Feng *Journal of Statistical Distributions and Applications* (2021) 8:8
https://doi.org/10.1186/s40488-021-00121-4

Journal of Statistical Distributions and Applications

REVIEW Open Access

A comparison of zero-inflated and hurdle models for modeling zero-inflated count data

Cindy Xin Feng

Correspondence: cindy.feng@dal.ca
Department of Community Health and Epidemiology, Faculty of Medicine, Dalhousie University, 5790 University Avenue, B3H 4R2 Halifax, Nova Scotia, Canada

Abstract

Counts data with excessive zeros are frequently encountered in practice. For example, the number of health services visits often includes many zeros representing the patients with no utilization during a follow-up time. A common feature of this type of data is that the count measure tends to have excessive zero beyond a common count distribution can accommodate, such as Poisson or negative binomial. Zero-inflated or hurdle models are often used to fit such data. Despite the increasing popularity of ZI and hurdle models, there is still a lack of investigation of the fundamental differences between these two types of models. In this article, we reviewed the zero-inflated and hurdle models and highlighted their differences in terms of their data generating processes. We also conducted simulation studies to evaluate the performances of both types of models. The final choice of regression model should be made after a careful assessment of goodness of fit and should be tailored to a particular data in question.

Keywords: Zero inflation, Hurdle model, Zero deflation, Model diagnosis

1 Introduction

In public health and epidemiology research, count data with a large proportion of zeros are often encountered. For example, in health services utilization study, the number of service utilization often includes a large number of zeros representing the patients with no utilization during the study period. In the substance abuse field, substances of interest are characterized by different frequencies of drug or alcohol use with a large number of patients reporting zero days of use during treatment. A common feature of this type of data is that the count measure tends to have excessive zero beyond a common count distribution can accommodate, such as Poisson or negative binomial. For example, in counting the number of responses to a disease, an individual may have no disease response because of the individual is immune or resistant to the disease. Previous research has shown that if excessive zero is not accounted for, unreasonable fit for both the zeros and nonzero counts will be resulted (Perumean-Chaney et al. 2013). Zero-inflated (ZI) (Lambert 1992) and hurdle models(Mullahy 1986; Heilbron 1994) have been developed to model zero-inflation when the regular count models such as Poisson or negative binomial are unrealistic. Both types of models have gained increasing popularities in many fields including public health services research (Neelon et al. 2010; Neelon et al. 2013; Neelon et al. 2016), substance abuse (DeSantis and Bandyopadhyay 2011; Buu et al. 2012), occupational injury (Yau and Lee 2001), medicine (Bohning et al. 1999; Rose et al. 2006), psychology (Atkins and Gallop 2007), public health (Yau and Lee 2001; Yau et al. 2003; YB 2002; Sharker et al. 2020), ecological and environmental studies (Agarwal et al. 2002; Rathbun and Fei 2006; Feng and Dean 2012; Feng 2020).

Despite the increasing popularity of ZI and hurdle models, the differences between these two types of models are understudied. The choice between the two types of models is often determined by comparing model fit statistics post-fitting both types of models. Among these studies, the conclusions are inconsistent. Some revealed that ZI and hurdle models are indistinguishable with respect to goodness of fit measures (Xu et al. 2015; Tüzen et al. 2018); whereas, some studies found the hurdle model had a better fit than the ZI model (Min and Agresti 2005; Sharker et al. 2020) and other empirical application found ZI model performs better than the hurdle model (Hu et al. 2011). It is therefore desired to identify the situations where hurdle models perform better than ZI and vice versa through simulation studies. Model comparison measures, such as Akaike information criterion (AIC) (Akaike et al. 1973; Akaike 2011) and Vuong's test (Vuong 1989) are used to compare the goodness of fit of these two types of models. Examination of residuals has been an important step to detect model misspecification and departure from the model assumption. Randomized quantile residuals (RQR) have been proposed by Dunn and Smyth (Dunn and Smyth 1996) for assessing the model fits for discrete outcome data. If the model is correctly specified, RQRs should be approximately normally distributed and the plot of RQRs against the predicted values should be randomly scattered without any discernible pattern(Dunn and Smyth 1996; Feng et al. 2020). Built on these works, we examine and compare the absolute fit (how well the model fits the data) of the ZI and hurdle models using RQRs.

The paper is organized as follows. In Section 2, we give a brief review of hurdle and ZI regression models. Section 3 reviewed model comparison strategies for examining model adequacy. Section 4 presented simulation studies to compare hurdle and ZI models. Concluding remarks are given in Section 5.

2 Statistical models

2.1 Zero-Inflated model

In a zero-inflated (ZI) model (Lambert 1992), zero observations have two different origins: "structural" and "sampling". The sampling zeros are from the usual Poisson or negative binomial (NB) distribution, which are assumed that were occurred by chance.

