Article

Sociological Methods & Research
1-36

DOI: 10.1177/0049124118782535
journals.sagepub.com/home/smr
SAGE

Generalized Inflated Discrete Models: A Strategy to Work with Multimodal Discrete Distributions

Tianji Cai[1], Yiwei Xia[1] and Yisu Zhou[2]

Abstract

Analysts of discrete data often face the challenge of managing the tendency of inflation on certain values. When treated improperly, such phenomenon may lead to biased estimates and incorrect inferences. This study extends the existing literature on single-value inflated models and develops a general framework to handle variables with more than one inflated value. To assess the performance of the proposed maximum likelihood estimator, we conducted Monte Carlo experiments under several scenarios for different levels of inflated probabilities under multinomial, ordinal, Poisson, and zero-truncated Poisson outcomes with covariates. We found that ignoring the inflations leads to substantial bias and poor inference of the inflations—not only for the intercept(s) of the inflated categories but other coefficients as well. Specifically, higher values of inflated probabilities are associated with larger biases. By contrast, the generalized inflated discrete models (GIDMs) perform well with unbiased estimates and satisfactory coverages even when the number of parameters that need to be estimated is quite large. We showed that model fit criteria, such as Akaike information criterion, could be

[1] Department of Sociology, University of Macau, Taipa, Macau, China
[2] Faculty of Education, University of Macau, Taipa, Macau, China

Corresponding Author:
Tianji Cai, Department of Sociology, University of Macau, Avenida da Universidade, Taipa, Macau, China.
Email: tjcai@umac.mo

used in selecting the appropriate specifications of inflated models. Lastly, the GIDM was implemented using large-scale health survey data as a comparison to conventional modeling approaches such as various Poisson and Ordered Logit models. We showed that the GIDM fits the data better in general. The current work provides a practical approach to analyze multimodal data that exists in many fields, such as heaping in self-reported behavioral outcomes, inflated categories of indifference and neutral in attitude surveys, large amounts of zero, and low occurrences of delinquent behaviors.

Keywords

multiple data inflations, generalized inflated discrete models, maximum likelihood estimator, probabilities of inflation, Monte Carlo experiments

Introduction

Many of the quantitative variables used in social science studies are discrete in nature. Respondents typically choose from a limited number of categories on ordinal scales, and researchers have long recognized that such data have the tendency to include inflated values. Data inflation may take various forms. It could be that respondents forget the information, or round to a nearby number of convenience (Crawford, Weiss, and Suchard 2015); it can also be the result of hiding one's ignorance in situations where face-saving is deemed important (Bagozzi and Mukherjee 2012). In other scenarios, measures on counts or summarized items, for example, the number of hospital visits and the delinquency scale, may naturally concentrate on values of zero or low occurrences such as one or two.

Depending on the assumptions about how the inflation was generated, two common modeling strategies are available to address the inflation on one value. The first strategy assumes the inflation is a form of data reporting error, referred to as "heaping," and then adds parametric components that correspond to rounding in the model (e.g., Heitjan and Rubin 1990; Pickering 1992), while the second strategy parameterizes the inflation as a result of the mixture of two distributions—a binary part for inflation and a regular part for outcome—and proposes inflated models (e.g., Lambert 1992; Hall 2001; Vieira, Hinde, and Demetrio 2010).

Yet, none of the approaches enable scholars to deal with scenarios of multiple data inflations resulting in multiple modes in empirical data distribution. Such scenarios are not uncommon in empirical research. Li and Hitt (2008) showed that the distribution of online consumer reviews for many products tends to be bimodal, with reviews split by two extremes such as one or five on a one-to-five scale. The bipolar distribution of reviews may be due to self-selection—people who have extreme experiences are more likely to leave a review (Li and Hitt 2008)—or it may be a case of fraudulent reviews driven by economic incentives (Luca and Zervas 2016). Sometimes clumping of values occurs because of the process of scale construction. For example, in a health survey that poses the question of how many days a respondent has smoked cigarettes over the past 30 days, the distribution of responses may concentrate on the values of 0 and 30, because some people never smoke and some smoke every day (Farrell, Fry, and Harris 2011). Likewise, in a sociological study of adolescents, self-reported weekly hours on unstructured socialization could have either extremely low or high values (Basner et al. 2007).

Although there is a growing demand of extending models that handle single inflation to situations of multiple inflations, studies on multimodal distributions or inflation on multiple values are sparse. A conventional approach analysts have commonly resorted to is employing the standard Multinomial or Poisson models, even when the proportions for certain values in observations exceed the predicted probabilities. However, this might lead to biased estimates and incorrect inferences (Lambert 1992). Bagozzi (2016) suggested that in a dyad-year study of international relationships, where the dyadic pairing produces country-year data that are used to indicate the general relationship between specific countries, that in most cases, the status quo tended toward "peace." Yet the prevalence of peace over other possible outcomes could be due to the lack of meaningful relationship status choices between the two countries or simply because the question was irrelevant because of geographic distance or a lack of political–economic interaction between the countries being considered. Including the mixed "peace" category as the baseline reference created a bias on the estimated effects in indeterminate directions and led to faulty inferences. Similarly, in an analysis of the number of children born within a family, the majority of cases were reported along a distribution of zero, one, and two. Poston and McKibben (2003) stated that the regular Poisson model underpredicted the number of children born at the values of zero and one. Although the zero-inflated Poisson (ZIP) model provided a more consistent prediction at the value zero, it still underpredicted at the value of two.

Drawn from item response theory, some of the more recent work has adopted a latent variable approach. For example, assuming the inflated responses is a mixture of multiple groups, a latent class membership and a latent group specified random effect can be used to account for differences in response style at both the group and individual levels (e.g., Finkelman et al. 2011; Magnus and Thissen 2017). However, the primary goal of those works is not to directly address the inflation on particular value(s), and most of the research has focused on Poisson outcomes; therefore, they are fundamentally different from the strategies we mentioned above, and a direct comparison of their work is beyond the scope of this study. Instead, to fill the gap, the current study aims to develop a general modeling strategy to handle multimodal discrete distributions, such as multinomial, cumulative logit (CL; ordered), Poisson, and zero-truncated Poisson outcomes. Potential implications for this strategy may include health economics, political science, psychology, criminology, sociology, and educational research.

This article is organized as follows: The second section outlines the general framework with detailed model specifications and estimation methods. In the third section, we set up several Monte Carlo experiments, and the results from the simulation experiments are reported. The fourth section shows an example of empirical data analysis. A conclusion and some perspectives are provided in the fifth section. Technical details and exemplary code are included in Online Appendix.

Generalized Inflated Discrete Models

Models that handle single inflated values such as zero have been proposed since the early 1990s. Lambert (1992) developed a ZIP model to address the extra zeros in count data. Assuming zero counts is generated from two processes, the ZIP model has two corresponding components when the outcome is equal to zero. The first component is the probability of a binary distribution that generates extra zeros, and the second is the probability of zero from a Poisson distribution. For example, for a nonnegative integer outcome Y_i with extra zeros, the probability mass function (PMF) could be written as:

$$p(Y_i = y_i|\lambda_i, \pi) = \begin{cases} \pi + (1-\pi) \times e^{-\lambda_i}, \text{if } y_i = 0 \\ (1-\pi) \times \dfrac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!}, \text{if } y_i > 0 \end{cases}, \tag{1}$$

where λ_i is the expected Poisson outcome for the ith individual, and π is the probability of extra zeros from a binary distribution, for example, logistic or probit.

The ZIP model has been further developed by many others, for instance, to deal with overdispersion and heterogeneity by incorporating a zero-inflated negative binomial model (e.g., Ridout, Hinde, and Demetrio 2001; Mwalili, Lesaffre, and Declerck 2008), adding additional random effects (e.g., Monod 2014), or both (e.g., Moghimbeigi et al. 2009), and mixing multiple groups (e.g., Lim, Li, and Yu 2014).

Lin and Tsai (2013) further extended the ZIP model to include inflations at both zero and the value k. To illustrate, a response Y_i with excessive values of zero and k, the PMF is defined as follows:

$$p(Y_i = y_i|\lambda_i, \pi, \phi) = \begin{cases} \pi + (1-\pi-\phi) \times e^{-\lambda_i}, \text{if } y_i = 0 \\ \phi + (1-\pi-\phi) \times \dfrac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!}, \text{if } y_i = k \\ (1-\pi-\phi) \times \dfrac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!}, \text{if } y_i > 0 \text{ and } y_i \neq k \end{cases}, \quad (2)$$

where π and ϕ is the probability of extra zeros and the value k, respectively.

Begum, Mallick, and Pal (2014) suggested a general framework of inflated values for Poisson outcomes. To simplify our notation, we drop the subscript for individuals and use the subscript to differentiate the inflated values. For example, suppose a discrete random variable Y has inflated probabilities at values $k_1, ..., k_m \in \{0, 1, 2, ...\}$, the PMF could be written as:

$$p(Y = k|\lambda, \pi_i, 1 \leq i \leq m) = \begin{cases} \pi_i + \left(1 - \sum_{i=1}^{m} \pi_i\right) \times p(k|\lambda), \text{if } k = k_1, ..., k_m \\ \left(1 - \sum_{i=1}^{m} \pi_i\right) \times p(k|\lambda), \text{if } k \neq k_i, 1 \leq i \leq m \end{cases}, \quad (3)$$

where $p(Y = k|\lambda)$ is a regular Poisson PMF with the parameter λ for $k = 0, 1, 2, ...$; and π_i is the probability of inflation at the value k_i with $1 \leq i \leq m$, and $\sum_{i=1}^{m} \pi_i \in (0, 1)$.

For other types of discrete outcomes, such as binary, multinomial, or ordinal, various single value inflated models were developed, including binary choice model with misclassification (Hausman, Abrevaya, and Scott-Morton 1998), zero-inflated Bernoulli model (Diop, Diop, and Dupuy 2016), zero-inflated binomial model (Hall 2001; Vieira, Hinde, and Demetrio 2010), zero-inflated ordered (ZIO) probit model (Harris and Zhao 2007), baseline or zero-inflated multinomial logit model (Bagozzi 2016; Diallo, Diop, and Dupuy 2017), and middle category inflated ordered model

(Bagozzi and Mukherjee 2012). Similar extensions have been made to incorporate inflations other than zero for multinomial or ordinal outcomes (e.g., Sweeney, Haslett, and Parnell 2017).

Following Begum et al. (2014), a further generalization could be made if the PMF is replaced by other discrete distributions, for example, multinomial, negative binomial, and so on. For instance, if $p(k|\theta)$ denotes a multinomial PMF with total K categories, then equation (1) is transformed into:

$$p(Y_i = k|\boldsymbol{\beta}_k,\ \pi_i, 1 \le i \le m) = \begin{cases} \pi_i + \left(1 - \sum_{i=1}^{m} \pi_i\right) \times \dfrac{e^{x_i \boldsymbol{\beta}_k}}{1 + \sum_{k=1}^{K} e^{x_i \boldsymbol{\beta}_k}}, \text{if } k = k_1, \ldots, k_m \\ \left(1 - \sum_{i=1}^{m} \pi_i\right) \times \dfrac{e^{x_i \boldsymbol{\beta}_k}}{1 + \sum_{k=1}^{K} e^{x_i \boldsymbol{\beta}_k}}, \text{if } k \ne k_i, 1 \le i \le m \end{cases}, \tag{4}$$

where $\boldsymbol{\beta}_k$ is the vector of parameters for the kth category in the multinomial distribution. Similarly, if a CL (ordered) PMF is specified, equation (1) could be expressed as follows:

$$p(Y_i \le k|\boldsymbol{\beta},\ \pi_i, 1 \le i \le m) = \begin{cases} \pi_i + \left(1 - \sum_{i=1}^{m} \pi_i\right) \times pr(Y_i \le k), \text{if } k = k_1, \ldots, k_m \\ \left(1 - \sum_{i=1}^{m} \pi_i\right) \times pr(Y_i \le k), \text{if } k \ne k_i, 1 \le i \le m \end{cases}, \tag{5}$$

where $pr(Y_i \le k) = \frac{1}{1+\exp(\alpha_k + \mathbf{x}_i\boldsymbol{\beta})} - \frac{1}{1+\exp(\alpha_{k-1} + \mathbf{x}_i\boldsymbol{\beta})}$ for $1 < k \le K$, and $pr(Y_i \le 1) = \frac{1}{1+\exp(\alpha_1 + \mathbf{x}_i\boldsymbol{\beta})}$. The probability of inflation at the value k_i, π_i, could also depend on covariates. For example, if a logit model is specified,

$$\pi_i = \frac{1}{1 + \exp(-\mathbf{z}_i\boldsymbol{\gamma})},$$

where $\mathbf{z}_i$ and $\boldsymbol{\gamma}$ is the vector of predictors for the ith observation and the vector of corresponding parameters, respectively. A probit model could be derived if a probit function is specified for π_i.

