomial model due to excess zeros. The zero-inflated negative binomial (ZINB) likewise assumes a degenerate distribution at zero with probability π but now considers the more flexible negative binomial with parameters (θ, λ_*) with probability $1 - \pi$. Accordingly, the ZINB pmf has the form,

$$P(Y = y) = \left[\pi + (1-\pi)\left(\frac{\theta}{\lambda_* + \theta}\right)^{\theta}\right]^{u} \times \left[(1-\pi)\binom{\theta + y - 1}{y}\left(\frac{\theta}{\lambda_* + \theta}\right)^{\theta}\left(\frac{\lambda_*}{\lambda_* + \theta}\right)^{y}\right]^{1-u}$$

where u is the indicator function and $y = 0, 1, 2, \ldots$. The ZINB model has been applied in various fields [10–12], and includes the ZIP and zero inflated geometric (ZIG) distributions as special cases [13]. The ZINB model with fixed effects and random effects can be found in [14,15], respectively. Yau et al. [16] discussed ZINB mixed regression models, and Ridout et al. [17] developed a score test for the ZINB.

In addition to zero, however, some data may contain inflated counts for an additional count value $k > 0$ as a result of multiple effects (e.g. the study design). For example, research questionnaire studies frequently produce data containing zero- and k-inflated counts typically as a result of the way the questions were asked or the responses provided. A study investigating the frequency of pap smear tests in women over a six-year period, for example, had large numbers of women who never had a pap smear, and many who had pap smears on an annual basis; thus, the survey resulted in large frequencies at zero and six [18]. Another source for inflation is the nature of the response. For example, consider the study that counts the number of days per week a subject felt depressed in a sample that consists of depressed and non-depressed subjects. For several non-depressed subjects, the count will be zero while many depressed individuals will report the count to equal seven; the data will therefore likely have 0 and 7 counts inflated. Lin and Tsai [19] describe a survey where adults were asked about the number of cigarettes they consume on a given day. The responses tend to be none or one pack consisting of 20 cigarettes, hence the data result in inflated frequencies at 0 and 20. Lin and Tsai [19] proposed a zero- and k-inflated Poisson (ZkIP) regression model to analyse such data, and used the non-linear optimization method to obtain the maximum likelihood estimates (MLEs) and standard errors (SEs). Sheth-Chandra [20] also introduced two forms of ZkIP models (known as doubly-inflated Poisson (DIP) models), among them the ZkIP form given in [19] which is the same as the

second DIP model proposed in [20]. The ZkIP model has the pmf

$$P(Y=y)=\begin{cases}\pi_1+\pi_3 e^{-\lambda} & \text{when } y=0\\ \pi_2+\pi_3\dfrac{\lambda^k e^{-\lambda}}{k!} & \text{when } y=k\\ \pi_3\dfrac{\lambda^y e^{-\lambda}}{y!} & \text{when } y\geq 1,\ y\neq k,\end{cases} \tag{2}$$

where $0\leq\pi_i\leq 1$, $i=1,2,3$ such that $\sum_{i=1}^{3}\pi_i=1$; clearly, the ZkIP contains the ZIP model when $\pi_2=0$ and the Poisson model if $\pi_1=\pi_2=0$. A special case of the ZkIP model is the zero- and one-inflated Poisson model (ZOIP). Zhang et al. [21] studied the properties and inference for the parameters of the ZOIP distribution in the absence of covariates while its inference was described by Alshkaki [22]. A Bayesian approach for the ZOIP model was examined by Tang et al. [23].

While the ZkIP model accounts for dispersion in count data due to the presence of two inflated count frequencies, there still remains the potential presence of underlying dispersion that needs to be addressed. More broadly, data dispersion of varying types can exist depending on what data features are likewise addressed during data modelling [24]. The NB distribution can only accommodate data over-dispersion, yet there remains the need for a distribution that can accommodate data that are under-dispersed. The Conway–Maxwell–Poisson (CMP) distribution [25] is a two-parameter extension (λ,ν) of the Poisson distribution that can accommodate data over- or under-dispersion, where λ is the usual rate parameter under the Poisson model, and ν is the dispersion parameter. Recent works have studied its statistical properties [26] while others have utilized its framework to establish a flexible regression model [27] and zero-inflated analogs to address excess zeros [28,29].

Recognizing the unaddressed matter of potential count inflation at two values (say 0 and $k>0$) and the substantial need to still address underlying data dispersion that may not be directly recognized in the presence of these inflated counts, this work extends several aforementioned models to introduce a zero- and k-inflated CMP (ZkICMP) model, an extension of ZICMP [28], and constructs an associated regression model. The paper proceeds as follows. Section 2 describes CMP and ZkICMP in greater detail. Section 3 introduces the ZkICMP regression model, including discussion regarding parameter estimation, hypothesis testing, model selection, and diagnostics. Sections 4 and 5 demonstrate the ZkICMP model flexibility, respectively, via simulated and real data examples. These illustrations show that the ZkICMP can accommodate data over-, equi-, or under-dispersion, and its regression model form can further allow for varying parameter values associated with covariates via an appropriate link function. Section 6 concludes the manuscript with discussion.

2. Background: motivating distributions

This section introduces the reader to the Conway–Maxwell–Poisson (CMP) distribution, a flexible two-parameter distribution that accounts for data dispersion present in count data. The CMP model (described in Section 2.1) serves as the motivating distribution toward developing the zero- and k-inflated CMP analog (in Section 2.2), and the zero- and k-inflated CMP regression in Section 3.

2.1. CMP distribution

A Conway–Maxwell–Poisson (CMP) distributed random variable Y has the pmf

$$P(Y = y) = \frac{\lambda^y}{(y!)^\nu Z(\lambda, \nu)}, \quad \text{for } y = 0, 1, 2, \ldots \tag{3}$$

where $\lambda = E(Y^\nu) > 0$ is easily recognized as the rate parameter under the special case ($\nu = 1$) of the Poisson model, $\nu \geq 0$ is the dispersion parameter, and $Z(\lambda, \nu) = \sum_{j=0}^{\infty} \frac{\lambda^j}{(j!)^\nu}$ is the normalizing constant [25,26]. The dispersion parameter $\nu = 1$ corresponds to equi-dispersion such that the CMP distribution reduces to the Poisson distribution, while $\nu <$ ($>$)1 indicates data over-dispersion (under-dispersion). The CMP contains two other well-known special cases, namely the geometric distribution with success probability $(1 - \lambda)$ when $\nu = 0$ and $\lambda < 1$, and the Bernoulli distribution with mean $\lambda/(1 + \lambda)$ when $\nu \to \infty$. The CMP moment generating function (mgf) is given by $M_Y(t) = Z(\lambda e^t, \nu)/Z(\lambda, \nu)$; it can be used to derive the raw moments of the distribution. In particular, the mean and variance of the CMP distribution and their approximations are

$$E(Y) = \lambda \frac{\partial \log Z}{\partial \lambda} \approx \lambda^{1/\nu} - \frac{\nu - 1}{2\nu}, \quad \text{and}$$
$$V(Y) = \frac{\partial E(Y)}{\partial \log \lambda} \approx \frac{1}{\nu} \lambda^{1/\nu},$$

where the approximations are good when $\nu \leq 1$ or $\lambda > 10^\nu$. See Sellers et al. [30] for more details.

2.2. Zero- and k-inflated CMP distribution

An appropriate model for under- or over-dispersed count data with excessive zeros is the zero-inflated Conway–Maxwell–Poisson (ZICMP) distribution that was introduced by Sellers and Raim [28]. This generalization of the zero-inflated Poisson (ZIP) distribution is a mixture of the degenerate distribution at zero with probability π_1 and a CMP(λ, ν) distribution with probability $(1 - \pi_1)$. The resulting pmf is

$$P(Y = y) = \begin{cases} \pi_1 + (1 - \pi_1) \dfrac{1}{Z(\lambda, \nu)} & \text{when } y = 0 \\ (1 - \pi_1) \dfrac{\lambda^y}{(y!)^\nu Z(\lambda, \nu)} & \text{when } y \geq 1, \end{cases} \tag{4}$$

where $0 \leq \pi_1 \leq 1$. The special cases of ZICMP include the zero-inflated Poisson (ZIP) when $\nu = 1$, zero-inflated geometric (ZIG) when $\nu = 0$ and $\lambda < 1$, and 'zero-inflated Bernoulli (ZIB)' (i.e. a Bernoulli distribution with adjusted success probability) when ν converges to infinity. The ZICMP model has been used in psychology [28], health [31], and agriculture [32] research. The statistical software SAS [33] includes CMP and ZICMP in their COUNTREG and HPCOUNTREG procedures, while the `COMPoissonReg` [34] package in R [35] allows for (ZI)CMP regression.

We generalize this model to establish a count distribution containing inflated frequencies both at 0 and a count value $k > 0$. We consider a random variable Y defined by a mixture

of three distributions: the degenerate distribution at 0 with probability π_1, the degenerate distribution at count $k > 0$ with probability π_2, and the CMP(λ, ν) distribution with probability $\pi_3 = 1 - \pi_1 - \pi_2$. The resulting pmf has the form

$$P(Y = y) = \begin{cases} \pi_1 + \pi_3 \dfrac{1}{Z(\lambda, \nu)} & \text{when } y = 0 \\ \pi_2 + \pi_3 \dfrac{\lambda^k}{(k!)^\nu Z(\lambda, \nu)} & \text{when } y = k \\ \pi_3 \dfrac{\lambda^y}{(y!)^\nu Z(\lambda, \nu)} & \text{when } y \geq 1,\ y \neq k, \end{cases} \tag{5}$$

where $\pi_1 + \pi_2 + \pi_3 = 1$ and $\pi_i \geq 0$. The distribution (5) is an extension of the ZICMP [28] and we will refer to it as the zero- and k-inflated CMP (ZkICMP) distribution. The flexibility of the CMP model allows for the ZkICMP to likewise contain three special case distributions and serve as a bridge between them: when $\nu = 1$, (5) reduces to the zero- and k-inflated Poisson (ZkIP) distribution; for $\nu = 0$ and $\lambda < 1$, the ZkICMP is the zero- and k-inflated geometric (ZkIG) distribution; and as $\nu \to \infty$, and the ZkICMP distribution converges to a 'zero- and k-inflated Bernoulli (ZkIB)' distribution, i.e. a Bernoulli distribution with success probability $\pi_2 + (\lambda\pi_3)/(1 + \lambda)$. The mean and variance of the ZkICMP distribution are

$$\begin{aligned} E(Y) &= k\pi_2 + \pi_3\lambda \frac{\partial \log Z}{\partial \lambda}, \text{and} \\ V(Y) &= \pi_2 \left(k^2(1 - \pi_2) - 2k\pi_3\lambda \frac{\partial \log Z(\lambda, \nu)}{\partial \lambda} \right) \\ &\quad + \pi_3 \left(\frac{\partial E(Y)}{\partial \log \lambda} + (1 + \pi_3) \left(\lambda \frac{\partial \log Z(\lambda, \nu)}{\partial \lambda} \right)^2 \right), \end{aligned}$$

where $\pi_3 = 1 - \pi_1 - \pi_2$.

3. Zero- and *k*-inflated CMP regression model

Consider a vector $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ of n independent counts where y_i has a ZkICMP$(\pi_1, \pi_2, \lambda_i, \nu)$ distribution, $i = 1, \ldots, n$. We further assume that the varying parameter $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_n)$ has a log-linear relationship $\log(\lambda_i) = \mathbf{x_i}^T\boldsymbol{\beta}$ with a p dimensional covariate vector $\mathbf{x_i} = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$ through the corresponding coefficient vector, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$; we set $x_{i1} = 1$ for the regression model with an intercept. Given this construct, the likelihood and log-likelihood functions of the observed sample are

$$\begin{aligned} L_{\text{obs}}(\pi_1, \pi_2, \boldsymbol{\lambda}, \nu | \mathbf{y}) &= \prod_{i:y_i=0} \left(\pi_1 + \pi_3 \frac{1}{Z(\lambda_i, \nu)} \right) \prod_{i:y_i=k} \left(\pi_2 + \pi_3 \frac{{\lambda_i}^k}{(k!)^\nu Z(\lambda_i, \nu)} \right) \\ &\quad \times \prod_{i:y_i \neq 0,k} \left(\pi_3 \frac{{\lambda_i}^{y_i}}{(y_i!)^\nu Z(\lambda_i, \nu)} \right) \\ &= \prod_{i:y_i=0} (\pi_1 + \pi_3 p_{0i}) \prod_{i:y_i=k} (\pi_2 + \pi_3 p_{ki}) \prod_{i:y_i \neq 0,k} (\pi_3 p_{y_i}), \end{aligned} \tag{6}$$

$$\ell_{\text{obs}}(\pi_1, \pi_2, \boldsymbol{\lambda}, \nu|\mathbf{y}) = \sum_{i:y_i=0} \log(\pi_1 + \pi_3 p_{0i}) + \sum_{i:y_i=k} \log(\pi_2 + \pi_3 p_{ki}) + \sum_{i:y_i\neq 0,k} (\log \pi_3 + \log p_{y_i}), \tag{7}$$

where $p_{y_i} \doteq \lambda_i^{y_i}/[(y_i!)^{\nu} Z(\lambda_i, \nu)]$ for $y_i = 0, 1, \ldots, k, \ldots$. To use optimization routines for obtaining the maximum likelihood estimates (MLEs), we reparametrize the constants as

$$\log\left(\frac{\pi_1}{\pi_3}\right) = \gamma, \quad \log\left(\frac{\pi_2}{\pi_3}\right) = \delta, \quad \text{and} \quad \log(\nu) = \eta. \tag{8}$$

One can likewise elaborate the models in Equation (8) to associate the respective relations to covariates of interest, however for simplicity and ease of discussion, we maintain them as unknown constants. Similarly, this model can be extended to allow for variable dispersion (see, for example, [24] or [29]); for ease of interpretation and discussion, however, we maintain a constant dispersion assumption.