The general structure of a ZI model is given as:

$$P(Y_i = y_i) = \begin{cases} \pi_i + (1-\pi_i)p(y_i = 0; \mu_i) & y_i = 0, \\ (1-\pi_i)p(y_i; \mu_i) & y_i > 0, \end{cases} \tag{1}$$

which consists of a degenerate distribution at zero and an untruncated count distribution with a vector of parameters μ_i. If the count distribution follows a Poisson distribution, the *zero inflated Poisson model (ZIP)* is given by:

$$P(Y_i = y_i) = \begin{cases} \pi_i + (1-\pi_i)\, e^{-\mu_i} & \text{if } y_i = 0, \\ (1-\pi_i)\frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} & \text{if } y_i > 0 \end{cases}, \tag{2}$$

where μ_i is the mean of the standard Poisson distribution. ZI model is also fumulated as a latent variable model with an unobserved Bernouli random variable z_i (Lambert 1992):

$$z_i = \begin{cases} 1 & \text{if } y_i \text{ is structural zero,} \\ 0 & \text{if } y_i \sim \text{Poisson}(\mu_i), \end{cases} \tag{3}$$

which establishes

$$E(y_i) = E\left(E(y_i|z_i)\right) = (1-\pi_i)\mu_i \tag{4}$$

$$Var(y_i) = E(Var(y_i|z_i)) + Var(E(y_i|z_i)) = (1-\pi_i)\mu_i\,(1+\mu_i\pi_i)\,. \tag{5}$$

For modeling the count component of a ZI model, Poisson regression assumes the conditional mean equals to the conditional variance, which may not be valid in some situations. If data have greater conditional variance than is assumed under the Poisson model, overdispersion would occur, which may be due to population heterogeneity or clustering, omission of important covariates in the model, or the presence of outliers (Cox 1983; Dean 1992; Dean and Lundy 2016; Payne et al. 2017). Negative binomial (NB) regression could be then used to model overdispersed Poisson count data. The *zero inflated negative binomial model (ZINB)* is then given by:

$$P(Y_i = y_i) = \begin{cases} \pi_i + (1-\pi_i)\left[\left(\frac{r}{\mu_i+r}\right)^r\right] & \text{if } y_i = 0, \\ (1-\pi_i)\frac{\Gamma(y_i+r)}{\Gamma(r)y_i!}\left(\frac{\mu_i}{\mu_i+r}\right)^{y_i}\left(\frac{r}{\mu_i+r}\right)^r & \text{if } y_i > 0 \end{cases}, \tag{6}$$

where μ_i is the mean of the NB model, π_i is the probability of a structural zero, r is the dispersion parameter, Γ is the gamma function. The mean and variance of the ZINB are then given by $E(y_i) = (1-\pi_i)\mu_i$ and $Var(y_i) = (1-\pi_i)\mu_i(1+\mu_i/r+\pi_i\mu_i)$. As r goes to infinity, the ZINB reduces to the ZIP model. Therefore, small values of r indicate overdispersion. In many applications, it is common to assume that the parameters μ_i and π_i depend on vectors of explanatory variables $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$. Covariates can be associated with the probability of a structural zero, π_i, as well as the mean function μ_i of the count model. Generally, π_i is modeled with a logistic regression and μ_i is modeled as a log-linear regression. The ZI model can be written as,

$$\log(\mu_i) = \boldsymbol{x}_i^T\boldsymbol{\alpha}, \quad \text{logit}(\pi_i) = \boldsymbol{z}_i^T\boldsymbol{\beta} \tag{7}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are regression coefficients for the covariates $\boldsymbol{x}_i^T$ and $\boldsymbol{z}_i^T$. Note that the explanatory variables describing the μ_i do not need to be the same as those describing π_i. The ZINB model allows for added flexibility compared to the ZIP model. It allows for over-dispersion arising from excess zeros and heterogeneity in the Poisson component, whereas the ZIP model only accommodates over-dispersion from excess zeroes.

2.2 Hurdle model

In contrast to ZI models, hurdle models (Mullahy 1986; Heilbron 1994) can be viewed as a two-component mixture model consisting of a zero mass and the positive observations component following a truncated count distribution, such as truncated Poisson or truncated NB distribution.

Let Y_i denote the response of the ith observation, $i = 1, \cdots, n$, where n denote the total number of observations. The general structure of a hurdle model is given by

$$P(Y_i = y_i) = \begin{cases} p_i & y_i = 0, \\ (1-p_i)\frac{p(y_i;\mu_i)}{1-p(y_i=0;\mu_i)} & y_i > 0, \end{cases} \tag{8}$$

where p_i is the probability of a subject belonging to the zero component; $p(y_i; \mu_i)$ represents a probability mass function (PMF) for a regular count distribution with a vector of parameters μ_i and $p(y_i = 0; \mu_i)$ is the distribution evaluated at zero. It can be seen that the positive count is governed by a regular counts distribution as the PMF divided by 1 minus the PMF of this regular counts distribution evaluated at zero.

For example, if the count distribution follows a Poisson distribution, the probability distribution for the *hurdle Poisson model* is written as:

$$P(Y_i = y_i) \; = \begin{cases} p_i & y_i = 0, \\ (1-p_i)\frac{e^{-\mu_i}\mu_i^{y_i}/y_i!}{1-e^{-\mu_i}} & y_i > 0 \end{cases} . \tag{9}$$

Alternatively, the non-zero count component can follow other distributions to account for overdispersion and NB distribution is the most commonly used. The *HNB* model is then given by:

$$P(Y_i = y_i) \; = \begin{cases} p_i & y_i = 0, \\ \frac{1-p_i}{1-\left(\frac{r}{\mu_i+r}\right)^r}\frac{\Gamma(y_i+r)}{\Gamma(r)y_i!}\left(\frac{\mu_i}{\mu_i+r}\right)^{y_i}\left(\frac{r}{\mu_i+r}\right)^r & y_i > 0 \end{cases} , \tag{10}$$

Similar as a ZI model, covariates can enter the probability of a zero p_i and the mean function μ_i for a hurdle model. Hence, the hurdle model can be written as:

$$\log(\mu_i) = \boldsymbol{x}_i^T\boldsymbol{\alpha}, \quad \text{logit}(p_i) = \boldsymbol{z}_i^T\boldsymbol{\beta} \tag{11}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the regression coefficients for the covariates $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$, respectively.