Once the model is specified, the full likelihood function can be constructed. Suppose the random variable Y follows a multinomial distribution with a total of five categories and has the inflated probabilities at the values 1 and 3. Defined $\boldsymbol{\theta} = (\boldsymbol{\gamma}_1', \boldsymbol{\gamma}_3', \boldsymbol{\beta}_1', \boldsymbol{\beta}_2', \boldsymbol{\beta}_3', \boldsymbol{\beta}_4')'$. The log-likelihood function of $\boldsymbol{\theta}$ for predictors $\boldsymbol{X}$ and $\boldsymbol{Z}$ can be specified as:

$$
\begin{aligned}
\log\ L(\boldsymbol{\theta}) \;=\; & \sum_{s=1}^{n}\Bigg\{ I(y_s=1)\times\log\left(\pi_1+(1-\pi_1-\pi_3)\times\frac{e^{\mathbf{X}_s\boldsymbol{\beta}_1}}{h_s(\boldsymbol{\beta})}\right) \\
& +I(Y_s=2)\times\Big(\log(1-\pi_1-\pi_3)+\mathbf{X}_s\boldsymbol{\beta}_2-\log(h_s(\boldsymbol{\beta}))\Big) \\
& +I(Y_s=3)\times\log\left(\pi_3+(1-\pi_1-\pi_3)\times\frac{e^{\mathbf{X}_s\boldsymbol{\beta}_3}}{h_s(\boldsymbol{\beta})}\right) \\
& +I(Y_s=4)\times\Big(\log(1-\pi_1-\pi_3)+\mathbf{X}_s\boldsymbol{\beta}_4-\log(h_s(\boldsymbol{\beta}))\Big) \\
& +I(Y_s=5)\times\Big(\log(1-\pi_1-\pi_3)+\mathbf{X}_s\boldsymbol{\beta}_5-\log(h_s(\boldsymbol{\beta}))\Big)\Bigg\}+c,
\end{aligned}
\tag{6}
$$

where $I(y_s = k_i)$ is an indicator function for $y_s = k_i$, $h_s(\boldsymbol{\beta}) = 1 + \sum_{k=1}^{4} e^{\mathbf{X}_s\boldsymbol{\beta}_k}$, and c is a constant that does not depend on the vector of the unknown parameters $\boldsymbol{\theta}$ for $s = 1,\ldots,n$. The inflation probabilities π_1 and π_3 depend on $\boldsymbol{Z}$ via a logit link function specified as:

$$
\pi_1 = \frac{1}{1+\exp(-\mathbf{Z}_s\gamma_1)} \text{ and } \pi_3 = \frac{1}{1+\exp(-\mathbf{Z}_s\gamma_3)}.
$$

The maximum likelihood estimator of $\boldsymbol{\theta}$ is the solution of the score equations

$$
\frac{\partial \log L(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = 0. \tag{7}
$$

The information matrix can be obtained by taking the second derivatives of the log likelihood. The detailed derivations for the above example and a general k-category multinomial logit model are given in Online Appendix A. Various methods have been proposed to find the solutions for the unknown parameters, for example, method of moments (Begum et al. 2014), direct maximum likelihood (Bagozzi 2016; Begum et al. 2014; Diallo et al. 2017), and maximum likelihood via the expectation−maximization algorithm (Begum et al. 2014; Su et al. 2013). Furthermore, Diallo et al. (2017) provided a rigorous investigation of the maximum likelihood estimator in terms of the identifiability, existence, consistency, and asymptotic normality under classical regularity conditions.

The implementation of the generalized inflated models usually requires users to construct the log-likelihood function and its gradient and then solve the score function by a Newton–Raphson type of algorithm. A direct implementation of maximum likelihood for the five-category example above using the commercial package SAS/IML (1990) is provided in Online Appendix B. To facilitate further use of the generalized inflated models, we also provided an implementation for the direct maximum likelihood method in the commercial package SAS by using PROC NLMIXED (SAS 2013; see Online Appendix C for details). The PROC NLMIXED offers great flexibility in specifying various likelihood functions and powerful capabilities to conduct numerical computations. Not only are standard single-category inflated models (Voronca, Egede, and Gebregziabher 2014) allowed, but multicategory inflated discrete ones as well. In addition, further extension can be easily made to manage clustering due to longitudinal or hierarchical data structures by adding a random effect in the log-likelihood function above.

Monte Carlo Experiments

To evaluate the performance of the maximum likelihood estimator derived above, we conducted a series of Monte Carlo experiments for the multinomial, CL (ordered), Poisson, and zero-truncated Poisson models with two inflated values, for example, at 1 and 3. Four independent variables X_1 to X_4 were generated from uniform $U(2,5)$, normal $N(1,1.5)$, exponential $\varepsilon(1)$, and Bernoulli distributions $B(.3)$, accordingly. Although the covariates in both the inflated and the regular model parts could be identical without altering the results, to ensure exclusion restrictions and gain possible enhancements on the precision of parameter estimates, we varied the set of predictors by the outcomes (Bagozzi and Mukherjee 2012; Harris and Zhao 2007). For example, in the experiment of multinomial models, for the categories "1" and "3," X_1 to X_3 were included with different coefficients; for the category "2," X_2 to X_4 were used; and all X_1 to X_4 were enclosed for the category "4."[1]

We considered two scenarios: Case (i)—the probabilities of inflations are fixed and Case (ii)—the probabilities of inflations are covariate-dependent. For Case (i), the inflation probabilities take the value of .05 to .20 by a step of .05. For Case (ii), to see whether the estimation is stable, two additional random variables from normal $N(-1,1)$ and Bernoulli $B(.3)$ were included for the inflation probabilities π_1 and π_3, respectively.

The parameter vector of the inflated values was chosen to make the average of inflation probabilities within each sample as .05, .10, .15, and .20, accordingly. Each of experiments was replicated 500 times with a sample size of 3,000. The exemplary code for Case (i) is given in Online

Appendix B (SAS/IML implementation) and C (NLMIXED implementation). The results and the rest of codes are available upon request.

In addition to the true model and naive model, both of which ignore the inflations, it is helpful to consider a scenario in which researchers don't know precisely which categories are inflated. We also estimated models where the inflated categories were incorrectly specified, for example, at values 1 and 4.

The quality of estimates is evaluated by using the standardized bias, the root standardized mean square error (RSMSE), and the coverage rate (CR). The *SB* for parameter θ is defined as:

$$\mathrm{SB}(\hat{\theta}) = E(\hat{\theta} - \theta)/\theta \approx \frac{1}{N}\left(\frac{1}{\theta}\sum_{s=1}^{N}(\hat{\theta}_s - \theta)\right). \quad (8)$$

The RSMSE for parameter θ is calculated using the following formula:

$$\mathrm{RSMSE}(\hat{\theta}) = \sqrt{E(\hat{\theta} - \theta)^2/\theta^2} \approx \sqrt{\frac{1}{N}\left(\frac{1}{\theta^2}\sum_{s=1}^{N}(\hat{\theta}_s - \theta)^2\right)}, \quad (9)$$

where N denotes the number of replicates; $\hat{\theta}_s$ refers to the estimated value of the parameter θ from the sample s. The CR is calculated as the percentage of the true parameter that falls within the 95 percent confidence region for each of replicated samples.

$$\mathrm{CR} = \frac{\sum_{s=1}^{N} I(\hat{\theta}_s - 1.96 \times SE(\hat{\theta}_s) \le \theta \le \hat{\theta}_s + 1.96 \times SE(\hat{\theta}_s))}{N},$$

where $SE(\hat{\theta}_s)$ is the estimated standard error for $\hat{\theta}_s$ from the sample s.

Results

Figure 1 presents the comparison of the estimates between the generalized inflated multinomial (GIM) model and the naive multinomial model for the simulated data for an outcome from "1" to "5," with inflation at values "1" and "3." With the reference group at "5," the second subscript is used to differentiate the categories, for instance, β_{01} to β_{04} are estimates for the intercept with categorical data ranging from "1" to "4"; β_{11} to β_{14} are estimates for the first independent variable, and so on. Since only category "4" has four predictors, β_{44} refers the coefficent of X_4 for category "4."

Using the last category "5" as the reference, all estimates obtained from the GIM are unbiased with nearly 100% CR. The CRs were plotted on the

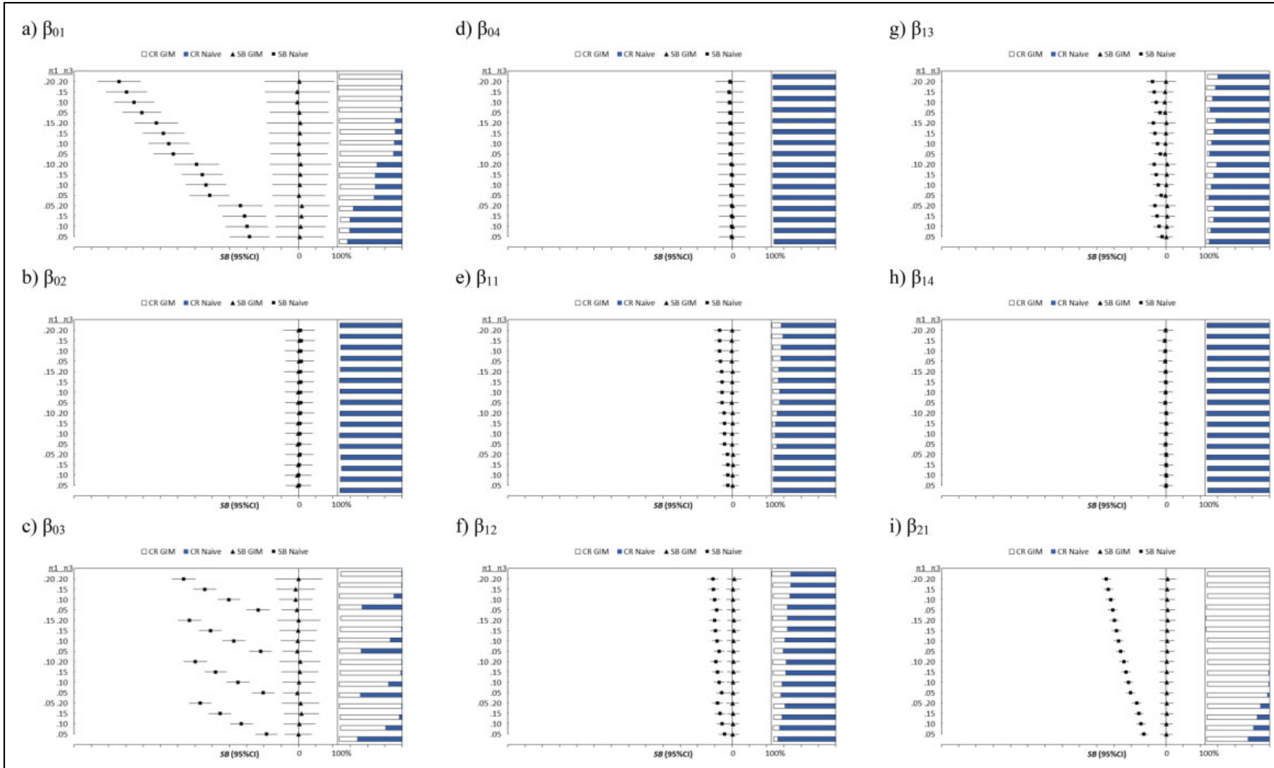

Figure 1. Plots of the SB, RSMSE, and CR of the generalized inflated multinomial model plotted against the fixed π_1 and π_3. The dot represents the SB, with the RSMSE as the error bar. The horizontal bar represents the CR at each level of π_1 and π_3. SB = standardized bias; RSMSE = root standardized mean square error; CR = coverage rate.