When $\pi_2 = 0$, Equation (6) reduces to

$$L_{\text{obs}}(\pi_1, \boldsymbol{\lambda}, \nu \mid \mathbf{y}) = \prod_{i:y_i=0} \left(\pi_1 + (1-\pi_1)\frac{1}{Z(\lambda_i, \nu)}\right) \prod_{i:y_i>0} \left((1-\pi_1)\frac{\lambda_i^{y_i}}{(y_i!)^{\nu} Z(\lambda_i, \nu)}\right),$$

which is the likelihood function of the ZICMP model [28]. Further when $\pi_1 = 0$ and $\pi_2 = 0$, Equation (6) becomes $L_{\text{obs}}(\boldsymbol{\lambda}, \nu \mid \mathbf{y}) = \prod_{i=1}^{n} \frac{\lambda_i^{y_i}}{(y_i!)^{\nu} Z(\lambda_i, \nu)}$, which is the likelihood function of the CMP distribution [26].

3.1. Estimation of ZkICMP parameters

In this section, we discuss the maximum likelihood (ML) estimation of the ZkICMP regression model parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}, \gamma, \delta, \nu)$. The dimension p of the regression coefficient vector $\boldsymbol{\beta}$ depends on the number of covariates included in the model. We assume k is data-driven and known. The parameters γ and δ quantify the zero and k inflations, respectively, and ν is the dispersion parameter in the ZkICMP model.

Substituting $\pi_1 = e^{\gamma}/(1 + e^{\gamma} + e^{\delta})$ and $\pi_2 = e^{\delta}/(1 + e^{\gamma} + e^{\delta})$ into Equation (7), the loglikelihood can be rewritten as

$$\ell_{\text{obs}}(\boldsymbol{\theta}) = \sum_{i:y_i=0} \log(e^{\gamma} + p_{0i}) + \sum_{i:y_i=k} \log(e^{\delta} + p_{ki}) + \sum_{i:y_i\neq 0,k} \log p_{y_i} - n(\log(1 + e^{\gamma} + e^{\delta})). \tag{9}$$

We obtain the following score functions by taking the partial derivatives of (9) with respect to the parameters where, for notational convenience, we let $Z \doteq Z(\lambda_i, \nu)$:

$$\frac{\partial \ell_{\text{obs}}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} = \left(-\sum_{i:y_i=0} \frac{p_{0i}}{e^{\gamma} + p_{0i}} \frac{1}{Z}\frac{\partial Z}{\partial \lambda_i} + \sum_{i:y_i=k} \frac{p_{ki}}{e^{\delta} + p_{ki}} \left(\frac{k}{\lambda_i} - \frac{1}{Z}\frac{\partial Z}{\partial \lambda_i}\right)\right) \lambda_i \mathbf{x_i} + \sum_{i:y_i\neq 0,k} \left(\frac{y_i}{\lambda_i} - \frac{1}{Z}\frac{\partial Z}{\partial \lambda_i}\right) \lambda_i \mathbf{x_i}$$

$$\frac{\partial \ell_{\text{obs}}(\boldsymbol{\theta})}{\partial \gamma} = \sum_{i:y_i=0} \frac{e^{\gamma}}{e^{\gamma} + p_{0i}} - \frac{ne^{\gamma}}{1 + e^{\gamma} + e^{\delta}}$$

$$\frac{\partial \ell_{\text{obs}}(\boldsymbol{\theta})}{\partial \delta} = \sum_{i:y_i=k} \frac{e^{\delta}}{e^{\delta} + p_{ki}} - \frac{ne^{\delta}}{1 + e^{\gamma} + e^{\delta}}$$

$$\begin{aligned}\frac{\partial \ell_{\text{obs}}(\boldsymbol{\theta})}{\partial \nu} = &- \sum_{i:y_i=0} \frac{p_{0i}}{e^{\gamma} + p_{0i}} \frac{1}{Z} \frac{\partial Z}{\partial \nu} - \sum_{i:y_i=k} \frac{p_{ki}}{e^{\delta} + p_{ki}} \left(\log(k!) + \frac{1}{Z} \frac{\partial Z}{\partial \nu} \right) \\ &- \sum_{i:y_i \neq 0,k} \left(\log(y_i!) + \frac{1}{Z} \frac{\partial Z}{\partial \nu} \right). \end{aligned} \tag{10}$$

The score equations (10) do not have closed form solutions, however they can be solved numerically using routines for solving non-linear equations. Alternatively, we can optimize the likelihood function (9) directly. The statistical computing tool R [35] includes optimization functions such as `nlminb` and `optim` whose routines perform optimization via box-constraint quasi-newton (L-BGFS-B), conjugate gradient (CG), or simulated annealing algorithm (SANN). We use `nlminb` which implements a type of Newton–Raphson procedure. The Poisson regression coefficient estimates serve as the initial values for $\boldsymbol{\beta}$. The initial values for γ and δ are obtained from the sample proportions of zeros and k's while we set $\nu = 1$ as the initial value for the dispersion parameter (in accordance with the Poisson initial assumption). We did not encounter any convergence problems using these initial values.

Under standard regularity conditions according to Cramèr's theorem, the MLEs are asymptotically normal with covariance matrix given by inverse of the observed Fisher information. For the ZkICMP model, the observed Fisher information is given by

$$\mathcal{I}_{\text{obs}} = \left[-\frac{\partial^2 \ell_{\text{obs}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right]. \tag{11}$$

The elements of $\mathcal{I}_{\text{obs}}$ can be obtained taking the partial derivatives of the score functions (10); derivations are given in the Appendix. We use these formulas to compute (11) and obtain the SE's of the MLEs via the inverse of the observed Fisher information matrix.

3.2. Hypothesis testing

Testing the impact of the jth covariate on the count response is equivalent to testing $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$. This hypothesis test is straightforward and can be done using the standard Wald statistic, $z = \widehat{\beta}_j / SE(\widehat{\beta}_j)$ which is asymptotically standard normal under the null hypothesis. We are likewise interested in testing $H_0 : \nu = 1$ vs. $H_1 : \nu \neq 1$ to determine whether significant data dispersion exists such that the ZkIP distribution is not an appropriate zero- and k-inflated count model; in particular, $H_1 : \nu > 1$ indicates under-dispersion, and $H_1 : \nu < 1$ implies over-dispersion. Once again, this hypothesis can be tested using the Wald test statistic, $z = (\widehat{\nu} - 1)/SE(\widehat{\nu})$ which is asymptotically standard normal. An alternative approach for testing $H_0 : \nu = 1$ is the likelihood ratio test (LRT)

statistic given by

$$-2\log\Lambda = -2\log\frac{L_{\text{obs}}(\widetilde{\beta},\,\widetilde{\gamma},\,\widetilde{\delta},\,\nu=1)}{L_{\text{obs}}(\widehat{\beta},\widehat{\gamma},\widehat{\delta},\widehat{\nu})}, \tag{12}$$

which is asymptotically distributed as chi-square with one degree of freedom, where $\widetilde{\beta}$, $\widetilde{\gamma}$ and $\widetilde{\delta}$ denote the MLEs under the ZkIP model and $\widehat{\beta}$, $\widehat{\gamma}$, $\widehat{\delta}$, $\widehat{\nu}$ are the MLEs under the ZkICMP model.

The (Z(k)I)CMP models are nested in that the CMP is a special case of ZICMP which is a special case of ZkICMP. Accordingly, one can use the LRT to test the significance of the nested models, i.e. whether the ZkICMP model can be simplified by the ZICMP model or whether the ZICMP model can be reduced to the CMP model. To see if significantly inflated frequency exists at count $k>0$, we test the hypotheses $H_0:\pi_2=0$ vs. $H_1:\pi_2>0$; rejecting the null hypothesis implies that we should forego the ZICMP for the ZkICMP model while failing to reject H_0 implies that the ZICMP model suffices. Similarly, significant zero inflation is tested by the hypotheses, $H_0:\pi_1=0$ vs. $H_1:\pi_1>0$. Since $0\le\pi_i\le 1$ $(i=1,2)$, the null hypothesis $H_0:\pi_i=0$ corresponds to testing a parameter value on the boundary. In this case, the regularity conditions are not met. The standard asymptotic theory for the LRT statistic thus is not applicable and instead its asymptotic distribution is a mixture of χ^2 distributions [37,38].

4. Data simulation examples

This section illustrates the flexibility of the ZkICMP regression through simulated data examples that contain significant data dispersion and inflated frequencies at both zero and $k>0$. For model comparison, the CMP, NB, and Poisson base models along with their zero- (and k-)inflated analogs are considered to describe each given count dataset with inflated frequencies. The variance of the negative binomial (NB) distribution, after reparametrization, can be written as $\lambda+r\lambda^2$, where λ denotes the mean. If $r=0$ then the NB distribution simplifies to the Poisson distribution. The parameter estimates and SEs of the ZkICMP, ZICMP models were computed in R using the non-linear optimization methods mentioned in Section 3.1 while the results of the CMP, Poisson and NB models were obtained using the SAS [33] software count regression (COUNTREG) and generalized linear model (GENMOD) procedures, respectively; the results of the ZkINB, ZINB, ZkIP and ZIP models were obtained from the finite mixture model (FMM) procedure in SAS. For model selection and to prevent data over-fitting, we consider the Akaike Information Criterion (AIC) [39] to select the model that best fits the data, selecting the model with the minimum AIC and apply the Burnham and Anderson [36] approach for model comparison via AIC differences, $\Delta_i=\text{AIC}_i-\min(\text{AIC})$, to determine the level of empirical support for Model i; see Table 1 for details. We further compare models via their respective goodness-of-fit (GOF) statistics and absolute error (ABE), where the model with the smallest GOF and ABE are likewise considered optimal.

The first simulated data is from the ZkICMP distribution tailored to study correct specification. And the second simulation study is designed to study mis-specification. Here the data are generated from a ZkIP distribution involving a varying rate parameter associated with two explanatory variables to illustrate the model flexibility in a regression format. In

Table 1. Guidelines for measuring the empirical support level of a model as determined by AIC difference, $\Delta_i = \text{AIC}_i - \min(\text{AIC})$ [36].

Δ_i	Empirical support level for model i
0–2	Substantial
4–7	Considerably less
> 10	Essentially none

these simulated examples, zeros and ks are inflated because π_1 and π_2 are positive. Various sample sizes ($n = 200, 500, 1000, 2000$) are considered for this study.

4.1. Simulation I

For the first data simulation study, we use the true parameter values $\lambda = 3$, $\nu = 1.5$, $\pi_1 = 0.4$ and $\pi_2 = 0.1$ for the ZkICMP distribution. The simulated data are under-dispersed (since $\nu > 1$) and inflated at zero ($\pi_1 = 0.4$) and $k = 2$ ($\pi_2 = 0.1$). We consider the Poisson, NB, and CMP models, as well as their zero-inflated (ZI) and zero-and-k-inflated (ZkI) analogs for model comparisons. The results of our analyses are summarized in Tables 2 and 3, comparing the models via their respective log-likelihoods, AIC values, GOF, and ABE values.

Based on the AIC results provided in Table 2, the ZkICMP model is either selected as the best model or demonstrates substantial support for consideration in comparison to the best model as determined by the minimum AIC. Further, the distinction between the best models strengthens in favour of uniquely selecting the ZkICMP model as the sample size increases. At $n = 200$, the ZkIP model has the smallest AIC but both the ZkINB and the ZkICMP models have AIC difference values, $\Delta < 2$. Meanwhile, at $n = 500$ and 1000, the ZkICMP already demonstrates itself to attain the minimum AIC, this time with the ZkIP and ZkINB models demonstrating substantial support as defined by Burnham and Anderson [36]. At $n = 2000$, the ZkICMP distinguishes itself from the other models, where the next best models are the ZkIP and ZkINB have AIC differences $\Delta \approx 10$ which, according to Burnham and Anderson [36], demonstrate essentially no empirical support for these models in comparison to the ZkICMP, thus the ZkICMP is uniquely the best model.

Table 2 further illustrates how accounting for excess zeroes and inflated frequency at $k = 2$ influence the detected amount and type of data dispersion in a dataset. Among the various CMP models, the CMP parameters are consistently estimated as $\widehat{\nu} < 1$ thus giving the impression of apparent data over-dispersion. The ZICMP and ZkICMP models, however, each determine $\widehat{\nu} > 1$ thus detecting data under-dispersion when accounting for inflated frequency at 0 in the ZICMP (and k in the ZkICMP) model. These simulations and resulting analyses illustrate the impact of mixture distributions on the detected amount and type of dispersion. Sellers and Shmueli [24] discuss that the true direction of data dispersion can be masked by other influences that impacting the perceived dispersion in both size and direction. This data simulation demonstrates that what initially appears as an over-dispersed dataset does so because of the inflated frequencies, and the ZkICMP correctly recognizes the underlying data under-dispersion, when accounting for inflated frequencies.

