Health Serv Outcomes Res Method (2018) 18:175–214
https://doi.org/10.1007/s10742-018-0183-6

Are marginalized two-part models superior to non-marginalized two-part models for count data with excess zeroes? Estimation of marginal effects, model misspecification, and model selection

Xueyan Liu[1] · Bo Zhang[2] · Li Tang[1] · Zhiwei Zhang[3] · Ning Zhang[4] · Jeroan J. Allison[2] · Deo Kumar Srivastava[1] · Hui Zhang[1]

Received: 8 November 2017 / Revised: 28 March 2018 / Accepted: 25 May 2018 /
Published online: 5 June 2018

Abstract The marginalized two-part models, including the marginalized zero-inflated Poisson and negative binomial models, have been proposed in the literature for modelling cross-sectional healthcare utilization count data with excess zeroes and overdispersion. The motivation for these proposals was to directly capture the overall marginal effects and to avoid post-modelling effect calculations that are needed for the non-marginalized conventional two-part models. However, are marginalized two-part models superior to non-marginalized two-part models because of their structural property? Is it true that the marginalized two-part models can provide direct marginal inference? This article aims to answer these questions through a comprehensive investigation. We first summarize the existing non-marginalized and marginalized two-part models and then develop marginalized hurdle Poisson and negative binomial models for cross-sectional count data with abundant zero counts. Our interest in the investigation lies particularly in the (average) marginal effect and (average) incremental effect and the comparison of these effects. The estimators of these effects are presented, and variance estimators are derived by using delta methods and Taylor series approximations. Though the marginalized models attract attention because of the alleged convenience of direct marginal inference, we provide evidence for the impact of model misspecification of the marginalized models over the conventional models, and provide evidence for the importance of goodness-of-fit evaluation and model selection in

✉ Bo Zhang
Bo.Zhang@umassmed.edu

✉ Hui Zhang
Hui.Zhang@stjude.org

[1] Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, TN 38105, USA

[2] Department of Quantitative Health Sciences, University of Massachusetts Medical School, Worcester, MA 01605, USA

[3] Department of Statistics, University of California at Riverside, Riverside, CA 92521, USA

[4] Department of Health Policy and Promotion, School of Public Health and Health Sciences, University of Massachusetts, Amherst, MA 01003, USA

differentiating between the marginalized and non-marginalized models. An empirical analysis of the German Socioeconomic Panel data is presented.

Keywords Count data · Excess zeroes · Two-part models · Marginal inference · Marginalization · Healthcare utilization

1 Introduction

Count data collected in healthcare utilization studies exhibit remarkable features, including excess zeroes from the non-users of healthcare facilities, overdispersion, and multi-modality due to between-subject heterogeneity (Cameron and Trivedi 2005, 2013). The conventional two-part count models, such as zero-inflated Poisson and negative binomial models and hurdle Poisson and negative binomial models, have long been used to accommodate these features when analyzing healthcare utilization data in health economics and health services research. Recently, several marginalized two-part count models were proposed in the literature. These marginalized models were largely promoted because the models can allegedly provide "direct" marginal inference, whereas their non-marginalized counterparts cannot do so. This article is devoted to determine whether it is true that the marginalized two-part models are superior to non-marginalized two-part models for count data with excess zeroes because of the claimed advantage of direct marginal inference.

Investigational studies in healthcare utilization in health economics and health services research often set up their primary outcome as the number of usages of healthcare facilities, such as visits to primary care doctors or emergency department and days of hospitalization after surgeries. A consequence of this is that the outcomes are count observations expressed numerically as non-negative integers. Excessive zeroes occur when the studies involve participants that do not use any healthcare facilities during the study period. To account for excess zeroes in count data, Lambert (1992) first introduced the zero-inflated Poisson (ZIP) models as a two-part mixture model that combined a regular Poisson model with a latent binary distribution that governs the probability of generating structural zeroes and generating the Poisson counts. Since then, the ZIP models have been one of the most popular models for count data with excess zeroes (Winkelmann 2008) and have been extended to multivariate settings (Li et al. 1999) and models with random effects (Hall 2000). Recently, Long et al. (2014) modified the ZIP models and developed the marginalized ZIP (MZIP) models by specifying linear predictors for the overall mean of the count variable rather than using a linear predictor for the mean of Poisson component in the ZIP models. The MZIP models were claimed to be able to provide overall marginal effect inference while accommodating the mechanism of mixture of a random population and a degenerate component and also avoiding the misuse of conditional mean as the population mean.

A natural extension of the ZIP models is the zero-inflated negative binomial (ZINB) models proposed by Greene (1994), in which the Poisson model in the ZIP models is replaced by a negative binomial model and the component for structural zeroes remains. When equality of mean and variance fails even after structural zeroes are split, the ZINB models would be a better choice than the ZIP models. Ridout et al. (2001) indicated a serious bias of parameter estimates by the ZIP modelling if the nonzero counts are overdispersed in relation to the ZIP models, and they provided a score test for testing the ZIP models against the ZINB models. As a parallel proposal with Long et al. (2014), Preisser et al. (2016) introduced the marginalized ZINB (MZINB) models and justified the MZINB models by comparing parameter estimates with the ZIP and MZIP models from fitting simulated MZINB data. The rationale behind the MZINB models is identical to that of MZIP models in terms of seeking instant marginal inference. The difference is that the MZINB models specified the negative binomial distribution, rather than the Poisson distribution, to account for additional overdispersion.

Closely related to these zero-inflated models are the hurdle models that were originally proposed by Cragg (1971) and formally presented by Mullahy (1986). The hurdle models are dichotomous models combining a binary distribution of probing the count below or above the hurdle with a truncated count model above the hurdle. Hurdle-at-zero models are the most common hurdle models, among which the hurdle Poisson (Mullahy 1986) (HP) and hurdle negative binomial (HNB) models developed by Pohlmeier and Ulrich (1995) are the top choices in empirical analysis. Because of the complete separation of zero counts from the population of positive counts, hurdle models can accommodate count data with either zero-inflation or zero-deflation and either underdispersion or overdispersion based on the underlying count distributions. Although Kassahun et al. (2014) and Tabb et al. (2016) explored the marginalized hurdle models for panel count data, no research in the literature includes formal discussion of marginalized hurdle models for cross-sectional count data with excess zeroes.

The primary objective of this article is to rectify the previous misleading statement on the marginalized two-part models over their non-marginalized counterparts in characterizing the count data with excess zeroes. This article thoroughly defines and derives the (average) marginal and incremental effects of a covariate with respect to the overall marginal mean of count outcomes with excess zeroes in the context of four non-marginalized two-part models (the ZIP, ZINB, HP, and HNB models) and four marginalized two-part models (the MZIP, MZINB, marginalized hurdle Poisson, and marginalized hurdle negative binomial models). Among these models, it is the first time that the marginalized hurdle Poisson (MHP) and marginalized hurdle negative binomial (MHNB) models are formally proposed for cross-sectional data. Estimators and variance estimators are developed for the (average) marginal and incremental effect in each of the models. The derived effects and their estimators demonstrate that both types of models, either non-marginalized or marginalized, can provide marginal inference on the overall marginal mean of count outcomes with excess zeroes. The marginalized models have simplified marginal and incremental effects, but there is not any extra computational burden in estimating the effects by using the conventional models. Instead of promoting the use of marginalized two-part models, we emphasize that the two types of models should be taken as parallel competitors and that unjustified faith in either type of model will result in model misspecification bias in statistical inference, including the inference of marginal means. Comprehensive numerical studies were conducted and are reported in this article to illustrate the consequences on statistical analysis when the marginalized two-part models are misused for the data that are generated from the non-marginalized two-part models and vice versa. Substantial biases were observed in statistical inference in the numerical studies when either type of model was mistakenly replaced by its counterpart. We propose a solution to the possible misuse of either type of model, which is to conduct rigorous model comparison and selection by using the information reflected in the observed data. Simulation studies were conducted and are reported to investigate the three model comparison and selection criteria: the effect-specific mean square error criterion (Dow and Norton 2003), the information criteria, and the Vuong's closeness test (Vuong 1989). The studies verify that the information criteria can best select among the two types of models regardless of the magnitude of sample size. Although the performance of the mean square error criterion is acceptable, the Vuong's closeness test is not an ideal tool for distinguishing the non-marginalized and marginalized two-part models.

This article is organized as follows: Sect. 2 introduces the definitions of marginal and incremental effects and their average effects; Sect. 3 reviews the two zero-inflated models (i.e. ZIP and ZINB models) and their marginalized peers (i.e. MZIP and MZINB models), including their effect estimation; Sect. 4 discusses the two hurdle models (i.e. HP and HNB models) and the marginalized hurdle models (i.e. MHP and MHNB models) with effect estimation; Sect. 5 derives the variance estimators of marginal effects, incremental effects, and their average effects in these two-part models; Sect. 6 provides a thorough discussion on the question of superiority of marginalized two-part models over non-marginalized two-part models; Sect. 7 presents our simulation studies for comparison between ZIP and MZIP and between HNB and MHNB; Sect. 8 discusses model selection via effect estimates and shows further comparisons between paired models based on our simulation results; Sect. 9 reports an empirical analysis of German Socioeconomic Panel data using these models.

2 Marginal effects and average marginal effects

Let y be a count response variable (dependent variable) that takes the value of either a positive integer or zero. Let $x = (x_1, x_2, \dots, x_J)$ be a vector of J covariates (independent variables). Denote $\mu(x) = \mathrm{E}\,(y|x)$ the expected value of y, then the *marginal effect* (Greene 2002), or *partial effect*, of the jth covariate x_j on the expected overall outcome is defined as

$$\eta_j(x) = \frac{\partial \mu(x)}{\partial x_j} = \frac{\partial\,\mathrm{E}\,(y|x)}{\partial x_j}, \tag{1}$$

where $j = 1, 2, \dots, J$. The marginal effect allows us to quantify the marginal change in the expected overall outcome when covariate x_j changes by a small amount while holding other covariates $x_{(-j)} = (x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_J)'$ constant. The marginal effect is a function of both unknown parameters and covariates, and is evaluated at a particular combination of covariate values, say $x = x^{(0)}$, with the parameter estimates. Another quantity of interest in health economics and health services research is average marginal effect. Note that the marginal effect (1) represents the effect of the subpopulation that satisfies $x = x^{(0)}$. This subpopulation may be a small or even negligible portion of the entire population. When the study objective is to assess the marginal effect on the outcomes in the entire population, the expected value of the marginal effect over the population distribution of all covariates is then the primary interest. This is quantified by the *average marginal effect* that is defined as

$$\mathrm{E}\,\{\eta_j(x)\} = \mathrm{E}\left\{\frac{\partial \mu(x)}{\partial x_j}\right\},$$

in which the expectation is taken with respect to $x = (x_1, x_2, \dots, x_J)$.

When x_j is a categorical covariate that represents multiple levels or experimental groups, the quantity of interest is usually the *incremental effect* (Greene 2002; Basu and Rathouz 2005). The incremental effect is defined as

$$\pi_j(x) = \mu(x_j = l_2, x_{(-j)}) - \mu(x_j = l_1, x_{(-j)}),$$

in which l_1 and l_2 are two levels of covariate x_j. The incremental effect measures the difference in the expected overall outcome at the two levels of x_j while holding other covariates $x_{(-j)}$ constant. When x_j is binary that takes values 1 or 0, the incremental effect from level 0 to level 1 is

$$\pi_j(x) = \mu(x_j = 1, x_{(-j)}) - \mu(x_j = 0, x_{(-j)}).$$

The average incremental effect is defined as

$$\mathrm{E}\ \{\pi_j(x)\} = \ \mathrm{E}\ \{\mu(x_j = l_2, x_{(-j)}) - \mu(x_j = l_1, x_{(-j)})\},$$

in which the expectation is taken with respect to $x_{(-j)} = (x_1, x_2, \dots, x_{(j-1)}, x_{(j+1)}, \dots, x_J)$.

3 Estimation of marginal effects: zero-inflated models and marginalized zero-inflated models

3.1 Zero-inflated Poisson and negative binomial models

The zero-inflated Poisson (ZIP) model (Lambert 1992) for the count data with excess zeroes is a mixture of constant zeroes and a standard Poisson model. For the ith outcome y_i, $i = 1, 2, \dots, n$, the ZIP model is given by

$$y_i = \begin{cases} 0 & \text{if} \quad c_i = 1; \\ y_i^* & \text{if} \quad c_i = 0, \end{cases}$$

in which c_i is a Bernoulli variable with mean $\psi_i = P(c_i = 1)$ and $y_i^* \sim$ Poisson (μ_i) with a probability mass function (pmf) $g(y_i^*|\mu_i) = e^{-\mu_i}\mu_i^{y_i^*}/y_i^*!$. The marginal pmf of y_i in the ZIP model is

$$f(y_i) = \begin{cases} \psi_i + (1-\psi_i)e^{-\mu_i}, & \text{for} \ \ y_i = 0, \\ (1-\psi_i)e^{-\mu_i}\mu_i^{y_i}/y_i!, & \text{for} \ \ y_i = 1, 2, 3, \dots, \end{cases}$$

with E $(y_i) = \mu_i(1-\psi_i)$ and var $(y_i) = \mu_i(1-\psi_i)(1+\mu_i\psi_i)$. In contrast to the standard Poisson model, the overdispersion in the ZIP model is measured by var$(y_i)/$ E $(y_i) = 1 + \psi_i\mu_i$. To further characterize the dependence of y_i on the covariates, Lambert (1992) constructed a ZIP model as

$$\ln(\mu_i) = x_i'\beta \qquad \text{and} \qquad \mathrm{logit}(\psi_i) = \ln\left(\frac{\psi_i}{1-\psi_i}\right) = z_i'\gamma, \tag{2}$$

in which $x_i = (x_{i0} \equiv 1, x_{i1}, x_{i2}, \dots, x_{iJ_1})'$ and $z_i = (z_{i0} \equiv 1, z_{i1}, z_{i2}, \dots, z_{iJ_2})'$ are two vectors of covariates that may or may not overlap with each other, $\beta = (\beta_0, \beta_1, \dots, \beta_{J_1})'$ is the vector of regression coefficients for the Poisson process and $\gamma = (\gamma_0, \gamma_1, \dots, \gamma_{J_2})'$ is the vector of regression coefficients for the excess zeroes, and β_0 and γ_0 are regression intercepts. Let $\theta = (\beta', \gamma')'$ denote the vector that contains all unknown parameters in the ZIP model, then the log-likelihood function of the model is

$$\begin{aligned} \ell(\theta|y, x, z) = & \sum_{i=1,\dots,n;\, y_i=0} \ln(e^{z_i'\gamma} + e^{-e^{x_i'\beta}}) + \sum_{i=1,\dots,n;\, y_i>0} (y_i x_i'\beta - e^{x_i'\beta} - \ln y_i!) \\ & - \sum_{i=1}^{n} \ln(1 + e^{z_i'\gamma}). \end{aligned} \tag{3}$$

In (3), the observed data are represented by the collection of $y = (y_1, y_2, \dots, y_n)'$, $x = (x_1', x_2', \dots, x_n')'$, and $z = (z_1', z_2', \dots, z_n')'$. Estimates of the unknown parameters in the

ZIP model can be obtained by maximizing (3) using numerical optimization methods. Lambert (1992) also derived the joint probability density function of y_i and c_i and the Expectation-Maximization (EM) algorithm for maximizing the complete log-likelihood function.