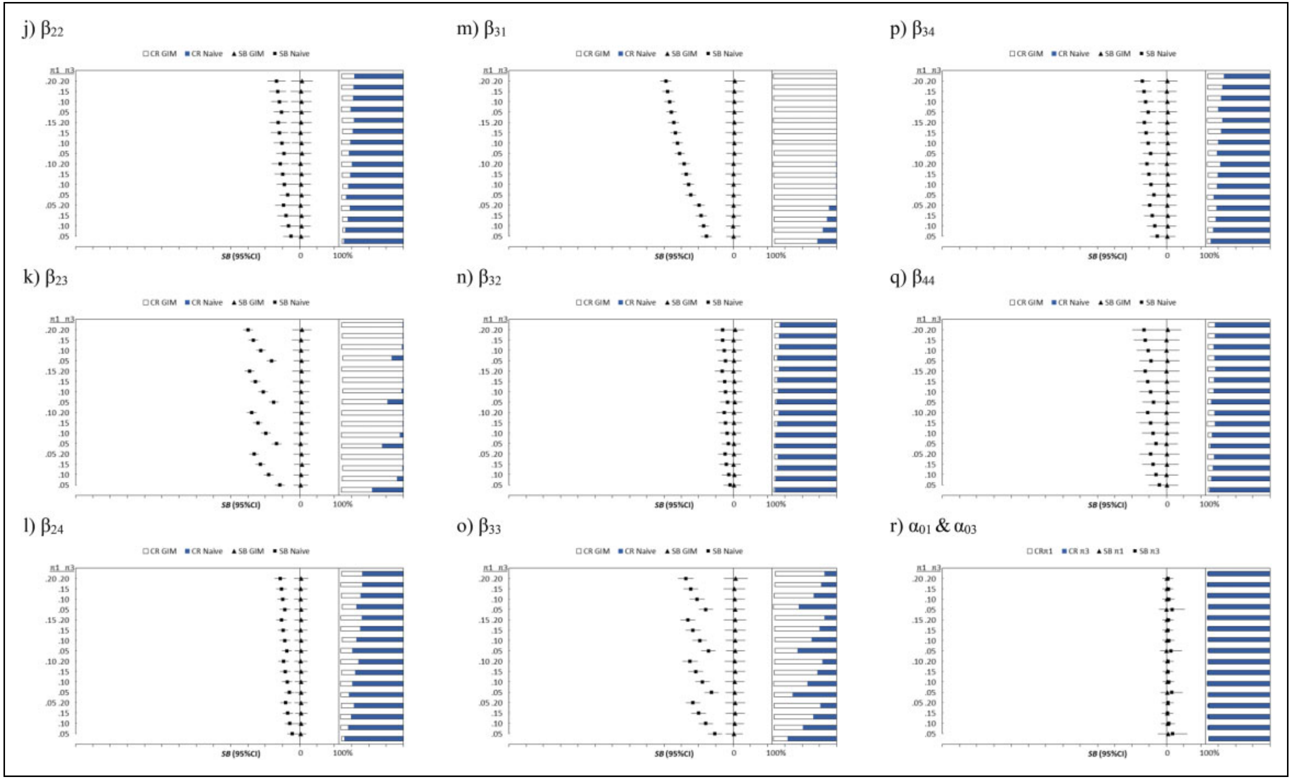

Figure 1. (continued).

right side of each of charts, with the CR from the GIM as the base (white bar) overlapped with the CR from the naive model (blue bar). As expected, the naive multinomial models produce biased estimates of low CRs on the intercepts β_{01} for the category "1" and β_{03} for the category "3" as shown in panels (a) and (c). Specifically, a higher value of π_1 or π_3 is associated with a larger bias and a lower CR. In addition, the naive multinomial models not only yield negatively biased intercepts β_{01} and β_{03} but also under estimate all parameters for the outcome of "1" and "3," namely, β_{11}, β_{21}, and β_{31} (panels e, i, and m) for the cateory "1," and β_{13}, β_{23}, and β_{33} (panels g, k, and o) for the category "3." For the other categories, such as 2 and 4, some of the estimates produced by the naive multinomial models overlap with those obtained from the GIM model, although slight negative biases are still observed, especially when probabilities of inflation increase. In terms of the RSMSE, the two models perform very similarly for the outcomes of "2" and "4," while the naive multinomial models give smaller RSMSE for the inflated categories of "1" and "3."

Figure 2 summarizes the performances of the generalized inflated cumulative logit (GICL) model and the naive CL (ordered) models for an outcome of "1" to "5" with inflation at the values of "1" and "3." Unlike the multinomial case, the naive CL (ordered) models give biased estimates and low CRs for all parameters. Similar to the naive multinomial models, the bias increases with the size of inflated probabilities π_1 and π_3 increases for the naive CL models, and the intercepts of the categories of "2" and "4" are less biased with higher CRs than the intercepts of the categories of "1" and "3." Compared to the naive CL (ordered) models, the GICL models yield unbiased estimates, almost 100% CRs, and similar sizes of the RSMSE for all of the parameters included.

Both Figures 3 and 4 show the exact pattern of biases and CRs for the Poisson and zero-truncated Poisson models: (1) the naive models produce biased estimates and poor CRs but low RSMSEs, (2) the size of bias increases and the CR decreases dramatically as the probabilities of inflation increases, especially for π_3, and (3) the generalized inflated Poisson (GIP) or zero-truncated Poisson (GIZTP) models yield unbiased estimates, high CRs, and small sizes of RSMSE for all of the paramters.

In summary, the estimates obtained from the naive models are not satisfactory with regard to the biasness and the CR—for example, the biases could be as high as 200 percent for intercepts of the inflated categories in the GIM when the inflation probabilities reach a .20 level. Interpretations based on naive estimators could distort the underlying data-generation mechanism. The generalized inflated models, on the

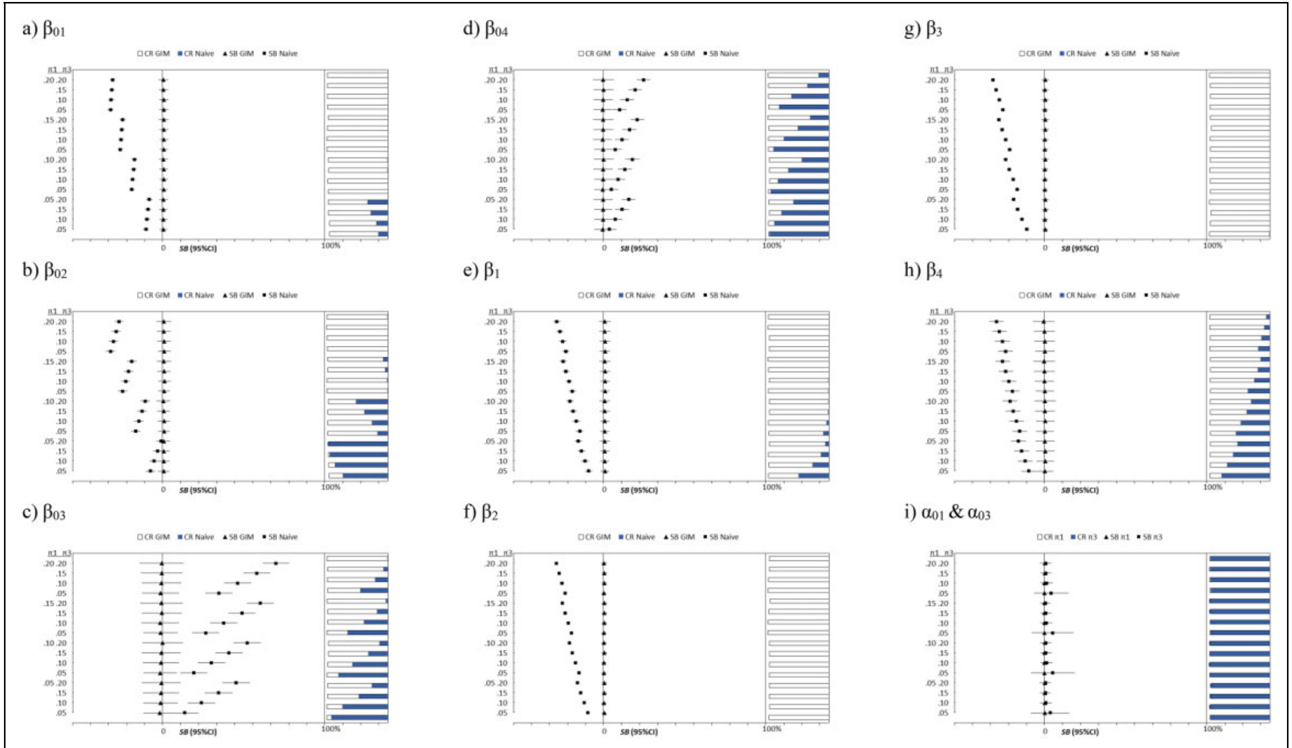

Figure 2. Plots of the SB, RSMSE, and CR of the generalized cumulative logit model plotted against the fixed π_1 and π_3. The dot represents the SB, with the RSMSE as the error bar. The horizontal bar represents the CR at each level of π_1 and π_3. SB = standardized bias; RSMSE = root standardized mean square error; CR = coverage rate.

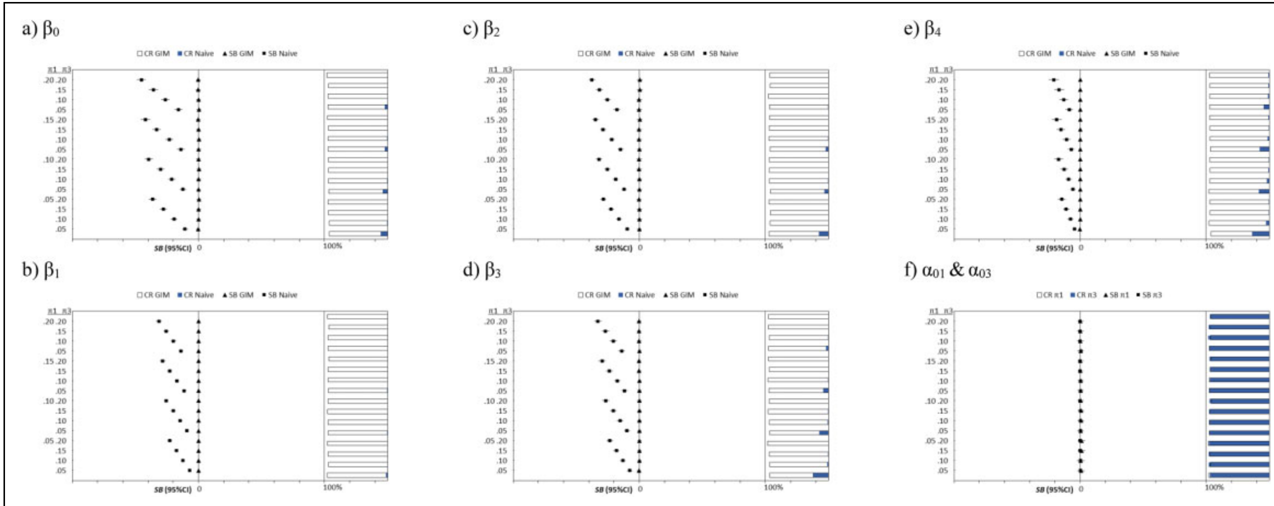

Figure 3. Plots of the SB, RSMSE, and CR of the generalized inflated Poisson model plotted against the fixed π_1 and π_3. The dot represents the SB, with the RSMSE as the error bar. The horizontal bar represents the CR at each level of π_1 and π_3. SB = standardized bias; RSMSE = root standardized mean square error; CR = coverage rate.