Table 2. Estimates, standard errors (in parentheses), and model diagnostics (log-likelihood, and AIC) for simulated data example I.

n	Parameters	ZkICMP	ZICMP	CMP	ZkINB	ZINB	NB	ZkIP	ZIP	Poisson
2000	$\widehat{\lambda}$	3.3416	7.6842	0.7449	1.8049	1.7538	1.1750	1.8049	1.7538	1.1750
		(0.6286)	(1.2603)	(0.0285)	(0.0531)	(0.0455)	(0.0242)	(0.0531)	(0.0455)	(0.0242)
	$\widehat{\nu}$	1.5633	2.3812	0.3642	–	–	–	–	–	–
		(0.1734)	(0.1559)	(0.0429)						
	$\widehat{\pi}_1$	0.4040	0.4259	–	0.3637	0.3300	–	0.3637	0.3300	–
		(0.0157)	(0.0122)		(0.0130)	(0.0149)		(0.0130)	(0.0149)	
	$\widehat{\pi}_2$	0.1114	–	–	0.1359	–	–	0.1359	–	–
		(0.0161)			(0.0082)			(0.0082)		
	$\widehat{r}$	–	–	–	< 0.0001	< 0.0001	0.4803	–	–	–
					–		(0.0545)			
	$\log L_{\text{obs}}$	−2788.86	−2811.44	−2957.01	−2794.85	−2860.37	−2971.44	−2794.87	−2860.37	−3044.38
	AIC	5585.72	5628.87	5918.01	5595.70	5724.73	5946.88	5595.73	5724.73	6090.77
1000	$\widehat{\lambda}$	2.8471	7.4459	0.7781	1.5904	1.5912	1.1230	1.5905	1.5912	1.1230
		(0.8582)	(1.8122)	(0.0434)	(0.0706)	(0.0617)	(0.0335)	(0.0706)	(0.0617)	(0.0335)
	$\widehat{\nu}$	1.5488	2.4968	0.4616	–	–	–	–	–	–
		(0.2862)	(0.2407)	(0.0667)						
	$\widehat{\pi}_1$	0.3788	0.4138	–	0.3282	0.2943	–	0.3282	0.2943	–
		(0.0255)	(0.0178)		(0.0197)	(0.0226)		(0.0197)	(0.0226)	
	$\widehat{\pi}_2$	0.1080	–	–	0.1331	–	–	0.1331	–	–
		(0.0234)			(0.0124)			(0.0124)		
	$\widehat{r}$	–	–	–	< 0.0001	< 0.0001	0.3576	–	–	–
					–		(0.0708)			
	$\log L_{\text{obs}}$	−1369.53	−1378.40	−1437.91	−1371.60	−1402.54	−1444.30	−1371.62	−1402.54	−1465.19
	AIC	2747.06	2762.81	2879.82	2749.20	2809.07	2892.60	2749.23	2809.07	2932.38
500	$\widehat{\lambda}$	3.7276	7.0113	0.7215	1.8874	1.8242	1.1920	1.8873	1.8242	1.1920
		(1.3399)	(2.2265)	(0.0542)	(0.1071)	(0.0927)	(0.0488)	(0.1071)	(0.0927)	(0.0488)
	$\widehat{\nu}$	1.6161	2.2362	0.3133	–	–	–	–	–	–
		(0.3284)	(0.2956)	(0.0818)						
	$\widehat{\pi}_1$	0.4160	0.4310	–	0.3756	0.3466	–	0.3756	0.3466	–
		(0.0300)	(0.0246)		(0.0261)	(0.0290)		(0.0261)	(0.0290)	
	$\widehat{\pi}_2$	0.0929	–	–	0.1199	–	–	0.1199	–	–
		(0.0316)			(0.0159)			(0.0159)		
	$\widehat{r}$	–	–	–	< 0.0001	< 0.0001	0.5580	–	–	–
					–		(0.1153)			
	$\log L_{\text{obs}}$	−705.10	−709.33	−746.68	−707.10	−720.10	−750.31	−707.12	−720.10	−773.82
	AIC	1418.19	1424.67	1497.36	1420.20	1444.20	1504.63	1420.24	1444.20	1549.64

(*continued*).

Table 2. Continued.

n	Parameters	ZkICMP	ZICMP	CMP	ZkINB	ZINB	NB	ZkIP	ZIP	Poisson
200	$\widehat{\lambda}$	2.8648	7.6011	0.8323	1.7307	1.6960	1.2150	1.7307	1.6960	1.2150
		(1.7051)	(3.9083)	(0.1029)	(0.1618)	(0.1383)	(0.0779)	(0.1618)	(0.1383)	(0.0779)
	$\widehat{\nu}$	1.4642	2.4206	0.4751	–	–	–	–	–	–
		(0.5506)	(0.4956)	(0.1419)						
	$\widehat{\pi}_1$	0.3607	0.3924	–	0.3213	0.2836	–	0.3213	0.2836	–
		(0.0538)	(0.0388)		(0.0409)	(0.0481)		(0.0409)	(0.0481)	
	$\widehat{\pi}_2$	0.1284	–	–	0.1496	–	–	0.1496	–	–
		(0.0514)			(0.0287)			(0.0287)		
	$\widehat{r}$	–	–	–	≤ 0.0001	≤ 0.0001	0.3338	–	–	–
					–		(0.1456)			
	$\log L_{\text{obs}}$	−282.10	−284.86	−297.85	−282.50	−289.98	−299.25	−282.49	−289.98	−303.58
	AIC	572.19	575.71	599.71	571.00	583.96	602.50	570.99	583.96	609.15
	No. of param.	4	3	2	4	3	2	3	2	1

Note: The true values of the parameters are $\lambda = 3$, $\nu = 1.5$, $\pi_1 = 0.4$, $\pi_2 = 0.1$ and the points of inflation are 0 and $k = 2$.

Table 3. Frequency comparisons associated with simulated data example I, where ABE denotes absolute error, χ^2 is the goodness-of-fit statistic, and $k = 2$.

n	Count	Observed	ZkICMP	ZICMP	CMP	ZkINB	ZINB	NB	ZkIP	ZIP	Poisson
2000	0	892	892.00	892.06	783.60	892.00	892.01	787.81	892.02	891.97	617.64
	1	281	280.54	309.35	583.71	297.10	406.83	591.73	297.13	406.84	725.72
	2	540	540.00	456.28	337.80	540.00	356.72	328.96	539.94	356.75	426.36
	3	188	190.29	256.28	168.65	161.32	208.54	161.48	161.32	208.56	166.99
	4	75	72.80	72.56	75.83	72.79	91.45	74.01	72.79	91.44	49.05
	5	19	19.65	12.08	31.43	26.28	32.08	32.48	26.28	32.07	11.53
	6	5	3.99	1.30	12.19	7.90	9.38	13.83	7.90	9.37	2.26
	ABE		6.60	193.47	653.10	55.18	363.57	675.78	55.27	363.56	889.89
	χ^2		0.37	50.71	04.40	8.44	145.45	327.94	8.44	145.40	449.23
1000	0	438	438.00	437.99	391.39	438.00	438.06	389.04	438.00	438.04	325.30
	1	169	168.49	180.08	304.54	174.62	228.71	311.71	174.64	228.71	365.31
	2	272	272.00	237.56	172.08	272.00	181.95	169.53	271.98	181.97	205.12
	3	82	85.15	113.87	80.63	73.62	96.51	77.66	73.63	96.51	76.78
	4	32	28.32	26.61	33.09	29.27	38.40	32.24	29.28	38.39	21.56
	5	6	6.67	3.56	12.25	9.31	12.22	12.56	9.31	12.22	4.84
	6	1	1.18	0.30	4.17	2.47	3.24	4.67	2.47	3.24	0.91
	ABE		8.19	85.93	293.94	21.51	179.19	308.96	21.53	179.15	392.80
	χ^2		0.69	18.96	129.55	3.44	68.12	139.99	3.44	68.10	172.04
500	0	226	226.00	226.00	197.82	226.00	226.02	200.50	226.01	226.01	151.81
	1	68	67.05	73.52	142.73	72.12	96.16	143.53	72.11	96.16	180.95
	2	128	128.00	109.41	82.88	128.00	87.70	80.04	128.00	87.70	107.85
	3	46	51.49	65.76	42.38	42.82	53.33	40.41	42.81	53.33	42.85
	4	26	20.43	20.77	19.81	20.20	24.32	19.34	20.20	24.32	12.77
	5	6	5.65	3.98	8.63	7.63	8.88	8.95	7.62	8.87	3.04
	ABE		12.36	51.13	160.47	14.73	80.36	164.19	14.74	80.35	226.63
	χ^2		2.14	11.85	70.76	2.48	28.82	75.77	2.48	28.81	127.34
200	0	83	83.00	83.00	72.85	83.00	83.00	72.13	83.01	83.00	59.34
	1	31	31.13	34.31	60.63	32.45	44.57	62.35	32.45	44.57	72.10
	2	58	58.00	48.72	36.30	58.00	37.79	35.94	58.00	37.80	43.80
	3	20	18.53	25.92	17.93	16.20	21.37	17.27	16.20	21.37	17.74
	4	4	6.98	6.87	7.72	7.01	9.06	7.47	7.01	9.06	5.39
	5	4	1.89	1.06	2.99	2.43	3.07	3.02	2.43	3.07	1.31
	ABE		6.67	24.33	68.28	9.83	41.13	71.46	9.84	41.13	85.30
	χ^2		3.73	12.77	31.23	3.27	18.13	33.30	3.27	18.12	43.64

Note: The models considered are the base Poisson, negative binomial, and Conway–Maxwell–Poisson distributions, along with their zero-inflated and zero-and-k-inflated analogs.

While one should use caution when comparing parameter estimates across all models, some results nonetheless provide insights regarding model comparisons. The ZkICMP estimated parameters are close to their true values even for a small sample size, and get closer to the true value as the sample size increases. The two models ZkIP and ZkINB capture both the inflations at zero and at 2 even for a sample of size 200. The estimates of π_1 and π_2 for these models are close to the true values. The ZICMP is likewise able to capture the inflation at zero even for a small sample size; $\widehat{\pi}_1 = 0.3924$ is very close to the true value, 0.4.

The ZkINB model fails to detect the data under-dispersion; for all sample sizes, the estimate $\widehat{r}$ of the dispersion parameter is approximately 0. This is expected because the ZkINB model can only address data over-dispersion. Given that the simulated data are under-dispersed, ZkINB with $\widehat{r}$ is approximately 0 implies that the best the ZkINB model can do is to conduct itself in a manner akin to the ZkIP model. Notice that the associated estimated ZkIP and ZkINB parameters and SEs equal each other for all n; the same is true for the ZIP and ZINB models, as well as the Poisson and NB models.

Table 4. Testing zero inflation for simulated data example I.

n		ZICMP vs. CMP	ZIP vs. Poisson
2000	Likelihood ratio	291.14	368.02
	p-value	< 0.0001	< 0.0001
1000	Likelihood ratio	119.02	125.30
	p-value	< 0.0001	< 0.0001
500	Likelihood ratio	74.70	107.44
	p-value	< 0.0001	< 0.0001
200	Likelihood ratio	25.98	27.20
	p-value	< 0.0001	< 0.0001

Notes: The test is $H_0 : \pi_1 = 0$ against $H_1 : \pi_1 > 0$ with significance level, $\alpha = 0.05$. The asymptotic distribution is $0.5\chi_0^2 + 0.5\chi_1^2$.

Table 5. Testing k inflation for simulated data example I.

n		ZkICMP vs. ZICMP	ZkIP vs. ZIP
2000	Likelihood ratio	45.16	851.44
	p-value	< 0.0001	< 0.0001
1000	Likelihood ratio	17.74	392.36
	p-value	< 0.0001	< 0.0001
500	Likelihood ratio	8.46	192.46
	p-value	0.0018	< 0.0001
200	Likelihood ratio	5.52	27.20
	p-value	0.0094	< 0.0001

Notes: The test is $H_0 : \pi_2 = 0$ against $H_1 : \pi_2 > 0$ with significance level, $\alpha = 0.05$ and $k = 2$. The asymptotic distribution is $0.5\chi_0^2 + 0.5\chi_1^2$.

The GOF and ABE results are displayed in Table 3. The ZINB is unable to capture the under-dispersion and is reduced to ZIP, thus the expected frequencies under these distributions are almost equal. Similarly, the expected frequencies under negative binomial and Poisson distribution are very close. The ZkICMP, ZkINB and ZkIP models are again shown to be optimal when $n = 200$ and 500 in that the resulting GOF and ABE values are small, however these values remain small only for ZkICMP model as n increases.

The results from the respective LRTs for pairwise comparisons between the models are given in Tables 4 and 5. These tables show that, for all sample sizes, the ZkICMP is better than ZICMP, which is better than CMP. Similarly, the ZkIP is better than ZIP, which is better than Poisson for all sample sizes.

In summary, the data is generated from ZkICMP model. The Poisson and CMP models are inappropriate as they are unable to capture the inflations. The ZIP model accounts for the inflation at 0 but fails to account for the under-dispersion and inflation at k. In most cases, the ZkIP and ZkINB captures the inflation at zero and k, but fails to capture the underlying under-dispersion in the data. The ZICMP model allows the flexibility of capturing inflation at zero along with under-dispersion thus it outperforms against the other competing models. But it fails to capture the inflation at k thus does not have minimum AIC. As expected, the ZkICMP model has minimum AIC and gives data a good fit.

4.2. Simulation II

In this simulation study, we construct a zero-and-k-inflated Poisson regression model where $k = 2$, $\pi_1 = 0.2$, $\pi_2 = 0.4$, and $\log(\boldsymbol{\lambda}) = \beta_0 + \beta_1 \boldsymbol{x}_1 + \beta_2 \boldsymbol{x}_2$ where $\beta_0 = 0.7$, $\beta_1 = -0.01$, $\beta_2 = -0.04$, and $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are generated, respectively, from a standard normal and

Bernoulli(0.5) distribution. The MLEs and associated SEs obtained from assuming the various considered count models are reported in Table 6 for the various sample sizes while the resulting GOF test and ABE results are provided in Table 7.

We observe that, for all the sample sizes in the simulated data the ZIP model underestimates ($\widehat{\pi}_1 \approx 0.1$) the inflation at zero. The ZINB model behaves in a similar way as the ZIP model, while, the ZICMP model overestimates ($\widehat{\pi}_1 > 0.2$) the inflation at zero. For $n = 200$, the ZkIP, ZkINB models capture the inflation at zero ($\widehat{\pi}_1 = 0.18$), and at $k = 2$ ($\widehat{\pi}_2 = 0.45$). The ZkICMP does slightly better ($\widehat{\pi}_1 = 0.19, \widehat{\pi}_2 = 0.45$) than ZkIP and ZkINB models. While, on increasing the sample size the estimated values of the inflation at zero and 2 get closer to the true value for all the three models (ZkIP, ZkINB and ZkICMP).

This study and the results in Table 6 also illustrate that the ZkICMP performs comparably with other best models for all n. The best models are the ZkIP and the ZkINB, while the ZkICMP model is one that demonstrates substantial support relative to the ZkIP and ZkINB models because $\Delta < 2$ for the ZkICMP model for all n. Further, given the respective coefficient estimates and SEs, these results are reasonably close to their respective true values. The ZkIP and ZkINB models again produce nearly identical parameter estimates and $\widehat{r} \approx 0$. In this example, these results again make sense as any (Z(k)I)NB model contains its (Z(k)I)P model as a limiting case. Given that the simulated data arise from a ZkIP model, the results obtained in Table 6 affirms that these three models are optimal.