An extension of the ZIP model is the zero-inflated negative binomial (ZINB) model that assumes the count data with excess zeroes are observed from a mixture of constant zeroes and a standard negative binomial model. For the ith outcome y_i, the ZINB model is given by

$$y_i = \begin{cases} 0 & \text{if } \ c_i = 1; \\ y_i^* & \text{if } \ c_i = 0, \end{cases}$$

in which c_i is still a Bernoulli variable with mean $\psi_i = P(c_i = 1)$ but y_i follows a negative binomial distribution NegBin (μ_i, α) with a pmf

$$g(y_i|\mu_i, \alpha) = \frac{\Gamma(y_i + \alpha)}{\Gamma(\alpha)\Gamma(y_i + 1)} \left(\frac{\alpha}{\alpha + \mu_i}\right)^{\alpha} \left(\frac{\mu_i}{\alpha + \mu_i}\right)^{y_i}, \quad y_i = 0, 1, 2, \ldots.$$

The marginal pmf of y_i in the ZINB model is

$$f(y_i) = \begin{cases} \psi_i + (1 - \psi_i)\left(\dfrac{\alpha}{\alpha + \mu_i}\right)^{\alpha}, & \text{for } \ y_i = 0, \\ (1 - \psi_i)\dfrac{\Gamma(y_i + \alpha)}{\Gamma(\alpha)\Gamma(y_i + 1)}\left(\dfrac{\alpha}{\alpha + \mu_i}\right)^{\alpha}\left(\dfrac{\mu_i}{\alpha + \mu_i}\right)^{y_i}, & \text{for } \ y_i = 1, 2, \ldots, \end{cases} \tag{4}$$

with E $(y_i) = \mu_i(1 - \psi_i)$ and var $(y_i) = \mu_i(1 - \psi_i)\big(1 + \mu_i/\alpha + \mu_i\psi_i\big)$. The negative binomial model for y_i^* in the count component of the ZINB model represents the overdispersed y_i^* in that var $(y_i^*)/$ E $(y_i^*) = 1 + \mu_i/\alpha$. The overdispersion in the ZINB model as a whole is measured by var $(y_i)/$ E $(y_i) = 1 + \mu_i(1 + \alpha\psi_i)/\alpha$. To describe the dependence of y_i on the covariates, the ZINB model specifies that

$$\ln(\mu_i) = x_i'\beta \qquad \text{and} \qquad \text{logit}(\psi_i) = z_i'\gamma,$$

as in (2). Let $\theta = (\beta', \gamma', \alpha)'$ denote the vector that contains all unknown parameters in the ZINB model, then the log-likelihood function of the model is

$$\begin{aligned} \ell(\theta|x, z, y) = &- \sum_{i=1}^{n} \ln(1 + e^{z_i'\gamma}) + \sum_{i=1,\ldots,n;\, y_i=0} \ln\left\{e^{z_i'\gamma} + \left(\frac{\alpha}{\alpha + e^{x_i'\beta}}\right)^{\alpha}\right\} \\ &+ \sum_{i=1,\ldots,n;\, y_i>0} \left\{ \sum_{j=0}^{y_i-1} \ln(j + \alpha) - \big(\alpha + y_i\big)\ln\big(\alpha + e^{x_i'\beta}\big) \right. \\ &\left. + \alpha\ln\alpha + y_i x_i'\beta - \ln y_i! \right\}. \end{aligned}$$

The marginal and incremental effects of the ZIP and ZINB models can be derived according to the definitions given in Sect. 2 and the modelling framework of the ZIP model. The marginal expectation of the response y_i in both the ZIP and the ZINB models possesses an identical expression as below:

$$\text{E}\,(y_i|x_i, z_i) = \mu_i(1 - \psi_i) = \frac{e^{x_i'\beta}}{1 + e^{z_i'\gamma}}.$$

Here, to derive the marginal and incremental effects of the two models, we investigate the scenario, in which the first J_0 covariates in x_i are duplicated in z_i; that is, we assume that covariate x_{ij} in the vector x_i and covariate z_{ij} in the vector z_i are identical covariates for $j = 1, 2, \ldots, J_0$ with $J_0 \leq J_1$ and $J_0 \leq J_2$. Then, the marginal effect $\eta_j(x_i, z_i, \theta)$ of covariate x_{ij}, or z_{ij}, with respect to the overall mean of response y_i in the ZIP and ZINB models is

$$\eta_j(x_i, z_i, \theta) = \frac{\partial \,\mathrm{E}\,(y_i|x_i, z_i)}{\partial x_{ij}} = \frac{e^{x_i'\beta}}{(1 + e^{z_i'\gamma})^2}\{\beta_j + e^{z_i'\gamma}(\beta_j - \gamma_j)\}.$$

When x_{ij} is a categorical covariate, the incremental effect $\pi_j(x_{i(-j)}, z_{i(-j)}, \theta)$ from level l_1 to level l_2 in x_{ij} with respect to y_i is

$$\begin{aligned}\pi_j(x_{i(-j)}, z_{i(-j)}, \theta) =&\ \mathrm{E}\,(y_i|x_{i(-j)}, z_{i(-j)}, x_{ij} = l_2, \theta) - \ \mathrm{E}\,(y_i|x_{i(-j)}, z_{i(-j)}, x_{ij} = l_1, \theta)\\ =&\ \frac{e^{x_{i(-j)}'\beta_{(-j)} + l_2\beta_j}}{1 + e^{z_{i(-j)}'\gamma_{(-j)} + l_2\gamma_j}} - \frac{e^{x_{i(-j)}'\beta_{(-j)} + l_1\beta_j}}{1 + e^{z_{i(-j)}'\gamma_{(-j)} + l_1\gamma_j}}.\end{aligned}$$

The estimates of the marginal and incremental effects $\hat{\eta}_j(x_i, z_i, \hat{\theta})$ and $\hat{\pi}_j(x_{i(-j)}, z_{i(-j)}, \hat{\theta})$ are obtained by substituting the unknown parameters θ in $\eta_j(x_i, z_i, \theta)$ and $\pi_j(x_{i(-j)}, z_{i(-j)}, \theta)$ with their maximization likelihood estimates $\hat{\theta}$. The average marginal and average incremental effects of covariate x_{ij} with respect to the overall mean of response y_i in the ZIP and ZINB models are

$$\begin{aligned}\bar{\eta}_j(\theta) =&\ \mathrm{E}_{\,x,z}(\eta_j(x, z, \theta)) = \int \eta_j(x, z, \theta)\, dF_{x,z}(x, z)\\ =&\int \frac{e^{x'\beta}}{(1 + e^{z'\gamma})^2}\{\beta_j + e^{z'\gamma}(\beta_j - \gamma_j)\}\, dF_{x,z}(x, z)\end{aligned}$$

and

$$\begin{aligned}\bar{\pi}_j(\theta) =&\ \mathrm{E}_{\,x_{(-j)},z_{(-j)}}(\pi_j(x_{(-j)}, z_{(-j)}, \theta)) = \int \pi_j(x_{(-j)}, z_{(-j)}, \theta)\, dF_{x_{(-j)},z_{(-j)}}(x_{(-j)}, z_{(-j)}),\\ =&\int \left(\frac{e^{x_{(-j)}'\beta_{(-j)} + l_2\beta_j}}{1 + e^{z_{(-j)}'\gamma_{(-j)} + l_2\gamma_j}} - \frac{e^{x_{(-j)}'\beta_{(-j)} + l_1\beta_j}}{1 + e^{z_{(-j)}'\gamma_{(-j)} + l_1\gamma_j}}\right) dF_{x_{(-j)},z_{(-j)}}(x_{(-j)}, z_{(-j)}),\end{aligned}$$

in which $F_{x,z}(x, z)$ and $F_{x_{(-j)},z_{(-k)}}(x_{(-j)}, z_{(-k)})$ are the joint cumulative density functions of (*x*, *z*) and $(x_{(-j)}, z_{(-k)})$, respectively. Estimators of the average marginal and average incremental effects are given by averaging the estimated marginal and incremental effects that are evaluated at the observed data:

$$\hat{\bar{\eta}}_j(\hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \hat{\eta}_j(x_i, z_i, \hat{\theta}) \qquad \text{and} \qquad \hat{\bar{\pi}}_j(\hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \hat{\pi}_j(x_{i(-j)}, z_{i(-j)}, \hat{\theta}). \tag{5}$$

3.2 Marginalized zero-inflated Poisson and negative binomial models

The formulas of the (average) marginal and (average) incremental effects in the ZIP and ZINB models are complex as shown in Sect. 3.1. Especially, both the ZIP and the ZINB models cannot provide direct marginal inference on the overall mean of the response due to the fact that these models does not connect a linear predictor of covariates directly to the overall marginal mean. Long et al. (2014) and Preisser et al. (2016) proposed marginalized versions of the ZIP and ZINB models, named marginalized zero-inflated Poisson (MZIP) model and marginalized zero-inflated negative binomial (MZINB) model, respectively. The MZIP model (Long et al. 2014) still assumes that the zero-inflated count outcome $y_i = 0$ when $c_i = 1$ and $y_i = y_i^*$ when $c_i = 0$, in which the binary variable $c_i \sim$ Bernoulli (ψ_i) and $y_i^* \sim$ Poisson (μ_i) with a pmf $g(y_i^*|\mu_i) = e^{-\mu_i}\mu_i^{y_i^*}/y_i^*!$. However, instead of specifying a linear model for the log of μ_i as in (2), the MZIP model assumes that the overall mean of the outcome is directly associated with a linear predictor of covariates:

$$\ln(v_i) = x_i'\beta \qquad \text{and} \qquad \text{logit}(\psi_i) = z_i'\gamma, \tag{6}$$

in which $v_i =$ E (y_i). With $\theta = (\beta', \gamma')'$, the log-likelihood function of the MZIP model is

$$\begin{aligned}\ell(\theta|y,x,z) = &-\sum_{i=1}^{n} \ln(1+e^{z_i'\gamma}) + \sum_{i=1,\dots,n;\, y_i=0} \ln\left\{e^{z_i'\gamma} + e^{-e^{x_i'\beta}(1+e^{z_i'\gamma})}\right\} \\ &+ \sum_{i=1,\dots,n;\, y_i>0} \left\{y_i x_i'\beta + y_i \ln(1+e^{z_i'\gamma}) - e^{x_i'\beta}(1+e^{z_i'\gamma}) - \ln y_i!\right\}.\end{aligned}$$

The MZINB model (Preisser et al. 2016) assumes that the zero-inflated count outcome $y_i = 0$ when $c_i = 1$ and $y_i = y_i^*$ when $c_i = 0$, in which $c_i \sim$ Bernoulli (ψ_i) and $y_i^* \sim$ NegBin (μ_i) with a pmf described in (4). To get a direct marginal interpretation on the overall mean of the response, as in the MZIP model, the same two regression equations are constructed in the MZINB model

$$\ln(v_i) = x_i'\beta \qquad \text{and} \qquad \text{logit}(\psi_i) = z_i'\gamma,$$

in which $v_i =$ E (y_i). With $\theta = (\beta', \gamma', \alpha)'$, the log-likelihood function of the MZINB model is

$$\begin{aligned}\ell(\theta|y,x,z) = &-\sum_{i=1}^{n} \ln(1+e^{z_i'\gamma}) + \sum_{i=1,\dots,n;\, y_i=0} \ln\left[e^{z_i'\gamma} + \left\{\frac{\alpha}{\alpha + e^{x_i'\beta}(1+e^{z_i'\gamma})}\right\}^{\alpha}\right] \\ &+ \sum_{i=1,\dots,n;\, y_i>0} \left[\sum_{j=0}^{y_i-1} \ln(\alpha+j) - (\alpha+y_i)\ln\{\alpha + e^{x_i'\beta}(1+e^{z_i'\gamma})\}\right] \\ &+ \sum_{i=1,\dots,n;y_i>0} \left\{\alpha\ln\alpha + y_i x_i'\beta + y_i \ln(1+e^{z_i'\gamma}) - \ln y_i!\right\}.\end{aligned}$$

The specification of the MZIP and the MZINB models leads to

$$\text{E }(y_i|x_i, z_i) = e^{x_i'\beta}. \tag{7}$$

This concise representation on the overall mean response results in the simplified formulas of the marginal and incremental effects for the MZIP and the MZINB models. It can be derived that, for these models, the marginal and incremental effects of covariate x_{ij}, or z_{ij}, with respect to the overall mean of response y_i are

$$\eta_j(x_i, z_i, \theta) = \frac{\partial \text{ E }(y_i|x_i, z_i)}{\partial x_{ij}} = \beta_j e^{x_i'\beta} \tag{8}$$

and

$$\begin{aligned}\pi_j(x_{i(-j)}, z_{i(-j)}, \theta) = &\ \mathrm{E}\,(y_i | x_{i(-j)}, z_{i(-j)}, x_{ij} = l_2, \theta) - \mathrm{E}\,(y_i | x_{i(-j)}, z_{i(-j)}, x_{ij} = l_1, \theta) \\ = &\ e^{x'_{i(-j)}\beta_{(-j)} + l_2\beta_j} - e^{x'_{i(-j)}\beta_{(-j)} + l_1\beta_j},\end{aligned} \tag{9}$$

respectively. The estimates of the marginal and incremental effects are

$$\hat{\eta}_j(x_i, z_i, \hat{\theta}) = \hat{\beta}_j e^{x'_i\hat{\beta}} \qquad \text{and} \qquad \hat{\pi}_j(x_{i(-j)}, z_{i(-k)}, \hat{\theta}) = e^{x'_{i(-j)}\hat{\beta}_{(-j)} + l_2\hat{\beta}_j} - e^{x'_{i(-j)}\hat{\beta}_{(-j)} + l_1\hat{\beta}_j}. \tag{10}$$

The average marginal effect and average incremental effect of covariate x_{ij} with respect to the overall mean of response y_i in the MZIP and the MZINB models are

$$\bar{\eta}_j(\theta) = \mathrm{E}_{x,z}(\eta_j(x, z, \theta)) = \int \beta_j e^{x'\beta} dF_{x,z}(x, z) \tag{11}$$

and

$$\begin{aligned}\bar{\pi}_j(\theta) = &\ \mathrm{E}_{x_{(-j)}, z_{(-k)}}(\pi_j(x, z, \theta)) \\ = &\int \left(e^{x'_{(-j)}\beta_{(-j)} + l_2\beta_j} - e^{x'_{(-j)}\beta_{(-j)} + l_1\beta_j} \right) dF_{x_{(-j)}, z_{(-k)}}(x_{(-j)}, z_{(-k)}),\end{aligned} \tag{12}$$

respectively. Estimators of these average marginal and average incremental effects are again given by averaging the estimated marginal and incremental effects evaluated at the observed data:

$$\hat{\bar{\eta}}_j(\hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \hat{\eta}_j(x_i, z_i, \hat{\theta}) \qquad \text{and} \qquad \hat{\bar{\pi}}_j(\hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \hat{\pi}_j(x_{i(-j)}, z_{i(-j)}, \hat{\theta}). \tag{13}$$

4 Estimation of marginal effects: hurdle models and marginalized hurdle models

4.1 Hurdle Poisson and negative binomial models

Hurdle models (Mullahy 1986) characterize the statistical processes that generate observations below the hurdle and above the hurdle. Hurdle models are two-component models, in which one component is a dichotomous model for a latent binary variable indicating outcomes below or above the hurdle and another component is, when the hurdle at zero is crossed, a truncated model for outcomes above the hurdle. In the hurdle models, a Bernoulli binary variable c_i with a mean of ψ_i is combined with a zero-truncated count variable y_i^* with a zero-truncated pmf

$$\tilde{g}(y_i^*) = \frac{g(y_i^*)}{1 - g(0)}, \quad y_i^* = 1, 2, 3, \ldots, \tag{14}$$

yielding the outcome y_i through the mechanism

$$y_i = \begin{cases} 0 & \text{if } c_i = 1; \\ y_i^* & \text{if } c_i = 0, \end{cases}$$

in which $g(y_i^*)$ that has support over the nonnegative integers including zero is a pmf before zero truncation. The marginal pmf of y_i in the hurdle model is

$$f(y_i) = \begin{cases} \psi_i, & \text{for } y_i = 0, \\ \dfrac{1-\psi_i}{1-g(0)} g(y_i), & \text{for } y_i = 1, 2, 3, \ldots . \end{cases}$$

The mean and variance of y_i in the hurdle models are

$$\begin{aligned} \mathrm{E}\,(y_i) &= \frac{1-\psi_i}{1-g(0)}\mu_i, \\ \mathrm{var}\,(y_i) &= \frac{1-\psi_i}{1-g(0)}\sigma_i^2 + \frac{(1-\psi_i)(\psi_i - g(0))}{(1-g(0))^2}\mu_i^2, \end{aligned}$$

in which μ_i and σ_i^2 are the mean and variance, respectively, of the pmf $g(y_i^*)$. In the hurdle models, the zero observations are below the hurdle and the positive counts are assumed to be produced from the zero-truncated count model when above the hurdle. Because the zero and positive count data are completely separated by the two parts of the models, the hurdle models can be used to fit both zero-inflated count data and zero-deflated count data. The zero inflation or deflation is determined by the magnitude of $1 - \psi_i$ and $1 - g(0)$ or, equivalently, the magnitude of ψ_i and $g(0)$. The overdispersion in the hurdle models is measured by $\dfrac{\mathrm{var}\,(y_i)}{\mathrm{E}\,(y_i)} = \dfrac{\sigma_i^2}{\mu_i} + \dfrac{\psi_i - g(0)}{1-g(0)}\mu_i$.

Conventional hurdle models include hurdle Poisson (HP) model and hurdle negative binomial (HNB) model. The HP model is constructed by specifying $g(y_i^*)$, the pmf before zero truncation in (14), to be the pmf of Poisson (μ_i). As such, the marginal pmf of y_i in the HP model is

$$f(y_i) = \begin{cases} \psi_i, & \text{for } y_i = 0, \\ \dfrac{1-\psi_i}{1-e^{-\mu_i}} \cdot \dfrac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}, & \text{for } y_i = 1, 2, 3, \ldots . \end{cases}$$

The HNB model is constructed by specifying $g(y_i^*)$ to be the pmf of NegBin (μ_i, α). The marginal pmf of y_i in the HNB model is

$$f(y_i) = \begin{cases} \psi_i, & \text{for } y_i = 0, \\ \dfrac{1-\psi_i}{1-\{\alpha/(\alpha+\mu_i)\}^\alpha} \cdot \dfrac{\Gamma(y_i+\alpha)}{\Gamma(\alpha)\Gamma(y_i+1)} \cdot \left(\dfrac{\alpha}{\alpha+\mu_i}\right)^\alpha \left(\dfrac{\mu_i}{\alpha+\mu_i}\right)^{y_i}, & \text{for } y_i = 1, 2, 3, \ldots . \end{cases}$$

To characterize the dependence of y_i on the covariates, the HP and HNB models set up two regression models as in (2):

$$\ln(\mu_i) = x_i'\beta \qquad \text{and} \qquad \mathrm{logit}(\psi_i) = z_i'\gamma.$$

Denote the parameter vector in the HP and HNB models by $\theta = (\beta', \gamma')'$ and $\theta = (\beta', \gamma', \alpha)'$, respectively, then the log-likelihood function of the HP model is

$$\begin{aligned}\ell(\theta|y,x,z) = & \sum_{i=1,\ldots,n;\, y_i=0} z_i'\gamma - \sum_{i=1}^{n} \ln(1+e^{z_i'\gamma}) \\ & + \sum_{i=1,\ldots,n;\, y_i>0} \left\{ y_i x_i'\beta - \ln(e^{e^{x_i'\beta}} - 1) - \ln y_i! \right\},\end{aligned}$$

and the log-likelihood function of the HNB model is

$$\begin{aligned}\ell(\theta|y,x,z) = & \sum_{i=1,\ldots,n;\, y_i=0} z_i'\gamma - \sum_{i=1}^{n} \ln(1+e^{z_i'\gamma}) \\ & + \sum_{i=1,\ldots,n;\, y_i>0} \left\{ \sum_{j=0}^{y_i-1} \ln(\alpha+j) - \ln y_i! + \alpha \ln \alpha + y_i x_i'\beta \right\} \\ & - \sum_{i=1,\ldots,n;\, y_i>0} \left[\ln\left\{ 1 - \left(\frac{\alpha}{\alpha + e^{x_i'\beta}} \right)^{\alpha} \right\} + (\alpha + y_i)\ln(\alpha + e^{x_i'\beta}) \right].\end{aligned}$$