2.3 Hurdle model versus zero-Inflated model

In general, ZI and hurdle models differ based on their conceptualization of the zeros and interpretation of model parameters. A ZI model (Lambert 1992) assumes that zero counts result from a mixture of two distributions, one where subjects always produce zero counts, which are often called "structural zeros" or "excessive zeros". Subjects who are exposed to the outcome but did not or did not report the experience of the outcome during the study period, are termed as "sampling zeros". The rationale for differentiating the zeros into two groups is that excessive zeros are often due to the existence of a subpopulation of subjects who are not at risk for certain outcomes during the study period. For example, when modeling the count of certain high-risk behaviors, some participants may score zero because they are not at risk for such health-risk behavior; these are the structural zeros since they cannot exhibit such high-risk behaviors. Other participants who are at risk may score zero because they did not exhibit such high-risk behaviors during the study period. The likelihood of being from either population is estimated with a zero-inflation probability component, while the counts in the second population of the user group are modeled by an ordinary count distribution, such as a Poisson or negative binomial (NB) distribution. In contrast, a hurdle model (Mullahy 1986; Heilbron 1994) assumes all zero data are from one "structural" source with one part of the model being a binary model for modeling whether the response variable is zero or positive, and another part using a truncated model, such as a truncated Poisson or a truncated NB distribution for the positive data. For example, in healthcare utilization studies, the zero part involves the decision of seeking care, and the positive component determines how frequent the utilization among the user's group. Below details of the difference between hurdle and zero-inflated models in terms of their ability to handle zero deflation and differences in the generating process for excessive zeros versus sampling zeros.

2.3.1 Ability to handle zero deflation

Another important difference between hurdle and ZI models is their capacity to handle zero deflation (fewer zeros than expected by the data-generating process). ZI models are not able to handle zero-deflation at any level of a factor and will result in parameter estimates of infinity for the logistic component, whereas hurdle models can handle zero-deflation (Min and Agresti 2005).

As shown in Eq. (3), ZI models are only suitable for handling zero inflation, since the probability of observing zeros in a ZI model is always greater than the probability of sampling zeros, i.e.,

$$\pi_i + (1-\pi_i)p(y_i = 0; \mu_i) > p(y_i = 0; \mu_i). \tag{12}$$

In contrast, hurdle model is not only able to handle zero inflation, but also suitable for modelling zero deflation. As shown in Eq. (8), when the probability of observing any zeros is greater than the probability of observing sampling zeros, i.e., $p_i > p(y_i = 0; \mu_i)$ the data are zero inflated; whereas, when $p_i < p(y_i = 0; \mu_i)$, the data are zero-deflated. For example, when the true model is a HNB model, zero deflation occurs when

$$\frac{\exp\left(\boldsymbol{z}_i^T\boldsymbol{\beta}\right)}{1+\exp\left(\boldsymbol{z}_i^T\boldsymbol{\beta}\right)} < \left(\frac{r}{\exp\left(\boldsymbol{x}_i^T\boldsymbol{\alpha}\right)+r}\right)^r, \tag{13}$$

which indicates that in zero-inflated count data, zero deflation could still occur at specific levels of covariates. Therefore, it is plausible that the hurdle model outperforms the counterpart ZI model, as the percentage of zero deflation across all the data points increases. As shown in Eq. (13), percentage of zero deflation depends on the mean structures for both the logistic and log-linear components. For illustration, we simulate data from a HNB model as follows,

$$\text{logit}(p_i) = \beta_0 + \beta_1 x_i,\ \log(\mu_i) = \alpha_0 + \alpha_1 x_i \tag{14}$$

where x_i is a Bernoulli random variable with probability parameter 0.5. We also consider another scenario when x_i is generated from a standard normal distribution $N(0,1)$. We set sample size as $n = 300$, the intercept for both the zero and truncated counts components as $\beta_0 = \alpha_0 = 1$ to ensure the data are overall zero inflated. The regression coefficients of x_i for the zero (β_1) and positive counts components (α_1) are set as -2 to 2 at an increment of 0.02. Percentage of zero deflation across all the data points is then calculated as:

$$\sum_{i=1}^{n} I\left\{\frac{\exp(\beta_0+\beta_1 x_i)}{1+\exp(\beta_0+\beta_1 x_i)} < \left(\frac{r}{\exp(\alpha_0+\alpha_1 x_i)+r}\right)^r\right\}/n \tag{15}$$

where $I\{\cdot\}$ is an indicator variable.