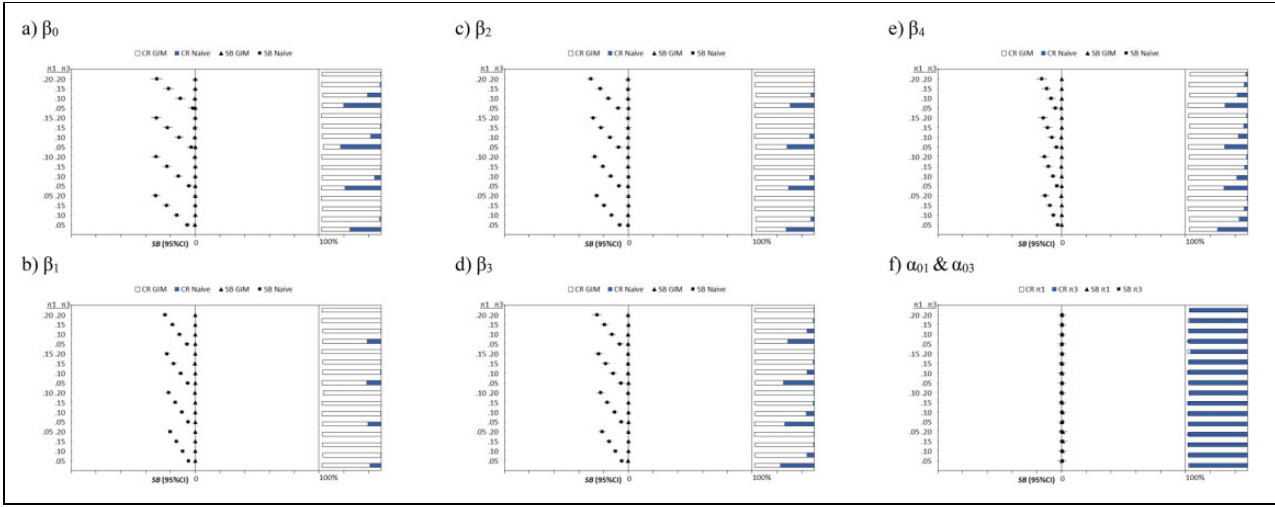

Figure 4. Plots of the SB, RSMSE, and CR of the generalized inflated zero-truncated Poisson model plotted against the fixed π_1 and π_3. The dot represents the SB, with the RSMSE as the error bar. The horizontal bar represents the CR at each level of π_1 and π_3. SB = standardized bias; RSMSE = root standardized mean square error; CR = coverage rate.

other hand, show superiority over the naive models, for example, unbiased estimates (usually less than 5 percent), high CRs, and small RSMSEs.

For Case (ii) where the probabilities of inflation depend on covariates, only the estimates produced by the generalized inflated models are reported because none of the naive models give unbiased estimates. For the multinomial outcomes (Table 1), none of the biases are beyond 5 percent, and most of CRs are higher than 95 percent. In terms of precision, there is no difference between the estimates obtained from the inflation equations and those from the multinomial outcomes. The size of π_1 or π_3 is not correlated to the size of bias for any parameters in the model, and none of parameters show high RSMSE. The same pattern can be found in Tables 2–4, which presents the bias, RSMSE, and CR for the GICL, the GIP, and the GIZTP models, accordingly.

Based on the evaluation for the estimates obtained from the direct maximum likelihood method, the Newton–Raphson type of algorithm seems to provide an efficient way to estimate the generalized inflated models for a moderate sample size of 3,000.

Since the performance of the misspecified models is only slightly better than the naive models with biased estimates and poor CR,[2] to save space, only fit indices for the experiments are given. Begum et al. (2014) suggested using Akaike information criterion (AIC) to find appropriate inflated models among all possible combinations of inflated values based on empirical distribution. Table 5 gives the average AIC for the naive, the misspecified, and the true models by the type of models and the level of inflation. For all experiments, the correct specified models yield the lowest average AIC value. Compared to the naive models, the misspecified ones usually have lower AIC except for a few cases, for example, (i) GIM and (ii) GICL when the inflation probability π_1 is low at the .05 level. According to the results, AIC is a useful tool for finding the appropriate specification of inflation parts.

Applications in Health Study

To illustrate the implementation of the generalized inflated models, we fitted a GIP model using the Wave 1 data from the National Longitudinal Study of Adolescent Health (Add Health) study, which is a longitudinal study of a nationally representative sample of adolescents in grades 7–12 in the United States during the 1994–1995 school year (K. M. Harris et al. 2009). The Add Health study was designed as a stratified two-stage cluster sampling in which

Table 1. Simulation Results of the Generalized Inflated Multinomial Model for Case (ii).

		π_1	.05				.10				.15				.20			
Parameter		π_3	.05	.10	.15	.20	.05	.10	.15	.20	.05	.10	.15	.20	.05	.10	.15	.20
α_1	α_{01}	SB	.037	.034	.023	.026	.013	.007	.008	.016	.006	.007	.010	.012	.004	.007	.012	.005
		RSMSE	.042	.052	.037	.042	.030	.027	.026	.028	.028	.027	.028	.028	.032	.030	.030	.032
		CV	.938	.938	.946	.954	.938	.964	.962	.960	.946	.964	.936	.952	.960	.966	.966	.934
	α_{11}	SB	.014	.050	.036	.049	.015	.021	.030	.039	.016	.026	.008	.017	.008	.004	.013	.003
		RSMSEE	.241	.164	.139	.146	.090	.087	.080	.087	.066	.065	.067	.072	.059	.058	.059	.058
		CV	.960	.964	.968	.946	.950	.952	.960	.954	.952	.960	.942	.952	.944	.962	.950	.964
	α_{21}	SB	.076	.069	.052	.076	.031	.034	.023	.036	.018	.021	.020	.029	.017	.017	.029	.035
		RSMSEE	.104	.100	.091	.120	.054	.057	.056	.063	.042	.045	.047	.050	.038	.039	.042	.045
		CV	.934	.940	.918	.914	.948	.940	.942	.936	.962	.946	.948	.954	.948	.956	.960	.958
α_3	α_{03}	SB	.028	.012	.007	.002	.029	.008	.008	.010	.023	.015	.008	−.009	.043	.015	−.003	.003
		RSMSE	.042	.030	.032	.035	.041	.029	.029	.034	.039	.031	.028	.031	.055	.028	.027	.030
		CV	.950	.946	.948	.936	.946	.954	.954	.956	.964	.944	.942	.938	.940	.958	.932	.940
	α_{13}	SB	.037	.008	.010	.008	.022	.012	.019	.016	.042	.034	.000	−.013	.095	.021	−.013	.003
		RSMSE	.138	.081	.068	.053	.125	.083	.065	.056	.130	.087	.064	.057	.150	.084	.069	.061
		CV	.952	.966	.938	.960	.964	.944	.958	.954	.960	.946	.956	.944	.950	.950	.942	.932
	α_{23}	SB	.060	.039	.034	.018	.076	.043	.024	.018	.072	.047	.023	.014	.128	.049	.026	.021
		RSMSE	.086	.063	.050	.039	.099	.062	.047	.042	.098	.063	.050	.045	.145	.064	.055	.050
		CV	.944	.926	.936	.948	.922	.936	.948	.938	.920	.936	.944	.934	.918	.962	.938	.936

(continued)

Table 1. (continued)

Parameter		π_1	.05				.10				.15				.20			
		π_3	.05	.10	.15	.20	.05	.10	.15	.20	.05	.10	.15	.20	.05	.10	.15	.20
β_1	β_{01}	SB	.001	.004	−.010	−.006	−.009	−.015	.024	.017	−.024	.001	−.008	.002	−.017	.000	−.035	−.013
		RSMSE	.147	.150	.154	.162	.154	.160	.168	.177	.162	.176	.182	.192	.175	.184	.187	.188
		CV	.952	.940	.952	.938	.954	.944	.946	.930	.954	.940	.940	.936	.946	.948	.938	.954
	β_{11}	SB	.001	.000	−.003	.001	−.003	−.005	.006	.006	−.007	.000	−.001	.008	−.002	.004	.001	.006
		RSMSE	.038	.039	.038	.041	.039	.040	.043	.044	.041	.044	.045	.048	.042	.045	.046	.046
		CV	.930	.940	.956	.944	.942	.942	.930	.942	.942	.930	.924	.944	.946	.962	.950	.950
	β_{21}	SB	.014	.009	.010	.007	.010	.014	.016	.009	.010	.015	.004	.004	.009	.007	−.005	.002
		RSMSE	.035	.039	.039	.039	.039	.039	.041	.040	.040	.042	.041	.043	.042	.040	.043	.047
		CV	.944	.920	.926	.932	.910	.940	.936	.958	.948	.946	.940	.942	.944	.962	.950	.938
	β_{31}	SB	.017	.018	.020	.022	.017	.019	.025	.019	.022	.025	.015	.021	.022	.022	.012	.019
		RSMSE	.041	.041	.044	.046	.043	.045	.048	.050	.047	.049	.051	.050	.052	.052	.050	.054
		CV	.932	.958	.936	.934	.952	.946	.932	.924	.940	.932	.944	.954	.938	.934	.958	.944
	β_{02}	SB	.025	.011	.014	.019	.009	.009	.007	.019	.008	.001	.009	.032	.024	.003	.031	.016
		RSMSE	.073	.069	.074	.075	.070	.070	.076	.078	.074	.077	.077	.078	.078	.079	.078	.077
		CV	.948	.960	.948	.952	.958	.958	.956	.950	.936	.954	.944	.962	.928	.942	.962	.962
β_2	β_{12}	SB	.006	.006	.009	.009	.005	.009	.001	.009	.007	−.001	.001	.012	.004	.005	.010	.014
		RSMSE	.034	.037	.034	.037	.036	.035	.036	.038	.034	.038	.038	.039	.038	.039	.039	.040
		CV	.950	.934	.954	.948	.940	.960	.958	.948	.958	.946	.948	.946	.944	.950	.960	.948
	β_{22}	SB	−.005	.002	.003	−.001	.004	.005	−.009	−.003	.011	−.003	.001	−.007	−.002	.008	−.008	.001
		RSMSE	.049	.050	.050	.054	.049	.050	.051	.055	.050	.051	.055	.056	.053	.054	.057	.055
		CV	.948	.952	.948	.946	.942	.934	.958	.940	.946	.952	.956	.948	.944	.946	.948	.956
	β_{32}	SB	.004	.007	−.003	.015	.011	.001	−.001	.023	.013	.004	.005	.015	.017	.028	.028	.018
		RSMSE	.043	.043	.049	.046	.046	.045	.048	.052	.047	.048	.052	.051	.051	.051	.052	.055
		CV	.938	.966	.930	.962	.936	.956	.950	.958	.950	.940	.944	.962	.954	.948	.954	.958

(continued)

Table 1. (continued)