The marginal and incremental effects of the HP and HNB models are considerably complex. The marginal expectation of the response y_i in the HP model is

$$\mathrm{E}\,(y_i|x_i, z_i) = \frac{e^{x_i'\beta}}{(1+e^{z_i'\gamma})(1-e^{-e^{x_i'\beta}})}.$$

It can be derived that the marginal effect $\eta_j(x_i, z_i, \theta)$ of covariate x_{ij}, or z_{ij}, with respect to the overall mean of response y_i in the HP model is

$$\begin{aligned}\eta_j(x_i, z_i, \theta) = & \frac{e^{x_i'\beta + e^{x_i'\beta}}}{(1+e^{z_i'\gamma})^2 (e^{e^{x_i'\beta}} - 1)^2} \\ & \cdot \left[\left(e^{e^{x_i'\beta}} - 1 \right) \left\{ \beta_j + (\beta_j - \gamma_j) e^{z_i'\gamma} \right\} - \beta_j e^{x_i'\beta} (1 + e^{z_i'\gamma}) \right],\end{aligned}$$

When x_{ij} is a categorical covariate, the incremental effect $\pi_j(x_{i(-j)}, z_{i(-j)}, \theta)$ from level l_1 to level l_2 in x_{ij} with respect to y_i is

$$\begin{aligned}
&\pi_j(x_{i(-j)}, z_{i(-j)}, \theta)\\
&\quad = \mathrm{E}\,(y_i|x_{i(-j)}, z_{i(-j)}, x_{ij} = l_2, \theta) - \mathrm{E}\,(y_i|x_{i(-j)}, z_{i(-j)}, x_{ij} = l_1, \theta)\\
&\quad = \frac{e^{x'_{i(-j)}\beta_{(-j)}+l_2\beta_j+e^{x'_{i(-j)}\beta_{(-j)}+l_2\beta_j}}}{\{1+e^{z'_{i(-j)}\gamma_{(-j)}+l_2\gamma_j}\}\{e^{x'_{i(-j)}\beta_{(-j)}+l_2\beta_j}-1\}}\\
&\qquad - \frac{e^{x'_{i(-j)}\beta_{(-j)}+l_1\beta_j+e^{x'_{i(-j)}\beta_{(-j)}+l_1\beta_j}}}{\{1+e^{z'_{i(-j)}\gamma_{(-j)}+l_1\gamma_j}\}\{e^{x'_{i(-j)}\beta_{(-j)}+l_1\beta_j}-1\}}.
\end{aligned}$$

The average marginal and average incremental effects of covariate x_{ij} with respect to the overall mean of response y_i in the HNB model are consequently

$$\bar{\eta}_j(\theta) = \int \frac{e^{x'\beta+e^{x'\beta}}}{(1+e^{z'\gamma})^2(e^{e^{x'\beta}}-1)^2} \cdot \left[\left(e^{e^{x'\beta}}-1\right)\left\{\beta_j+(\beta_j-\gamma_j)e^{z'\gamma}\right\} - \beta_j e^{x'\beta}(1+e^{z'\gamma})\right] dF_{x,z}(x,z)$$

and

$$\begin{aligned}
\bar{\pi}_j(\theta) = \int &\left[\frac{e^{x'_{(-j)}\beta_{(-j)}+l_2\beta_j+e^{x'_{(-j)}\beta_{(-j)}+l_2\beta_j}}}{\{1+e^{z'_{(-j)}\gamma_{(-j)}+l_2\gamma_j}\}\{e^{x'_{(-j)}\beta_{(-j)}+l_2\beta_j}-1\}}\right.\\
&\left. - \frac{e^{x'_{(-j)}\beta_{(-j)}+l_1\beta_j+e^{x'_{(-j)}\beta_{(-j)}+l_1\beta_j}}}{\{1+e^{z'_{(-j)}\gamma_{(-j)}+l_1\gamma_j}\}\{e^{x'_{(-j)}\beta_{(-j)}+l_1\beta_j}-1\}}\right]\\
&\cdot dF_{x_{(-j)},z_{(-j)}}(x_{(-j)}, z_{(-j)}).
\end{aligned}$$

The marginal expectation of the response y_i in the HNB model is

$$\mathrm{E}\,(y_i|x_i, z_i) = \frac{e^{x'_i\beta}}{(1+e^{z'_i\gamma})\left\{1-\left(\frac{\alpha}{\alpha+e^{x'_i\beta}}\right)^{\alpha}\right\}}.$$

Then, the (average) marginal and (average) incremental effects in the HNB model are

$$
\begin{aligned}
\eta_j(x_i, z_i, \theta) = \frac{\partial \, \mathrm{E}\,(y_i|x_i, z_i)}{\partial x_{ij}} = {} & \frac{e^{x_i'\beta}}{(1+e^{z_i'\gamma})^2 \left\{ 1 - \left(\frac{\alpha}{\alpha + e^{x_i'\beta}} \right)^{\alpha} \right\}^2} \\
& \cdot \left[\left\{ 1 - \left(\frac{\alpha}{\alpha + e^{x_i'\beta}} \right)^{\alpha} \right\} \{ \beta_j + (\beta_j - \gamma_j) e^{z_i'\gamma} \} \right. \\
& \left. - \beta_j e^{x_i'\beta} \left(\frac{\alpha}{\alpha + e^{x_i'\beta}} \right)^{\alpha+1} (1 + e^{z_i'\gamma}) \right],
\end{aligned}
$$

$$
\begin{aligned}
\pi_j(x_{i(-j)}, z_{i(-j)}, \theta) = {} & \mathrm{E}\,(y_i|x_{i(-j)}, z_{i(-j)}, x_{ij} = l_2, \theta) - \mathrm{E}\,(y_i|x_{i(-j)}, z_{i(-j)}, x_{ij} = l_1, \theta) \\
= {} & \frac{e^{x_{i(-j)}'\beta_{(-j)} + l_2\beta_j}}{(1 + e^{z_{i(-j)}'\gamma_{(-j)} + l_2\gamma_j}) \left\{ 1 - \left(\frac{\alpha}{\alpha + e^{x_{i(-j)}'\beta_{(-j)} + l_2\beta_j}} \right)^{\alpha} \right\}} \\
& - \frac{e^{x_{i(-j)}'\beta_{(-j)} + l_1\beta_j}}{(1 + e^{z_{i(-j)}'\gamma_{(-j)} + l_1\gamma_j}) \left\{ 1 - \left(\frac{\alpha}{\alpha + e^{x_{i(-j)}'\beta_{(-j)} + l_1\beta_j}} \right)^{\alpha} \right\}},
\end{aligned}
$$

$$
\begin{aligned}
\bar{\eta}_j(\theta) = \int & \frac{e^{x'\beta}}{(1+e^{z'\gamma})^2 \left\{ 1 - \left(\frac{\alpha}{\alpha + e^{x'\beta}} \right)^{\alpha} \right\}^2} \\
& \cdot \left[\left\{ 1 - \left(\frac{\alpha}{\alpha + e^{x'\beta}} \right)^{\alpha} \right\} \{ \beta_j + (\beta_j - \gamma_j) e^{z'\gamma} \} \right. \\
& \left. - \beta_j e^{x'\beta} \left(\frac{\alpha}{\alpha + e^{x'\beta}} \right)^{\alpha+1} (1 + e^{z'\gamma}) \right] dF_{x,z}(x, z),
\end{aligned}
$$

and

$$
\begin{aligned}
\bar{\pi}_j(\theta) = \int & \left[\frac{e^{x_{(-j)}'\beta_{(-j)} + l_2\beta_j}}{(1 + e^{z_{(-j)}'\gamma_{(-j)} + l_2\gamma_j}) \left\{ 1 - \left(\frac{\alpha}{\alpha + e^{x_{(-j)}'\beta_{(-j)} + l_2\beta_j}} \right)^{\alpha} \right\}} \right. \\
& \left. - \frac{e^{x_{(-j)}'\beta_{(-j)} + l_1\beta_j}}{(1 + e^{z_{(-j)}'\gamma_{(-j)} + l_1\gamma_j}) \left\{ 1 - \left(\frac{\alpha}{\alpha + e^{x_{(-j)}'\beta_{(-j)} + l_1\beta_j}} \right)^{\alpha} \right\}} \right] dF_{x_{(-j)}, z_{(-j)}}(x_{(-j)}, z_{(-j)}),
\end{aligned}
$$

respectively. The estimates of the marginal and incremental effects in the two models can be obtained by substituting the unknown parameters in the effects with their maximum likelihood estimates. Estimators of the average marginal and average incremental effects are obtained by averaging the estimated marginal and incremental effects that are evaluated at the observed data.

4.2 Marginalized hurdle Poisson and negative binomial models

It is straightforward to construct marginalized hurdle Poisson and negative binomial models for cross-sectional count data with excess zero. However, it has not been officially reported in the literature, although Tabb et al. (2016) proposed marginalized random-effects hurdle Poisson and negative binomial models for panel count data. The marginalized hurdle models assume, as in the hurdle models in Sect. 4.1, that the zero-inflated count outcome $y_i = 0$ when $c_i = 1$ and $y_i = y_i^*$ when $c_i = 0$, in which the binary variable $c_i \sim$ Bernoulli (ψ_i) and y_i^* follows a zero-truncated distribution with a pmf $\tilde{g}(y_i^*) = \dfrac{g(y_i^*)}{1 - g(0)}$, $y_i^* = 1, 2, 3, \ldots$. To achieve the goal of making direct inference on the overall mean of the outcome y_i, the marginalized hurdle models specify as in (6) that

$$\ln(v_i) = x_i'\beta \qquad \text{and} \qquad \operatorname{logit}(\psi_i) = z_i'\gamma,$$

in which $v_i =$ E (y_i). The marginalized hurdle Poisson (MHP) model can be constructed by assigning $g(y_i^*)$, the pmf before zero truncation, to be the pmf of Poisson (μ_i), and the marginalized hurdle negative binomial (MHNB) model is constructed by assigning $g(y_i^*)$, the pmf before zero truncation, to be the pmf of NegBin (μ_i, α).

It can be derived that, for the MHP model, the log-likelihood function is

$$\begin{aligned}\ell(\theta, \mu | y, x, z) = &\sum_{i=1,\ldots,n;\, y_i=0} \left\{ z_i'\gamma - \ln(1 + e^{z_i'\gamma}) \right\} \\ &+ \sum_{i=1,\ldots,n;\, y_i>0} \left\{ x_i'\beta - \ln y_i! + (y_i - 1)\ln \mu_i - \mu_i \right\},\end{aligned} \tag{15}$$

$$\text{subject to} \qquad e^{x_i'\beta} = \frac{\mu_i}{(1 + e^{z_i'\gamma})(1 - e^{-\mu_i})}, \qquad i = 1, 2, \ldots, n, \tag{16}$$

in which $\theta = (\beta', \gamma')'$ and $\mu = (\mu_1, \mu_2, \ldots, \mu_n)$. For the MHNB model, the log-likelihood function is

$$\begin{aligned}\ell(\theta, \mu | y, x, z) = &\sum_{i=1,\ldots,n;\, y_i=0} \left\{ z_i'\gamma - \ln(1 + e^{z_i'\gamma}) \right\} \\ &+ \sum_{i=1,\ldots,n;\, y_i>0} \left\{ \sum_{j=0}^{y_i-1} \ln(j + \alpha) - \ln y_i! + x_i'\beta \right\} \\ &+ \sum_{i=1,\ldots,n;\, y_i>0} \left\{ (y_i - 1)\ln \mu_i + \alpha \ln \alpha - (\alpha + y_i)\ln(\alpha + \mu_i) \right\},\end{aligned} \tag{17}$$

$$\text{subject to} \qquad e^{x_i'\beta} = \frac{\mu_i}{(1 + e^{z_i'\gamma})[1 - \{\alpha/(\alpha + \mu_i)\}^{\alpha}]}, \qquad i = 1, 2, \ldots, n, \tag{18}$$

in which $\theta = (\beta', \gamma', \alpha)'$ and $\mu = (\mu_1, \mu_2, \ldots, \mu_n)$. The maximum likelihood estimates are obtained in the MPH and MHNB models by numerically solving $\hat{\theta} = \max_\theta \ell(\theta)$ in (15) and (17) but subject to (16) and (18), respectively.

The specification of the MHP and MHNB models leads to $\mathrm{E}\,(y_i|x_i, z_i) = e^{x_i'\beta}$, which is identical to the expression in (7) for the MZIP and MZINB models. Therefore, the (average) marginal and (average) incremental effects for the MHP and MHNB models, and their estimates, are given by (8)–(13).

5 Variance estimation of marginal effects

Asymptotic variances of the estimated marginal effects and average marginal effects can be derived using the delta method and Taylor series approximations. Note that the parameters θ in the models summarized in Sects. 3 and 4 are estimated by maximizing their log-likelihood functions. Under regular conditions, $\hat{\theta} \xrightarrow{P} \theta$ as $n \to \infty$ and

$$\hat{\theta} \xrightarrow{D} N(\theta, [I_n(\theta)]^{-1}),$$

where $I_n(\theta) = -\,\mathrm{E}\left(\frac{\partial^2 \ell(\theta)}{\partial \theta^2}\right)$ is the Fisher information matrix and $[I_n(\theta)]^{-1} = (1/n)[I_1(\theta)]^{-1} \xrightarrow{P} 0$ as $n \to \infty$. The observed Fisher information matrix is $I_n(\hat{\theta}) = -\left.\frac{\partial^2 \ell(\theta)}{\partial \theta^2}\right|_{\theta=\hat{\theta}}$. By the delta method, variances of the estimated marginal and incremental effects $\hat{\eta}_j(x_i, z_i, \hat{\theta})$ and $\hat{\pi}_j(x_{i(-j)}, z_{i(-j)}, \hat{\theta})$, as continuously differentiable functions of the parameters, can be estimated by

$$\widehat{\mathrm{var}}_{\,\hat{\eta}_j}(x_i, z_i, \hat{\theta}) = \left(\nabla_\theta \hat{\eta}_j(x_i, z_i, \hat{\theta})\right)' \left[I_n(\hat{\theta})\right]^{-1} \left(\nabla_\theta \hat{\eta}_j(x_i, z_i, \hat{\theta})\right) \tag{19}$$

and

$$\widehat{\mathrm{var}}_{\,\hat{\pi}_j}(x_{i(-j)}, z_{i(-j)}, \hat{\theta}) = \left(\nabla_\theta \hat{\pi}_j(x_{i(-j)}, z_{i(-j)}, \hat{\theta})\right)' \left[I_n(\hat{\theta})\right]^{-1} \left(\nabla_\theta \hat{\pi}_j(x_{i(-j)}, z_{i(-j)}, \hat{\theta})\right), \tag{20}$$

respectively.

To derive the variance estimator of the average marginal effect $\hat{\bar{\eta}}_j(\hat{\theta})$, the multivariate Taylor's theorem is applied for $\hat{\bar{\eta}}_j(\hat{\theta})$ with respect to $\hat{\theta}$ at the true value θ:

$$\hat{\bar{\eta}}_j(\hat{\theta}) = \hat{\bar{\eta}}_j(\theta) + \left(\nabla_\theta \hat{\bar{\eta}}_j(\theta)\right)'(\hat{\theta} - \theta) + h_1(\tilde{\theta})(\hat{\theta} - \theta),$$

where $\tilde{\theta}$ is some value between θ and $\hat{\theta}$, and $\lim_{\hat{\theta} \to \theta} h_1(\tilde{\theta}) = 0$ in probability. Thus,

$$\begin{aligned}
\mathrm{var}\,(\hat{\bar{\eta}}_j(\hat{\theta})) = {} & \mathrm{var}\,(\hat{\bar{\eta}}_j(\theta)) + \mathrm{var}\left(\left(\nabla_\theta \hat{\bar{\eta}}_j(\theta)\right)'(\hat{\theta} - \theta)\right) \\
& + 2\ \mathrm{cov}\left(\hat{\bar{\eta}}_j(\theta), \left(\nabla_\theta \hat{\bar{\eta}}_j(x_i, z_i, \theta)\right)'(\hat{\theta} - \theta)\right) \\
& + \mathrm{var}\left(h_1(\tilde{\theta})(\hat{\theta} - \theta)\right) + 2\ \mathrm{cov}\left(\hat{\bar{\eta}}_j(\theta)), h_1(\tilde{\theta})(\hat{\theta} - \theta)\right) \\
& + 2\ \mathrm{cov}\left(\left(\nabla_\theta \hat{\bar{\eta}}_j(\theta)\right)'(\hat{\theta} - \theta), h_1(\tilde{\theta})(\hat{\theta} - \theta)\right).
\end{aligned} \tag{21}$$

The first term on the right-hand side of (21) is estimated by $\widehat{\text{var}}\,(\hat{\hat{\eta}}_j(\theta)) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\hat{\eta}_j(x_i, z_i, \hat{\theta}) - \hat{\hat{\eta}}_j(\hat{\theta}) \right)^2$. For the second term, the delta method gives

$$\left(\left(\nabla_\theta \hat{\hat{\eta}}_j(\theta) \right)' (\hat{\theta} - \theta) \right) \xrightarrow{D} N\left(0, \left(\nabla_\theta \hat{\hat{\eta}}_j(\theta) \right)' [I_n(\theta)]^{-1} \left(\nabla_\theta \hat{\hat{\eta}}_j(\theta) \right) \right), \tag{22}$$

which implies E $\left(\left(\nabla_\theta \hat{\hat{\eta}}_j(\theta) \right)' (\hat{\theta} - \theta) \right) \xrightarrow{P} 0$, as $n \to \infty$. Therefore, the second term on the right-hand side of (21) is estimated by $\left(\nabla_\theta \hat{\hat{\eta}}_j(\hat{\theta}) \right)' [I_n(\hat{\theta})]^{-1} \left(\nabla_\theta \hat{\hat{\eta}}_j(\hat{\theta}) \right)$, in which $\nabla_\theta \hat{\hat{\eta}}_j(\hat{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta \hat{\eta}_j(\hat{\theta})$. In addition, the consistency of $\hat{\theta}$, the normality in (22), the fact that $\lim_{\hat{\theta} \to \theta} h_1(\tilde{\theta}) = 0$ as $n \to \infty$, and the Slutsky's Theorem together indicate the remaining four terms in (21) approach 0 in probability as $n \to \infty$. Therefore, the estimator of var $(\hat{\hat{\eta}}_j(\hat{\theta}))$ is

$$\begin{aligned} \widehat{\text{var}}\,(\hat{\hat{\eta}}_j(\hat{\theta})) = & \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\hat{\eta}_j(x_i, z_i, \hat{\theta}) - \frac{1}{n} \sum_{i=1}^{n} \hat{\hat{\eta}}_j(\hat{\theta}) \right)^2 \\ & + \left(\frac{1}{n} \sum_{i=1}^{n} \nabla_\theta \hat{\eta}_j(x_i, z_i, \hat{\theta}) \right)' [I_n(\hat{\theta})]^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \nabla_\theta \hat{\eta}_j(x_i, z_i, \hat{\theta}) \right). \end{aligned} \tag{23}$$