The CMP model interprets the simulated data (at all sample sizes) as approximately equi-dispersed. Accordingly, the resulting CMP parameter estimates are approximately equal to the corresponding Poisson estimates (and $\widehat{\nu} \approx 1$ for all n), and their respective log-likelihood values are approximately equal with their AIC difference less than 2. These results likewise demonstrate substantial support that the CMP model performs in a manner comparable to the Poisson model. Table 7 shows that the ZkICMP has the smallest ABE and GOF measures, except for sample size 500, and thus provides a good fit for the simulated data.

Tables 8 and 9 provide the pairwise model comparisons testing for significant inflation at 0 (Table 8) and $k = 2$ (Table 9). Pairwise comparisons using the LRTs for the simulated data reveal that the CMP model fits better than the Poisson, ZICMP fits better than ZIP, and ZkICMP fits better than ZICMP model. Thus, using LRT, we conclude that the CMP models fit significantly better than their Poisson counterparts and, among the CMP models, the ZkICMP outperforms the other models (as expected). The results from the AIC criterion concur with the LRT.

In summary the response variable is from ZkIP distribution. We notice that the results of the ZkINB and ZkIP models are comparable. In the ZkINB model, the estimates, SEs, log likelihood and AIC are similar to the ZkIP model. Both the ZkIP and ZkINB models capture the inflations at zero and k, but ZkINB fails to capture the underlying equi-dispersion in the data. For $\nu = 1$ the ZkICMP is same as ZkIP which is seen in the estimates and comparable AIC values. We obtain similar results when we increase the sample size to 1000 and 2000. Notably, for large sample sizes, not only do the ZkICMP and ZkIP outperform the other models, but their respective parameter estimates get closer to the true values.

5. Applications

We use the ZkICMP model to analyse two real datasets: (1) an example where the data contain excess counts of 0 and $k = 1$; (2) an example containing inflated frequencies at 0 and

Table 6. Estimates, standard errors (in parentheses), and model diagnostics (log-likelihood, and AIC) for simulated data example II.

n	Parameters	ZkICMP	ZICMP	CMP	ZkINB	ZINB	NB	ZkIP	ZIP	Poisson
2000	$\widehat{\beta}_0$	0.6498	3.2177	0.5159	0.7089	0.5890	0.4798	0.7089	0.5890	0.4798
		(0.1614)	(0.1681)	(0.0489)	(0.0407)	(0.0287)	(0.0247)	(0.0407)	(0.0287)	(0.0247)
	$\widehat{\beta}_1$	−0.0109	−0.0083	−0.0097	−0.0104	−0.0080	−0.0093	−0.0104	−0.0080	−0.0093
		(0.0280)	(0.0322)	(0.0185)	(0.0285)	(0.0195)	(0.0182)	(0.0285)	(0.0195)	(0.0182)
	$\widehat{\beta}_2$	−0.0354	−0.0387	−0.0207	−0.0361	−0.0196	−0.0199	−0.0361	−0.0196	−0.0199
		(0.0555)	(0.0625)	(0.0361)	(0.0564)	(0.0378)	(0.0354)	(0.0564)	(0.0378)	(0.0354)
	$\widehat{\nu}$	0.9494	3.4908	1.0682	–	–	–	–	–	–
		(0.1336)	(0.1609)	(0.0509)						
	$\widehat{\pi}_1$	0.1956	0.2473	–	0.1994	0.1036	–	0.1994	0.1036	–
		(0.0178)	(0.0099)		(0.0141)	(0.0137)		(0.0141)	(0.0137)	
	$\widehat{\pi}_2$	0.4004	–	–	0.3988	–	–	0.3988	–	–
		(0.0175)			(0.0175)			(0.0175)		
	$\widehat{r}$	–	–	–	< 0.0001	< 0.0001	0.0000	–	–	–
					–		(0.0002)			
	$\log L_{\text{obs}}$	−2689.65	−2932.62	−3126.00	−2689.70	−3098.95	−3126.57	−2689.70	−3098.95	−3126.57
	AIC	5391.29	5875.23	6259.00	5389.40	6205.90	6259.13	5389.40	6205.90	6259.13
1000	$\widehat{\beta}_0$	0.7880	3.6653	0.5116	0.6203	0.5483	0.4429	0.6203	0.5483	0.4429
		(0.2593)	(0.2571)	(0.0695)	(0.0611)	(0.0422)	(0.0363)	(0.0611)	(0.0422)	(0.0363)
	$\widehat{\beta}_1$	0.0272	0.0323	0.0073	0.0259	0.0087	0.0068	0.0259	0.0087	0.0068
		(0.0424)	(0.0470)	(0.0258)	(0.0405)	(0.0266)	(0.0249)	(0.0405)	(0.0266)	(0.0249)
	$\widehat{\beta}_2$	−0.0246	−0.0160	−0.0265	−0.0241	−0.0207	−0.0247	−0.0241	−0.0207	−0.0247
		(0.0854)	(0.0955)	(0.0528)	(0.0818)	(0.0544)	(0.0510)	(0.0818)	(0.0544)	(0.0510)
	$\widehat{\nu}$	1.1497	4.0420	1.1180	–	–	–	–	–	–
		(0.2261)	(0.2538)	(0.0750)						
	$\widehat{\pi}_1$	0.2087	0.2599	–	0.1975	0.1019	–	0.1975	0.1019	–
		(0.0251)	(0.0141)		(0.0207)	(0.0202)		(0.0207)	(0.0202)	
	$\widehat{\pi}_2$	0.3795	–	–	0.3848	–	–	0.3848	–	–
		(0.0248)			(0.0247)			(0.0247)		
	$\widehat{r}$	–	–	–	< 0.0001	< 0.0001	< 0.0001	–	–	–
	$\log L_{\text{obs}}$	−1326.84	−1416.87	−1530.00	−1327.05	−1519.65	−1531.73	−1327.05	−1519.65	−1531.73
	AIC	2665.68	2843.75	3069.00	2664.10	3047.30	3069.47	2664.10	3047.30	3069.47

(*continued*).

Table 6. Continued.

n	Parameters	ZkICMP	ZICMP	CMP	ZkINB	ZINB	NB	ZkIP	ZIP	Poisson
500	$\widehat{\beta}_0$	0.9494	3.2343	0.5824	0.7600	0.6193	0.5204	0.7600	0.6193	0.5204
		(0.3235)	(0.3306)	(0.0991)	(0.0766)	(0.0561)	(0.0483)	(0.0766)	(0.0561)	(0.0483)
	$\widehat{\beta}_1$	−0.0227	−0.0190	−0.0220	−0.0218	−0.0174	−0.0206	−0.0218	−0.0174	−0.0206
		(0.0564)	(0.0618)	(0.0370)	(0.0539)	(0.0377)	(0.0358)	(0.0539)	(0.0377)	(0.0358)
	$\widehat{\beta}_2$	−0.0680	−0.0876	−0.0288	−0.0630	−0.0307	−0.0269	−0.0630	−0.0307	−0.0269
		(0.1130)	(0.1218)	(0.0718)	(0.1076)	(0.0738)	(0.0695)	(0.1076)	(0.0738)	(0.0695)
	$\widehat{\nu}$	1.1609	3.4337	1.1060	–	–	–	–	–	–
		(0.2686)	(0.3110)	(0.1015)						
	$\widehat{\pi}_1$	0.1953	0.2314	–	0.1851	0.0924	–	0.1851	0.0924	–
		(0.0298)	(0.0194)		(0.0269)	(0.0265)		(0.0269)	(0.0265)	
	$\widehat{\pi}_2$	0.3891	–	–	0.3945	–	–	0.3945	–	–
		(0.0348)			(0.0345)			(0.0345)		
	$\widehat{r}$	–	–	–	< 0.0001	< 0.0001	< 0.0001	–	–	–
					–					
	$\log L_{\text{obs}}$	−684.02	−739.93	−787.93	−684.20	−782.50	−788.49	−684.20	−782.50	−788.49
	AIC	1380.04	1489.87	1584.00	1378.40	1573.00	1582.98	1378.40	1573.00	1582.98
200	$\widehat{\beta}_0$	0.8657	4.6873	0.5405	0.7669	0.6130	0.5378	0.7669	0.6130	0.5378
		(0.6023)	(0.6457)	(0.1593)	(0.1217)	(0.0855)	(0.0733)	(0.1217)	(0.0855)	(0.0733)
	$\widehat{\beta}_1$	−0.0025	−0.0397	0.0338	−0.0015	0.0182	0.0288	−0.0015	0.0182	0.0288
		(0.0910)	(0.1066)	(0.0575)	(0.0887)	(0.0556)	(0.0529)	(0.0887)	(0.0556)	(0.0529)
	$\widehat{\beta}_2$	−0.4424	−0.3937	−0.2374	−0.4341	−0.1930	−0.2021	−0.4341	−0.1930	−0.2021
		(0.2338)	(0.2395)	(0.1269)	(0.2233)	(0.1219)	(0.1153)	(0.2233)	(0.1219)	(0.1153)
	$\widehat{\nu}$	1.0843	4.9268	1.2796	–	–	–	–	–	–
		(0.5050)	(0.6322)	(0.1775)						
	$\widehat{\pi}_1$	0.1902	0.2482	–	0.1841	0.0763	–	0.1841	0.0763	–
		(0.0591)	(0.0307)		(0.0471)	(0.0454)		(0.0471)	(0.0454)	
	$\widehat{\pi}_2$	0.4493	–	–	0.4517	–	–	0.4517	–	–
		(0.0564)			(0.0560)			(0.0560)		
	$\widehat{r}$	–	–	–	< 0.0001	< 0.0001	< 0.0001	–	–	–
	$\log L_{\text{obs}}$	−248.77	−272.21	−301.17	−248.80	−301.15	−302.51	−248.80	−301.15	−302.51
	AIC	509.54	554.43	610.34	507.60	610.30	611.02	507.60	610.30	611.02
	No. of param.	6	5	4	6	5	4	5	4	3

Table 7. Frequency comparisons associated with simulated data example II, where ABE denotes absolute error, χ^2 is the goodness-of-fit statistic, and $k = 2$.

n	Count	Observed	ZkICMP	ZICMP	CMP	ZkINB	ZINB	NB	ZkIP	ZIP	Poisson
2000	0	508	508.05	507.92	400.57	508.09	508.18	403.92	508.07	508.13	403.87
	1	219	219.60	325.80	663.78	217.75	536.74	645.76	217.75	536.75	645.78
	2	1015	1015.00	710.35	525.04	1014.99	479.12	516.71	1015.02	479.16	516.75
	3	142	142.62	376.95	270.67	145.21	286.10	276.95	145.22	286.12	276.96
	4	76	72.17	73.31	102.15	72.67	127.81	110.87	72.67	127.80	110.86
	5	24	29.36	6.51	30.20	28.91	45.46	35.32	28.90	45.45	35.31
	6	10	9.89	0.30	7.22	9.45	13.30	9.22	9.45	13.29	9.21
	7	6	2.94	0.01	1.50	2.69	3.40	2.11	2.69	3.39	2.11
	ABE		13.62	682.35	1210.47	16.65	1077.07	1214.94	16.66	1076.99	1214.96
	χ^2		4.39	4996.35	867.73	5.15	894.00	876.94	5.16	893.85	876.87
1000	0	264	264.04	264.04	209.43	264.06	264.05	214.98	264.05	264.03	214.95
	1	120	120.07	160.35	344.58	122.20	277.59	330.50	122.20	277.60	330.51
	2	497	497.00	377.09	261.16	496.97	237.52	254.00	496.98	237.54	254.02
	3	72	72.92	173.93	126.68	69.44	136.31	131.00	69.44	136.32	131.00
	4	32	31.72	24.56	43.73	31.44	57.82	49.82	31.44	57.81	49.82
	5	10	10.88	1.43	11.90	11.61	19.81	15.32	11.61	19.80	15.31
	6	5	2.99	0.04	2.63	3.54	5.63	3.91	3.53	5.63	3.91
	ABE		4.20	283.19	585.67	8.50	517.69	585.74	8.47	517.65	585.75
	χ^2		1.43	783.68	402.76	0.97	419.73	412.82	0.98	419.66	412.78
500	0	119	119.03	118.98	103.89	119.04	119.07	95.08	119.03	119.05	95.07
	1	54	53.28	79.51	183.31	54.84	133.29	157.84	54.83	133.29	157.85
	2	254	254.00	178.58	133.29	253.98	121.84	130.98	253.99	121.85	130.99
	3	38	41.66	100.29	58.00	39.37	74.57	72.82	39.37	74.58	72.82
	4	22	20.99	21.08	17.45	20.55	34.25	30.36	20.55	34.25	30.36
	5	11	8.05	2.05	3.83	8.46	12.35	9.86	8.46	12.35	9.86
	6	2	2.73	0.11	0.76	3.18	4.07	3.02	3.18	4.07	3.01
	ABE		9.10	174.99	298.09	7.42	263.77	296.11	7.41	263.74	296.12
	χ^2		1.66	149.25	226.23	1.36	214.05	209.31	1.36	214.01	209.29
200	0	50	50.40	50.00	51.70	50.46	50.62	43.56	50.46	50.61	43.56
	1	19	21.32	28.48	76.62	21.51	57.62	65.54	21.51	57.62	65.54
	2	109	109.01	83.51	48.70	109.01	48.31	50.80	109.01	48.31	50.80
	3	17	13.03	37.15	20.21	12.68	28.88	28.27	12.68	28.88	28.27
	4	1	7.65	4.83	6.18	7.58	13.87	12.55	7.58	13.86	12.55
	5	2	3.18	0.19	1.35	3.27	5.11	4.29	3.27	5.11	4.29
	6	2	0.62	0.00	0.17	0.68	1.17	0.92	0.67	1.17	0.92
	ABE		15.91	62.77	130.48	16.46	128.62	137.36	16.46	128.61	137.37
	χ^2		10.76	2459.82	142.55	10.57	121.45	118.28	10.57	121.43	118.27

Note: The models considered are the base Poisson, negative binomial, and Conway–Maxwell–Poisson distributions, along with their zero-inflated and zero-and-k-inflated analogs.