The left panel of Fig. 1 displays the percentage of zero deflation as a function of the regression coefficients (β and α) in the two model components when the data are simulated from a HNB model with a binary covariate simulated from a Bernoulli distribution with probability parameter 0.5. As displayed, when the regression coefficients for the

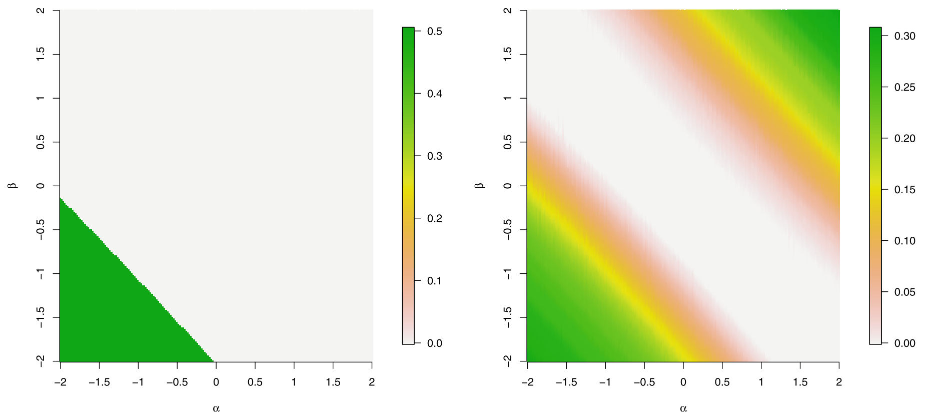


Fig. 1 Percentage of zero deflation over all data points when the data are simulated from a HNB model of sample size $n = 300$. In the left panel, the covariate x is a binary variable simulated from a Bernoulli distribution with probability parameter 0.5. In the right panel, the covariate is a continuou variable simulated from a standard normal distribution

logistic component (β) and log-linear components (α) are below zero, more than 50% of the data are zero-deflated, shown as the green shaded areas in the bottom left corner.

To further demonstrate at what level of the covariate, zero deflation may occur, Fig. 2 plots the probability of being a zero, the probability of being a sampling zero, and their differences, i.e., probability of being a zero minus the probability of being a sampling zero against the covariate when the regression coefficients for the zero (β_1) and the truncated counts component (α_1) are set as -2, -1.5, -1, -0.1, 0.1, 1, 1.5 and 2. As shown in Fig. 2, when the covariate equals zero, no zero deflation would occur. Specifically, the probability of being a zero is $\exp(1)/(1 + \exp(1)) \approx 0.75$ and the probability of being a sampling zero from the NB model is $(r/\exp(1) + r)^r \approx 0.24$, when $r = 1.2$. Therefore, when the covariate is zero, the probability of being zero is always greater than the probability of being a sampling zero in this setting. Nevertheless, when the binary covariate equals one, zero deflation occurs when α_1 and β_1 approach to -2, since the probability of being zero is $\exp(-1)/(1 + \exp(-1)) \approx 0.27$ and the probability of being a sampling zero is $(r/\exp(1) + r)^r \approx 0.73$; that is, the probability of being zero is less than the probability of being a sampling zero. In this illustrative example, the covariate was generated with 50% of ones, so the probability of zero deflation would be approximately equal to 50% when the regression coefficients α and β belong to the bottom left corner of Fig. 1. Similarly, if q% of one's in the covariate, we would expect q% of zero deflation when α_1 and β_1 belong to the bottom left corner of Fig. 1.

The right panel of Fig. 1 displays the percentage of zero deflation as a function of the regression coefficients in the two model components in the scenario when the data are simulated from a HNB model with a continuous covariate generated from a standard normal distribution. As displayed, as the regression coefficients become larger in magnitude in the same direction, the percentage of zero deflation increases. For example, when $\beta = -2$ and $\alpha = -2$ or when $\beta = 2$ and $\alpha = 2$, the percentage of zero-deflation is above 30%. However, when the regression coefficients are in different sign, for example, $\beta = 2$ and $\alpha = -2$ or $\beta = -2$ and $\alpha = 2$, the percentage of zero deflation tends to be low.

To further explore at what level of covariate zero deflation may occur, Fig. 3 plots the probability being a zero, the probability being sampling zeros, and their differences