It can be derived similarly that the estimator of var $(\hat{\hat{\pi}}_j(x_i, z_i, \hat{\theta}))$ is

$$\begin{aligned} \widehat{\text{var}}\,(\hat{\hat{\pi}}_j(\hat{\theta})) = & \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\hat{\pi}_j(x_{i(-j)}, z_{i(-j)}, \hat{\theta}) - \frac{1}{n} \sum_{i=1}^{n} \hat{\hat{\pi}}_j(\hat{\theta}) \right)^2 \\ & + \left(\frac{1}{n} \sum_{i=1}^{n} \nabla_\theta \hat{\pi}_j(x_{i(-j)}, z_{i(-j)}, \hat{\theta}) \right)' \left[I_n(\hat{\theta}) \right]^{-1} \\ & \times \left(\frac{1}{n} \sum_{i=1}^{n} \nabla_\theta \hat{\pi}_j(x_{i(-j)}, z_{i(-j)}, \hat{\theta}) \right). \end{aligned} \tag{24}$$

Variance estimators (19), (20), (23), and (24) involve gradients of the marginal and incremental effects $\nabla_\theta \eta_j(x_i, z_i, \theta)$ and $\nabla_\theta \pi_j(x_{i(-j)}, z_{i(-j)}, \theta)$ that need to be derived specifically for each of the models in Sects. 3 and 4. For the marginalized models, i.e. the MZIP, MZINB, MHP, and MHNB models, the gradients of their marginal effects and incremental effects are

$$\nabla_\theta \eta_j(x_i, z_i, \theta) = \beta_j e^{x_i' \beta} \sum_{m=0}^{J_1} x_{im} u_{(m+1)} + e^{x_i' \beta} u_{(j+1)},$$

$$\begin{aligned} \nabla_\theta \pi_j(x_{i(-j)}, z_{i(-j)}, \theta) = & \left[e^{x_{i(-j)}' \beta_{(-j)} + l_2 \beta_j} - e^{x_{i(-j)}' \beta_{(-j)} + l_1 \beta_j} \right] \sum_{m=0, \neq j}^{J_1} x_{im} u_{(m+1)} \\ & + \left[l_2 e^{x_{i(-j)}' \beta_{(-j)} + l_2 \beta_j} - l_1 e^{x_{i(-j)}' \beta_{(-j)} + l_1 \beta_j} \right] u_{(j+1)}, \end{aligned}$$

where $u_{(m)}$ is a unit vector with 1 in the mth component and 0 in others. The length of $u_{(m)}$ is $(J_1 + J_2 + 2)$ for the MZIP and MHP models and is $(J_1 + J_2 + 3)$ for the MZINB and MHNB models. The gradients of the marginal effects and incremental effects for the non-marginalized models, i.e. the ZIP, ZINB, HP, and HNB models, are considerably complex and are reported in "Appendix".

6 Superiority of marginalized two-part models over non-marginalized two-part models: true or false?

Several previous articles Long et al. (2014), Preisser et al. (2016) promoted the use of marginalized two-part models over the traditional non-marginalized two-part models for the count data with excess zeroes. It has been argued that the marginalized two-part models can provide "direct" marginal inference, which gives an impression that these models are superior to the non-marginalized two-part models. Is this really true?

Marginal inference and interpretation The discussion in Sects. 3 and 4 reveals that both types of models, either marginalized or non-marginalized, can provide marginal inference through marginal effects on the overall mean of count outcomes with excess zeroes. The difference is that estimators and variance estimators of the marginal effects that are derived from the non-marginalized models are a little more complex than the marginalized models. For example, the marginal effect of a covariate x_{ij} with respect to the overall marginal mean derived from the ZIP and ZINB models is $\frac{\partial \mathrm{E}\,(y_i|x_i, z_i)}{\partial x_{ij}} = \frac{e^{x_i'\beta}}{(1 + e^{z_i'\gamma})^2}\{\beta_j + e^{z_i'\gamma}(\beta_j - \gamma_j)\}$, whereas this marginal effect is $\frac{\partial \mathrm{E}\,(y_i|x_i, z_i)}{\partial x_{ij}} = e^{x_i'\beta}$ in the MZIP and MZNB models. However, our numerical studies in Sect. 7 show that the numerical implementations of the marginal effect estimators for the two types of models are both convenient and computationally fast.

Furthermore, the argument that the marginalized two-part models can provide direct marginal inference is actually not precise. In health economics and health services research, what is concerned is the marginal effect on the overall mean of a response variable, not on any transformation of the overall mean. Only when a linear predictor is directly connected to the expectation of responses (e.g., $\mathrm{E}\,(y_i|x_i) = x_i'\beta$) can a direct marginal inference be made through the regression coefficients and the marginal effects $\frac{\partial \mathrm{E}\,(y_i|x_i)}{\partial x_{ij}} = \beta_j$. For the marginalized two-part models, it is obvious that the direct marginal inference can be made only for the logarithmic scale of the overall mean of the response $\frac{\partial \log\ \mathrm{E}\,(y_i|x_i, z_i)}{\partial x_{ij}} = \beta_j$ but not on the original scale.

Model misspecification and model selection The discussion in Sects. 3 and 4 reveals that the marginalized two-part models possess a linear representation in the logarithmic scale of the overall mean of outcomes but have a non-linear representation in the logarithmic scale of the mean of positive outcomes (in the marginalized hurdle models) or positive outcomes with some zeroes (in the marginalized zero-inflated models). In contrast, the non-marginalized two-part models possess a linear representation in the logarithmic scale of the mean of positive outcomes or positive outcomes with some zeroes but have a non-linear representation on the other side. Therefore, the marginalized or non-marginalized two-part models are indeed two parallel competitors in modelling count data with excess zeroes, and neither type of model is superior to the other. It would be problematic to, by default, believe that the linear representation should be imposed to any side. Model misspecification is always an issue when the assumed model is not true or is not close to the truth. In Sect. 7, we report the simulation studies that we conducted to show the consequences of model misspecification when fitting a marginalized model to the data generated from its non-marginalized counterpart and vice versa. Because of the bias that may be induced by model misspecification, it is recommended that data analysts follow formal model selection procedures to choose between the marginalized and conventional two-part models while making inference with the models. In Sect. 8, three model selection criteria are investigated to examine the performance of each of them in this particular setting.

7 Model misspecification: theories and numerical studies

In this section, we report the results gathered from the simulation studies that were conducted to investigate the impact of model misspecification on marginal effects estimation in the conventional and marginalized two-part models for zero-inflated count data. The investigation on model misspecification was restricted to two scenarios: (1) the underlying true model that generates the simulated data is a conventional two-part model, but the corresponding marginalized two-part model is fit to the data, and (2) the underlying true model is a marginalized two-part model, but the corresponding conventional two-part model is fit.

7.1 Theories on model misspecification

For either a marginalized or a nonmarginalized two-part model, consider the zero-inflated response variable y with its true probability density function $g(y)$. Let $\{f(y;\theta), \theta \in \Theta\}$ be a parametric family of probability density functions that may be misspecified for y. White (1982) showed that, under suitable regularity conditions, there exists a $\theta^* \in \Theta$ such that the quasi-maximum likelihood estimator $\hat{\theta}^{(n)} = \underset{\theta\in\Theta}{\text{argmax}}\, \frac{1}{n}\sum_{i=1}^{n} \log f(y_i;\theta)$ almost surely converges to θ^*, in which θ^* minimizes the Kullback–Leibler distance between $g(y)$ and $f(y;\theta)$:

$$I(g(x) : f(y;\theta)) = \text{E}_g\left\{\log \frac{g(y)}{f(y;\theta)}\right\}.$$

In addition, asymptotic normality holds for $\hat{\theta}^{(n)}$ as

$$\sqrt{n}(\hat{\theta}^{(n)} - \theta^*) \xrightarrow{\text{a.s.}} N(0, V(\theta^*))$$

and $V_n(\hat{\theta}^{(n)}) \xrightarrow{\text{a.s.}} V(\theta^*)$, where $V_n(\hat{\theta}^{(n)}) = A_n^{-1}(\hat{\theta}^{(n)}) B_n(\hat{\theta}^{(n)}) A_n(\hat{\theta}^{(n)})$, $V(\theta^*) = A^{-1}(\theta^*)B(\theta^*)A(\theta^*)$, and

$$A_n(\theta) = \left\{\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \log f(y_i;\theta)}{\partial\theta_k \partial\theta_l}\right\}, \quad B_n(\theta) = \left\{\frac{1}{n}\sum_{i=1}^{n} \frac{\partial \log f(y_i;\theta)}{\partial\theta_k}\frac{\partial \log f(y_i;\theta)}{\partial\theta_l}\right\},$$

$$A(\theta) = E\left(\left\{\frac{\partial^2 \log f(y;\theta)}{\partial\theta_k \partial\theta_l}\right\}\right), \quad B(\theta) = E\left(\left\{\frac{\partial \log f(y,\theta)}{\partial\theta_k}\frac{\partial \log f(y;\theta)}{\partial\theta_l}\right\}\right).$$

Table 1 True values, estimates, standard deviations (SD), standard errors (SE), and biases of the average marginal effects and average incremental effects for the true ZIP models and misspecified MZIP models

n	True μ value	$\beta = (\beta_0, \beta_1, \beta_2)$ value	Estimand	True value	True ZIP				Misspecified MZIP			
					Est.	SD	SE	Bias	Est.	SD	SE	Bias
100	4.697	(1.0, − 1, .10)	$\bar{\eta}_1(\theta)$	− 4.068	− 4.030	0.746	0.739	− 0.038	− 3.907	0.749	0.815	− 0.161
			$\bar{\pi}_2(\theta)$	1.156	1.135	0.571	0.558	0.021	1.102	0.618	0.601	0.054
	6.760	(− 2, 2.0, 2.0)	$\bar{\eta}_1(\theta)$	1.889	1.967	1.560	1.455	− 0.078	2.847	1.908	1.837	− 0.958
			$\bar{\pi}_2(\theta)$	2.490	2.501	1.342	1.307	− 0.011	3.100	1.723	1.587	− 0.610
	9.795	(2.0, 1.0, − 5)	$\bar{\eta}_1(\theta)$	1.179	1.159	0.672	0.654	0.020	1.862	0.689	0.720	− 0.683
			$\bar{\pi}_2(\theta)$	0.242	0.240	0.835	0.847	0.002	− 0.202	0.823	0.828	0.444
	12.882	(1.0, .50, 2.0)	$\bar{\eta}_1(\theta)$	0.455	0.473	0.816	0.811	− 0.018	1.480	0.723	0.749	− 1.025
			$\bar{\pi}_2(\theta)$	10.454	10.393	1.586	1.645	0.061	10.910	1.614	1.687	− 0.456
	15.051	(.50, 1.8, 1.0)	$\bar{\eta}_1(\theta)$	3.834	4.086	2.421	2.320	− 0.252	5.660	2.589	2.676	− 1.826
			$\bar{\pi}_2(\theta)$	4.638	4.723	2.053	1.999	− 0.085	5.288	2.197	2.146	− 0.650
500	4.702	(1.0, − 1, .10)	$\bar{\eta}_1(\theta)$	− 4.074	− 4.077	0.334	0.334	0.003	− 3.865	0.322	0.368	− 0.209
			$\bar{\pi}_2(\theta)$	1.156	1.150	0.236	0.243	0.006	1.052	0.260	0.254	0.104
	6.663	(− 2, 2.0, 2.0)	$\bar{\eta}_1(\theta)$	1.880	1.904	0.669	0.637	− 0.024	2.561	0.796	0.737	− 0.681
			$\bar{\pi}_2(\theta)$	2.497	2.496	0.608	0.575	0.001	2.899	0.680	0.631	− 0.402
	9.798	(2.0, 1.0, − 5)	$\bar{\eta}_1(\theta)$	1.181	1.174	0.290	0.293	0.007	1.867	0.301	0.319	− 0.686
			$\bar{\pi}_2(\theta)$	0.243	0.255	0.363	0.377	− 0.012	− 0.214	0.363	0.368	0.457
	12.890	(1.0, .50, 2.0)	$\bar{\eta}_1(\theta)$	0.456	0.461	0.351	0.359	− 0.005	1.513	0.304	0.330	− 1.057
			$\bar{\pi}_2(\theta)$	10.451	10.425	0.720	0.738	0.026	10.876	0.718	0.756	− 0.425
	14.990	(.50, 1.8, 1.0)	$\bar{\eta}_1(\theta)$	3.825	3.830	1.027	0.991	− 0.005	5.467	1.251	1.182	− 1.642
			$\bar{\pi}_2(\theta)$	4.650	4.609	0.914	0.871	0.041	5.181	1.056	0.952	− 0.531

Table 1 (continued)

n	True μ value	$\beta = (\beta_0, \beta_1, \beta_2)$ value	Estimand	True value	True ZIP				Misspecified MZIP			
					Est.	SD	SE	Bias	Est.	SD	SE	Bias
1000	4.709	(1.0, − 1, .10)	$\bar{\eta}_1(\theta)$	− 4.082	− 4.079	0.234	0.237	− 0.003	− 3.852	0.231	0.262	− 0.230
			$\bar{\pi}_2(\theta)$	1.157	1.150	0.167	0.171	0.007	1.042	0.181	0.178	0.115
	6.769	(− 2, 2.0, 2.0)	$\bar{\eta}_1(\theta)$	1.888	1.877	0.453	0.443	0.011	2.554	0.539	0.525	− 0.666
			$\bar{\pi}_2(\theta)$	2.510	2.494	0.420	0.406	0.016	2.916	0.471	0.454	− 0.406
	9.780	(2.0, 1.0, − 5)	$\bar{\eta}_1(\theta)$	1.181	1.180	0.202	0.208	0.001	1.877	0.206	0.227	− 0.696
			$\bar{\pi}_2(\theta)$	0.242	0.256	0.250	0.267	− 0.014	− 0.220	0.248	0.260	0.462
	12.927	(1.0, .50, 2.0)	$\bar{\eta}_1(\theta)$	0.454	0.461	0.244	0.254	− 0.007	1.528	0.220	0.233	− 1.074
			$\bar{\pi}_2(\theta)$	10.450	10.460	0.505	0.522	− 0.010	10.900	0.524	0.535	− 0.450
	15.191	(.50, 1.8, 1.0)	$\bar{\eta}_1(\theta)$	3.836	3.844	0.737	0.706	− 0.008	5.590	0.963	0.869	− 1.754
			$\bar{\pi}_2(\theta)$	4.668	4.652	0.638	0.622	0.016	5.284	0.761	0.695	− 0.616

If either a marginalized or a non-marginalized two-part model is correctly specified (i.e., its corresponding counterpart is misspecified), there exists a $\theta^{(0)} \in \Theta$ such that $f(y;\theta^{(0)}) = g(y)$ and the quasi-maximum likelihood estimator becomes the maximum likelihood estimator and $\theta^* = \theta^{(0)}$ with the inverse of the Fisher's information matrix as the asymptotic covariance matrix estimator. When the model is misspecified, the standard errors for $\hat{\theta}^{(n)}$ should be obtained from the sandwich estimator $V_n(\hat{\theta}^{(n)})$.

7.2 True ZIP models versus misspecified MZIP models

In the first simulation study, a total number of 500 data sets with three sample sizes $n = 100$, 500, and 1000 were generated from the ZIP model, in which the linear predictors were specified as $\ln(\mu_i) = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2$ and $\text{logit}(\psi_i) = \gamma_0 + z_{i1}\gamma_1 + z_{i2}\gamma_2$ with one continuous covariate $x_{i1} = z_{i1} \sim N(0, 1)$ and one binary covariate $x_{i2} = z_{i2} \sim \text{Bernoulli}(0.5)$. When generating the simulated data sets, five combinations of $\beta = (\beta_0, \beta_1, \beta_2)'$ (see Table 1 for details of the combinations) were considered such that the average values of μ_i's range from approximately 4–15. We fixed $\gamma = (\gamma_0, \gamma_1, \gamma_2) = (0.5, 1, -1)$ to maintain the average value of ψ_i's (i.e., the average percentage of structural zeroes) to be around the intermediate value of 50%. Both the ZIP and MZIP models were then fit to each of the simulated data sets, and the true average marginal effect $\bar{\eta}_1(\theta)$ of x_1 and the true average incremental effects $\bar{\pi}_2(\theta)$ of x_2 given by the two models, as well as their estimates $\hat{\bar{\eta}}_1(\hat{\theta})$ and $\hat{\bar{\pi}}_2(\hat{\theta})$, were calculated.