Table 8. Testing zero inflation for simulated data example II.

n		ZICMP vs. CMP	ZIP vs. Poisson
2000	Likelihood ratio	386.76	55.24
	p-value	< 0.0001	< 0.0001
1000	Likelihood ratio	226.26	24.16
	p-value	< 0.0001	< 0.0001
500	Likelihood ratio	96.00	11.98
	p-value	< 0.0001	0.0003
200	Likelihood ratio	57.92	2.72
	p-value	< 0.0001	0.05

Notes: The test is $H_0 : \pi_1 = 0$ against $H_1 : \pi_1 > 0$ with significance level, $\alpha = 0.05$. The asymptotic distribution is $0.5\chi_0^2 + 0.5\chi_1^2$.

$k = 5$. Both data examples originate from the National Health and Nutrition Examination Survey (NHANES). Along with the model comparison procedures discussed in Section 4, we conduct residual analysis via randomized quantile residuals as advised by Dunn and

Table 9. Testing k inflation for simulated data example II.

n		ZkICMP vs. ZICMP	ZkIP vs. ZIP
2000	Likelihood ratio	485.94	818.50
	p-value	< 0.0001	< 0.0001
1000	Likelihood ratio	180.06	385.20
	p-value	< 0.0001	< 0.0001
500	Likelihood ratio	111.82	196.60
	p-value	< 0.0001	< 0.0001
200	Likelihood ratio	46.88	104.70
	p-value	< 0.0001	< 0.0001

Notes: The test is $H_0 : \pi_2 = 0$ against $H_1 : \pi_2 > 0$ with significance level, $\alpha = 0.05$ where $k = 2$. The asymptotic distribution is $0.5\chi^2_0 + 0.5\chi^2_1$.

Smyth [40] and analogously considered by Sellers and Raim [28] for goodness-of-fit with the ZICMP models.

5.1. Drugs data

In this example, the response count variable is the number of joints/pipes of a drug smoked by the adults without a prescription from a doctor. Subjects aged between 18 and 59 were asked two questions in a survey, the first question being 'Have you ever used marijuana or hashish?'. A negative response is recorded as a count value of zero. Meanwhile, if the response is positive, a follow-up question asks 'How many joints/pipes did you smoke in a day?' and the response is recorded. Four covariates are included in the survey – BMI, age, gender, and family income (as measured by the ratio to poverty level). While over 4000 people were surveyed, we utilize the complete data stemming from 2481 subjects. The mean and variance of the count response were 0.70 and 1.20, respectively, thus we preliminarily detect possible data over-dispersion to be present; however, as noted in [24], we are cautious not to be mislead by this initial perception regarding apparent dispersion. The percentage of people who never smoked was 64.05% and, among the adults who smoked, 15.12% did so on an average of one joint/day, while 20.83% smoked more than a joint per day. Clearly, the 0 and 1 counts have high frequencies, thus consideration of ZkI count models is reasonable. For completeness, we consider the Poisson, NB, and CMP base models along with their ZI and ZkI analogs as described in Section 4. Table 10 provides the parameter estimates and SEs for all of the considered models, while Table 11 provides their frequency comparisons, GOF statistics, and respective ABE values.

The ZkICMP model is optimal as shown by the minimum AIC (5432.85) while the next best model as determined by AIC is the ZICMP ($\Delta = 8.81$); see Table 10. Thus, even the ZICMP is shown to have empirical support that is considerably small relative to the ZkICMP model. The ZkICMP regression model estimates a substantial relative frequency of zeros (0.6389) and ones (0.1084), respectively, such that the remaining distribution is recognized as under-dispersed with $\widehat{\nu} = 3.8416$. In contrast, the other ZkI models estimate $\widehat{\pi}_2 \approx 0$ which implies that the other models do not detect an excess frequency of ones. The ZICMP regression meanwhile estimates the relative frequency of zeros $\widehat{\pi}_1 = 0.5637$ and likewise detects modest potential under-dispersion ($\widehat{\nu} = 1.4968$). Meanwhile, by not accounting for excess zeros (or ones), the drugs data are perceived to be over-dispersed as noted by the CMP ($\widehat{\nu} = 0.0019$) and NB ($\widehat{r} = 1.3895$) regressions. We see from the

Table 10. Estimates, standard errors (in parentheses), and model diagnostics (log-likelihood, and AIC) for the drugs data example.

Parameters	ZkICMP	ZICMP	CMP	ZkINB	ZINB	NB	ZkIP	ZIP	Poisson
Intercept	4.5274*	1.1522*	−0.7962*	0.5678*	0.5678*	−0.1791	0.5678*	0.5678*	−0.1647*
	(0.7977)	(0.2196)	(0.0635)	(0.0984)	(0.0984)	(0.1176)	(0.0984)	(0.0984)	(0.0834)
Age	−0.0163*	−0.0093*	−0.0059*	−0.0081*	−0.0081*	−0.0099*	−0.0081*	−0.0081*	−0.0103*
	(0.0050)	(0.0028)	(0.0015)	(0.0024)	(0.0024)	(0.0028)	(0.0024)	(0.0024)	(0.0020)
Income	−0.1854*	−0.1038*	−0.0368*	−0.0787*	−0.0787*	−0.0643*	−0.0787*	−0.0787*	−0.0637*
	(0.0390)	(0.0224)	(0.0120)	(0.0181)	(0.0181)	(0.0219)	(0.0181)	(0.0181)	(0.0157)
Gender	0.4945*	0.4881*	0.3357*	0.4444*	0.4444*	0.5707*	0.4444*	0.4444*	0.5675*
	(0.1223)	(0.0719)	(0.0391)	(0.0609)	(0.0609)	(0.0693)	(0.0609)	(0.0609)	(0.0501)
$\widehat{\gamma}$	0.9279	0.2562	–	0.0447	0.0447	–	0.0447	0.0447	–
	(0.0745)	(0.0738)		(0.0609)	(0.0609)		(0.0609)	(0.0609)	
$\widehat{\delta}$	−0.8456	–	–	−16.1765	–	–	−16.5838	–	–
	(0.1867)			(377.00)			(462.14)		
$\widehat{\nu}$	3.8416*	1.4968*	0.0019	–	–	–	–	–	–
	(0.5674)	(0.1608)	–						
$\widehat{r}$	–	–	–	0.0000	< 0.0001	1.3895*	–	–	–
				–	–	(0.1151)			
$\widehat{\pi}_1$	0.6389	0.5637	–	0.5112	0.5112	–	0.5112	0.5112	–
	(0.0172)	(0.0182)		(0.0152)	(0.0152)		(0.0152)	(0.0152)	
$\widehat{\pi}_2$	0.1084	–	–	0.0000	–	–	0.0000	–	–
	(0.0181)			(< 0.0001)			(< 0.0001)		
No. of param.	7	6	5	7	6	5	6	5	4
log L_{obs}	−2709.42	−2714.83	−2805.00	−2720.25	−2720.25	−2797.45	−2720.25	−2720.25	−3001.96
AIC	5432.85	5441.66	5619.00	5452.50	5450.50	5604.89	5452.50	5450.50	6011.93

*Statistically significant regression and dispersion parameters (at the $\alpha = 0.05$ significance level).

Table 11. Frequency comparisons for the drugs data example.

Count	Observed	ZkICMP	ZICMP	CMP	ZkINB	ZINB	NB	ZkIP	ZIP	Poisson
0	1589	1589.01	1593.65	1508.79	1596.65	1596.65	1564.87	1596.63	1596.63	1295.35
1	375	383.99	383.89	581.62	405.11	405.11	524.04	405.12	405.12	822.19
2	250	276.54	316.05	250.53	290.02	290.02	231.51	290.04	290.04	325.35
3−5	206	201.14	167.99	114.23	157.14	157.14	111.98	157.15	157.15	98.25
>5	61	76.40	82.86	67.57	87.88	87.88	69.12	87.87	87.87	33.01
ABE		55.80	139.46	385.70	153.51	153.51	293.80	153.51	153.51	951.95
χ^2		5.98	28.39	152.03	31.21	31.21	124.13	31.21	31.21	469.17

Z(k)ICMP models, however, that the apparent over-dispersion recognized by the CMP model is clearly attributed to the excess frequencies at zero (and, perhaps, one). Thus, there appears to be some measure of data under-dispersion that surfaces when accounting for (at least) excess zeros in the drugs dataset, as discussed in [24].

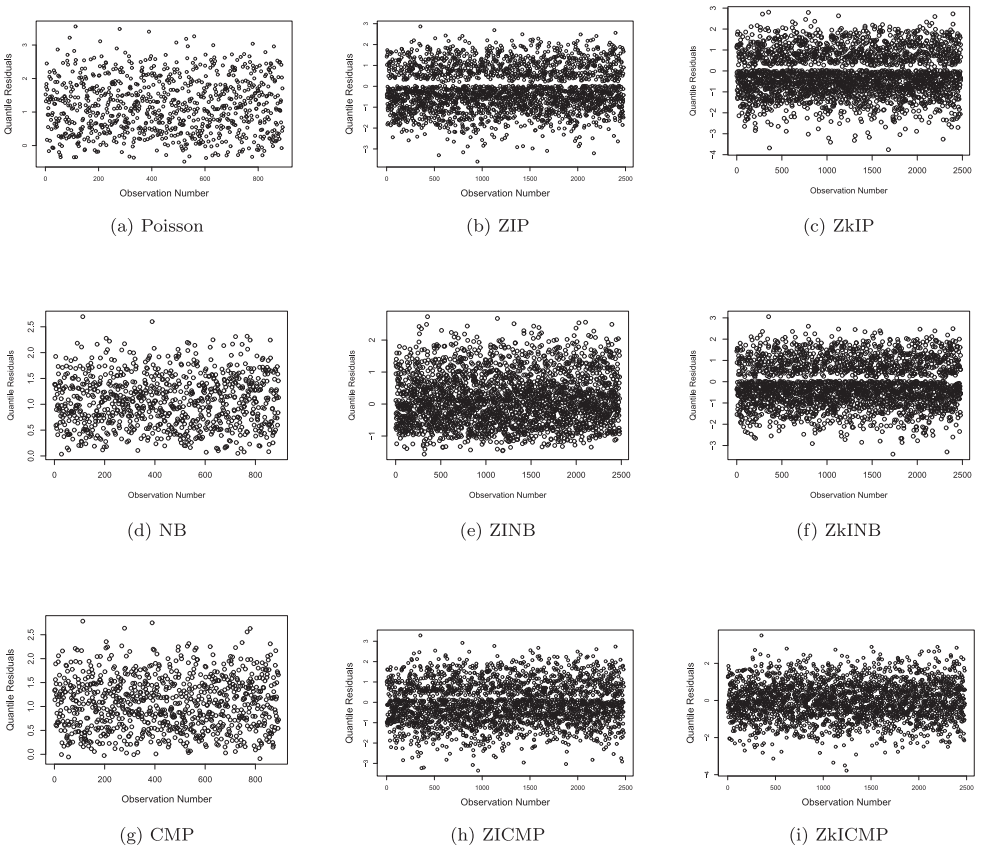
The apparent data under-dispersion when considering Z(k)I models is further recognized when considering the ZkINB, ZINB, ZkIP, and ZIP regressions. The Z(k)INB regression simplifies to the Z(k)IP model when $r = 0$, since NB becomes Poisson in this case. (Z(k)I)Poisson and (Z(k)I)NB models cannot adequately accommodate under-dispersed data, hence Table 10 shows that they produce near equal coefficient estimates since all assume (at best) data equi-dispersion (i.e. $\widehat{r} = 0$ for the Z(k)INB), while $\widehat{\pi}_1 = 0.5112$ for all four models and $\widehat{\pi}_2 = 0$ for both the ZkINB and ZkIP models.

Thus, among these four models, the ZIP is preferred since it has the least number of parameters and the smallest AIC. Note, however, that these models still produce potentially biased results as none of them properly detect the underlying data under-dispersion that is apparent when compensating for excess zeros.

Table 11 provides the observed vs. expected frequencies determined from any of the considered models. We observe that the ZIP and ZINB models do capture fair amount of inflation at zero. The ZICMP model gives a better fit than these two models but not as good as ZkICMP. Through the resulting ABE and GOF measures, we again see that the ZkICMP model is the best performer.

Consider the hypothesis test that assumes the appropriateness of the ZkIP model and asks if significant data dispersion exists such that one should instead consider a ZkICMP regression. Accordingly, we consider $H_0 : \nu = 1$ vs. $H_0 : \nu \neq 1$ and find that the LRT statistic in (12) from Section 3.2 reports $-2\log\Lambda = 21.66$ (p-value < 0.0001), and thus can infer that statistically significant data dispersion exists such that the ZkICMP model is more appropriate than ZkIP for these data. Similarly, the LRT test determines that the ZICMP regression fits better than ZIP. As noted above, the COUNTREG procedure in SAS reports the estimated dispersion parameter $\widehat{\nu} = 0.0019$ for the CMP model, but gives very small covariances for the dispersion parameter and we cannot obtain the SE. The ZICMP estimated dispersion parameter is $\widehat{\nu} = 1.50$ while it is $\widehat{\nu} = 3.84$ for the ZkICMP model and both values are statistically significant. These results confirm the initially detected under-dispersion when accounting for potentially inflated frequencies at zero (and one) is, in fact, statistically significant.