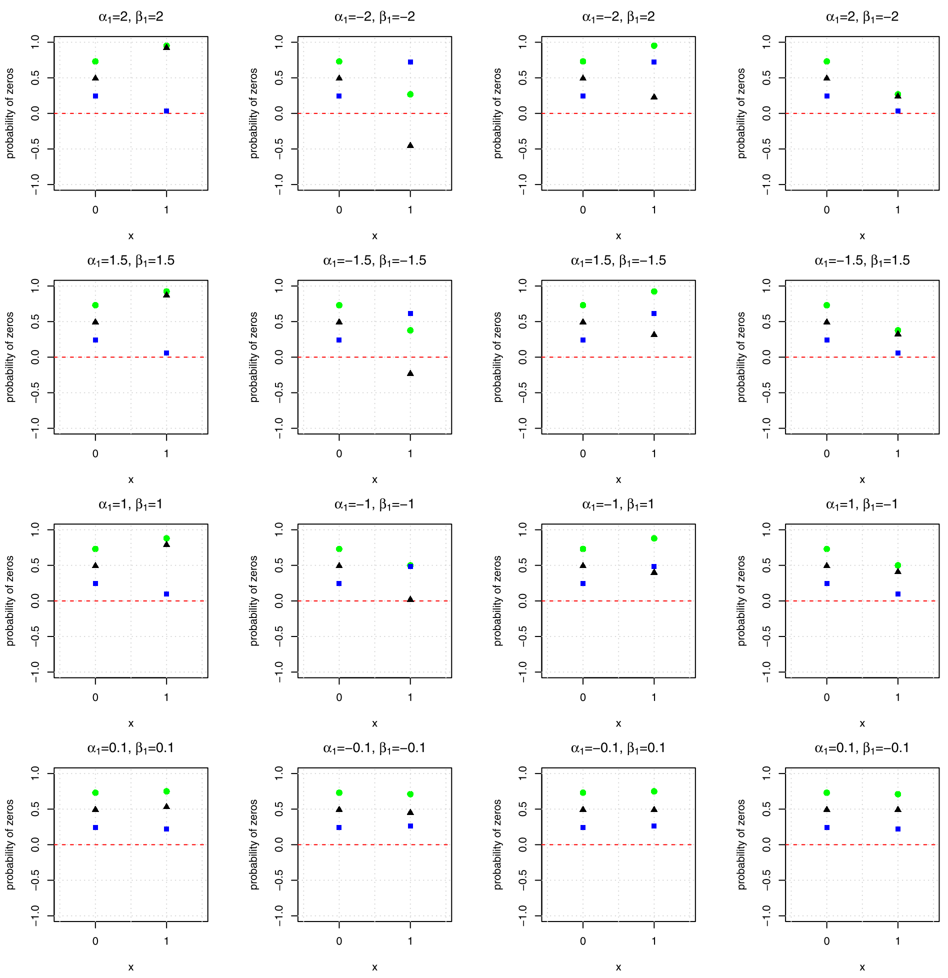


Fig. 2 Probabilities of observing a zero (green), a sampling zero (blue) and their differences (black) against the covariate when the data are simulated from a HNB model with a binary covariate of sample size $n = 300$. The intercepts for both the zero and truncated counts components are set as 1. The regression coefficients of the covariate for the zero (β_1) and the truncated counts component (α_1) are set as -2, -1.5, -1, -0.1, 0.1, 1, 1.5 and 2

against the covariate when the regression coefficients for the logistic component (β_1) and the log-linear component (α_1) are set as -2, -1.5, -1, -0.1, 0.1, 1, 1.5 and 2. As shown in the Figure, when the regression coefficients for the logistic and log-linear components are equal to 2, zero deflation occurs when the covariate x is roughly below -0.5. In this case, the means of the logistic and log-linear components are negative, resulting in a small chance of observing zeros but a large chance of observing sampling zeros. Similarly, when α_1 and β_1 are equal to -2, zero deflation is observed when the covariate x is above 0.5. However, when the signs of α_1 and β_1 are opposite, or both are at a relatively smaller magnitude, zero deflation will be less likely to occur.

In the circumstances when there is no zero deflation at any level of the covariates, ZI model can be rewritten as a hurdle model. To illustrate this, suppose a simple hurdle model is written as follows,

$$\text{logit}(p_i) = \beta_0 + \beta_1 x_i \;\; \log(\mu_i) = \alpha_0 + \alpha_1 x_i \tag{16}$$

where x_i follows a standard normal distribution $N(0,1)$. The counterpart ZI model is expressed as

$$\text{logit}(\pi_i) = \beta_0^* + \beta_1^* x_i \;\; \log(\mu_i) = \alpha_0^* + \alpha_1^* x_i \tag{17}$$

The connection between ZI and hurdle models can be built through equating the probability of observing zeros in the data, i.e.,

$$\pi_i + (1-\pi_i)p(0;\mu_i) = p_i \tag{18}$$

$$\pi_i = \frac{p_i - p(0;\mu_i)}{1-p(0;\mu_i)}. \tag{19}$$

When $x_i = 0$, $p_i = e^{\beta_0}/\left(1+e^{\beta_0}\right)$, so

$$\beta_0^* = \text{logit}(\pi_i) = \text{logit}\left(\frac{e^{\beta_0}/\left(1+e^{\beta_0}\right) - p(0;\mu_i)}{1-p(0;\mu_i)}\right). \tag{20}$$