Table 1 reports the true value of the average marginal and incremental effects, the mean and standard deviation of effect estimates, and the mean of SEs given by the ZIP and MZIP models from fitting the 500 simulated data sets in three sample sizes. The results in Table 1 demonstrate that the estimated average marginal and incremental effects obtained from the underlying true model, the ZIP model, are unbiased across all combinations of $\beta = (\beta_0, \beta_1, \beta_2)'$ and sample sizes. The finite-sample bias of the estimates of average marginal and incremental effects given by the misspecified MZIP model is larger than the one given by the true model, though the bias usually does not exceed two times of standard errors. The simulation results in Table 1 also show that, for both models, the average SE of the average marginal and incremental effects is close to the corresponding standard deviation. In addition, the average SE evidently shrinks as the sample size increases from 100, 500, to 1000. This piece of evidence verifies that the variance estimation procedure, which we derived in Sect. 5 based upon the asymptotic properties of marginal and incremental effects, are valid for the finite samples.

7.3 True MZIP models versus misspecified ZIP models

The investigational plan of the remaining three simulation studies is comparable to the first study in Sect. 7.2. In the second simulation study, 500 data sets were produced with sample sizes $n = 100$, 500, and 1000 from the MZIP model by specifying the linear predictors as $\ln(\nu_i) = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2$ and $\text{logit}(\psi_i) = \gamma_0 + z_{i1}\gamma_1 + z_{i2}\gamma_2$ with $x_{i1} = z_{i1} \sim N(0, 1)$ and $x_{i2} = z_{i2} \sim \text{Bernoulli}(0.5)$. Note that $\ln(\nu_i) = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2$ is the linear predictor for marginal expectation of the counts including zeroes, instead of positive counts. As such, the expectation of positive counts satisfies $\mu_i = \nu_i/(1 - \psi_i) = e^{\beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2}(1 + e^{\gamma_0 + z_{i1}\gamma_1 + z_{i2}\gamma_2})$. Five combinations of $\beta = (\beta_0, \beta_1, \beta_2)'$ (see Table 2) were used for data generation, and the averages of the resulting μ_i's range from approximately 3 to 11. The parameter $\gamma = (\gamma_0, \gamma_1, \gamma_2) = (0.5, 1, -1)$ is also fixed yielding an average of ψ_i's being around the

Table 2 True values, estimates, standard deviations (SD), standard errors (SE), and biases of the average marginal effects and average incremental effects for the true MZIP models and misspecified ZIP models

n	True μ value	$\beta = (\beta_0, \beta_1, \beta_2)$ value	Estimand	True value	True MZIP				Misspecified ZIP			
					Est.	SD	SE	Bias	Est.	SD	SE	Bias
100	3.374	(.50, – .5, -– 5)	$\bar{\eta}_1(\theta)$	– 0.751	– 0.742	0.198	0.196	– 3.326	– 0.744	0.213	0.191	– 3.324
			$\bar{\pi}_2(\theta)$	– 0.735	– 0.733	0.386	0.391	– 0.002	– 0.722	0.399	0.403	– 0.013
	5.762	(.50, – 1, .50)	$\bar{\eta}_1(\theta)$	– 3.577	– 3.543	0.620	0.606	– 0.034	– 3.459	0.645	0.534	– 0.118
			$\bar{\pi}_2(\theta)$	1.756	1.738	0.686	0.620	0.018	1.668	0.715	0.623	0.088
	7.853	(1.0, – 1, .10)	$\bar{\eta}_1(\theta)$	– 4.697	– 4.663	0.760	0.765	0.034	– 4.559	0.816	0.687	– 0.138
			$\bar{\pi}_2(\theta)$	0.469	0.445	0.709	0.741	0.024	0.380	0.829	0.806	0.089
	9.162	(– .5, 1.0, .50)	$\bar{\eta}_1(\theta)$	1.310	1.508	0.781	0.793	– 0.198	1.243	0.812	0.746	0.067
			$\bar{\pi}_2(\theta)$	0.643	0.695	0.559	0.555	– 0.052	0.750	0.596	0.588	– 0.107
	10.742	(1.0, .50, -.5)	$\bar{\eta}_1(\theta)$	1.235	1.343	0.634	0.603	– 0.108	1.082	0.712	0.628	0.153
			$\bar{\pi}_2(\theta)$	– 1.209	– 1.244	0.772	0.761	0.035	– 1.032	0.865	0.837	– 0.177
500	3.382	(.50, – .5, -– 5)	$\bar{\eta}_1(\theta)$	– 0.750	– 0.750	0.079	0.083	0.000	– 0.751	0.087	0.083	0.001
			$\bar{\pi}_2(\theta)$	– 0.735	– 0.746	0.177	0.171	0.011	– 0.745	0.187	0.180	0.010
	5.770	(.50, – 1, .50)	$\bar{\eta}_1(\theta)$	– 3.586	– 3.585	0.274	0.271	– 0.001	– 3.540	0.302	0.240	– 0.046
			$\bar{\pi}_2(\theta)$	1.758	1.734	0.259	0.264	0.024	1.708	0.291	0.275	0.050
	7.857	(1.0, – 1, .10)	$\bar{\eta}_1(\theta)$	– 4.703	– 4.709	0.341	0.338	0.006	– 4.664	0.376	0.308	– 0.039
			$\bar{\pi}_2(\theta)$	0.470	0.461	0.330	0.310	0.009	0.446	0.383	0.355	0.024
	9.015	(– .5, 1.0, .50)	$\bar{\eta}_1(\theta)$	1.314	1.336	0.298	0.311	– 0.022	1.151	0.322	0.318	0.163
			$\bar{\pi}_2(\theta)$	0.644	0.641	0.247	0.227	0.003	0.705	0.273	0.250	– 0.061
	10.731	(1.0, .50, -.5)	$\bar{\eta}_1(\theta)$	1.235	1.248	0.258	0.257	– 0.013	1.056	0.321	0.277	0.179
			$\bar{\pi}_2(\theta)$	– 1.210	– 1.234	0.336	0.333	0.024	– 1.075	0.400	0.376	– 0.135

Table 2 (continued)

n	True μ value	$\beta = (\beta_0, \beta_1, \beta_2)$ value	Estimand	True value	True MZIP				Misspecified ZIP			
					Est.	SD	SE	Bias	Est.	SD	SE	Bias
1000	3.381	(.50, − .5, -− 5)	$\bar{\eta}_1(\theta)$	− 0.750	− 0.751	0.057	0.058	0.001	− 0.752	0.066	0.058	0.002
			$\bar{\pi}_2(\theta)$	− 0.735	− 0.733	0.122	0.121	− 0.002	− 0.731	0.129	0.127	− 0.004
	5.778	(.50, − 1, .50)	$\bar{\eta}_1(\theta)$	− 3.594	− 3.591	0.183	0.192	− 0.003	− 3.555	0.209	0.170	− 0.039
			$\bar{\pi}_2(\theta)$	1.761	1.754	0.190	0.186	0.007	1.741	0.206	0.194	0.020
	7.862	(1.0, − 1, .10)	$\bar{\eta}_1(\theta)$	− 4.710	− 4.720	0.237	0.240	0.010	− 4.689	0.265	0.218	− 0.021
			$\bar{\pi}_2(\theta)$	0.471	0.460	0.226	0.217	0.011	0.460	0.265	0.250	0.011
	9.139	(− .5, 1.0, .50)	$\bar{\eta}_1(\theta)$	1.321	1.327	0.211	0.218	− 0.006	1.151	0.243	0.226	0.170
			$\bar{\pi}_2(\theta)$	0.647	0.656	0.162	0.160	− 0.009	0.721	0.187	0.177	− 0.074
	10.817	(1.0, .50, -.5)	$\bar{\eta}_1(\theta)$	1.236	1.234	0.182	0.180	0.002	1.058	0.224	0.197	0.178
			$\bar{\pi}_2(\theta)$	− 1.211	− 1.215	0.234	0.234	0.004	− 1.073	0.275	0.267	− 0.138

Table 3 True values, estimates, standard deviations (SD), standard errors (SE), and biases of the average marginal effects and average incremental effects for the true HNB models and misspecified MHNB models

n	True μ value	$\beta = (\beta_0, \beta_1, \beta_2)$ value	Estimand	True value	True HNB				Misspecified MHNB			
					Est.	SD	SE	Bias	Est.	SD	SE	Bias
100	2.796	(.80, − .5, .20)	$\bar{\eta}_1(\theta)$	− 1.489	− 1.496	0.400	0.400	0.007	− 1.562	0.480	0.486	0.073
			$\bar{\pi}_2(\theta)$	1.021	1.044	0.536	0.565	− 0.023	1.112	0.576	0.610	− 0.091
	4.523	(− .5, -5, 2.5)	$\bar{\eta}_1(\theta)$	− 2.467	− 2.428	0.732	0.737	− 0.039	− 2.403	0.772	0.818	− 0.064
			$\bar{\pi}_2(\theta)$	5.445	5.377	1.166	1.182	0.068	5.254	1.274	1.197	0.191
	6.731	(2.0, − .5, -.5)	$\bar{\eta}_1(\theta)$	− 3.322	− 3.258	0.831	0.870	− 0.064	− 3.288	0.908	1.013	− 0.034
			$\bar{\pi}_2(\theta)$	− 0.476	− 0.407	1.231	1.201	− 0.069	− 0.249	1.307	1.261	− 0.227
	8.197	(2.2, .50, − 5)	$\bar{\eta}_1(\theta)$	− 0.156	− 0.109	0.560	0.564	− 0.047	0.314	0.592	0.640	− 0.470
			$\bar{\pi}_2(\theta)$	0.159	0.130	1.058	1.118	0.029	− 0.222	1.080	1.107	0.381
	10.949	(1.5, .50, 1.2)	$\bar{\eta}_1(\theta)$	0.149	0.244	0.963	0.911	− 0.095	0.438	0.924	0.947	− 0.289
			$\bar{\pi}_2(\theta)$	6.897	6.985	1.982	1.871	− 0.088	6.567	1.833	1.722	0.330
500	2.799	(.80, − .5, .20)	$\bar{\eta}_1(\theta)$	− 1.491	− 1.494	0.173	0.176	0.003	− 1.570	0.198	0.217	0.079
			$\bar{\pi}_2(\theta)$	1.020	1.036	0.256	0.254	− 0.016	1.107	0.266	0.273	− 0.087
	4.520	(− .5, -5, 2.5)	$\bar{\eta}_1(\theta)$	− 2.465	− 2.459	0.342	0.332	− 0.006	− 2.479	0.358	0.368	0.014
			$\bar{\pi}_2(\theta)$	5.446	5.421	0.532	0.539	0.025	5.393	0.608	0.549	0.053
	6.726	(2.0, − .5, -.5)	$\bar{\eta}_1(\theta)$	− 3.321	− 3.304	0.378	0.390	− 0.017	− 3.359	0.427	0.461	0.038
			$\bar{\pi}_2(\theta)$	− 0.478	− 0.467	0.567	0.545	− 0.011	− 0.315	0.604	0.572	− 0.163
	8.198	(2.2, .50, − 5)	$\bar{\eta}_1(\theta)$	− 0.157	− 0.157	0.264	0.242	0.000	0.250	0.265	0.272	− 0.407
			$\bar{\pi}_2(\theta)$	0.159	0.130	0.492	0.497	0.029	− 0.242	0.478	0.486	0.401
	10.954	(1.5, .50, 1.2)	$\bar{\eta}_1(\theta)$	0.150	0.185	0.384	0.388	− 0.035	0.380	0.368	0.396	− 0.230
			$\bar{\pi}_2(\theta)$	6.896	6.866	0.856	0.830	0.030	6.391	0.802	0.757	0.505

Table 3 (continued)

n	True μ value	$\beta = (\beta_0, \beta_1, \beta_2)$ value	Estimand	True value	True HNB				Misspecified MHNB			
					Est.	SD	SE	Bias	Est.	SD	SE	Bias
1000	2.799	(.80, – .5, .20)	$\bar{\eta}_1(\theta)$	– 1.492	– 1.495	0.120	0.124	0.003	– 1.571	0.137	0.153	0.079
			$\bar{\pi}_2(\theta)$	1.020	1.022	0.178	0.179	– 0.002	1.091	0.184	0.193	-0.071
	4.518	(– .5, -5, 2.5)	$\bar{\eta}_1(\theta)$	– 2.464	– 2.459	0.237	0.234	– 0.005	– 2.496	0.251	0.261	0.032
			$\bar{\pi}_2(\theta)$	5.447	5.441	0.371	0.382	0.006	5.464	0.399	0.391	– 0.017
	6.722	(2.0, – .5, -.5)	$\bar{\eta}_1(\theta)$	– 3.319	– 3.314	0.267	0.275	– 0.005	– 3.367	0.298	0.325	0.048
			$\bar{\pi}_2(\theta)$	– 0.479	– 0.480	0.381	0.387	0.001	– 0.320	0.408	0.405	– 0.159
	8.210	(2.2, .50, – 5)	$\bar{\eta}_1(\theta)$	– 0.157	– 0.160	0.178	0.170	0.003	0.242	0.186	0.190	– 0.399
			$\bar{\pi}_2(\theta)$	0.160	0.148	0.368	0.351	0.012	– 0.216	0.365	0.342	0.376
	10.957	(1.5, .50, 1.2)	$\bar{\eta}_1(\theta)$	0.150	0.172	0.270	0.272	-0.022	0.374	0.262	0.278	– 0.224
			$\bar{\pi}_2(\theta)$	6.895	6.861	0.619	0.587	0.034	6.386	0.561	0.535	0.509

Table 4 True values, estimates, standard deviations (SD), standard errors (SE), and biases of the average marginal effects and average incremental effects for the true MHNB models and misspecified HNB models

n	True μ values	$\beta = (\beta_0, \beta_1, \beta_2)$ values	Estimand	True value	True MHNB				Misspecified HNB			
					Est.	SD	SE	Bias	Est.	SD	SE	Bias
100	4.040	(– 1, .50, 2.0)	$\bar{\eta}_1(\theta)$	1.136	1.100	0.496	0.614	0.036	1.239	0.954	0.949	– 0.103
			$\bar{\pi}_2(\theta)$	3.028	2.977	0.814	0.934	0.051	3.157	1.214	1.231	– 0.129
	4.412	(– .5, – .2, 2.0)	$\bar{\eta}_1(\theta)$	– 0.516	0.613	17.127	1.101	– 1.129	– 0.427	0.469	0.407	– 0.089
			$\bar{\pi}_2(\theta)$	3.954	4.398	7.175	1.131	– 0.444	4.066	0.927	0.974	– 0.112
	7.663	(1.5, – .5, – .5)	$\bar{\eta}_1(\theta)$	– 2.043	– 2.004	0.752	0.739	– 0.039	– 2.103	0.714	0.686	0.060
			$\bar{\pi}_2(\theta)$	– 1.997	– 2.178	1.407	1.364	0.181	– 2.251	1.438	1.402	0.254
	9.340	(1.0, – .8, .50)	$\bar{\eta}_1(\theta)$	– 3.950	– 3.980	1.194	1.197	0.030	– 3.650	1.014	0.933	– 0.300
			$\bar{\pi}_2(\theta)$	2.423	2.381	1.492	1.465	0.042	2.074	1.420	1.383	0.349
	12.886	(1.5, .50, .50)	$\bar{\eta}_1(\theta)$	3.353	3.194	1.581	1.650	0.159	2.722	1.901	1.785	0.631
			$\bar{\pi}_2(\theta)$	3.287	3.047	1.897	1.923	0.240	3.901	2.344	2.387	– 0.614
500	4.033	(– 1, .50, 2.0)	$\bar{\eta}_1(\theta)$	1.134	1.076	0.170	0.251	0.058	1.147	0.381	0.388	– 0.013
			$\bar{\pi}_2(\theta)$	3.028	2.980	0.283	0.397	0.048	3.098	0.525	0.527	– 0.070
	4.469	(– .5, – .2, 2.0)	$\bar{\eta}_1(\theta)$	– 0.519	– 0.547	0.118	0.150	0.028	– 0.407	0.198	0.170	– 0.112
			$\bar{\pi}_2(\theta)$	3.954	3.932	0.354	0.372	0.022	4.060	0.384	0.429	– 0.106
	7.599	(1.5, – .5, – .5)	$\bar{\eta}_1(\theta)$	– 2.039	– 2.022	0.327	0.318	– 0.017	– 2.122	0.318	0.294	0.083
			$\bar{\pi}_2(\theta)$	– 1.997	– 2.042	0.577	0.600	0.045	– 2.111	0.587	0.620	0.114
	9.367	(1.0, – .8, .50)	$\bar{\eta}_1(\theta)$	– 3.960	– 4.027	0.520	0.533	0.067	– 3.692	0.453	0.415	– 0.268
			$\bar{\pi}_2(\theta)$	2.424	2.405	0.634	0.655	0.019	2.081	0.623	0.622	0.343
	12.897	(1.5, .50, .50)	$\bar{\eta}_1(\theta)$	3.357	3.297	0.684	0.726	0.060	2.760	0.912	0.804	0.597
			$\bar{\pi}_2(\theta)$	3.290	3.253	0.826	0.855	0.037	4.132	1.070	1.084	– 0.842

Table 4 (continued)

n	True μ values	$\beta = (\beta_0, \beta_1, \beta_2)$ values	Estimand	True value	True MHNB				Misspecified HNB			
					Est.	SD	SE	Bias	Est.	SD	SE	Bias
1000	4.041	(− .5, − .2, 2.0)	$\bar{\eta}_1(\theta)$	1.135	1.087	0.121	0.172	0.048	1.140	0.267	0.273	− 0.005
			$\bar{\pi}_2(\theta)$	3.031	2.993	0.189	0.277	0.038	3.102	0.365	0.373	− 0.071
	4.476	(− .5, − .2, 2.0)	$\bar{\eta}_1(\theta)$	− 0.518	− 0.528	0.097	0.103	0.010	− 0.411	0.143	0.119	− 0.107
			$\bar{\pi}_2(\theta)$	3.953	3.948	0.249	0.264	0.005	4.074	0.279	0.304	− 0.121
	7.598	(1.5, − .5, − .5)	$\bar{\eta}_1(\theta)$	− 2.039	− 2.041	0.239	0.224	0.002	− 2.141	0.236	0.208	0.102
			$\bar{\pi}_2(\theta)$	− 1.997	− 2.015	0.417	0.424	0.018	− 2.087	0.427	0.438	0.090
	9.382	(1.0, − .8, .50)	$\bar{\eta}_1(\theta)$	− 3.962	− 3.991	0.370	0.372	0.029	− 3.668	0.324	0.290	− 0.294
			$\bar{\pi}_2(\theta)$	2.426	2.416	0.440	0.462	0.010	2.095	0.435	0.439	0.331
	13.005	(1.5, .50, .50)	$\bar{\eta}_1(\theta)$	3.360	3.317	0.457	0.509	0.043	2.807	0.611	0.573	0.553
			$\bar{\pi}_2(\theta)$	3.294	3.280	0.618	0.604	0.014	4.163	0.740	0.771	− 0.869

intermediate value of 50%. The ZIP and MZIP models were subsequently fit to each of the simulated data sets, and the true average marginal effects $\bar{\eta}_1(\theta)$ of x_1 and true average incremental effects $\bar{\pi}_2(\theta)$ of x_2, as well as their estimates $\hat{\bar{\eta}}_1(\hat{\theta})$ and $\hat{\bar{\pi}}_2(\hat{\theta})$, were computed.