We also consider the LRT test as described in Section 3.2 to test for significant inflation probabilities. The LRT test statistic for $H_0 : \pi_1 = 0$ vs. $H_1 : \pi_1 > 0$ (i.e. comparing the CMP and ZICMP models) is $-2\log\Lambda = 180.34$ (p-value < 0.0001), implying that the



(a) Poisson (b) ZIP (c) ZkIP

(d) NB (e) ZINB (f) ZkINB

(g) CMP (h) ZICMP (i) ZkICMP

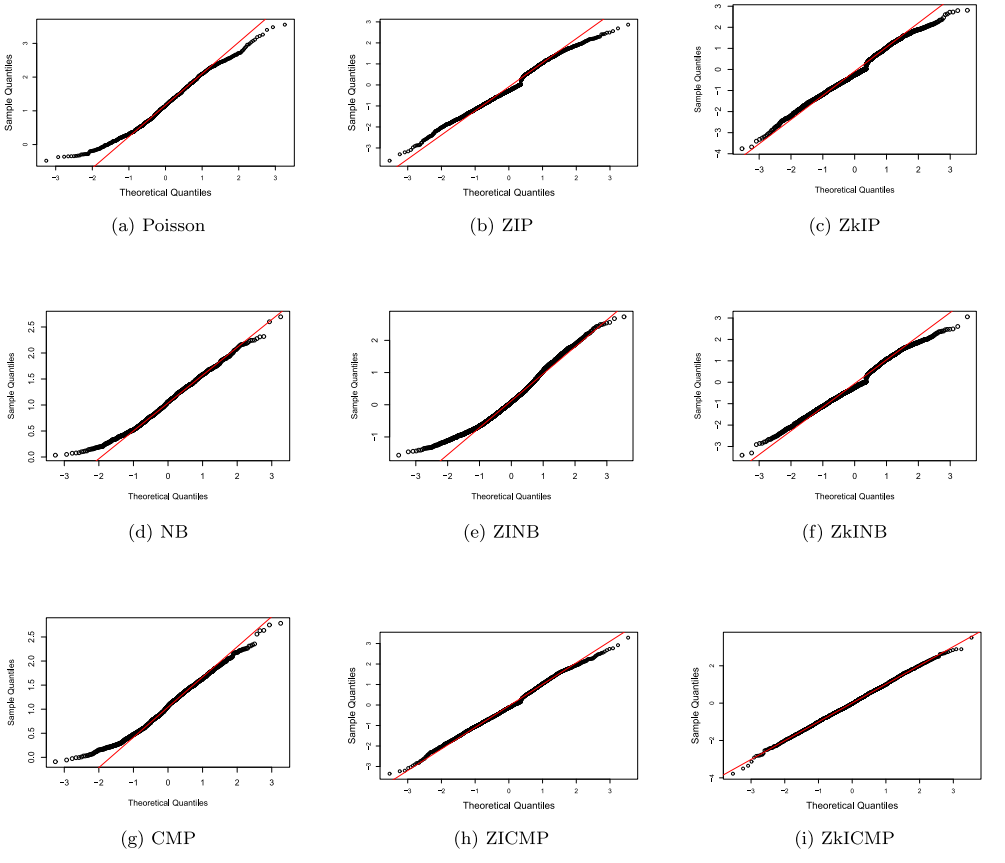
Figure 1. Randomized quantile residual plots for the drugs data example, where the rows reflect the results for the various models (Poisson, negative binomial, and Conway–Maxwell–Poisson, respectively), and the columns contain the respective base distribution, its zero-inflated version, and its zero-and-*k*-inflated analog. (a) Poisson, (b) ZIP, (c) ZkIP, (d) NB, (e) ZINB, (f) ZkINB, (g) CMP, (h) ZICMP, (i) ZkICMP.

ZICMP model is statistically significantly better than the CMP model (i.e. the inflation at zero is significant). Meanwhile, the LRT test for $H_0 : \pi_2 = 0$ vs. $H_1 : \pi_2 > 0$ (i.e. comparing a ZICMP and ZkICMP regression) also shows that the ZkICMP is better than ZICMP ($-2 \log \Lambda = 10.82; p-\text{value} = 0.0005$) and thus that there are statistically significantly inflated frequencies both at zero and one.

To select the best fit model, we further compare the residuals plots associated with the various models; see Figures 1 and 2. The best model is ZkICMP as the residual plot looks completely random and the QQ plot has most of the quantiles agreeing with the standard normal quantiles (apart from some deviation which might be due to few outlier observations).

5.2. Exercise data

In this example, the response variable is the number of times a subject did vigorous or moderate activities in a week. The variable is constructed using the following four questions on



(a) Poisson (b) ZIP (c) ZkIP

(d) NB (e) ZINB (f) ZkINB

(g) CMP (h) ZICMP (i) ZkICMP

Figure 2. QQ plots for the drugs data example, where the rows reflect the results for the various models (Poisson, negative binomial, and Conway–Maxwell–Poisson, respectively), and the columns contain the respective base distribution, its zero-inflated version, and its zero-and-*k*-inflated analog. (a) Poisson, (b) ZIP, (c) ZkIP, (d) NB, (e) ZINB, (f) ZkINB, (g) CMP, (h) ZICMP, (i) ZkICMP.

the questionnaire survey: (1) 'Have you done vigorous activity during the week?', (2) 'Have you done moderate activity during the week?', and (3) 'How many days did you do vigorous activity in the week?', and (4) 'How many days did you do moderate activity during the week?' The response variable is taken as zero if the answer to the first two questions is negative. Otherwise the response variable is taken as the sum of the values obtained for Questions (3) and (4).

While there were several relevant covariates for consideration, we considered only those variables containing complete data, namely age, body mass index (BMI), body weight,[1] ratio of family income to poverty (Ratio), gender, average systolic (Avg. Sys.) blood pressure (BP), and average diastolic (Avg. Dias.) BP. The variables age, ratio and gender were provided in the demographic file of the NHANES survey, while BMI and weight are listed as 'BMXWT' and 'BMXBMI', respectively, in the body measure file. The average BP data were obtained by averaging the four readings of the examined subjects, where the readings were provided in the blood pressure file under the examination data section in NHANES.

The survey respondents (across both genders) were between 12 and 80 years old; 6122 subjects were included in this data analysis. Among these subjects, 62.15% never did any activity, while 7.87% subjects did vigorous or moderate (or both types of) exercise five times a week, thus 0 and 5 occur with high frequencies in the data. Additional exploratory data analysis shows that the number of times subjects exercised (at least moderately) varied between 0 and 13 with the sample mean and variance, respectively, reported at 2.03 and 10.18 thus demonstrating apparent data over-dispersion, and the observed count frequencies at 0 and 5 are more than that expected under a Poisson regression model.

Table A1 in the appendix reports the estimated coefficients and SEs, log-likelihood, and AIC for the respective models when considering all initial covariates, noting that the covariates associated with age and average systolic BP are not statistically significant for most of the considered (i.e. the ZI and ZkI) models. Thus, we refit all of the considered models removing these covariates from the respective models. Table 12 shows that the ZkICMP optimally models the exercise data with the AIC equaling 19245.75; the model with the smallest AIC difference is ZkINB ($\Delta = 39.55$), demonstrating that none of the other models offer any empirical support favouring an alternate model [36].

The CMP dispersion estimate $\widehat{\nu} = 0.0009$ is close to zero; the SAS software failed to provide a corresponding SE. This result is consistent with our initially detected over-dispersion. When accounting for inflated frequencies at 0 (and 5), we still detect apparent over-dispersion (although at a reduced level). The ZICMP model estimates $\widehat{\nu} = 0.4605$ when accounting for excess zeros, while the ZkICMP regression produces $\widehat{\nu} = 0.3817$ when accounting for inflated frequencies at both 0 and 5. In all cases, $\widehat{\nu} < 1$ indicating the presence of statistically significant data over-dispersion.[2] The estimate $\widehat{r} = 0.2108$ for the ZkINB model is likewise statistically significant, further supporting the presence of data over-dispersion. For all of the models, the covariates exhibit similar relation with the number of times a subject did activity/activities in a week. The covariates BMI, gender (male) and average diastolic blood pressure have positive relation with the response variable whereas the variable ratio of family income to poverty has a negative relation.

Table 12 shows that the AIC values for the CMP models are smaller than their Poisson analogs for all (whether non-inflated or (single or double) inflated) cases. In all cases, the appropriate LRT test shows that the (Z(k)I)CMP model is statistically significantly better than the corresponding (Z(k)I)Poisson model. This makes sense because the CMP models have the ability to capture underlying data over-dispersion. Table 12 further shows that the respective AIC values of the inflated models is less than that of their non-inflated counterpart models, with the ZkICMP model reporting the smallest AIC (19245.75). We also compare the observed and expected frequencies by calculating the sum of the absolute errors (ABE) as well as the Pearson chi-square statistic; see Table 13. The ZkICMP model has an ABE of 614.79 and a GOF value equaling 434.72; these values are the smallest among the competing models. Thus the ZkICMP regression is the best model among those considered.

For a postmortem analysis, we plot the randomized residuals as described in [28]; see Figure 3. If the model fits the data correctly, these residuals should exhibit a random behaviour. The residual plots of the models appear to be random, in particular, for the ZkINB and ZkICMP models. The values are not concentrated near a point 0, or a band (-1 to 1). Most of the residual values of the ZkINB and ZkICMP models lie between -3 and 3. The sample quartiles of Poisson, ZIP, ZkIP, NB and CMP models do not agree with

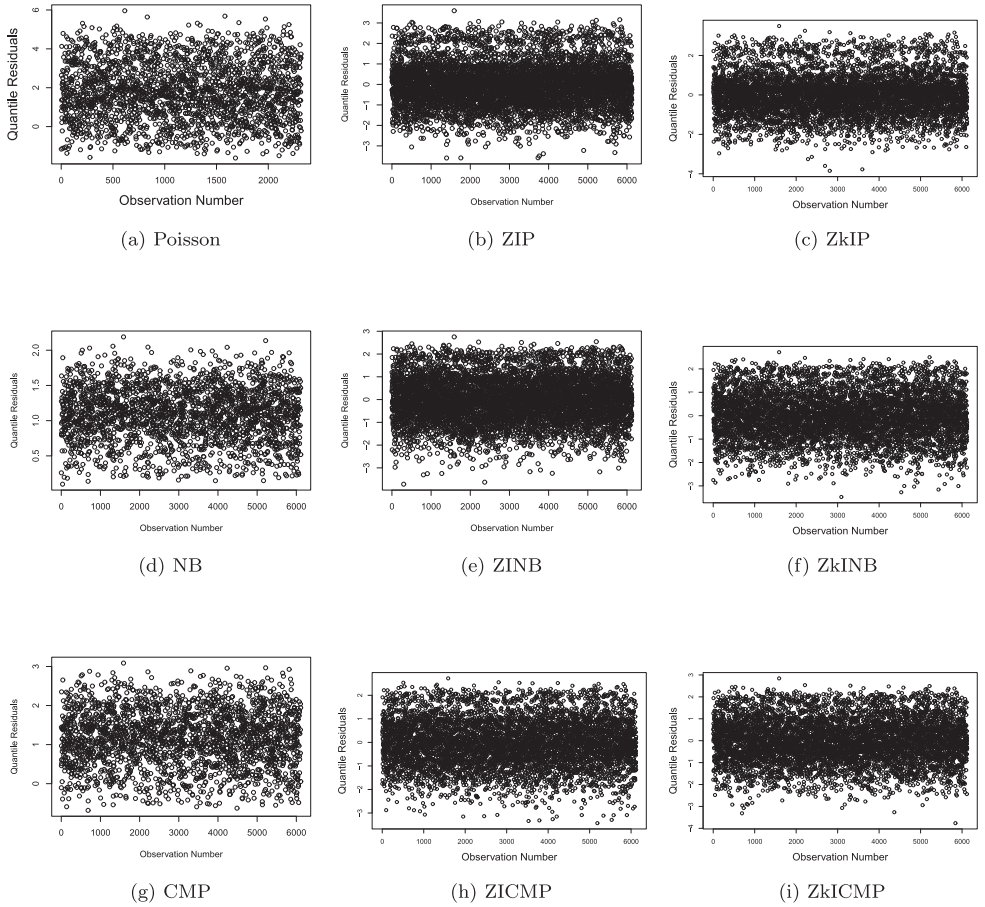
Table 12. Estimates, standard errors (in parentheses), and model diagnostics (log-likelihood, and AIC) for the exercise data example, after removing age and average systolic blood pressure from the model.

Parameters	ZkICMP	ZICMP	CMP	ZkINB	ZINB	NB	ZkIP	ZIP	Poisson
Intercept	0.3215*	0.4644*	−0.6683*	1.1956*	1.2475*	−0.1366	1.2856*	1.2993*	−0.1191*
	(0.0566)	(0.0567)	(0.0312)	(0.0915)	(0.0809)	(0.1753)	(0.0596)	(0.0582)	(0.0557)
BMI	0.0022*	0.0022*	0.0026*	0.0048*	0.0042*	0.0090*	0.0038*	0.0037*	0.0081*
	(0.0010)	(0.0010)	(0.0007)	(0.0021)	(0.0019)	(0.0042)	(0.0013)	(0.0013)	(0.0012)
Ratio	−0.0122*	−0.0121*	−0.0162*	−0.0243*	−0.0214*	−0.0549*	−0.0200*	−0.0191*	−0.0500*
	(0.0044)	(0.0045)	(0.0032)	(0.0091)	(0.0081)	(0.0176)	(0.0060)	(0.0058)	(0.0056)
Gender	0.0632*	0.0625*	0.1254*	0.1206*	0.1068*	0.3852*	0.0967*	0.0934*	0.3830*
	(0.0142)	(0.0142)	(0.0106)	(0.0291)	(0.0258)	(0.0563)	(0.0190)	(0.0186)	(0.0183)
Avg. Dias.	0.0023*	0.0023*	0.0024*	0.0046*	0.0041*	0.0076*	0.0040*	0.0038*	0.0075*
	(0.0006)	(0.0006)	(0.0004)	(0.0011)	(0.0010)	(0.0022)	(0.0008)	(0.0007)	(0.0007)
$\widehat{\gamma}$	0.5191	0.4319	–	0.5460	0.4557	–	0.5322	0.4866	–
	(0.0298)	(0.0276)		(0.0291)	(0.0270)		(0.0285)	(0.0265)	
$\widehat{\delta}$	−2.2581	–	–	−2.2714			−3.0767	–	–
	(0.1092)			(0.1130)			(0.2413)		
$\widehat{\nu}$	0.3817*	0.4605*	0.0009	–			–	–	–
	(0.0220)	(0.0221)	–						
$\widehat{r}$	–	–	–	0.2108*	0.1608*	4.2723*	–	–	–
				(0.0162)	(0.0125)	(0.1258)			
$\widehat{\pi}_1$	0.6034	0.6063	–	0.6101	0.6120	–	0.6194	0.6193	–
	(0.0071)	(0.0066)		(0.0063)	(0.0064)		(0.0064)	(0.0062)	
$\widehat{\pi}_2$	0.0375	–	–	0.0365			0.0168	–	–
	(0.0039)			(0.0015)			(0.0015)		
No. of param.	8	7	6	8	7	6	7	6	5
log L_{obs}	−9614.88	−9678.00	−11661.00	−9634.65	−9634.65	−10431.26	−9886.15	−9896.45	−17628.99
AIC	19245.75	19369.99	23335.00	19285.30	19399.60	20874.52	19786.30	19804.90	35267.97

*Statistically significant regression and dispersion parameters (at the $\alpha = 0.05$ significance level).