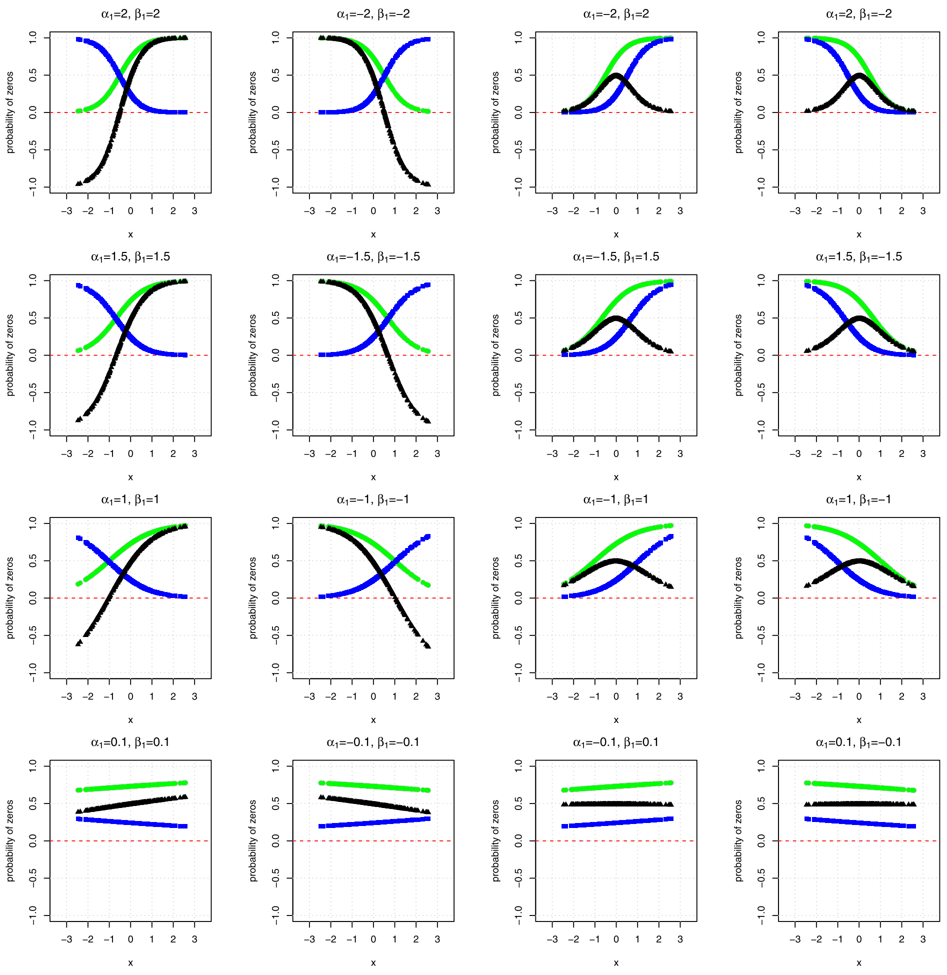


Fig. 3 Probabilities of observing a zero (green), a sampling zero (blue) and their differences (black) against the covariate, when the data are simulated from a HNB model with a continuous covariate of sample size $n = 300$. The intercepts for both the zero and truncated counts components are set as 1. The regression coefficients of the covariate for the zero (β_1) and the truncated counts component (α_1) are set as -2, -1.5, -1, -0.1, 0.1, 1, 1.5 and 2

When $x_i = 1, p_i = e^{\beta_0+\beta_1}/\left(1+e^{\beta_0+\beta_1}\right)$, so

$$\beta_1^* = \text{logit}(\pi_i) - \beta_0^* = \text{logit}\left(\frac{e^{\beta_0+\beta_1}/\left(1+e^{\beta_0+\beta_1}\right) - p(0;\mu_i)}{1-p(0;\mu_i)}\right) - \beta_0^*. \tag{21}$$

For example, for a HNB model of sample size $n = 1000$. Suppose $\beta_0 = \beta_1 = \alpha_0 = \alpha_1 = 1$. The intercept and regression coefficient for the ZINB model are then about $\beta_0^* = 0.599$ and $\beta_1^* = 1.288$. The second part of the hurdle model has the same parameters as the second part of the ZI model.

2.3.2 Generating processes for excessive zeros versus sampling zeros

Although the hurdle model is able to handle zero deflation at any level of the covariates, it treats all the zeros generated from the same processes; whereas, the ZI model allows for two data generating processes for zeros depending on the mean structures of the logistic and log-linear components. As a result, it should be expected that the ZI model outperforms the hurdle model when the data generating processes for the excessive zeros and sampling zeros differ to some extent.

To measure the difference in probability of a binary variable being an excessive zero versus being a sampling zero, we employed the standardized difference for a binary random variable (Austin 2009), which is defined as

$$d_i = \frac{\pi_{i1} - \pi_{i2}}{\sqrt{\frac{(\pi_{i1}(1-\pi_{i1})+\pi_{i2}(1-\pi_{i2})}{2}}}, \tag{22}$$

where π_{i1} and π_{i2} denote the probability of the underlying Bernoulli distribution of the binary variable, i.e., the probability of being an excessive zero and sample zero, respectively. It should be noted that there is no accepted threshold for the standardized difference to indicate the presence of meaningful imbalance (Austin 2009). Nevertheless, a α% confidence interval for d_i: $d_i \pm z_{\alpha/2} \times \sigma(d_i)$, where $\sigma(d_i) = \sqrt{2 + d_i^2/4}$ for comparing the means of Bernoulli variables (Hedges and Olkin 1985) was applied to approximately determine the differences in the generating process of sampling zeros and structural zeros. In this circumstance, the direction of the comparison is not of interest, but rather the magnitude of the differences, i.e., $|d_i|$. As a result, at α% level of significance, if the absolute value of the standardized difference $|d_i|$ exceeds $z_{\alpha\%} \times \sigma(d_i)$, we regard there is strong evidence of the probabilities of being an excessive zero and sampling zero are substantially different. To quantify the overall distances between the probability being an excessive zero vs. sampling zero across all data points, we calculate the mean of $|d_i| > z_{\alpha\%} \times \sigma(d_i)$.

To illustrate the impact of this standardized distance measure on the model fit performance between ZINB and hurdle models, we simulate data from a ZINB model with the mean structures as follows:

$$\text{logit}(\pi_{i1}) = \beta_0 + \beta_1 x_i, \ \log(\mu_i) = \alpha_0 + \alpha_1 x_i \tag{23}$$

where x_i is a Bernoulli random variable with probability of event as 0.5. In another simulation scenario, x is generated from a standard normal distribution $N(0,1)$. We set sample size as $n = 300$, the intercept for both the zero and truncated counts components as $\beta_0 = \alpha_0 = 1$ and the regression coefficients of x_i for the zero (β_1) and positive counts components (α_1) as -2 to 2 at an increment of 0.02. In this illustrative example, $\pi_{i1} = \frac{\exp(\beta_0+\beta_1 x_i)}{1+\exp(\beta_0+\beta_1 x_i)}$ and $\pi_{i2} = p(y=0;\mu_i) = \left(\frac{r}{\exp(\alpha_0+\alpha_1 x_i)+r}\right)^r$.