Table 2 reports the true value of the average marginal and incremental effects, the mean and standard deviation of their estimates, and the mean of SEs given by the ZIP and MZIP models from fitting the simulation data sets. In Table 2, the estimated average marginal and incremental effects obtained from the underlying true model, the MZIP model, are still unbiased as expected. The misspecified ZIP model across all combinations of $\beta = (\beta_0, \beta_1, \beta_2)'$ and the three sample sizes provides the effect estimates with larger bias. For both models, the average SE of the average marginal and incremental effects is close to the corresponding standard deviations and decreases as the sample size increases from 100, 500, to 1000, which verified the validity of the variance estimation procedure in Sect. 5 for the finite samples.

7.4 True HNB models versus misspecified MHNB models

In the third simulation study, 500 data sets with sample sizes $n = 100$, 500, and 1000 were simulated from the HNB model with $\ln(\mu_i) = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2$ and $\text{logit}(\psi_i) = \gamma_0 + z_{i1}\gamma_1 + z_{i2}\gamma_2$, in which $x_{i1} = z_{i1} \sim N(0, 1)$ and $x_{i2} = z_{i2} \sim \text{Bernoulli}(0.5)$. Five combinations of $\beta = (\beta_0, \beta_1, \beta_2)'$ (see Table 3) were examined, such that the averages of μ_i's vary from approximately 3 to 11. The fixed parameter $\gamma = (\gamma_0, \gamma_1, \gamma_2) = (0.5, 1, -1)$ provides an average of ψ_i's being around 50%. The scale parameter was fixed at $\alpha = 1.5$ for all data sets. After data generation, the HNB and MHNB models were fit to each of the simulated data sets. We calculated for each model the true average marginal effect $\bar{\eta}_1(\theta)$ of x_1 and true average incremental effects $\bar{\pi}_2(\theta)$ of x_2, as well as their estimates $\hat{\bar{\eta}}_1(\hat{\theta})$ and $\hat{\bar{\pi}}_2(\hat{\theta})$. The simulation results were reported in Table 3. Evidently, effect estimates from the true HNB model have smaller bias than the ones from the misspecified MHNB model.

7.5 True MHNB models versus misspecified HNB models

In the fourth simulation study, we produced 500 data sets with sample sizes $n = 100$, 500, and 1000 from the MHNB model using the linear predictors as $\ln(\nu_i) = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2$ and $\text{logit}(\psi_i) = \gamma_0 + z_{i1}\gamma_1 + z_{i2}\gamma_2$ with $x_{i1} = z_{i1} \sim N(0, 1)$ and $x_{i2} = z_{i2} \sim \text{Bernoulli}(0.5)$. We still considered five combinations of $\beta = (\beta_0, \beta_1, \beta_2)'$ (see Table 4) yielding the averages of μ_i's varying from approximately 4–13. As in the third simulation study, $\gamma = (\gamma_0, \gamma_1, \gamma_2) = (0.5, 1, -1)$ and $\alpha = 1.5$ were fixed when generating simulation data. Then, the MHNB and HNB models were fit to each of simulated data sets. Table 4 presents the results of the average marginal and incremental effects in terms of the true value, the mean and standard deviation of their estimates, and the mean of SEs given by the true and misspecified models. It is observed that the behavior of the effect estimates and the SEs of average marginal and incremental effects is as same as in the previous simulation studies.

The conclusion from the four back-to-back simulation studies is straightforward. No matter which type of model, the marginalized or conventional two-part model, is fit to the data, the estimates of marginal effects will be biased as long as the model is misspecified. The marginalized two-part models do not have any advantage to reduce such type of bias in estimating marginal effects of a covariate with respect to the expected outcomes.

7.6 Robustness

The results from the above numerical studies are consistent with the presented theories in Sect. 7.1, in that the misspecified models have larger bias than the true models in the maximum likelihood estimation. Cross-comparison of the estimation biases produced by the misspecified ZIP and MZIP models reveals that the misspecified ZIP models induce smaller biases than the misspecified MZIP models (see the results on biases in Tables 1, 2). This indicates that the maximum likelihood estimators of the MZIP models are less robust to model misspecification than the maximum likelihood estimators given by the ZIP models, which would be even worse if compared with Poisson models (Staub and Winkelmann 2013). The results on biases show that there is not significant difference in robustness on the maximum likelihood estimators given by the HNB and MHNB models with respect to model misspecification (see the results on biases Tables 3, 4).

8 Model selection via marginal effects

When the primary interest of data analysis lies in estimating the (average) marginal or incremental effect of a covariate with respect to the expected outcomes, the empirical mean square error (MSE) criterion (Dow and Norton 2003; Madden 2008) can be used for selecting the best model among the candidate models to estimate the effects. Suppose the goal of data analysis is to precisely estimate an incremental effect $\pi_j(x_{i(-j)}, z_{i(-j)}, \theta)$ or an average incremental effect $\bar{\pi}_j(\theta)$ subject to a change of x_j. The MSE of an effect estimator $\hat{\pi}_j(x_{i(-j)}, z_{i(-j)}, \hat{\theta})$ or $\hat{\bar{\pi}}_j(\hat{\theta})$ is equal to the variance of the estimators plus the square of its bias:

$$
\begin{aligned}
\mathrm{MSE}\big[\hat{\pi}_j(x_{i(-j)}, z_{i(-j)}, \hat{\theta})\big] &= \mathrm{var}\,\big[\hat{\pi}_j(x_{i(-j)}, z_{i(-j)}, \hat{\theta})\big] + \mathrm{Bias}^2\big[\hat{\pi}_j(x_{i(-j)}, z_{i(-j)}, \hat{\theta})\big] \\
&= \mathrm{var}\,\big[\hat{\pi}_j(x_{i(-j)}, z_{i(-j)}, \hat{\theta})\big] \\
&\quad + \big[\hat{\pi}_j(x_{i(-j)}, z_{i(-j)}, \hat{\theta}) - \pi_j(x_{i(-j)}, z_{i(-j)}, \theta)\big]^2
\end{aligned} \tag{25}
$$

and

$$
\begin{aligned}
\mathrm{MSE}\big[\hat{\bar{\pi}}_j(\hat{\theta})\big] &= \mathrm{var}\,\big[\hat{\bar{\pi}}_j(\hat{\theta})\big] + \mathrm{Bias}^2\big[\hat{\bar{\pi}}_j(\hat{\theta})\big] \\
&= \mathrm{var}\,\big[\hat{\bar{\pi}}_j(\hat{\theta})\big] + \big[\hat{\bar{\pi}}_j(\hat{\theta}) - \bar{\pi}_j(\theta)\big]^2.
\end{aligned} \tag{26}
$$

The MSE criterion selects the candidate model with the minimum MSE as the best model to estimate the corresponding marginal or incremental effect. Because in practice the true effects $\pi_j(x_{i(-j)}, z_{i(-j)}, \theta)$ and $\bar{\pi}_j(\theta)$ are unknown in (25) and (26), the empirical MSEs (EMSEs)

$$
\begin{aligned}
\mathrm{EMSE}\big[\hat{\pi}_j(x_{i(-j)}, z_{i(-j)}, \hat{\theta})\big] &= \mathrm{var}\,\big[\hat{\pi}_j(x_{i(-j)}, z_{i(-j)}, \hat{\theta})\big] \\
&\quad + \Big[\hat{\pi}_j(x_{i(-j)}, z_{i(-j)}, \hat{\theta}) - \hat{\pi}_j^c(x_{i(-j)}, z_{i(-j)}, \hat{\theta}^c)\Big]^2
\end{aligned} \tag{27}
$$

and

$$
\mathrm{EMSE}\big[\hat{\bar{\pi}}_j(\hat{\theta})\big] = \mathrm{var}\,\big[\hat{\bar{\pi}}_j(\hat{\theta})\big] + \big[\hat{\bar{\pi}}_j(\hat{\theta}) - \hat{\bar{\pi}}_k^c(\hat{\theta}^c)\big]^2. \tag{28}
$$

Table 5 Rates of selecting the true ZIP model over the misspecified MZIP model given by AIC/BIC and EMSE and rejection rates given by Vuong's test

Sample size	$\beta = (\beta_0, \beta_1, \beta_2)$ values	AIC/BIC rate (%)	EMSE$_{\hat{\hat{\eta}}_1}$ rate (%)	EMSE$_{\hat{\bar{\pi}}_2}$ rate (%)	Vuong's test rejection rate (%)
100	(1.0, – 1, .10)	78.40	89.80	66.60	8.00
	(– 2, 2.0, 2.0)	61.41	50.51	49.09	0.00
	(2.0, 1.0, – .5)	71.80	68.40	52.40	1.00
	(1.0, .50, 2.0)	81.80	68.00	56.40	13.00
	(.50, 1.8, 1.0)	69.00	62.40	46.40	0.00
500	(1.0, – 1, .10)	98.00	82.60	66.00	53.20
	(– 2, 2.0, 2.0)	78.20	70.40	59.40	3.00
	(2.0, 1.0, – .5)	97.80	91.00	74.60	40.00
	(1.0, .50, 2.0)	99.00	97.00	62.40	66.20
	(.50, 1.8, 1.0)	92.20	84.20	62.20	20.00
1000	(1.0, – 1, .10)	99.80	88.20	68.00	82.80
	(– 2, 2.0, 2.0)	90.40	79.80	67.60	16.60
	(2.0, 1.0, – .5)	99.20	96.80	81.40	74.20
	(1.0, .50, 2.0)	100.0	99.40	67.40	93.60
	(.50, 1.8, 1.0)	98.20	90.60	71.60	51.80

Table 6 Rates of selecting the true MZIP model over the misspecified ZIP model given by AIC/BIC and EMSE and rejection rates given by Vuong's test

Sample size	$\beta = (\beta_0, \beta_1, \beta_2)$ values	AIC/BIC rate (%)	EMSE$_{\hat{\hat{\eta}}_1}$ rate (%)	EMSE$_{\hat{\bar{\pi}}_2}$ rate (%)	Vuong's test rejection rate (%)
100	(.50, – 5, – .5)	79.00	52.20	64.40	3.80
	(.50, – 1, .50)	83.60	28.00	60.80	9.00
	(1.0, – 1, .10)	90.40	34.60	72.00	15.80
	(– .5, 1.0, .50)	62.40	65.40	64.60	0.20
	(1.0, 50, – .5)	75.20	70.20	74.20	4.40
500	(.50, – 5, – .5)	96.80	53.20	69.00	42.00
	(.50, – 1, .50)	99.80	32.60	66.80	69.60
	(1.0, – 1, .10)	100.0	40.20	74.40	86.80
	(– .5, 1.0, .50)	88.20	73.00	72.80	15.00
	(1.0, 50, – .5)	98.00	79.00	75.60	52.20
1000	(.50, – 5, – .5)	100.0	57.40	68.80	76.80
	(.50, – 1, .50)	99.80	35.80	63.00	95.40
	(1.0, – 1, .10)	100.0	37.00	75.80	100.00
	(– .5, 1.0, .50)	96.00	74.80	77.80	38.40
	(1.0, 50, – .5)	100.0	80.40	78.60	85.80

are used to accomplish the mission of model selection. In practice, the true effect in (27) and (28) is replaced by the estimated effect $\hat{\pi}_j^c(x_{i(-j)}, z_{i(-j)}, \hat{\theta}^c)$ or $\hat{\bar{\pi}}_k^c(\hat{\theta}^c)$ from a pre-specified model. Dow and Norton (2003) illustrated the use of the MSE criterion, through a Monte

Table 7 Rates of selecting the true HNB model over the misspecified MHNB model given by AIC/BIC and EMSE and rejection rates given by Vuong's test

Sample size	$\beta = (\beta_0, \beta_1, \beta_2)$ values	AIC/BIC rate (%)	EMSE$_{\hat{\eta}_1}$ rate (%)	EMSE$_{\hat{\pi}_2}$ rate (%)	Vuong's test rejection rate (%)
100	(.80, – .5, .20)	71.37	87.30	74.80	5.24
	(– .5, – .5, 2.5)	68.76	72.12	66.67	11.32
	(2.0, – .5, -– 5)	68.75	87.50	71.77	8.06
	(2.2, .50, – .5)	67.48	71.55	48.78	6.10
	(1.5, .50, 1.2)	75.27	53.58	38.83	14.75
500	(.80, – .5, .20)	86.80	84.00	71.60	23.20
	(– .5, – .5, 2.5)	84.11	73.84	66.20	20.72
	(2.0, – .5, -– 5)	90.60	90.60	71.00	32.20
	(2.2, .50, – .5)	86.36	81.41	66.12	21.90
	(1.5, .50, 1.2)	90.91	57.50	54.77	37.73
1000	(.80, – .5, .20)	95.00	80.60	67.20	37.60
	(– .5, – .5, 2.5)	92.60	76.20	65.60	29.80
	(2.0, – .5, -– 5)	95.40	89.40	67.00	45.20
	(2.2, .50, – .5)	94.99	87.89	68.69	37.37
	(1.5, .50, 1.2)	95.79	66.09	62.13	55.69

Table 8 Rates of selecting the true MHNB model over the misspecified HNB model given by AIC/BIC and EMSE and rejection rates given by Vuong's test

Sample size	$\beta = (\beta_0, \beta_1, \beta_2)$ values	AIC/BIC rate (%)	EMSE$_{\hat{\eta}_1}$ rate (%)	EMSE$_{\hat{\pi}_2}$ rate (%)	Vuong's test rejection rate (%)
100	(– 1, .50, 2.0)	43.56	84.66	76.44	15.57
	(– .5, – .2, 2.0)	67.34	60.41	70.81	12.72
	(1.5, – .5, – .5)	65.73	33.67	56.11	3.81
	(1.0, – .8, .50)	62.33	22.24	40.48	3.41
	(1.5, .50, .50)	62.13	77.66	80.00	4.47
500	(– 1, .50, 2.0)	68.26	97.61	93.41	27.19
	(– .5, – .2, 2.0)	96.09	71.95	76.78	47.59
	(1.5, – .5, – .5)	91.80	32.20	60.00	20.80
	(1.0, – .8, .50)	89.40	45.40	53.80	16.60
	(1.5, .50, .50)	87.30	84.58	81.63	12.47
1000	(– 1, .50, 2.0)	85.47	97.23	94.46	60.93
	(– .5, – .2, 2.0)	99.15	75.00	77.56	84.83
	(1.5, – .5, – .5)	98.00	42.00	60.00	45.60
	(1.0, – .8, .50)	95.40	53.40	60.80	35.60
	(1.5, .50, .50)	93.04	85.32	80.10	29.60

Carlo example, for selecting between sample selection models and two-part models for corner solutions in semicontinuous data. This MSE criterion was referred as "an empirical MSE test" by Dow and Norton (2003). The competitors of the MSE criterion include the information criteria, such as the Akaike information criterion (AIC) and Bayesian information criterion (BIC), and the Vuong's closeness test. The Vuong's closeness test (Vuong 1989), or the Vuong's test, is a likelihood-ratio-based test for examining whether two non-nested models are equally close to the true data generating process. The Vuong's test statistic is $V = \sqrt{n}(\bar{m}/s_m)$, in which $\bar{m} = \frac{1}{n}\sum_{i=1}^{n} m_i$, $s_m = \frac{1}{n}\sum_{i=1}^{n}(m_i - \bar{m})^2$, and $m_i = \ell_{i,0}(\theta) - \ell_{i,1}(\theta)$ is the ith term for observation i in the log-likelihood $\ell_0(\theta)$ under the null hypothesis model minus the ith term in the log-likelihood $\ell_1(\theta)$ under the alternative model. The Vuong's test statistic asymptotically follows a standard normal distribution under the hypothesis that the two models are equivalent, and a standard Z-test is then applied.

To examine the performance of the above model selection criteria, subsequent numerical studies were conducted following the simulation studies in Sect. 7. In each of the outlined simulation studies in Sect. 7, we collected the observed likelihood values and calculated the AIC and BIC values of the pair of true and misspecified models that were fit to each of the 500 simulated data sets. The best model was then selected based upon the AIC and BIC criteria for each pair of models. Note that the AIC and BIC values are identical through the four simulation studies, because the investigated true and misspecified models had the same number of unknown parameters. With the collected observed likelihood values, the Vuong's test with a significance level of 5% was also conducted for each pair of true and misspecified models with the misspecified model in the null hypothesis. The EMSEs of the average marginal and incremental effects in each pair of the models were calculated, in which the true effect estimates were taken to calculate the bias term. Tables 5, 6, 7 and 8 report the rates of selecting the true model over the misspecified model given by the AIC, BIC, and EMSE criteria among 500 simulation data sets. The tables also report rejection rates given by the Vuong's test for each of the combinations of parameters and sample sizes. Among these criteria, the selection rates of AIC and BIC are the highest for most cases. The rates of selecting a true model given by AIC and BIC are mostly over 90% when the sample size $n = 1000$ and more than 80% for $n = 500$. Even for $n = 100$, these rates are usually higher than 60%. There is not a clear trend of increase for the select rates of the EMSE criterion along with the increase of sample size. These rates mostly range from about 50–90%, but can be as low as 22% when the sample size is small. These rates in general do not exhibit a reliable pattern but acceptable. The Vuong's test rejection rates do not perform well, especially for the sample sizes $n = 100$ and $n = 500$, but the rates do increase with the growth of sample size. The simulation studies show that the information criteria are reliable in distinguishing between the standard and misspecified two-part models for count data with excess zeroes, and the Vuong's test cannot differentiate the models if the sample size is not large enough. The MSE criterion might be acceptable to be effect-specific model selection approach; however, it should be noted that in practice, its performance can be dramatically influenced by the hypothesis of which model is consistent and therefore can be used to calculate the bias term.