		π_1	.05				.10				.15				.20			
Parameter		π_3	.05	.10	.15	.20	.05	.10	.15	.20	.05	.10	.15	.20	.05	.10	.15	.20
β_3	β_{03}	SB	.003	.028	.027	.022	.023	.005	.012	.017	.019	.005	.006	.027	.008	.012	.020	.034
		RSMSE	.080	.087	.094	.098	.079	.088	.093	.099	.083	.089	.097	.107	.083	.093	.105	.116
		CV	.964	.956	.954	.972	.954	.962	.956	.966	.956	.960	.958	.952	.956	.958	.940	.936
	β_{13}	SB	.002	.014	.014	.006	.010	.000	.001	.003	.007	.000	−.001	.006	.003	.006	.007	.014
		RSMSE	.040	.040	.042	.045	.038	.042	.044	.045	.040	.042	.045	.048	.040	.044	.047	.049
		CV	.946	.950	.958	.956	.950	.960	.966	.966	.948	.962	.962	.968	.960	.944	.956	.972
	β_{23}	SB	.009	.014	.013	.016	.018	.016	.010	.011	.017	.008	.010	.022	.013	.006	.016	.017
		RSMSE	.038	.040	.041	.043	.040	.039	.042	.045	.039	.041	.043	.046	.039	.043	.045	.050
		CV	.950	.950	.946	.956	.942	.956	.962	.956	.948	.948	.952	.960	.946	.938	.948	.948
	β_{33}	SB	.001	.006	.011	.010	.008	.016	.004	.011	.014	.002	.009	.006	.008	.006	.003	.007
		RSMSE	.048	.050	.051	.055	.050	.049	.051	.056	.050	.051	.056	.060	.053	.054	.059	.063
		CV	.966	.956	.964	.966	.944	.966	.974	.966	.962	.956	.952	.948	.960	.956	.946	.946
β_4	β_{04}	SB	.007	.006	.007	.009	.003	−.006	.019	.015	.002	.016	.014	.016	.010	.017	.017	.025
		RSMSE	.071	.070	.074	.077	.068	.076	.079	.084	.073	.079	.081	.080	.079	.084	.080	.084
		CV	.952	.960	.940	.944	.964	.944	.940	.928	.938	.950	.942	.956	.936	.942	.952	.962
	β_{14}	SB	.006	.006	.010	.011	.004	.000	.013	.009	.006	.011	.009	.016	.011	.011	.017	.019
		RSMSE	.038	.038	.039	.039	.037	.039	.041	.042	.039	.041	.041	.041	.040	.042	.041	.043
		CV	.950	.932	.944	.956	.954	.944	.936	.936	.936	.944	.944	.952	.944	.944	.958	.950
	β_{24}	SB	.002	.010	.010	.004	.011	.009	−.001	.005	.011	−.001	.002	.010	.004	.004	.007	.014
		RSMSE	.032	.034	.033	.036	.034	.033	.036	.036	.033	.037	.037	.039	.036	.037	.039	.040
		CV	.950	.948	.952	.944	.946	.952	.946	.948	.952	.936	.946	.938	.946	.938	.938	.930
	β_{34}	SB	.000	.002	−.002	−.006	.004	.005	−.005	.006	.000	−.005	.008	−.006	−.005	.007	−.007	.003
		RSMSE	.045	.048	.048	.050	.048	.046	.048	.051	.048	.048	.052	.054	.051	.051	.054	.054
		CV	.964	.940	.956	.952	.938	.954	.972	.964	.958	.960	.964	.952	.946	.954	.952	.944
	β_{44}	SB	.007	.024	−.002	.005	.024	.005	−.010	−.007	−.002	−.007	.005	.014	.004	.006	.019	.035
		RSMSE	.066	.070	.075	.080	.070	.076	.075	.082	.073	.075	.083	.088	.079	.083	.089	.093
		CV	.962	.952	.946	.930	.956	.946	.946	.948	.958	.956	.934	.946	.938	.938	.944	.948

Note: All results are based on 500 simulated samples with the sample size 3,000. SB = standardized bias; RSMSE = root standardized mean square error; CR = coverage rate.

Table 2. Simulation Results of the Generalized Inflated Cumulative (Ordered) Logit Model for Case (ii).

		π_1	.05				.10				.15				.20			
Parameter		π_3	.05	.10	.15	.20	.05	.10	.15	.20	.05	.10	.15	.20	.05	.10	.15	.20
α_1	α_{01}	SB	.026	.030	.021	0.139	.012	.007	.011	.018	−.001	.005	.010	.007	.002	.011	.001	−.001
		RSMSE	.036	.040	.033	0.681	.025	.026	.026	.025	.027	.026	.025	.024	.029	.029	.029	.028
		CV	.952	.958	.956	0.966	.966	.972	.956	.962	.950	.950	.946	.956	.968	.950	.942	.948
	α_{11}	SB	.041	.054	.028	0.211	.012	.000	.018	.052	−.018	.005	.020	.028	.009	.028	.010	.011
		RSMSE	0.138	0.136	0.126	1.007	0.080	0.079	0.081	0.083	0.066	0.067	0.066	0.065	0.056	0.054	0.054	0.055
		CV	.962	.962	.956	0.958	.956	.964	.960	.962	.946	.940	.942	.948	.950	.964	.960	.954
	α_{21}	SB	.029	.041	.037	0.353	.016	.016	.020	.045	.007	.013	.022	.028	.013	.021	.018	.012
		RSMSE	0.068	0.088	0.074	1.745	0.048	0.047	0.051	0.053	0.038	0.041	0.042	0.042	0.030	0.034	0.035	0.038
		CV	.940	.928	.930	0.936	.942	.950	.954	.958	.942	.946	.942	.956	.964	.958	.948	.942
α_3	α_{03}	SB	.029	.004	−.002	−0.007	.017	.007	.000	−.001	.026	.008	.001	.001	.033	.011	.001	.002
		RSMSE	.043	.030	.029	0.031	.039	.030	.028	.033	.041	.028	.028	.028	.070	.029	.025	.027
		CV	.940	.940	.928	0.950	.954	.940	.940	.942	.944	.936	.934	.940	.936	.930	.942	.948
	α_{13}	SB	.043	.006	−.005	−0.005	.026	.011	.004	−.010	.053	.012	.000	−.008	.071	.015	−.009	.000
		RSMSE	.132	.076	.059	0.049	.129	.078	.058	.054	.134	.078	.062	.051	.210	.082	.063	.055
		CV	.962	.952	.950	0.950	.952	.938	.962	.944	.944	.940	.948	.938	.938	.940	.942	.954
	α_{23}	SB	.031	.014	.009	0.009	.033	.025	.020	.004	.049	.036	.015	.001	.052	.023	.004	.003
		RSMSE	.088	.052	.037	0.033	.081	.055	.043	.036	.088	.059	.047	.041	.116	.059	.049	.043
		CV	.922	.928	.956	0.954	.916	.936	.944	.946	.932	.926	.910	.946	.940	.930	.934	.936

(continued)

Table 2. (continued)

		π_1	.05				.10				.15				.20			
Parameter		π_3	.05	.10	.15	.20	.05	.10	.15	.20	.05	.10	.15	.20	.05	.10	.15	.20
β	β_{01}	SB	.005	−.011	.006	−0.005	−.013	−.015	−.004	−.002	.006	−.005	−.003	−.006	−.007	−.001	.001	−.009
		RSMSE	.047	.049	.049	0.053	.048	.050	.050	.057	.053	.051	.058	.060	.054	.059	.060	.065
		CV	.948	.948	.954	0.944	.960	.952	.970	.936	.944	.980	.942	.944	.962	.944	.948	.930
	β_{02}	SB	.005	−.027	.005	−0.014	−.028	−.033	−.014	−.009	.005	−.011	−.012	−.017	−.019	−.006	−.005	−.027
		RSMSE	.094	.096	.096	0.105	.094	.098	.100	.114	.101	.102	.116	.118	.108	.115	.116	.128
		CV	.940	.942	.954	0.938	.956	.946	.972	.936	.946	.980	.948	.934	.954	.942	.946	.924
	β_{03}	SB	−.015	.006	−.026	−0.009	.007	.017	−.005	−.019	−.018	−.005	−.016	−.022	.001	−.018	−.035	−.002
		RSMSE	.095	.096	.096	0.101	.095	.097	.103	.113	.099	.104	.115	.117	.104	.113	.116	.129
		CV	.940	.946	.948	0.956	.962	.960	.954	.938	.940	.960	.944	.946	.954	.946	.952	.938
	β_{04}	SB	−.005	.006	−.009	−0.002	.006	.012	−.001	−.009	−.005	−.001	−.007	−.007	.003	−.009	−.014	.003
		RSMSE	.047	.047	.048	0.050	.046	.049	.051	.056	.049	.051	.056	.057	.052	.056	.056	.063
		CV	.946	.950	.950	0.956	.964	.960	.954	.936	.934	.960	.948	.950	.950	.940	.954	.934
	β_1	SB	.008	−.001	.005	0.004	.000	−.005	.003	.008	.004	.001	.004	.004	.003	.006	.006	.004
		RSMSE	.026	.026	.026	0.027	.026	.027	.028	.029	.027	.028	.030	.032	.029	.030	.032	.033
		CV	.942	.952	.952	0.948	.948	.960	.956	.950	.944	.966	.952	.952	.946	.954	.942	.942
	β_2	SB	.004	.002	.003	0.003	.002	.002	.003	.003	.005	.001	.002	.003	.005	.001	.002	.009
		RSMSE	.013	.012	.013	0.013	.012	.014	.015	.014	.013	.015	.014	.017	.014	.014	.016	.017
		CV	.952	.962	.960	0.964	.962	.962	.936	.964	.956	.944	.956	.934	.956	.958	.950	.938
	β_3	SB	.007	.007	.004	0.010	.008	.006	.008	.010	.005	.005	.007	.007	.010	.007	.007	.011
		RSMSE	.015	.016	.017	0.018	.015	.017	.018	.018	.016	.018	.017	.019	.017	.018	.019	.019
		CV	.928	.958	.932	0.936	.968	.952	.958	.956	.950	.956	.948	.940	.946	.948	.942	.958
	β_4	SB	.013	.007	−.008	−0.010	.006	−.006	−.004	−.013	−.012	−.009	−.013	−.011	−.005	−.006	−.022	.016
		RSMSE	.054	.059	.057	0.064	.060	.058	.061	.065	.061	.062	.067	.070	.068	.067	.072	.076
		CV	.956	.940	.956	0.926	.938	.958	.944	.950	.950	.954	.952	.938	.944	.954	.942	.948

Note: All results are based on 500 simulated samples with the sample size 3,000. SB = standardized bias; RSMSE = root standardized mean square error; CR = coverage rate.

Table 3. Simulation Results of the Generalized Inflated Poisson Model for Case (ii).

		π_1	.05				.10				.15				.20			
Parameter		π_3	.05	.10	.15	.20	.05	.10	.15	.20	.05	.10	.15	.20	.05	.10	.15	.20
α_1	α_{01}	SB	.008	.009	.009	.008	.005	.002	.004	.002	.001	.002	.002	.002	−.001	.002	.000	−.001
		RSMSE	.021	.020	.019	.020	.017	.016	.017	.017	.018	.018	.019	.018	.022	.021	.021	.021
		CV	.950	.960	.958	.940	.968	.960	.944	.960	.952	.956	.950	.944	.952	.948	.950	.954
	α_{11}	SB	−.005	.005	.001	.011	.008	−.004	.005	.003	.008	.011	.000	−.001	−.001	.005	−.003	−.002
		RSMSE	.105	.105	.101	.090	.069	.065	.067	.065	.057	.054	.057	.055	.046	.047	.046	.049
		CV	.952	.940	.958	.968	.946	.948	.942	.942	.930	.944	.934	.936	.942	.936	.948	.940
	α_{21}	SB	.032	.023	.026	.022	.010	.012	.008	.018	.012	.008	.006	.014	.005	.005	.010	.012
		RSMSE	.056	.055	.055	.055	.038	.037	.037	.038	.030	.030	.031	.032	.026	.026	.027	.029
		CV	.938	.958	.946	.960	.952	.948	.954	.964	.952	.948	.944	.952	.946	.962	.950	.948
α_3	α_{03}	SB	.005	.001	−.002	−.003	.005	.005	.000	.005	.006	.003	.007	.005	.009	.009	.007	.004
		RSMSE	.017	.016	.017	.020	.018	.016	.016	.019	.016	.015	.016	.020	.017	.016	.016	.019
		CV	.950	.940	.948	.948	.946	.936	.958	.944	.970	.968	.956	.940	.944	.946	.948	.954
	α_{13}	SB	−.004	−.006	.000	−.010	.003	.020	−.006	.000	.010	−.010	.007	.006	.023	.013	.011	.001
		RSMSE	.077	.055	.044	.037	.076	.054	.041	.040	.073	.050	.044	.040	.076	.056	.046	.039
		CV	.944	.948	.950	.964	.970	.960	.962	.946	.970	.952	.948	.944	.950	.932	.942	.968
	α_{23}	SB	.005	.000	.003	−.001	.003	.010	−.001	−.003	.008	−.004	−.002	.002	.016	.006	.005	.002
		RSMSE	.040	.029	.024	.021	.040	.030	.026	.022	.041	.030	.025	.022	.044	.033	.027	.025
		CV	.958	.960	.956	.960	.960	.940	.944	.952	.948	.964	.954	.952	.932	.930	.946	.946