Table 13. Frequency comparisons for the exercise data example with BMI, ratio, gender and average diastolic as the covariates.

Count	Observed	ZkICMP	ZICMP	CMP	ZkINB	ZINB	NB	ZkIP	ZIP	Poisson
0	3805	3805.21	3805.15	2088.81	3804.32	3804.79	3639.16	3804.81	3805.01	937.38
1	181	181.86	176.08	1335.21	163.39	156.45	752.35	61.76	65.68	1577.03
2	252	236.90	250.72	891.23	240.91	252.65	415.28	157.45	167.19	1567.00
3	306	271.06	303.80	589.25	285.41	317.86	275.43	280.68	296.75	1045.61
4	200	272.14	315.67	397.18	285.17	329.36	199.66	362.90	382.43	569.43
5	482	482.05	297.58	270.74	482.14	304.25	151.84	482.47	398.57	268.11
6	143	222.79	263.40	201.68	222.16	262.14	122.63	342.38	357.61	132.05
7	254	180.38	211.40	121.33	176.13	205.98	93.64	256.97	266.86	34.12
8	79	148.93	170.96	99.89	141.95	162.99	80.76	186.59	192.23	16.62
9	40	113.97	127.16	69.97	107.70	120.30	66.84	116.01	118.63	6.73
10	229	84.71	90.93	50.85	79.84	86.07	56.37	65.72	66.61	1.48
11–12	90	58.44	59.95	33.97	55.89	57.73	46.79	31.98	32.20	0.30
13–14	61	42.67	41.56	26.35	41.53	40.96	40.74	16.61	16.51	0.08
ABE		614.79	838.31	4712.35	625.70	871.06	1706.98	1219.90	1159.24	7606.19
χ^2		434.72	563.45	4206.66	471.05	586.34	2094.44	1054.30	1217.99	119435.3

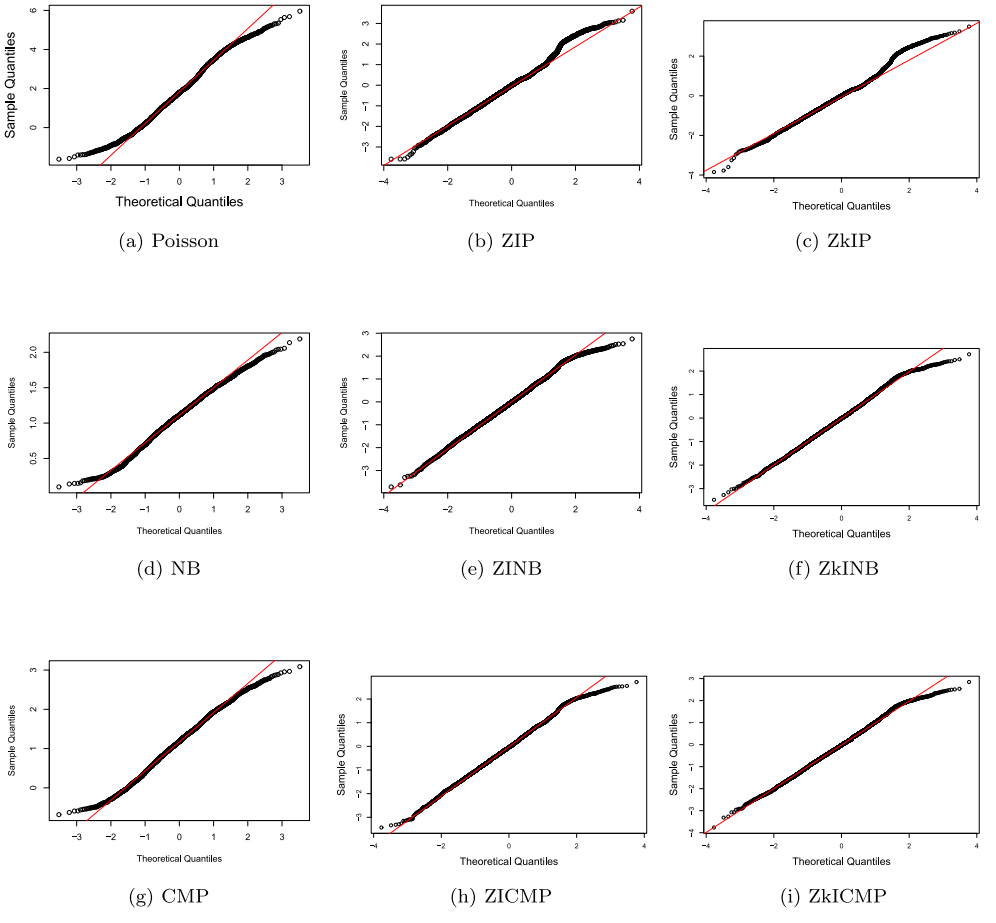


(a) Poisson (b) ZIP (c) ZkIP

(d) NB (e) ZINB (f) ZkINB

(g) CMP (h) ZICMP (i) ZkICMP

Figure 3. Randomized quantile residual plots for the exercise data example, where the rows reflect the results for the various models (Poisson, negative binomial, and Conway–Maxwell–Poisson, respectively), and the columns contain the respective base distribution, its zero-inflated version, and its zero-and-*k*-inflated analog. (a) Poisson, (b) ZIP, (c) ZkIP, (d) NB, (e) ZINB, (f) ZkINB, (g) CMP, (h) ZICMP, (i) ZkICMP.



(a) Poisson (b) ZIP (c) ZkIP

(d) NB (e) ZINB (f) ZkINB

(g) CMP (h) ZICMP (i) ZkICMP

Figure 4. QQ plots for the exercise data example, where the rows reflect the results for the various models (Poisson, negative binomial, and Conway–Maxwell–Poisson, respectively), and the columns contain the respective base distribution, its zero-inflated version, and its zero-and-k-inflated analog. (a) Poisson, (b) ZIP, (c) ZkIP, (d) NB, (e) ZINB, (f) ZkINB, (g) CMP, (h) ZICMP, (i) ZkICMP.

the normal quartiles. The ZINB and ZICMP QQ plot of residual quartiles does not provide a good fit with normal quartiles in the lower and upper tails, while the ZkINB and ZkICMP plots are mostly comparable (except for some differences in the tails). The ZkICMP QQ plot gives the best fit with a very small deviation from the straight line (Figure 4).

6. Discussion

This work introduces the reader to a zero- and k-inflated CMP regression model, and demonstrates its flexibility and ability to properly model both inflated frequencies at up to two count values, and additional data dispersion that remains when taking account of the excess frequency at 0 and $k > 0$. Simulated and real data examples illustrate model flexibility and superior performance when the response variable contains apparent data dispersion. As illustrated in [24], this work likewise demonstrates how addressing various contributors to data dispersion associate with the changing underlying dispersion measure.

The ZkICMP model is beneficial in that its form is less complex than other existing models yet it remains amenable and accommodating to various count data structures. Its form, however, accounts only for inflation at 0 and k, while it is possible to have more count values that are inflated. In such a situation, the generalized inflated models will be more appropriate or, for situations where count data are inflated at many different values, then one can consider alternatively utilizing a continuous distribution. Recall that, to test the inflation at 0 and k, the regularity conditions are not satisfied and the LRT statistic converges to a 50:50 chi-square distribution. This result not only applies to the ZkICMP model but is a general result to be used while testing for inflation at a boundary point.

We presented a model in which the covariates are linked to the Poisson mean parameter, λ_i. For illustrative purposes, we maintained constant assumptions for the dispersion and inflation components, however covariates can likewise be linked to these variables [2,24,28,29]. Accordingly, for example, one can reconsider (8) as $\log(\frac{\pi_{1i}}{\pi_{3i}}) = \mathbf{u_i}^T\boldsymbol{\gamma}$, $\log(\frac{\pi_{2i}}{\pi_{3i}}) = \mathbf{v_i}^T\boldsymbol{\delta}$, and $\log(\nu_i) = \mathbf{z_i}^T\boldsymbol{\eta}$, where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{p_1})^T$, $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_{p_2})^T$, and $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_{p_3})^T$. Depending on the number of covariates added to each component, the number of parameters to be estimated will increase. Variable selection techniques help to obtain a parsimonious model. The unknown parameters and their SEs can be estimated using optimization algorithms, however considering varying π_{1i}, π_{2i} and ν_i can make interpretations of the inflated and dispersion parameters a cumbersome task. The study of apparent dispersion is particularly enlightening as different levels and types of data dispersion may be (un)masked through consideration of a variable dispersion model [24], and we can expand our model to allow for such considerations as discussed here.

In this article, we use optimization routines to get the MLEs of the unknown parameters. Our future work involves the estimation of the parameters using the Expectation–Maximization (EM) method. The method ensures convergence and gives the estimates close to the MLE for all of the parameters. The EM algorithm gives the estimates of the unknown parameters but not the SEs. The SEs could be obtained using the approach given by Louis [41] or the bootstrap method [42].

Notes

1. The correlation between BMI and body weight was significantly high (0.9), thus we removed body weight from further consideration and analysis.
2. The CMP dispersion estimate $\widehat{\nu}$ is likewise clearly significant given its size, however it is not marked with an asterisk because it lacks an associated reported SE.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Kimberly F. Sellers http://orcid.org/0000-0001-6516-0548

References

[1] Cohen AC. Estimating the parameters of a modified Poisson distribution. J Amer Statist Assoc. 1960;55:139–143.

[2] Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics. 1992;34:1–14.
[3] Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the em algorithm. J R Stat Soc. 1977;39(1):1–38.
[4] Yau K, Lee A. Zero-inflated Poisson regression with random effects to evaluate an occupational injury prevention programme. Stat Med. 2001;20:2907–2920.
[5] Min Y, Agresti A. Random effect models for repeated measures of zero-inflated count data. Stat Model. 2005;5:1–19.
[6] Ghosh SK, Mukhopadhyay P, Lu JC. Bayesian analysis of zero-inflated regression models. J Statist Plann Inference. 2006;136:1360–1375.
[7] Agarwal DK, Gelfand AE, Citron-Pousty S. Zero-inflated models with application to spatial count data. Environ Ecol Stat. 2002;9:341–355.
[8] Saffari SE, Adnan R. Zero-inflated Poisson regression models with right censored count data. Matematika. 2011;27:21–29.
[9] Yang Y, Simpson DG. Conditional decomposition diagnostics for regression analysis of zero-inflated and left-censored data. Stat Methods Med Res. 2012;21:393–408.
[10] Shankar V, Milton J, Mannering F. Modeling accident frequencies as zero-altered probability processes: an empirical inquiry. Accid Anal Prev. 1997;29(6):829–837.
[11] Lee J, Mannering F. Impact of roadside features on the frequency and severity of run-off-roadway accidents: an empirical analysis. Accid Anal Prev. 2002;34:149–161.
[12] Kumara SSP, Chin HC. Modeling accident occurrence at signalized tee intersections with special emphasis on excess zeros. Traffic Inj Prev. 2003;4:53–57.
[13] Mullahy J. Specification and testing of some modified count data models. J Econ. 1986;33(3):341–365.
[14] Greene W. Accounting for excess zeros and sample selection in Poisson and negative binomial regression models. New York University, Leonard N. Stern School of Business, Department of Economics; 1994.
[15] Hall DB. Zero-inflated Poisson and binomial regression with random effects: a case study. Biometrics. 2000;56:1030–1039.
[16] Yau KKW, Wang K, Lee AH. Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros. Biom J. 2003;45:437–452.
[17] Ridout M, Hinde J, DeméAtrio CG. A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. Biometrics. 2001;57:219–223.
[18] Health Interview Survey National. 2015 National Health Interview Survey (NHIS) June 30, 2016 Public Use Data Release; 2019. Available from: https://www.cdc.gov/nchs/nhis/nhis_2015_data_release.htm
[19] Lin TH, Tsai MH. Modeling health survey data with excessive zero and *k* responses. Stat Med. 2012;32:1572–1583.
[20] Sheth-Chandra M, Chaganty N, Sabo R. A doubly-inflated poisson distribution and regression model. Cham: Springer; 2019. p. 131–145 (Chapter 7).
[21] Zhang C, Tian GL, Ng K. Properties of the zero-and-one inflated Poisson distribution and likelihood-based inference methods. Stat Interface. 2016;9:11–32.
[22] Alshkaki RSA. On the zero-one inflated Poisson distribution. Int J Stat Distrib Appl. 2016;2:42–48.
[23] Tang Y, Liu W, Xu A. Statistical inference for zero-and-one-inflated poisson models. Stat Theory Relat Fields. 2017;1:216–226.
[24] Sellers K, Shmueli G. Data dispersion: now you see it..now you don't. Commun Stat Theory Methods. 2013;42:1–14.
[25] Conway RW, Maxwell WL. A queuing model with state dependent service rates. J Ind Eng. 1962;12:132–136.
[26] Shmueli G, Minka TP, Kadane JB, et al. A useful distribution for fitting discrete data: revival of the Conway–Maxwell–Poisson distribution. R Stat Soc Ser C (Appl Stat). 2005;54:127–142.
[27] Sellers KF, Shmueli G. A flexible regression model for count data. Ann Appl Stat. 2010;4:943–961.