9 Application

The German Socioeconomic Panel (GSOEP) data (1984–1995) (Riphahn et al.2003) are used for empirical analysis with the four conventional two-part models and their corresponding marginalized models discussed in Sects. 3 and 4. The data were collected based on annual face-to-face individual or computer-assisted personal interviews with household members aged 16 or over living in Germany for comprehensive information to measure stability and change in living conditions Frick (2006).

The pooled subsample of the GSOEP data (1984–1994) includes 7293 German citizens, aged 26 through 65 Riphahn et al. (2003). After removing missing values, the subsample only includes years 1984–1988, 1991, and 1994 with 14,243 male observations and 13,083 female observations. The dependent variable is the number of doctor visits in the last 3 months right before the survey with 37.09% observations as zero and the mean across the whole sample is 3.18 with a standard deviation of 5.69. One key independent variable is the public indicator which divides people into the group mandatorily insured by the public insurance and the group voluntarily with the proportions of 88.57 versus 14.33%. Among those with coverage of public insurance, about 2.12% purchased add-on insurance which takes up 1.88% of the whole data. The add-on insurance indicator is another key covariate. The age and degree of health satisfaction (using integer scales 0–10 meaning bad to well) are the only two continuous covariates. All other independent variables including gender and years are converted to dummy variables.

9.1 Statistical modelling

In our statistical modelling, all independent variables are considered in both parts, that is,

$$\begin{aligned} x_i = z_i = \{&\text{female, age, health, public insurance, add-on insurance,} \\ &\text{year1985, year1986, year1987, year1988, year1991, year1994}\}, \end{aligned}$$

and the linear predictors $x_i'\beta$ and $z_i'\gamma$ and the link functions are specified as in Sections 3 and 4. However, the estimates of β and γ have different interpretations. The models involving negative binomial models contain an extra scale parameter α. All models have explicit log-likelihood functions except for the MHP and MHNB models. Their log-likelihood functions (15) and (17) are subject to nonlinear constraints (16) and (18). We used Newton-Raphson method for solving μ_i from these constraints at every iteration of maximizing the log-likelihood functions.

After fitting the data with the two-part models, we collected AIC, BIC, $\hat{\hat{\pi}}_{\text{public}}$, $\hat{\hat{\pi}}_{\text{add-on}}$, and the EMSE values of the two effects estimates, and conducted the Vuong's test for each pair of models. Regarding the EMSE values of $\hat{\hat{\pi}}_{\text{public}}$ and $\hat{\hat{\pi}}_{\text{add-on}}$, a pre-specified model must be selected for the true average incremental effect in (28), whereas the true model for real data is unpredicted, implying a parameter estimate consistency issue. Hence, for the purpose of comparison, both the conventional model and the corresponding marginalized model in each pair are selected as the pre-specified model and their estimated effects are used in (28), respectively.

9.2 Empirical results

Table 9 presents the results from fitting the two-part models. In general, all models provide positive and significant estimates $\hat{\beta}_{\text{public}}$ varying from 0.136 to 0.208, indicating that the participants covered by public insurance see doctors more frequently than private insurance cohort on a regular basis (for the zero-inflated models), or on need (for the hurdle models) or for the whole population (for the marginalized models). The MZINB model presents a non-significant negative estimate $\hat{\gamma}_{\text{public}}$ (-0.178), while other models show

Table 9 Parameter estimates (SEs) given by the non-marginalized and marginalized two-part models from fitting to the GSOEP data

Item	ZIP	MZIP	ZINB	MZINB	HP	MHP	HNB	MHNB
	Est. (SE)	Est. (SE)	Est. (SE)	Est. (SE)	Est. (SE)	Est. (SE)	Est. (SE)	Est. (SE)
β_0	2.137 (0.024)	1.780 (0.033)	1.993 (0.059)	1.842 (0.056)	2.143 (0.024)	1.702 (0.034)	1.997 (0.064)	1.754 (0.056)
β_{female}	0.124 (0.007)	0.251 (0.010)	0.169 (0.018)	0.298 (0.016)	0.122 (0.007)	0.247 (0.010)	0.149 (0.019)	0.297 (0.017)
β_{age}	0.006 (0.000)	0.009 (0.000)	0.008 (0.001)	0.010 (0.001)	0.006 (0.000)	0.010 (0.000)	0.007 (0.001)	0.011 (0.001)
β_{health}	− .171 (0.001)	− .232 (0.002)	− .208 (0.004)	− .251 (0.004)	− .171 (0.001)	− .230 (0.002)	− .209 (0.004)	− .248 (0.004)
β_{public}	0.141 (0.013)	0.199 (0.018)	0.136 (0.030)	0.163 (0.028)	0.138 (0.013)	0.208 (0.018)	0.143 (0.032)	0.192 (0.028)
$\beta_{\text{add-on}}$	− .107 (0.026)	− .032 (0.033)†	− .065 (0.060)†	− .014 (0.057)‡	− .106 (0.026)	− .045 (0.033)†	− .138 (0.066)	− .050 (0.055)†
β_{year1985}	0.004 (0.013)‡	− .015 (0.018)†	0.027 (0.034)†	0.035 (0.032)†	− .001 (0.013)‡	0.014 (0.018)†	0.033 (0.036)†	0.040 (0.033)†
β_{year1986}	0.097 (0.013)	0.115 (0.018)	0.134 (0.033)	0.156 (0.032)	0.092 (0.013)	0.142 (0.018)	0.130 (0.036)	0.159 (0.032)
β_{year1987}	0.056 (0.013)	0.067 (0.018)	0.076 (0.034)	0.106 (0.032)	0.051 (0.013)	0.096 (0.018)	0.086 (0.036)	0.114 (0.033)
β_{year1988}	− .144 (0.013)	− .078 (0.017)	− .169 (0.032)	− .056 (0.030)*	− .150 (0.013)	− .056 (0.017)	− .190 (0.035)	− .048 (0.030)†
β_{year1991}	− .220 (0.013)	− .100 (0.017)	− .240 (0.032)	− .086 (0.030)	− .226 (0.013)	− .081 (0.017)	− .297 (0.035)	− .076 (0.030)
β_{year1994}	0.120 (0.013)	0.197 (0.018)	0.149 (0.034)	0.217 (0.032)	0.115 (0.013)	0.219 (0.018)	0.144 (0.036)	0.226 (0.032)
γ_0	− .505 (0.100)	− 1.27 (0.095)	− 3.62 (0.344)	− 3.53 (0.355)	− 1.54 (0.099)	− 1.00 (0.089)	− 1.54 (0.099)	− 1.25 (0.096)
γ_{female}	− .604 (0.028)	− .462 (0.026)	−1.10 (0.083)	− .999 (0.084)	− .577 (0.027)	− .458 (0.024)	− .577 (0.027)	− .547 (0.026)
γ_{age}	−. 023 (0.001)	− .012 (0.001)	− .017 (0.003)	− .019 (0.003)	− .014 (0.001)	− .015 (0.001)	− .014 (0.001)	− .016 (0.001)
γ_{health}	0.236 (0.008)	0.252 (0.007)	0.509 (0.034)	0.495 (0.035)	0.326 (0.008)	0.257 (0.007)	0.326 (0.008)	0.301 (0.008)
γ_{public}	− .326 (0.042)	− .188 (0.040)	− .199 (0.095)	− .178 (0.098)*	− .168 (0.041)	− .236 (0.037)	− .168 (0.041)	− .226 (0.039)
$\gamma_{\text{add-on}}$	− .073 (0.106)†	− .308 (0.108)	− .600 (0.361)*	− .724 (0.463)†	− .299 (0.103)	− .180 (0.090)	− .297 (0.103)	− .187 (0.098)*
γ_{year1985}	0.069 (0.051)†	0.067 (0.045)†	−.053 (0.107)‡	0.042 (0.107)‡	− .038 (0.050)†	− .026 (0.042)‡	− .038 (0.050)†	− .025 (0.048)‡
γ_{year1986}	− .030 (0.051)‡	− .061 (0.045)†	− .110 (0.106)†	− .067 (0.107)‡	− .160 (0.050)	− .157 (0.042)	− .160 (0.050)	− .154 (0.048)
γ_{year1987}	− .008 (0.052)‡	− .033 (0.046)†	− .175 (0.111)†	− .085 (0.111)†	− .124 (0.050)	− .130 (0.043)	− .124 (0.050)	− .113 (0.048)
γ_{year1988}	−. 191 (0.050)	− .243 (0.046)	− .865 (0.140)	− .833 (0.152)	− .250 (0.048)	− .265 (0.041)	− .250 (0.048)	− .243 (0.046)
γ_{year1991}	− .366 (0.052)	− .476 (0.049)	−1.45 (0.197)	−1.90 (0.339)	− .406 (0.049)	− .424 (0.042)	− .405 (0.049)	− .399 (0.047)
γ_{year1994}	− .186 (0.053)	− .276 (0.048)	− .418 (0.119)	− .459 (0.127)	− .314 (0.052)	− .355 (0.044)	− .314 (0.052)	− .325 (0.050)
α	NA	NA	0.918 (0.019)	0.914 (0.019)	NA	NA	0.809 (0.022)	0.836 (0.023)

The superscript "‡" is used for estimates with nonsignificant p values > 0.5; "†" for p values between 0.1 and 0.5; "*" for p values between 0.05 and 0.1. Significant estimates with p values ≤ 0.05 are not marked by any superscripts

Table 10 Average incremental effect estimates (SEs, *p* values), the AIC, BIC, and EMSE values, and the Vuong's Z-statistics (*p* values) given by the non-marginalized and marginalized two-part models from fitting to the GSOEP data

	ZIP	MZIP	ZINB	MZINB	HP	MHP	HNB	MHNB
	Est. (SE, *p* value)	Est. (SE, *p* value)	Est. (SE, *p*-value)	Est. (SE, *p* value)	Est. (SE, *p* value)	Est. (SE, *p* value)	Est. (SE, *p* value)	Est. (SE, *p* value)
$\hat{\bar{\pi}}_{\text{public}}$	0.678 (0.252, 0.007)	0.581 (0.218, 0.008)	0.466 (0.486, 0.338)	0.492 (0.421, 0.242)	0.524 (0.205, 0.010)	0.603 (0.226, 0.008)	0.474 (0.387, 0.220)	0.570 (0.402, 0.157)
$\hat{\bar{\pi}}_{\text{add-on}}$	−. 264 (0.206,0.200)	− .100 (0.077, 0.197)	− .058 (0.129, 0.651)	− .046 (0.080, 0.569)	− .077 (0.105, 0.460)	− .139 (0.108, 0.200)	− .102 (0.325, 0.753)	−. 157 (0.248, 0.527)
AIC	152294.02	152372.01	115548.42	115592.58	152139.35	152379.30	115447.12	115494.77
BIC	152491.20	152569.19	115753.81	115797.97	152336.53	152576.47	115652.51	115700.16
$\text{EMSE}_{\text{C}}(\hat{\bar{\pi}}_{\text{public}})$	0.064	0.057	0.236	0.178	0.042	0.057	0.150	0.171
$\text{EMSE}_{\text{C}}(\hat{\bar{\pi}}_{\text{add-on}})$	0.043	0.033	0.017	0.007	0.011	0.016	0.106	0.064
$\text{EMSE}_{\text{M}}(\hat{\bar{\pi}}_{\text{public}})$	0.073	0.047	0.237	0.177	0.048	0.051	0.159	0.162
$\text{EMSE}_{\text{M}}(\hat{\bar{\pi}}_{\text{add-on}})$	0.070	0.006	0.017	0.007	0.015	0.012	0.109	0.061
Vuong's test	Z-stat. (*p* value)		Z-stat. (*p* value)		Z-stat. (*p* value)		Z-stat. (*p* value)	
	0.853 (0.394)		1.765 (0.077)		2.670 (0.008)		2.360 (0.018)	

significant negative estimates. This suggests that, under the MZINB models, there is not much difference of no regular doctor visits between public and private insurance cohorts, whereas under other models, there are substantial chances that public insurance cohort see doctors more regularly.

Coefficient estimates $\hat{\beta}_{\text{add-on}}$ and $\hat{\gamma}_{\text{add-on}}$ for add-on insurance are all negative but more diverse than for public insurance. Estimates of $\hat{\beta}_{\text{add-on}}$ are significant for the ZIP, HP and HNB models with larger magnitudes than the nonsignificant estimates for the MZIP, ZINB, MZINB, MHP and MHNB models. In terms of $\hat{\gamma}_{\text{add-on}}$, the MZIP, HP, MHP, and HNB models give significant estimates with magnitudes ranging from 0.180 to 0.308; whereas the ZIP, ZINB, MZINB, and MHNB models show non-significant estimates with magnitudes varying from 0.073 to 0.724.

Table 10 compares results from these models. All models provide positive incremental effect estimate $\hat{\bar{\pi}}_{\text{public}}$ and negative estimate $\hat{\bar{\pi}}_{\text{add-on}}$. The comparison of AIC and BIC support the HNB and MHNB models with smaller AIC and BIC values and the HNB model carries the smallest. Based on Vuong's test, the two models are significantly different in modelling the GSOEP data, indicating that the HNB model is the best one among these models for the GSEOP subsample. In terms of EMSEs of incremental effect estimates of public insurance and add-on insurance, we use notations of EMSE_C and EMSE_M for the EMSE values calculated based on the conventional and the marginalized models in each pair as the pre-specified model, respectively. The results show that the HNB model has smaller EMSE_C and EMSE_M values of $\hat{\bar{\pi}}_{\text{public}}$ than the MHNB model; however, its EMSE_C and EMSE_M values of $\hat{\bar{\pi}}_{\text{add-on}}$ are larger than the MHNB model due to the large SE of $\hat{\bar{\pi}}_{\text{add-on}}$. Regarding the incremental effect estimates $\hat{\bar{\pi}}_{\text{public}}$ and $\hat{\bar{\pi}}_{\text{add-on}}$, the results show that both effects are not significant to the overall healthcare utilization in terms of number of physician visits under both the HNB and MHNB models whereas the related parameter estimates are different stories, which seems to be a surprise to the initial motivation of the proposal of marginalized two-part models.

10 Conclusion

This article reviews four two-part models for cross-sectional count data with abundant zeroes (the ZIP, ZINB, HP, and HNB models) and two marginalized models (the MZIP and MZINB models) and proposes two other models (the MHP and MHNB models). We argue that the facility of marginalization of two-part models cannot be taken as a reason to choose marginalized models over the non-marginalized models to fit such data. Instead, appropriate model selection procedure should be followed to find the best model. In this article, we derive estimates and variance estimates of the (average) marginal effects and (average) incremental effects of covariates with respect to the overall mean outcomes for these two-part models. The average effect estimates given by the true models are unbiased in the simulation studies, and the irregular bias of average effect estimates given by the misspecified models is observed. Two pairs of non-marginalized and marginalized models are compared by using three model selection criteria in the simulation studies. The results confirm the reliability of the AIC and BIC criterion. In summary, despite marginalized two-part models can help in estimating overall marginal effects of covariates on the transformed expectation of count outcomes, this advantage should not be over-emphasized. Otherwise, model misspecification may lead to inaccurate statistical inference. When the two-part models include a large number of covariates, penalized maximum likelihood methods such as the least absolute shrinkage and selection operator, smoothly clipped absolute deviation (SCAD), or minimax concave penalty (MCP) are recommended to be applied to conduct variable selection. It has been shown that these methods can provide comparable estimation, but are more robust than the traditional stepwise variable selection in terms of variable selection (Wang et al. 2015).

Acknowledgements We sincerely thank the Editor, Associate Editor, and two anonymous reviewers for the valuable and insightful comments.

Funding This study is not funded by any specific grants.

Compliance with ethical standards

Conflict of interest All the authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

Informed consent Informed consent is not applicable, as the article does not contain any studies with human participants or animals performed by any of the authors.

Appendix: Gradients of marginal effects

Gradients of marginal effects in the ZIP and ZINB models

Recall that ZIP and ZINB have the identical expression of marginal expectation of response y_i: $\mathrm{E}\,(y_i|x_i, z_i) = \mu_i(1-\psi_i) = \dfrac{e^{x_i'\beta}}{1+e^{z_i'\gamma}}$ and as a consequence share the same marginal and incremental effect formulas. The difference is the parameter θ, where $\theta = (\beta', \gamma')'$ for ZIP models and $\theta = (\beta', \gamma', \alpha)'$ for ZINB models.

To simplify computation and notations, we introduce a pair of infinitely differentiable functions on the number line: $p^{\mathrm{ZIP}}(t) = e^t$ and $q(t) = \dfrac{1}{1+e^t}$ with $\dot{p}^{\mathrm{ZIP}}(t) = \ddot{p}^{\mathrm{ZIP}}(t) = p^{\mathrm{ZIP}}(t) = e^t$ and $\dot{q}(t) = -\dfrac{e^t}{(1+e^t)^2}$, $\ddot{q}(t) = -\dfrac{e^t}{(1+e^t)^2} \cdot \dfrac{1-e^t}{1+e^t}$, for $\forall\, t \in \mathbb{R}$. Even “ZIP” is used in the superscript of function p and its derivatives, their expressions are exactly the same for ZINB. The θ and superscript of p will not be restated again in this subsection. The following discussions are identical for both ZIP and ZINB unless indicated otherwise.