(continued)

Table 3. (continued)

		π_1	.05				.10				.15				.20			
Parameter		π_3	.05	.10	.15	.20	.05	.10	.15	.20	.05	.10	.15	.20	.05	.10	.15	.20
β	β_0	SB	−.001	−.001	.000	−.001	.000	.000	−.001	.000	.000	−.001	.000	−.001	−.001	−.001	−.001	−.001
		RSMSE	.008	.008	.008	.008	.008	.008	.008	.009	.008	.008	.008	.009	.008	.009	.009	.009
		CV	.944	.956	.946	.944	.952	.956	.960	.944	.956	.960	.950	.944	.948	.946	.960	.942
	β_1	SB	−.001	.000	−.001	−.001	.000	.000	−.001	−.001	.000	−.001	−.001	−.001	−.001	−.001	−.001	−.001
		RSMSE	.007	.007	.007	.008	.007	.007	.008	.008	.007	.007	.008	.008	.008	.008	.008	.008
		CV	.932	.952	.948	.954	.952	.958	.950	.950	.950	.948	.952	.960	.952	.958	.958	.938
	β_2	SB	−.001	.000	.000	−.001	.000	−.001	−.001	−.001	.000	−.001	−.001	.000	−.001	−.001	.000	.000
		RSMSE	.004	.004	.004	.005	.004	.004	.004	.004	.004	.004	.004	.005	.004	.004	.004	.005
		CV	.960	.954	.942	.934	.952	.940	.950	.958	.940	.946	.948	.948	.938	.958	.944	.952
	β_3	SB	.000	.000	.001	.001	.000	.001	.000	.001	.001	.000	.001	.000	.001	.001	.000	−.001
		RSMSE	.003	.003	.003	.003	.003	.003	.003	.004	.003	.003	.004	.004	.003	.004	.004	.004
		CV	.944	.932	.956	.962	.940	.954	.958	.950	.946	.958	.938	.948	.958	.950	.948	.954
	β_4	SB	.000	.000	.000	.000	−.001	.000	.001	.000	.000	.001	.000	.000	.001	.000	.000	−.001
		RSMSE	.009	.009	.010	.010	.009	.010	.011	.011	.010	.011	.011	.012	.010	.011	.012	.012
		CV	.944	.944	.942	.944	.944	.946	.940	.926	.940	.932	.942	.946	.948	.928	.946	.928

Note: All results are based on 500 simulated samples with the sample size 3,000. SB = standardized bias; RSMSE = root standardized mean square error; CR = coverage rate.

Table 4. Simulation Results of the Generalized Inflated Zero-truncated Poisson Model for Case (ii).

		π_1	.05				.10				.15				.20			
Parameter		π_3	.05	.10	.15	.20	.05	.10	.15	.20	.05	.10	.15	.20	.05	.10	.15	.20
α_1	α_{01}	SB	.006	.007	.008	.008	.003	.005	.009	.007	.007	.004	.004	.005	.010	.007	.003	.002
		RSMSE	.019	.019	.020	.019	.019	.020	.019	.019	.021	.021	.020	.018	.025	.022	.020	.021
		CV	.952	.954	.946	.954	.958	.948	.938	.940	.948	.948	.948	.944	.940	.954	.956	.932
	α_{11}	SB	.002	.002	.010	−.005	.004	.012	.020	.012	.020	.003	.007	−.003	.011	.017	−.005	−.013
		RSMSE	.064	.067	.075	.081	.055	.059	.065	.074	.054	.056	.060	.065	.052	.056	.060	.074
		CV	.952	.956	.950	.948	.958	.954	.946	.932	.936	.958	.962	.948	.944	.956	.960	.932
	α_{21}	SB	.006	.005	.008	.015	−.001	.005	.015	.014	.007	.006	.009	.009	.010	.006	.003	.003
		RSMSE	.034	.036	.042	.044	.029	.034	.035	.039	.026	.030	.033	.035	.026	.028	.030	.036
		CV	.952	.952	.948	.950	.948	.924	.954	.946	.940	.958	.948	.950	.952	.954	.958	.938
α_3	α_{03}	SB	.005	.001	.001	.006	.003	.003	.000	.001	.003	.004	.003	.006	.006	.003	.008	.006
		RSMSE	.014	.014	.015	.018	.015	.015	.016	.017	.015	.015	.015	.016	.015	.013	.014	.015
		CV	.954	.938	.962	.952	.960	.950	.944	.946	.968	.952	.954	.948	.962	.956	.952	.960
	α_{13}	SB	.010	.000	.004	.000	−.002	.000	−.012	.001	.007	.000	.007	.020	−.004	.006	.022	.022
		RSMSE	.042	.035	.034	.036	.044	.041	.040	.038	.049	.045	.041	.041	.053	.047	.045	.048
		CV	.950	.956	.960	.946	.970	.934	.948	.962	.948	.932	.960	.962	.958	.942	.944	.938
	α_{23}	SB	.006	.004	.001	.005	.009	.002	−.002	.001	.001	.005	.001	.005	.009	.010	.011	.007
		RSMSE	.025	.022	.020	.020	.029	.025	.025	.022	.032	.028	.026	.025	.036	.030	.029	.028
		CV	.952	.952	.954	.956	.940	.950	.940	.966	.942	.956	.954	.954	.946	.960	.960	.948

(continued)

Table 4. (continued)

Parameter		π_1 / π_3	.05 / .05	.05 / .10	.05 / .15	.05 / .20	.10 / .05	.10 / .10	.10 / .15	.10 / .20	.15 / .05	.15 / .10	.15 / .15	.15 / .20	.20 / .05	.20 / .10	.20 / .15	.20 / .20
β	β_0	SB	−.001	.000	.000	.001	−.001	.001	−.003	−.003	.000	−.003	−.001	−.002	.001	−.002	−.001	.000
		RSMSE	.010	.011	.011	.012	.011	.011	.012	.012	.011	.012	.012	.013	.012	.012	.013	.014
		CV	.958	.932	.958	.956	.930	.964	.956	.954	.962	.952	.946	.948	.970	.952	.952	.936
	β_1	SB	.000	.000	.000	−.001	.000	.000	−.003	−.002	.000	−.003	.000	−.002	−.001	−.002	−.002	.000
		RSMSE	.009	.010	.010	.010	.010	.010	.010	.011	.010	.010	.010	.011	.010	.011	.012	.012
		CV	.966	.932	.960	.966	.942	.956	.956	.952	.960	.960	.958	.960	.970	.960	.956	.938
	β_2	SB	−.001	.000	.000	.001	.000	.001	.000	−.001	.000	.000	−.001	−.001	.001	−.001	−.001	.000
		RSMSE	.005	.006	.006	.006	.006	.006	.006	.006	.006	.006	.006	.007	.006	.006	.007	.007
		CV	.944	.936	.948	.936	.938	.938	.936	.964	.944	.940	.940	.946	.944	.964	.940	.952
	β_3	SB	−.001	.001	.000	.001	.001	.000	.001	.000	.000	.000	−.001	.000	.002	.000	.000	−.001
		RSMSE	.004	.004	.004	.005	.004	.004	.005	.005	.004	.005	.005	.005	.005	.005	.005	.005
		CV	.960	.944	.958	.940	.954	.946	.958	.954	.954	.946	.946	.944	.946	.960	.946	.948
	β_4	SB	.001	.005	.002	.006	.004	.002	.005	.000	.000	.005	.000	.004	.004	−.001	.005	−.001
		RSMSE	.015	.016	.020	.020	.017	.020	.020	.023	.020	.020	.024	.028	.021	.023	.028	.030
		CV	.944	.954	.926	.960	.954	.938	.958	.958	.922	.962	.954	.950	.958	.954	.952	.962

Note: All results are based on 500 simulated samples with the sample size 3,000. SB = standardized bias; RSMSE = root standardized mean square error; CR = coverage rate.

Table 5. The Average AIC by Levels of Inflation Based on 500 Simulated Samples with the Sample Size 3,000.

	π_1	.05				.10				.15				.20			
Model	π_3	.05	.10	.15	.20	.05	.10	.15	.20	.05	.10	.15	.20	.05	.10	.15	.20
GIM (i)	Naive	13,092.0	13,158.8	13,137.1	13,028.2	13,034.7	13,114.1	13,093.9	12,990.8	12,900.0	12,983.7	12,969.5	12,867.6	12,694.0	12,784.8	12,775.1	12,668.6
	Incorrect	13,094.9	13,166.7	13,147.4	13,039.8	13,010.1	13,099.8	13,086.3	12,987.3	12,854.7	12,952.4	12,948.0	12,853.1	12,635.1	12,743.5	12,745.7	12,648.2
	Correct	13,058.5	13,115.5	13,084.1	12,968.1	12,975.4	13,051.1	13,026.4	12,920.3	12,820.8	12,905.9	12,891.1	12,789.5	12,603.3	12,699	12,692.1	12,587.7
GIM (ii)	Naive	7,854.2	7,892.4	7,885.1	7,833.5	7,830.3	7,869.3	7,875.7	7,834.4	7,758.7	7,811.9	7,829.2	7,798.8	7,646.5	7,708.9	7,737.0	7,725.0
	Incorrect	7,842.4	7,885.6	7,878.3	7,825.3	7,787.9	7,836.7	7,850.2	7,813.6	7,683.9	7,754.2	7,785.5	7,765.4	7,543.5	7,629.6	7,677.4	7,676.4
	Correct	7,800.1	7,804.9	7,759.3	7,670.7	7,742.7	7,752.3	7,728.9	7,659.1	7,637.1	7,668.4	7,664.1	7,609.2	7,495.2	7,544.0	7,554.3	7,522.8
GICL(i)	Naive	7,669.0	7,623.1	7,542.4	7,427.8	7,932.9	7,884.7	7,799.7	7,683.5	8,089.8	8,034.7	7,946.6	7,828.1	8,156.3	8,100.0	8,011.3	7,885.7
	Incorrect	7,656.6	7,611.7	7,530.5	7,413.6	7,837.3	7,799.8	7,721.8	7,610.4	7,924.4	7,886.8	7,812.3	7,704.7	7,935.0	7,903.2	7,834.2	7,724.0
	Correct	7,604.5	7,556.6	7,471.6	7,352.3	7,789.4	7,747.8	7,666.5	7,552.8	7,878.6	7,837.4	7,759.5	7,649.5	7,892.0	7,856.3	7,783.9	7,672.1
GICL(ii)	Naive	7,587.7	7,675.3	7,719.4	7,704.2	7,624.5	7,710.6	7,748.3	7,739.9	7,616.3	7,695.8	7,741.5	7,747.5	7,552.5	7,641.6	7,700.1	7,706.2
	Incorrect	7,604.6	7,696.8	7,742.7	7,723.8	7,588.9	7,691.0	7,738.6	7,734.4	7,523.9	7,628.9	7,692.5	7,709.1	7,406.3	7,529.1	7,612.2	7,634.7
	Correct	7,513.9	7,556.1	7,550.2	7,485.1	7,496.3	7,546.1	7,541.9	7,497.0	7,429.7	7,481.9	7,497.9	7,475.9	7,312.4	7,383.9	7,421.7	7,406.5
GIP (i)	Naive	10,215.2	10,658.0	11,051.2	11,393.5	10,433.4	10,861.7	11,236.9	11,543.8	10,633.2	11,039.9	11,384.0	11,670.5	10,802.8	11,183.2	11,501.6	11,732.5
	Incorrect	10,006.1	10,475.9	10,895.9	11,264.1	9,951.7	10,434.2	10,865.7	11,218.5	9,838.1	10,334.0	10,752.8	11,105.0	9,679.4	10,168.8	10,592.5	10,933.8
	Correct	9,731.6	9,824.6	9,819.5	9,743.1	9,672.3	9,771.4	9,773.3	9,692.7	9,556.6	9,659.5	9,658.3	9,584.8	9,391.9	9,494.4	9,499.6	9,412.3
GIP (ii)	Naive	11,721.7	12,165.0	12,357.7	12,445.1	11,853.5	12,195.8	12,377.1	12,466.5	11,867.7	12,180.3	12,351.9	12,406.4	11,831.0	12,130.6	12,262.7	12,332.8
	Incorrect	10,967.2	11,613.2	11,965.8	12,164.5	10,488.5	11,164.6	11,589.9	11,862.1	9,958.4	10,686.3	11,169.7	11,479.4	9,430.2	10,199.7	10,719.7	11,089.7
	Correct	9,433.5	9,209.1	8,868.0	8,498.9	9,081.4	8,990.9	8,794.9	8,535.7	8,668.8	8,716.9	8626.0	8,467.2	8,257.2	8,405.3	8,412.3	8,329.4
GIZTP (i)	Naive	9,002.5	9,415.0	9,769.7	10,099.5	9,165.8	9,552.0	9,915.1	10,236.4	9,275.8	9,673.3	10,028.7	10,302.9	9,363.9	9,758.2	10,065.5	10,336.4
	Incorrect	8,790.6	9,224.1	9,603.5	9,954.1	8,685.4	9,123.8	9,522.9	9,888.0	8,542.8	8,997.8	9,413.8	9,753.7	8,361.7	8,825.7	9,227.2	9,573.1
	Correct	8,570.1	8,690.5	8,725.2	8,685.0	8,468.0	8,598.0	8,632.0	8,601.2	8,323.5	8,451.9	8,496.2	8,455.8	8,132.5	8,272.7	8,307.7	8,270.4
GIZTP (ii)	Naive	10,312.9	10,849.0	11,143.2	11,287.4	10,367.7	10,849.7	11,120.4	11,294.0	10,314.1	10,774.0	11,060.5	11,232.1	10,181.0	10,669.0	10,947.6	11,111.2
	Incorrect	9,639.2	10,337.2	10,761.9	11,014.9	9,219.8	9,932.2	10,414.8	10,755.0	8,742.1	9,515.8	10,049.9	10,436.3	8,279.6	9,104.1	9,670.7	10,076.7
	Correct	8,345.8	8,236.1	8,003.3	7,710.7	8,016.5	8,036.7	7,922.4	7,748.6	7,652.5	7,782.3	7,777.6	7,686.5	7,278.7	7,517.8	7,587.9	7,563.9