[28] Sellers KF, Raim A. A flexible zero-inflated model to address data dispersion. Comput Stat Data Anal. 2016;99:68–80.
[29] Sellers KF, Young DS. Zero-inflated sum of Conway–Maxwell–Poissons (ziscmp) regression. J Stat Comput Simul. 2019;89(9):1649–1673.
[30] Sellers KF, Shmueli G, Borle S. The COM-Poisson model for count data: a survey of methods and applications. Appl Stoch Models Bus Ind. 2011;28:104–116.
[31] Choo-Wosoba H, Levy SM, Datta S. Marginal regression models for clustered count data based on zero-inflated Conway–Maxwell–Poisson distribution with applications. Biometrics. 2016;72:606–618.
[32] Barriga GD, Louzada F. The zero-inflated Conway–Maxwell–Poisson distribution: Bayesian inference, regression modeling and influence diagnostic. Stat Methodol. 2014;21:23–34.
[33] SAS Institute Inc. SAS/ETS(R) 13.1 User's Guide; 2014. Available from: http://www.sas.com/
[34] Sellers K, Lotze T, Raim A. COMPoissonReg: Conway–Maxwell Poisson (COM-Poisson) Regression, version 0.6.1; 2018. Available from: https://cran.r-project.org/web/packages/COMPoissonReg/index.html
[35] R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2014. Available from: http://www.R-project.org
[36] Burnham KP, Anderson DR. Model selection and multimodel inference. New York: Springer; 2002.
[37] Chant D. On asymptotic tests of composite hypotheses in nonstandard conditions. Biometrika. 1974;61:291–298.
[38] Shapiro A. Asymptotic distribution of test statistics in the analysis of moment structures under inequality constraints. Biometrika. 1985;72:133–144.
[39] Akaike H. A new look at the statistical model identification. IEEE Trans Automat Control. 1974;19:716–723.
[40] Dunn P, Smyth G. Randomized quantile residuals. J Comput Graph Stat. 1996;5:236–244.
[41] Louis TA. Finding the observed information matrix when using the em algorithm. J R Stat Soc. 1982;44(2):226–233.
[42] Efron B, Tibshirani R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Stat Sci. 1986;1(1):54–75.

Appendices

Appendix 1. Elements of the Fisher information matrix

The unknown parameters in the ZkICMP regression model are given by the vector $\boldsymbol{\theta} = (\boldsymbol{\beta}, \gamma, \delta, \nu)$. The elements of the Fisher information matrix (11) are given by

$$\begin{aligned}
\frac{-\partial^2 \ell_{\text{obs}}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{T}} = & -\sum_{i:y_i=0} \frac{p_{0i}}{e^{\gamma}+p_{0i}} \frac{1}{Z}\left(\frac{2\lambda_i}{Z}\left(\frac{\partial Z}{\partial \lambda_i}\right)^2 - \frac{p_{0i}}{e^{\gamma}+p_{0i}}\frac{\lambda_i}{Z}\frac{\partial Z}{\partial Z} - \frac{\partial \lambda_i}{\partial \lambda_i} - \lambda_i \frac{\partial^2 Z}{\partial {\lambda_i}^2}\right) \times \lambda_i \mathbf{x_i}\mathbf{x_i^T} \\
& - \sum_{i:y_i=k} \frac{p_{ki}}{e^{\delta}+p_{ki}}\left(\frac{e^{\delta}}{e^{\delta}+p_{ki}}\lambda_i\left(\frac{k}{\lambda_i} - \frac{1}{Z}\frac{\partial Z}{\partial \lambda_i}\right)^2\right)\lambda_i \mathbf{x_i}\mathbf{x_i^T} \\
& - \sum_{i:y_i=k} \frac{p_{ki}}{e^{\delta}+p_{ki}}\left(\lambda_i\left(\left(\frac{1}{Z}\frac{\partial Z}{\partial \lambda_i}\right)^2 - \frac{1}{Z}\frac{\partial^2 Z}{\partial {\lambda_i}^2}\right) - \frac{1}{Z}\frac{\partial Z}{\partial \lambda_i}\right)\lambda_i \mathbf{x_i}\mathbf{x_i^T} \\
& + \sum_{i:y_i\neq 0,k}\left(\lambda_i\left(\left(\frac{1}{Z}\frac{\partial Z}{\partial \lambda_i}\right)^2 - \frac{1}{Z}\frac{\partial^2 Z}{\partial {\lambda_i}^2}\right) - \frac{1}{Z}\frac{\partial Z}{\partial \lambda_i}\right)\lambda_i \mathbf{x_i}\mathbf{x_i^T} \\
\frac{-\partial^2 \ell_{\text{obs}}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta} \partial \gamma} = & -\sum_{i:y_i=0} \frac{e^{\gamma}}{(e^{\gamma}+p_{0i})^2} p_{0i}\lambda_i \mathbf{x_i}
\end{aligned}$$

$$\frac{-\partial^2 \ell_{\text{obs}}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta} \partial \delta} = \sum_{i:y_i=k} \frac{e^{\delta}}{(e^{\delta}+p_{ki})^2} p_{ki}(k-\lambda_i)\mathbf{x_i}$$

$$\begin{aligned}\frac{-\partial^2 \ell_{\text{obs}}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta} \partial \nu} &= \sum_{i:y_i=0} \frac{p_{0i}}{e^{\gamma}+p_{0i}} \frac{1}{Z}\left(\frac{\partial^2 Z}{\partial \nu \partial \lambda_i} - \frac{1}{Z}\frac{\partial Z}{\partial \nu}\frac{\partial Z}{\partial \lambda_i} - \frac{e^{\gamma}}{e^{\gamma}+p_{0i}}\frac{1}{Z}\frac{\partial Z}{\partial \nu}\frac{\partial Z}{\partial \lambda_i}\right)\lambda_i \mathbf{x_i} \\ &+ \sum_{i:y_i=k} \frac{p_{ki}}{e^{\delta}+p_{ki}}\left(\frac{1}{Z}\frac{\partial^2 Z}{\partial \nu \partial \lambda_i} - \frac{1}{Z}\frac{\partial Z}{\partial \nu}\frac{\partial Z}{\partial \lambda_i}\right)\lambda_i \mathbf{x_i} \\ &+ \sum_{i:y_i=k} \frac{p_{ki}}{e^{\delta}+p_{ki}}\left(\frac{e^{\delta}}{e^{\delta}+p_{ki}}\left(\log k! + \frac{1}{Z}\frac{\partial Z}{\partial \nu}\right)\left(\frac{k}{\lambda_i} - \frac{1}{Z}\frac{\partial Z}{\partial \lambda_i}\right)\right)\lambda_i \mathbf{x_i} \\ &+ \sum_{i:y_i\neq 0,k} \frac{1}{Z}\left(\frac{\partial^2 Z}{\partial \nu \partial \lambda_i} - \frac{1}{Z}\frac{\partial Z}{\partial \nu}\frac{\partial Z}{\partial \lambda_i}\right)\lambda_i \mathbf{x_i}\end{aligned}$$

$$\frac{-\partial^2 \ell_{\text{obs}}(\boldsymbol{\theta})}{\partial \gamma^2} = \frac{n e^{\gamma}(1+e^{\delta})}{(1+e^{\gamma}+e^{\delta})^2} - \sum_{i:y_i=0} \frac{e^{\gamma} p_{0i}}{(e^{\gamma}+p_{0i})^2}$$

$$\frac{-\partial^2 \ell_{\text{obs}}(\boldsymbol{\theta})}{\partial \gamma \partial \delta} = \frac{-n e^{\gamma+\delta}}{(1+e^{\gamma}+e^{\delta})^2}$$

$$\frac{-\partial^2 \ell_{\text{obs}}(\boldsymbol{\theta})}{\partial \gamma \partial \nu} = -\sum_{i:y_i=0} \frac{e^{\gamma} p_{0i}}{(e^{\gamma}+p_{0i})^2}\frac{1}{Z}\frac{\partial Z}{\partial \nu}$$

$$\frac{-\partial^2 \ell_{\text{obs}}(\boldsymbol{\theta})}{\partial \delta^2} = \frac{n e^{\delta}(1+e^{\gamma})}{(1+e^{\gamma}+e^{\delta})^2} - \sum_{i:y_i=k} \frac{e^{\delta} p_{ki}}{(e^{\delta}+p_{ki})^2}$$

$$\frac{-\partial^2 \ell_{\text{obs}}(\boldsymbol{\theta})}{\partial \delta \partial \nu} = -\sum_{i:y_i=k} \frac{e^{\delta} p_{ki}}{(e^{\delta}+p_{ki})^2}\left(\log k! + \frac{1}{Z}\frac{\partial Z}{\partial \nu}\right)$$

$$\begin{aligned}\frac{-\partial^2 \ell_{\text{obs}}(\boldsymbol{\theta})}{\partial \nu^2} &= \sum_{i:y_i=0} \frac{p_{0i}}{e^{\gamma}+p_{0i}}\frac{1}{Z}\left(\frac{\partial^2 Z}{\partial \nu^2} - \frac{p_{0i}+2e^{\gamma}}{e^{\gamma}+p_{0i}}\frac{1}{Z}\left(\frac{\partial Z}{\partial \nu}\right)^2\right) \\ &+ \sum_{i:y_i=k} \frac{p_{ki}}{e^{\delta}+p_{ki}}\left(\frac{1}{Z}\frac{\partial^2 Z}{\partial \nu^2} - \left(\frac{1}{Z}\frac{\partial Z}{\partial \nu}\right)^2 - \frac{e^{\delta}}{e^{\delta}+p_{ki}}\left(\log k! + \frac{1}{Z}\frac{\partial Z}{\partial \nu}\right)^2\right) \\ &+ \sum_{i:y_i\neq 0,k} \left(\frac{1}{Z}\frac{\partial^2 Z}{\partial \nu^2} - \left(\frac{1}{Z}\frac{\partial Z}{\partial \nu}\right)^2\right).\end{aligned}$$

These formulae were used to calculate the SEs of the MLEs of the parameters in ZkICMP regression models.

Appendix 2. Exercise data results with all considered covariates

Table A1 reports the estimated coefficients and SEs, log-likelihood, and AIC for all of the considered count models. This table reports that the covariates associated with age and average systolic BP are not statistically significant for most of the considered models (ZIP, ZkIP, ZkINB, ZICMP and ZkICMP).

Table A1. Estimates, standard errors (in parentheses), and model diagnostics (log-likelihood, and AIC) for the exercise data example.

Parameters	ZkICMP	ZICMP	CMP	ZkINB	ZINB	NB	ZkIP	ZIP	Poisson
Intercept	0.3419*	0.4814*	−0.6003*	1.2415*	1.2818*	0.0873	1.3109*	1.3221*	0.0483
	(0.0668)	(0.0673)	(0.0399)	(0.1194)	(0.1059)	(0.2321)	(0.0772)	(0.0755)	(0.0726)
Age	−0.0006	−0.0006	−0.0026*	−0.0010	−0.0008	−0.0090*	−0.0005	−0.0005	−0.0079*
	(0.0004)	(0.0004)	(0.0003)	(0.0009)	(0.0008)	(0.0017)	(0.0006)	(0.0006)	(0.0005)
BMI	0.0025*	0.0025*	0.0034*	0.0052*	0.0046*	0.0128*	0.0041*	0.0039*	0.0115*
	(0.0010)	(0.0010)	(0.0007)	(0.0022)	(0.0019)	(0.0043)	(0.0014)	(0.0013)	(0.0012)
Ratio	−0.0113*	−0.0112*	−0.0118*	−0.0229*	−0.0201	−0.0402*	−0.0191*	−0.0183*	−0.0372*
	(0.0045)	(0.0045)	(0.0032)	(0.0092)	(0.0082)	(0.0178)	(0.0060)	(0.0059)	(0.0056)
Gender	0.0641*	0.0635*	0.1253*	0.1225*	0.1086	0.3871*	0.0979*	0.0946*	0.3864*
	(0.0143)	(0.0144)	(0.0106)	(0.0296)	(0.02617)	(0.0568)	(0.0193)	(0.0188)	(0.0185)
Avg. Sys.	−0.0003	−0.0003	−0.0007*	−0.0006	−0.0004	−0.0019	−0.0003	−0.0003	−0.0016*
	(0.0005)	(0.0005)	(0.0004)	(0.0011)	(0.0010)	(0.0020)	(0.0007)	(0.0007)	(0.0007)
Avg. Dias.	0.0027*	0.0027*	0.0036*	0.0052*	0.0046*	0.0109*	0.0044*	0.0042*	0.0108*
	(0.0006)	(0.0006)	(0.0004)	(0.0013)	(0.0011)	(0.0023)	(0.0008)	(0.0008)	(0.0008)
$\widehat{\gamma}$	0.5169	0.4301	–	0.5451	0.4549	–	0.5321	0.4865	–
	(0.0298)	(0.0277)		(0.0291)	(0.0270)		(0.0286)	(0.0265)	
$\widehat{\delta}$	−2.2577	–	–	−2.2708			−3.0736	–	–
	(0.1090)			(0.1129)			(0.2406)		
$\widehat{\nu}$	0.3799*	0.4588*	0.0012	–			–	–	–
	(0.0220)	(0.0221)	–						
$\widehat{r}$	–	–	–	0.2114*	0.1611*	4.2086*	–	–	–
				(0.0163)	(0.0125)	(0.1245)			
$\widehat{\pi}_1$	0.6029	0.6059	–	0.6099	0.6118	–	0.6194	0.6193	–
	(0.0071)	(0.0066)		(0.0063)	(0.0064)		(0.0064)	(0.0062)	
$\widehat{\pi}_2$	0.0376	–	–	0.0365			0.0168	–	–
	(0.0039)			(0.0040)			(0.0015)		
No. of param.	10	9	8	10	9	8	9	8	7
log L_{obs}	−9613.15	−9676.54	−11609.00	−9633.60	−9691.85	−10411.49	−9885.30	−9895.70	−17466.45
AIC	19246.30	19371.08	23234.00	19287.20	19401.70	20838.98	19788.60	19807.40	34946.90

*Statistically significant regression and dispersion parameters (at the 5% significance level).