We then calculate the mean of $d_i, i = 1 \cdots, n$ as the measure of discrepancy of the data generating processes between the excessive zero and sampling zero.

Figure 4 displays the mean of d over varying values of the regression coefficients of the two model components when the data are simulated from a ZINB model with a binary covariate generated from a Bernoulli random variable with probability parameter 0.5 (left panel) and a ZINB model with a continuous covariate generated from a standard normal distribution (right panel). As shown in the left panel of Fig. 4, the difference in the probabilities of being an excessive zero versus being a sampling zero is negligible at various levels of the covariate effects; whereas, under the scenario with a continuous standard normal covariate (right panel of Fig. 4), the difference in the probabilities of observing an excessive zero versus observing a sampling zero manifests when the regression coefficients of the covariate in the logistic and log-linear components approach -2 or 2.

3 Model selection and goodness-of-fit statistic

3.1 Relative fit measures

Akaike information criterion (AIC) (Akaike et al. 1973; Akaike 2011) is used for comparing the model fits between hurdle and ZI models, which is computed as $AIC = -2\log(L) + 2q$, where L is the likelihood, and q is the number of parameters in the model. In general, the best fitting model yields the lowest AIC values. Another widely used tool for comparing ZI versus hurdle models is Vuong test (Vuong 1989), which compares the likelihood functions between two models. Let $f_1(y_i|\theta_1)$ and $f_2(y_i|\theta_2)$ denote the probability distribution functions of two models. Their likelihood functions should be nearly identical when the two models fit the data equally well. Differences between the likelihood functions indicate which model fits the data better. Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be the maximum likelihood estimate (MLE) of θ_1 and θ_2. The Vuong test is to compare the likelihood function at the MLE between the two models, that is $\rho_i = \log(f_1(y_i|\hat{\theta}_1)) - \log(f_2(y_i|\hat{\theta}_2))$. The Vuong test for comparing $f_1(y_i|\hat{\theta}_1)$ and $f_2(y_i|\hat{\theta}_2)$ is then defined as $V = \sqrt{n}\bar{\rho}/s_\rho$, where $\bar{\rho}$ and s_ρ is the mean and standard deviation of the vector of $\boldsymbol{\rho} = (\rho_1, \cdots, \rho_n)$.

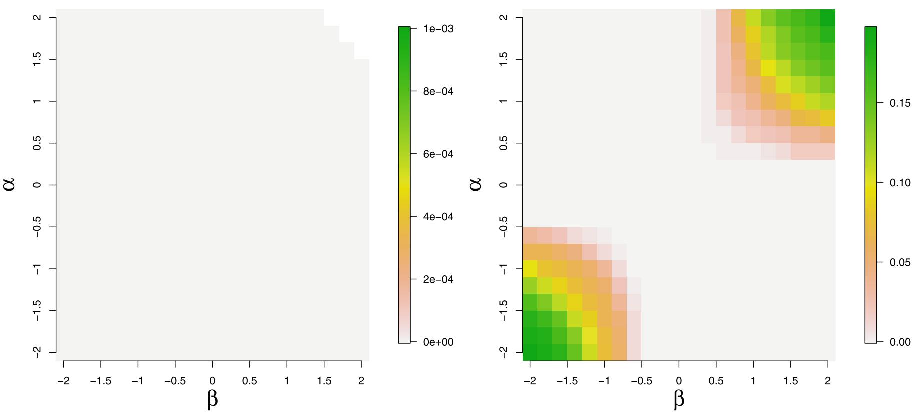


Fig. 4 Mean of the standardized differences of the probability of being an excessive zero and the probability of being a sampling zero when data are simulated from a ZINB model of sample size $n = 300$. In the left panel, the covariate is a binary variable from a Bernoulli random variable with probability parameter 0.5. In the right panel, the covariate is a continuous variable from a standard normal distribution

Under the null hypothesis, i.e., the two models fit the data equivalently, Vuong's statistic asymptotically follows a standard normal distribution. At the 5% level of significance, the critical value is 1.96, so if $V > 1.96$, the statistic favours the model in the numerator; whereas, if $V < -1.96$, the statistic favours the model in the denominator, and when $V \in (-1.96, 1.96)$, two models fit the data equally, with no preference given to either model.

3.2 Absolute fit measures

Model checking is an essential step of statistical modeling that ensures the assumptions are met for valid inference. Residual diagnosis is often used to assess how well the model captures the characteristics of the data; however, when data is discrete and skewed, standard residuals such as Pearson and deviance residuals do not follow a standard normal distribution. A tool that quantifies model fit on an easy-to-understand aligning with the scale of the traditional residuals used in normal regression, would be helpful to practitioners.