Note: AIC = Akaike information criterion; GIM = generalized inflated multinomial; GICL = generalized inflated cumulative logit; GIZTP = generalized inflated zero-truncated Poisson.

schools were selected first, and then individuals were selected within the selected schools. The Add Health study provides a rich set of information on respondents' social, economic, psychological, and physical well-being with contextual data on the family, neighborhood, community, school, friendships, peer groups, and romantic relationships.

Suppose we are interested in smoking behaviors. One measure available in the Add Health is the frequency of smoking, which was recoded from responses for the survey question that asked how many days the respondent smoked in the past 30 days. The responses varied from 0 to 30 and were heaped onto the values ending with 0 or 5. Although there was not a consensus on how to categorize the number of smoking days into patterns, for example, light smoking, moderate smoking, or heavy smoking (Schane, Ling, and Glantz 2010), we grouped the number of days respondents reported smoking cigarettes by intervals of 5, such as 0, 1–5, 6–10, 11–15, 16–20, 20–25, and 25+ (Bjartveit and Tverdal 2005). The distribution was highly skewed with two peaks at the "0" group (73.58%) and the "25+" group (11.60%), which suggests that there may exist inflations on both groups. Several predictors are included: age (continuous, ranged from 11 to 21), female (dummy, coded as 1 if *female*, 0 *otherwise*), race (dummy, coded as 1 if *African American*, 0 *otherwise*), repeated a grade (dummy, if repeated a grade or been held back a grade, 0 otherwise) and religiosity (dummy, if weekly attended religious services, 0 otherwise).

Table 6 reports the results obtained from four different models for the frequency of smoking. Besides the GIP, results from the naive Poisson model, ZIP model, ZIP model with random effect, and the GIP model with random effect are also provided. The naive Poisson model showed a strong positive effect (.227, $p < .001$) of "repeated a grade" on the frequency of smoking. We conducted the Vuong (1989) test and Clark (2007) test to see if a ZIP model was needed. Both the Vuong statistic (64.89, $p < .001$) and Clark statistic (7,897.00, $p < .001$) suggested that a ZIP model should be preferred over the naive Poisson model.

Three covariates were included in the logit of zero inflation—race, repeated a grade, and religiosity. The negative effect of "repeated a grade" suggests that those who repeated a grade are less likely to be in the "0" group. Specifically, the odds of being in the "0" group for one who repeated a grade is reduced by nearly 35 percent ($1 - \exp(-.442)$). Similar to the ZIP, the variables of black, repeated a grade, and religiosity were added to the logit for the inflation of the "25+" group for the GIP model. The strong positive coefficient shows that those who repeated a grade are 1.68 times ($\exp(.521)$)

Table 6. Comparison of Various Poisson Models for the Frequency of Smoking During the Past 30 Days.

Inflation	Parameter	Poisson	ZIP	ZIP/w R.E.	GIP	GIP/w R.E.
0	Intercept		.702 (.023)***	.685 (.023)***	.597 (.025)***	.567 (.026)***
	Black		.920 (.048)***	.918 (.049)***	.825 (.056)***	.808 (.058)***
	Repeated		−.422 (.038)***	−.421 (.039)***	−.433 (.043)***	−.427 (.044)***
	Religiosity		.577 (.036)***	.576 (.036)***	.521 (.041)***	.520 (.041)***
6	Intercept				−1.682 (.030)***	−1.687 (.030)***
	Black				−1.584 (.088)***	−1.582 (.089)***
	Repeated				.483 (.050)***	.480 (.050)***
	Religiosity				−.925 (.055)***	−.924 (.056)***
Poisson	Intercept	−2.164 (.070)***	0.276 (.083)***	0.302 (.093)**	−0.655 (.164)***	−0.785 (.182)***
	Age	.153 (.004)***	.069 (.005)***	.064 (.006)***	.082 (.010)***	.088 (.011)***
	Female	.017 (.014)	−.008 (.015)	−.023 (.015)	−.013 (.032)	−.017 (.032)
	Black	−1.096 (.025)***	−0.409 (.029)***	−0.396 (.032)***	−0.342 (.053)***	−0.381 (.059)***
	Repeated	.227 (.016)***	.026 (.017)	.036 (.018)*	−.053 (.039)	−.042 (.041)
	Religiosity	−.603 (.017)***	−.207 (.018)***	−.194 (.019)***	−.180 (.037)***	−.169 (.039)***
	σ_u^2			.033 (.008)***		.054 (.014)***
	−2LL	72,100	45,768	45,579	38,497	38,423
	AIC	72,112	45,788	45,601	38,525	38,453
	BIC	72,160	45,867	45,634	38,636	38,498
	N	20,424	20,424	20,424	20,424	20,424
	# of schools	146	146	146	146	146

Note: GIP = generalized inflated Poisson; ZIP = zero-inflated Poisson; AIC = Akaike information criterion; BIC = Bayesian information criterion; LL = loglikelihood.
†$p < .1$.
*$p < .05$.
**$p < .01$.
***$p < .001$.

as likely as those who did not repeat a grade to be in the "25+" group. Interestingly, the coefficient of "repeated a grade" in the Poisson part of the ZIP and the GIP models turned out to be not significant, which reveals that the positive effect of "repeated a grade" might be driven by the inflated parts. To address the possible clustering among the respondents within each of the schools, we added a random effect on school level for the ZIP and the GIP models. The significant variance of the random effect implies a clustering effect on the frequency of smoking within each of the schools.

Comparing the fit indices among the ZIP model, the ZIP with random effect, the GIP model, and the GIP model with random effect, the GIP model shows much smaller values of AIC and Bayesian information criterion (BIC) than the ZIP model, while the GIP model with random effect further improves the model fitting.

In the Add Health, another measure of smoking behaviors is the number of cigarettes smoked each day, which can be used as an example of the generalized inflated models for ordinal responses. Similar to the smoking frequency, we grouped the number of cigarettes smoked each day by an interval of 5, such as 0 (*never*), 1–5 (*light*, coded as 1), 6–10 (*mild*, coded as 2), 11–15 (*moderate*, coded as 3), 16–20 (*heavy*, coded as 4), and 20+ (*severe*, coded as 5; Farrell, Fry, and Harris 2003). Most of the responses concentrated on the *never* (74.31%) and *light* (16.06 percent) categories. Table 7 presents the results obtained from the CL (ordered) model, the ZIO logit model, and the GICL with inflations on *never* and *light* categories using the last group *severe* as reference. To be consistent with the previous Poisson models, the same set of independent variables for the regular model parts was used, but for the inflation parts, "black" was replaced by "female" to avoid numerical issues we encountered in the exploratory analysis. The value of AIC shows that the GICL is preferable among the three, while the ZIO would be preferred if BIC is used. It is worth noting that the estimated intercept for the *never* group is not significant for the CICL. This suggests that the model specification for the inflation parts or distribution assumption might not be appropriate.

Conclusion

Due to various reasons, variables from survey studies have inflations on certain values that may lead to biased estimates and incorrect inference if not treated properly. The current study integrated the existing literature on the single-value inflated models and developed a general framework to handle variables with more than one inflated value. We provided a general

Table 7. Comparison of Various Cumulative (Ordered) Logit Models for the Quantity of Smoking During the Past 30 Days.

Inflation	Parameter	CL	ZIO	CICL
"Never"	Intercept		.200 (.071)**	.075 (.113)
	Female		−.172 (.063)**	−.155 (.081)†
	Repeated		−.049 (.067)	−.008 (.088)
	Religiosity		.304 (.086)***	.440 (.118)***
"Light"	Intercept			−4.531 (2.212)*
	Female			.236 (.356)
	Repeated			.089 (.445)
	Religiosity			0.971 (1.844)
Ordered logit	Intercept "Never"	3.356 (.159)***	3.777 (.266)***	4.160 (.433)***
	Intercept "Light"	4.593 (.161)***	5.692 (.283)***	5.871 (.319)***
	Intercept "Mild"	5.188 (.162)***	6.408 (.287)***	6.589 (.327)***
	Intercept "Moderate"	5.705 (.164)***	6.987 (.289)***	7.169 (.331)***
	Intercept "Heavy"	6.617 (.170)***	7.949 (.294)***	8.132 (.337)***
	Age	−.165 (.010)***	−.298 (.018)***	−.308 (.021)***
	Female	−.025 (.033)	.191 (.072)**	.175 (.081)*
	Black	1.058 (.048)***	1.600 (.076)***	1.670 (.120)***
	Repeated	−.363 (.039)***	−.605 (.077)***	−.650 (.099)***
	Religiosity	.631 (.036)***	.697 (.095)***	.655 (.124)***
	N	20,225	20,225	20,225
	−2LL	33,600	33,386	33,369
	AIC	33,620	33,414	33,405
	BIC	33,700	33,525	33,547

Note: AIC = Akaike information criterion; BIC = Bayesian information criterion; LL = loglikelihood; CL = cumulative (ordered) logit model; ZIO = zero-inflated ordered logit model.
†$p < .1$.
*$p < .05$.
**$p < .01$.
***$p < .001$.

implementation with maximum likelihood estimation. To assess the performance of the maximum likelihood estimation, we conducted simulation experiments to evaluate the procedure for the multinomial, ordinal, Poisson, and zero-truncated Poisson outcomes under a range of scenarios, for example, different levels of inflated probabilities, and whether covariates are included.