Considering a continuous covariate x_{ij} in our regression models, we adopt the simplified notations: $p_i^{\mathrm{ZIP}}, \dot{p}_i^{\mathrm{ZIP}}, \ddot{p}_i^{\mathrm{ZIP}}, q_i, \dot{q}_i, \ddot{q}_i$ which are $p^{\mathrm{ZIP}}, \dot{p}^{\mathrm{ZIP}}, \ddot{p}^{\mathrm{ZIP}}, q, \dot{q}, \ddot{q}$ evaluated at $x_i'\beta$ and $z_i'\gamma$, respectively. Then, the marginal mean of y_i is $\mathrm{E}\,(y_i|x_i, z_i) = \mu_i(1-\psi_i) = p_i^{\mathrm{ZIP}} q_i$ and hence the marginal effect with respect to x_{ij} is $\eta_j(x_i, z_i, \theta) = \beta_j \dot{p}_i^{\mathrm{ZIP}} q_i + \gamma_j p_i^{\mathrm{ZIP}} \dot{q}_i$.

If the covariate x_j, or z_j, is categorical, to rewrite its incremental effect from level l_1 to l_2, the values of $p^{\mathrm{ZIP}}, \dot{p}^{\mathrm{ZIP}}, \ddot{p}^{\mathrm{ZIP}}$ at $x_{i(-j)}'\beta_{(-j)} + l_2\beta_j$ and $x_{i(-j)}'\beta_{(-j)} + l_1\beta_j$ will be denoted as $p_{2i}^{\mathrm{ZIP}}, \dot{p}_{2i}^{\mathrm{ZIP}}, \ddot{p}_{2i}^{\mathrm{ZIP}}$ and $p_{1i}^{\mathrm{ZIP}}, \dot{p}_{1i}^{\mathrm{ZIP}}, \ddot{p}_{1i}^{\mathrm{ZIP}}$, respectively; values of $q, \dot{q}, \ddot{q}$ at $z_{i(-j)}'\gamma_{(-j)} + l_2\gamma_j$ and

$z'_{i(-j)}\gamma_{(-j)} + l_1\gamma_j$ will be represented by $q_{2i}, \dot{q}_{2i}, \ddot{q}_{2i}$ and $q_{1i}, \dot{q}_{1i}, \ddot{q}_{1i}$, respectively. Then, the incremental effect with respect to x_{ij} is $\pi_j(x_{i(-j)}, z_{i(-j)}, \theta) = p_{2i}^{\text{ZIP}} q_{2i} - p_{1i}^{\text{ZIP}} q_{1i}$.

The gradients of marginal and incremental effects are

$$\begin{aligned}
&\nabla_\theta \eta_j(x_i, z_i, \theta) \\
&\quad = \left(\beta_j \ddot{p}_i^{\text{ZIP}} q_i + \gamma_j \dot{p}^{\text{ZIP}} \dot{q}\right) \sum_{m=0}^{J_1} x_{im} u_{(m+1)} + \dot{p}_i^{\text{ZIP}} q_i u_{(j+1)} \\
&\qquad + \left(\beta_j \dot{p}_i^{\text{ZIP}} \dot{q}_i + \gamma_j p_i^{\text{ZIP}} \ddot{q}_i\right) \sum_{m=0}^{J_2} z_{im} u_{(J_1+m+2)} + p_i^{\text{ZIP}} \dot{q}_i u_{(J_1+j+2)}, \\
&\nabla_\theta \pi_j(x_{i(-j)}, z_{i(-k)}, \theta) \\
&\quad = \left(\dot{p}_{2i}^{\text{ZIP}} q_{2i} - \dot{p}_{1i}^{\text{ZIP}} q_{1i}\right) \sum_{m=0,\neq j}^{J_1} x_{im} u_{(m+1)} + \left(l_2 \dot{p}_{2i}^{\text{ZIP}} q_{2i} - l_1 \dot{p}_{1i}^{\text{ZIP}} q_{1i}\right) \cdot u_{(j+1)} \\
&\qquad + \left(p_{2i}^{\text{ZIP}} \dot{q}_{2i} - p_{1i}^{\text{ZIP}} \dot{q}_{1i}\right) \sum_{m=0,\neq j}^{J_2} z_{im} u_{(J_1+m+2)} + \left(l_2 p_{2i}^{\text{ZIP}} \dot{q}_{2i} - l_1 p_{1i}^{\text{ZIP}} \dot{q}_{1i}\right) \cdot u_{(J_1+j+2)},
\end{aligned} \tag{29}$$

where $u_{(m)}$ is a unit vector of dimension $J_1 + J_2 + 2$ for ZIP and dimension $J_1 + J_2 + 3$ for ZINB with 1 in the mth component and 0 in others.

Gradients of marginal effects in the HP models

For HP models, we introduce functions: $p^{\text{HP}}(t) = \dfrac{e^{t+e^t}}{e^{e^t} - 1}$ and use the same q as ZIP. Then,

$$p^{\text{HP}}(t) = e^t + \sigma(t), \quad \dot{p}^{\text{HP}}(t) = e^t + \dot{\sigma}(t), \quad \ddot{p}^{\text{HP}}(t) = e^t + \ddot{\sigma}(t),$$

where $\sigma(t) = \dfrac{e^t}{e^{e^t} - 1}$, $\dot{\sigma}(t) = \sigma(t)\{1 - e^t - \sigma(t)\}$, $\ddot{\sigma}(t) = \dot{\sigma}(t)\{1 - e^t - 2\sigma(t)\} - e^t\sigma(t)$.

Using the similar notations for p, q and their derivatives as for ZIP and ZINB models in Sect. 1, the marginal mean is rewritten as $E(y_i|x_i, z_i) = p_i^{\text{HP}} q_i$, the marginal effect with respect to continuous covariate x_{ij} is $\eta_j(x_i, z_i, \theta) = \beta_j \dot{p}_i^{\text{HP}} q_i + \gamma_j p_i^{\text{HP}} \dot{q}_i$, and the incremental effect with respect to categorical covariate x_{ij} from level l_1 to level l_2 is $\pi_j(x_{i(-j)}, z_{i(-j)}, \theta) = p_{2i}^{\text{HP}} q_{2i} - p_{1i}^{\text{HP}} q_{1i}$, where $\theta = (\beta', \gamma')'$. Then, the formulas of gradients of marginal and incremental effects are in the same forms as ZIP models (29) with different layouts of p^{HP} and its derivatives $\dot{p}^{\text{HP}}$ and $\ddot{p}^{\text{HP}}$.

Gradients of marginal effects in the HNB models

The parameter in HNB models is $\theta = (\beta', \gamma', \alpha)'$, and we adopt the same q function in ZIP, ZINB, and HP models but define a new function p by $p^{\text{HNB}}(t, \alpha) = \dfrac{e^t}{1 - \rho(t, \alpha)}$, where $\rho(t, \alpha) = \tau^\alpha(t, \alpha)$, $\tau(t, \alpha) = \dfrac{\alpha}{\alpha + e^t}$, and $\alpha > 0$. We will use the same notations in terms of q

and its derivatives evaluated at $z_i'\gamma$, $z_{i(-j)}'\gamma_{(-j)} + l_2\gamma_j$ and $z_{i(-j)}'\gamma_{(-j)} + l_1\gamma_j$, as introduced in previous sections.

With simple computation, we can get derivatives of p^{HNB} with respect to t and $\alpha > 0$. In particular, $\dot{p}_t^{\mathrm{HNB}}(t,\alpha) = p^{\mathrm{HNB}}\big(1 - p^{\mathrm{HNB}}\rho\tau\big)$, $\ddot{p}_t^{\mathrm{HNB}}(t,\alpha) = \dot{p}_t^{\mathrm{HNB}}(t,\alpha)\big(1 - 2p^{\mathrm{HNB}}\rho\tau\big) + (\alpha+1)\big(p^{\mathrm{HNB}}\big)^2\rho\tau(1-\tau)$, $\dot{p}_\alpha^{\mathrm{HNB}}(t,\alpha) = p^{\mathrm{HNB}}\rho(\ln\tau + 1 - \tau)/(1-\rho)$, and $\ddot{p}_{t\alpha}^{\mathrm{HNB}}(t,\alpha) = \big\{\dot{p}_\alpha^{\mathrm{HNB}}(t,\alpha) \quad \cdot(1 - p^{\mathrm{HNB}}\rho\tau - p^{\mathrm{HNB}}\tau)\big\} - \big\{(p^{\mathrm{HNB}})^2\rho e^t/(\alpha + e^t)^2\big\}$, where $\tau = \tau(t,\alpha)$, $\rho = \rho(t,\alpha)$ for simplicity of notations, $\dot{\tau}_t(t,\alpha) = \tau(\tau-1)$, $\ddot{\tau}_{tt}(t,\alpha) = \tau(\tau-1)(2\tau-1)$, $\dot{\tau}_\alpha(t,\alpha) = \ e^t/(\alpha+e^t)^2$ $\dot{\rho}_t(t,\alpha) = \alpha\rho(\tau-1) = -e^t\rho\tau$, $\ddot{\rho}_{tt}(t,\alpha) = \alpha\rho(\tau-1)\{(\alpha+1)\tau - \alpha\} = \rho\tau^2e^t(e^t-1)$, $\dot{\rho}_\alpha(t,\alpha) = \rho(\ln\tau + 1 - \tau)$.

For functions $p^{\mathrm{HNB}}, \dot{p}_t^{\mathrm{HNB}}, \ddot{p}_{tt}^{\mathrm{HNB}}, \dot{p}_\alpha^{\mathrm{HNB}}, \ddot{p}_{t\alpha}^{\mathrm{HNB}}$ evaluated at fixed values of $(x_i'\beta, \alpha)$ are denoted by $p_i^{\mathrm{HNB}}, \dot{p}_{ti}^{\mathrm{HNB}}, \ddot{p}_{tti}^{\mathrm{HNB}}, \dot{p}_{\alpha i}^{\mathrm{HNB}}, \ddot{p}_{t\alpha i}^{\mathrm{HNB}}$, respectively. Values of $p^{\mathrm{HNB}}, \dot{p}_t^{\mathrm{HNB}}, \dot{p}_\alpha^{\mathrm{HNB}}$ at fixed values of $\big(x_{i(-j)}'\beta_{(-j)} + l_2\beta_j, \alpha\big)$ and $\big(x_{i(-j)}'\beta_{(-j)} + l_1\beta_j, \alpha\big)$ are denoted by $p_{2i}^{\mathrm{HNB}}, \dot{p}_{2ti}^{\mathrm{HNB}}, \dot{p}_{2\alpha i}^{\mathrm{HNB}}$, and $p_{1i}^{\mathrm{HNB}}, \dot{p}_{1ti}^{\mathrm{HNB}}, \dot{p}_{1\alpha i}^{\mathrm{HNB}}$, respectively.

By using p, q notations, the marginal mean of y_i can be rewritten as $E(y_i|x_i, z_i) = p_i^{\mathrm{HNB}}q_i$, the marginal effect with respect to continuous covariate x_{ij} is $\eta_j(x_i, z_i, \theta) = \beta_j\dot{p}_{ti}^{\mathrm{HNB}}q_i + \gamma_j p_i^{\mathrm{HNB}}\dot{q}_i$, and the incremental effect with respect to categorical covariate x_{ij} from level l_1 to level l_2 is $\pi_j(x_{i(-j)}, z_{i(-j)}, \theta) = p_{2i}^{\mathrm{HNB}}q_{2i} - p_{1i}^{\mathrm{HNB}}q_{1i}$.

Then, the formulas of gradients of marginal and incremental effects are in the same forms as ZIP models (29) with different layouts of p^{HP} and its derivatives $\dot{p}^{\mathrm{HP}}$ and $\ddot{p}^{\mathrm{HP}}$. The gradients of effects with respect to parameter θ are

$$
\begin{aligned}
&\nabla_\theta\eta_j(x_i, z_i, \theta)\\
&\quad = +\big(\beta_j\ddot{p}_{tti}^{\mathrm{HNB}}q_i + \gamma_j\dot{p}_{ti}^{\mathrm{HNB}}\dot{q}_i\big)\sum_{m=0}^{J_1} x_{im}u_{(m+1)}\dot{p}_{ti}^{\mathrm{HNB}}q_iu_{(j+1)}\\
&\qquad + \ \big(\beta_j\dot{p}_{ti}^{\mathrm{HNB}}\dot{q}_i + \gamma_j p_i^{\mathrm{HNB}}\ddot{q}_i\big)\sum_{m=0}^{J_2} z_{im}u_{(J_1+m+2)} + p_i^{\mathrm{HNB}}\dot{q}_iu_{(J_1+j+2)},\\
&\quad\ \nabla_\theta\pi_j(x_{i(-j)}, z_{i(-k)}, \theta)\\
&\quad = \big(\dot{p}_{2ti}^{\mathrm{HNB}}q_{2i} - \dot{p}_{1ti}^{\mathrm{HNB}}q_{1i}\big)\sum_{m=0,\neq j}^{J_1} x_{im}u_{(m+1)} + \big(l_2\dot{p}_{2ti}^{\mathrm{HNB}}q_{2i} - l_1\dot{p}_{1ti}^{\mathrm{HNB}}q_{1i}\big)\cdot u_{(j+1)}\\
&\qquad + \ \big(p_{2i}^{\mathrm{HNB}}\dot{q}_{2i} - p_{1i}^{\mathrm{HNB}}\dot{q}_{1i}\big)\sum_{m=0,\neq j}^{J_2} z_{im}u_{(J_1+m+2)} + \big(l_2p_{2i}^{\mathrm{HNB}}\dot{q}_{2i} - l_1p_{1i}^{\mathrm{HNB}}\dot{q}_{1i}\big)\cdot u_{(J_1+j+2)}\\
&\qquad + \ \big(\dot{p}_{2\alpha i}^{\mathrm{HNB}}q_{2i} - \dot{p}_{1\alpha i}^{\mathrm{HNB}}q_{1i}\big)u_{(J_1+J_2+3)},
\end{aligned}
$$

where $u_{(m)}$ is a unit vector of dimension $J_1 + J_2 + 3$ with 1 in the mth component and 0 in others.

References

Basu, A., Rathouz, P.J.: Estimating marginal and incremental effects on health outcomes using flexible link and variance function models. Biostatistics **6**, 93–109 (2005)

Cameron, A.C., Trivedi, P.K.: Microeconometrics: Methods and Applications. Cambridge University Press, Cambridge (2005)

Cameron, A.C., Trivedi, P.K.: Regression Analysis of Count Data. Cambridge University Press, Cambridge (2013)

Cragg, T.C.: Some statistical models for limited dependent variables with application to the demand for durable goods. Econometrica **39**, 829–844 (1971)

Dow, W., Norton, E.: Choosing between and interpreting the Heckit and two-part models for corner solutions. Health Serv. Outcomes Res. Methodol. **4**, 5–18 (2003)

Frick, J.R.: A General Introduction to the German Socio-Economic Panel Study (SOEP)-Design, Contents and Data Structure (Waves A-V, 1984–2005). Deutsches Institut für Wirtschaftsfor-schung, Berlin (2006)

Greene, W.H.: Accounting for excess zeroes and sample selection in Poisson and negative binomial regression models. NYU Working Paper No. EC-94-10: Department of Economics, New York University (1994). Available at SSRN https://ssrn.com/abstract=1293115

Greene, W.H.: Econometric Analysis, 5th edn. Prentice Hall, New Jersey (2002)

Hall, D.B.: Zero-inflated Poisson and binomial regression with random effects: a case study. Biometrics **56**, 1030–1039 (2000)

Kassahun, W., Neyens, T., Molenberghs, G., Faes, C., Verbeke, G.: Marginalized multilevel hurdle and zero-inflated models for overdispersed and correlated count data with excess zeroes. Stat. Med. **33**, 4402–4419 (2014)

Lambert, D.: Zero-inflated Poisson regression with an application to defects in manufacturing. Technometrics **34**, 1–4 (1992)

Li, C.S., Lu, J.C., Park, J., Kim, K., Brinkley, P.A., Peterson, J.P.: Multivariate zero-inflated Poisson models and their applications. J. Technometr. **41**, 29–38 (1999)

Long, L.D., Preisser, J.S., Herring, A.H., Golin, C.E.: A marginalized zero-inflated Poisson regression model with overall exposure effects. Stat. Med. **33**, 5151–5165 (2014)

Madden, D.: Sample selection versus two-part models revisited: the case of female smoking and drinking. J. Health Econ. **27**, 300–307 (2008)

Mullahy, J.: Specification and testing of some modified count data models. J. Econ. **33**, 341–365 (1986)

Pohlmeier, W., Ulrich, V.: An econometric model of the two-part decision making process in the demand for health care. J. Hum. Resour. **30**, 339–361 (1995)

Preisser, J.S., Das, K., Long, D.L., Divaris, K.: Marginalized zero-inflated negative binomial regression with application to dental caries. Stat. Med. **35**, 1722–1735 (2016)

Ridout, M., Hinde, J., Demetrio, C.G.B.: A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. Biometrics **57**, 219–223 (2001)

Riphahn, R., Wambach, A., Million, A.: Incentive effects in the demand for health care: a bivariate panel count data estimation. J. Appl. Econ. **18**, 387–405 (2003)

Staub, K., Winkelmann, R.: Consistent estimation of zero-inflated count models. Health Econ. **22**, 673–686 (2013)

Tabb, L.P., Tchetgen, E.J., Wellenius, G.A., Coull, B.A.: Marginalized zero-altered models for longitudinal count data. Stat. Biosci. **8**, 181–203 (2016)

Vuong, Q.H.: Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica **57**, 307–333 (1989)

Wang, Z., Ma, S., Wang, C.Y.: Variable selection for zero-inflated and overdispersed data with application to health care demand in Germany. Biometr. J. **57**(5), 867–884 (2015)

White, H.: Maximum likelihood estimationof misspecified models. Econometrica **50**, 1–25 (1982)

Winkelmann, R.: Econometric Analysis of Count Data, 5th edn. Springer, Berlin (2008)