We found substantial bias and poor inference for the naive models—not only for the intercept(s) of the inflated categories, but other coefficients as well. Specifically, with higher values of inflated probabilities, the naive models produced larger bias and lower CR—since the confidence intervals were too narrow to cover the true parameter. Although in many cases the biased estimates on the intercept(s) of the inflated categories might not be a major concern, ignoring the inflations introduces bias and leads to incorrect inferences for almost all the parameters included in a model. More importantly, doing so also distorts the mechanism that generates the data. Generally speaking, the maximum likelihood estimation performs well for all the models discussed in this study with unbiased estimates and satisfactory coverages, even when the number of parameters that need to be estimated is quite large.

Nevertheless, the proposed model has some limitations. First, to facilitate implementation, the inflated values need to be known in advance. Begum et al. (2014) proposed a three-step modeling strategy that estimates all possible combinations of the GIP models among the empirically observed values with high frequencies, and then, analysts evaluate the models by their goodness-of-fit indices, for example, χ^2 statistic and AIC. The estimates in the final model are assessed by asymptotic *t*-test statistics. Although the Vuong test or Clarke test has been widely used to evaluate between Poisson and ZIP models, a formal test is still needed for the generalized inflated models. Another possible strategy is to partition the sample to training and testing sets if the sample size is sufficiently large and use the testing set to evaluate whether the model specification on the inflations is reasonable. Secondly, to avoid numerical issues, cautions should be taken regarding the variables included in the logit of the inflated probabilities. For example, it has been suggested that at least one of the covariates included in the regular Poisson part needs to be excluded from the logit for the zero inflation for a ZIP model, or vice versa, to prevent the possible numerical difficulties (Diop, Diop, and Dupuy 2011; Staub and Winkelmann 2013). For a zero-inflated multinomial model, Diallo et al. (2017) proposed to run a variable selection using a Logistic model for a binary indicator of whether the response is zero and then use the resulting variables as candidates for the logit of zero-inflated probability. Acknowledged by Diallo et al., 2017 this procedure is not a precise one because some of the zeros actually belong to the multinomial part of the model. It is possible to mimic the procedure for each of the inflated values, but it would be more appropriate if a simultaneous selection procedure is developed. Furthermore, by taking advantage of the flexibility of the NLMIXED procedure, the current study demonstrates an example of

GIP with a random effect. Future work is desirable to extend it for multiple random effects or longitudinal structures. In addition, the current study adopted a mixed method that used a logit model for the inflated probabilities without further elaborating the mechanism of inflation. Previous studies have shown that the inflation or heaping on certain values may be due to rounding to nearby popular values (Crawford et al. 2015; Wang and Heitjan 2008). We hope that the current work can stimulate further developments that incorporate various measurement models.

Authors' Note

Views expressed are those of the authors.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was partially supported by the Multiple Year Research Grant (ref: MYRG2015-00005-FSS) funded by RDAO, University of Macau.

ORCID iD

Tianji Cai http://orcid.org/0000-0002-8962-2660
Yiwei Xia http://orcid.org/0000-0001-7360-732X

Supplemental Material

Supplementary material for this article is available online.

Notes

1. We replicated the experiment with identical set of predictors for both inflation parts and the outcome parts. The results are qualitative similar and are available upon request.
2. The estimates for the misspecified models are available upon request.

References

Bagozzi, Benjamin E. 2016. "The Baseline-inflated Multinomial Logit Model for International Relations Research." *Conflict Management and Peace Science* 33(2):174-97.

Bagozzi, Benjamin E. and Mukherjee Bumba. 2012. “A Mixture Model for Middle Category Inflation in Ordered Survey Responses.” *Political Analysis* 20(3): 369-86.

Basner, Mathias, Kenneth M. Fomberstein, Farid M. Razavi, Siobhan Banks, Jeffrey H. William, Roger R. Rosa, and David F. Dinges. 2007. “American Time Use Survey: Sleep Time and Its Relationship to Waking Activities.” *Sleep* 30: 1085-95.

Begum, Munni, Avishek Mallick, and Nabendu Pal. 2014. “A Generalized Inflated Poisson Distribution with Application to Modeling Fertility Data.” *Thailand Statistician* 12(2):135-59.

Bjartveit, K. and Tverdal, A. 2005. “Health Consequences of Smoking 1-4 Cigarettes per Day.” *Tobacco Control* 14(5):315-20.

Clarke, Kevin A. 2007. “A Simple Distribution-free Test for Nonnested Model Selection.” *Political Analysis* 15(3):347-63.

Crawford, Forrest W., Robert E. Weiss, and Marc A. Suchard. 2015. “Sex, Lies and Self-reported Counts: Bayesian Mixture Models for Heaping in Longitudinal Data via Birth-death Processes.” *The Annals of Applied Statistics* 9(2):572-96.

Diallo, Alpha Oumar, Aliou Diop, and Jean-François Dupuy. 2017. “Analysis of Multinomial Counts with Joint Zero-Inflation, with an Application to Health Economics.” *Journal of Statistical Planning and Inference*. doi: 10.1016/j.jspi.2017.09.005.

Diop, Aba, Aliou Diop, and Jean-François Dupuy. 2016. “Simulation-based Inference in a Zero-inflated Bernoulli Regression Model.” *Communications in Statistics—Simulation and Computation*, 45(10):3597-614.

Diop, Aba, Aliou Diop, and Jean-François Dupuy. 2011. “Maximum Likelihood Estimation in the Logistic Regression Model with a Cure Fraction.” *Electronic Journal of Statistics* 5:460-83.

Farrell, Lisa, Tim R. L. Fry, and Mark N. Harris. 2011. “‘A Pack a Day for Twenty Years’”: Smoking and Cigarette Pack Sizes.” *Applied Economics* 43(21): 2833-42.

Finkelman, Matthew D., Jennifer G. Green, Michael J. Gruber, and Alan M. Zaslavsky. 2011. “A Zero-and K-inflated Mixture Model for Health Questionnaire Data.” *Statistics in Medicine*, 30(9):1028-43.

Hall, Daniel. B. 2001. “Zero-inflated Poisson and Binomial Regression with Random Effects: A Case Study.” *Biometrics* 56:1030-39.

Harris, Kathleen M., Carolyn Tucker Halpern, Brett C. Haberstick, and Andrew Smolen. 2009. “The National Longitudinal Study of Adolescent Health: Research Design.” (http://www.cpc.unc.Edu/Projects/Addhealth/Design).

Harris, Mark N. and Xueyan Zhao. 2007. "A Zero-inflated Ordered Probit Model, with an Application to Modelling Tobacco Consumption." *Journal of Econometrics* 141(2):1073-99.

Hausman, Jerry A., Jason Abrevaya, and Fiona M. Scott-Morton. 1998. "Misclassification of the Dependent Variable in a Discrete-response Setting." *Journal of Econometrics* 87(2):239-69.

Heitjan, Daniel F. and Donald B. Rubin. 1990. "Inference from Coarse Data Via Multiple Imputation with Application to Age Heaping." *Journal of the American Statistical Association* 85:304-14.

Lambert, Diane. 1992. "Zero-inflated Poisson Regression, with an Application to Defects in Manufacturing." *Technometrics* 34(1):1-14.

Li, Xinxin and Lorin M. Hitt. 2008. "Self-selection and Information Role of Online Product Reviews." *Information Systems Research* 19(4):456-74.

Lim, Hwa Kyung, Wai Keung Li, and Philip L. H. Yu. 2014. "Zero-inflated Poisson Regression Mixture Model." *Computational Statistics & Data Analysis* 71: 151-58.

Lin, Ting Hsiang and Min-Hsiao Tsai, M. H. 2013. "Modeling Health Survey Data with Excessive Zero and K Responses." *Statistics in Medicine* 32(9):1572-83.

Luca, Michael and Georgios Zervas. 2016. "Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud." *Management Science* 62(12):3412-27.

Magnus, Brooke E. and David Thissen. 2017. "Item Response Modeling of Multivariate Count Data with Zero Inflation, Maximum Inflation, and Heaping." *Journal of Educational and Behavioral Statistics* 42(5):531-58.

Moghimbeigi, Abbas, Mohammad Reza Eshraghian, Kazem Mohammad, and Brian McArdle. 2009. "A Score Test for Zero-inflation in Multilevel Count Data." *Computational Statistics & Data Analysis* 53:1239-48.

Monod, Anthea. 2014. "Random Effects Modeling and the Zero-inflated Poisson Distribution." *Communications in* Statistics—Theory *and Methods* 43(4): 664-80.

Mwalili, Samuel M., Emmanuel Lesaffre, and Dominique Declerck. 2008. "The Zero-inflated Negative Binomial Regression Model with Correction for Misclassification: An Example in Caries Research." *Statistical Methods in Medical Research* 17(2):123-39.

Pickering, R. 1992. "Digit Preference in Estimated Gestational Age." *Statistics in Medicine* 11:1225-38.

Poston, Dudley and Sherry McKibben. 2003. "Using Zero-inflated Count Regression Models to Estimate the Fertility of U.S. Women." *Journal of Modern Applied Statistical Methods* 2(2). Retrieved (http://digitalcommons.wayne.edu/jmasm/vol2/iss2/10).

Ridout, Martin, John Hinde, and Clarice G. B. Demetrio. 2001. "A Score Test for Testing Zero Inflated Poisson Regression Model against Zero Inflated Negative Binomial Alternatives." *Biometrics* 57:219-23.

SAS. 1990. *SAS/IML Software: Usage and Reference, Version 6*. Cary, NC: SAS Institute.

SAS. 2013. "PROC NLMIXED: The NLMIXED Procedure: SAS/STAT(R) 9.2 User's Guide, Second Edition." Retrieved May 2, 2017 (https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#nlmixed_toc.htm).

Schane, Rebecca E., Pamela M. Ling, and Stanton A. Glantz. 2010. "Health Effects of Light and Intermittent Smoking." *Circulation* 121(13):1518-22.

Staub, Kevin E. and Rainer Winkelmann. 2013. "Consistent Estimation of Zero-inflated Count Models." *Health Economics* 22(6):673-86.

Su, Xiaogang, Juanjuan Fan, Richard A. Levine, Xianming Tan, and Arvind Tripathi. 2013. "Multiple-inflaction Poisson Model with L1 Regularization." *Statistica Sinica* 23(3):1071-90.

Sweeney, James, John Haslett, and Andrew C. Parnell, A.P. 2014. "Zero & N-inflated Binomial Distributions with Applications." arXiv:1407.0064v5.

Vieira, A. M. C., J. P. Hinde, and C. G. B. Demetrio. 2010. "Zero-inflated Proportion Data Models Applied to a Biological Control Assay." *Journal of Applied Statistics* 27(3): 373-89.

Voronca, Delia C., Leonard E Egede, and Mulugeta Gebregziabher. 2014. "Analysis of Zero Inflated Longitudinal Data Using PROC NLMIXED." Unpublished manuscript, North Carolina State University, Raleigh, NC.

Vuong, Quang H. 1989. "Likelihood Ratio Tests for Model Selection and Non-nested Hypotheses." *Econometrica* 57(2):307-33.

Wang, Hao and Daniel F. Heitjan. 2008. "Modeling Heaping in Self-reported Cigarette Counts." *Statistics in Medicine* 27(19):3789-804.

Author Biographies

Tianji Cai is an assistant professor in the Department of Sociology at the University of Macau. His research interests center on two areas: social mechanism of how biological and social factors influence behaviors and developing quantitative research methods. Specifically, he is interested in the integration of sociology with biological factors in the studies of sociological issues such as social and health behavior, stratification, and social networks. The current project in research methods focuses on the issues of sampling weights in multilevel/longitudinal models, for example, evaluating the effects of ignoring or incorporating sampling weights on the estimation of multilevel models under various sampling designs and developing methods to test the informativeness of the sampling weights.

Yiwei Xia is a PhD candidate in the Department of Sociology at the University of Macau. His research interests include quantitative methodology, juvenile delinquency, and substance abuse.

Yisu Zhou is an associate professor in the Faculty of Education at the University of Macau. He is interested in quantitative social sciences in general. His research expertise is educational policy in the greater China region. His past research projects use large-scale assessment data both internationally and domestically, covering issues such as the social environment of learning, teacher education and labor market, and social stratification in schools. He is currently working on social segregation in schools in Chinese societies.