Randomized quantile residual (RQR) was defined by Dunn and Smyth (1996) to diagnose counts models, such as Poisson or NB models, but has not been often used for diagnosing ZI or hurdle models. RQR has recently been demonstrated and evaluated through comprehensive simulation studies. The results showed that RQRs could be applied to diagnose regression models for scalar y_i provided that one can compute the CDF and PMF of the considered model (Feng et al. 2020). RQR can be defined as follows in general. Suppose we consider fitting a regression model with $F(y_i; \mu_i, \phi)$ denoting the CDF for a response variable y_i given a set of covariates x_i, where μ_i is typically a function of x_i, for example the conditional mean of y_i, whereas ϕ does not depend on x_i, for example dispersion parameter. Let $d(y_i; \mu_i, \phi)$ be the corresponding PMF of $F(y_i; \mu_i, \phi)$. If F is discrete, the estimated lower tail probability is randomized into a uniform random number, which is defined as a function with a random number u_i from the uniform distribution on $(0, 1]$ as an additional argument, $F^*(y_i; \hat{\mu}_i, \hat{\phi}, u_i) = F(y_i-; \hat{\mu}_i, \hat{\phi}) + u_i\, d(y_i; \hat{\mu}_i, \hat{\phi})$, where $F(y_i-; \hat{\mu}_i, \hat{\phi})$ is the lower limit of F at y_i, i.e., $\sup_{y<y_i} F(y; \hat{\mu}_i, \hat{\phi})$, the lower limit in the "gap" of $F(\cdot, \hat{\mu}_i, \hat{\phi})$ at y_i. Here we use right close interval for u_i only for mathematical convenient in our proof, which does not have practical implication. An alternative way to define the randomized lower tail probability as a uniform random number between $a = \sup_{y<y_i} F(y; \hat{\mu}_i, \hat{\phi})$ and $b = F(y_i; \hat{\mu}_i, \hat{\phi})$ (Dunn and Smyth 1996).

RQR for y_i is the standard normal quantile corresponding to the random lower tail probability with μ_i and ϕ estimated from the sample, $q_i = \Phi^{-1}(F^*(y_i; \hat{\mu}_i, \hat{\phi}, u_i)) = q(y_i; \hat{\mu}_i, \hat{\phi}_i, u_i)$, where Φ^{-1} is the quantile function of a standard normal distribution, and u_i is a random number uniformly distributed on $(0, 1]$. $F^*(y_i; \mu_i, \phi, u_i)$ can be converted to any other standard distribution as above. The normal distribution is chosen because most people are familiar with normal random variates with the so-called "empirical rules". For evaluating model goodness of fit, we can test the following hypotheses H_0: Model fits the data well and H_a: Model does not fit the data well, by examining the normality of RQR based on the Shapiro-Wilk (SW) normality test. Under the true model, the null hypothesis should not be rejected, so RQR should be normally distributed, i.e., the p-value of the SW test of RQRs should be greater than 5%. Whereas under the incorrectly specified model, the null hypothesis should be rejected, so RQR should not be normally distributed, i.e., the p-value of the SW test should be less than 5%.

4 Simulation studies

The simulation study is carried out to investigate the behavior of the hurdle versus ZI models. The simulation settings consist of model comparison using AIC and Vuong test as well as the overall model goodness of fit calculated as the SW normality test p-value for testing the normality of the RQR as described in Section 3.

4.1 Simulation settings

To compare the performance of hurdle and ZI models, we consider simulating data from (1) a HNB as the true model and (2) a ZINB as the true model. To highlight the model performance depending on the type of covariates included in the model, we incorporate different types of covariate in the model, i.e., (i) a binary covariate x simulated from a Bernoulli distribution $x \sim \text{Bern}(p)$ or (ii) a continuous covariate x simulated from a Normal distribution $x \sim N(0, 1)$. For each simulation scenario, we generated 200 random samples from the true model, and then both HNB and ZINB models are fitted to the simulated datasets with the covariate entering both the logistic and log-linear components of the models.

4.2 Factors considered

We varied the values of the following factors to investigate their influence on the performance of the model fits.

- **Strength of the covariate effects**: The strength of the association between the exposure and outcome, measured by β_1 and α_1. The values were set to -2, -1.5, -0.5, -0.1, 0.1, 0.5, 1.5 and 2 in the simulation.
- **Sample size**: To study the finite sample properties of the models, we considered sample sizes $n = 300, 500$, and 700.
- **Proportion of excess zeros**: The intercept for the logistic component, β_0 was set as 1 to control the percentage of zeros and ensure the simulated datasets are zero-inflated. The intercept of the log-linear component was set as one so that the mean of the counts is not too low. The overdispersion parameter for the NB model is set as 1.2. To confirm the simulated datasets are zero-inflated, we compared the ZINB (true) and NB models in terms of AICs and Vuong's test for each simulated dataset. The results showed that the true model outperformed the NB model in all simulated datasets.

4.3 Evaluation criteria

The performance of the two models is assessed by the relative fit measures and absolute fit measures as follows.

4.3.1 Relative fit measures

For relative fit measures, we used AIC to compare the true and misspecified models in terms of the percentage of the differences in the AICs for the misspecified model and true model are greater than 4 ($\%\Delta AIC > 4$) (Burnham and Anderson 2004) and the mean of the differences of AICs between the misspecified and true model ($\bar{\Delta}$AIC), where ΔAIC=AIC(W)-AIC(T), where W and T represent the wrong and true models, respectively. Both measurements should increase as the difference between the true and wrong models increases. We have also used Vuong's test to compare the ZI and HNB models by measuring the percentage of Vuong test p-value$<$ 5% over repeated samples.

4.3.2 Absolute fit measures

For absolute fit measures, we used the SW normality test to test the normality of RQR in terms of the type I error rates and power. The type I error rates are estimated using the proportion of datasets for which the null hypothesis (true model) is falsely rejected, i.e., the percentage of SW test p-value$<$ 5% for the true model over repeated samples. The power of the test is calculated as the proportion of datasets for which the alternative hypothesis (wrong model) is rejected, i.e., the percentage of SW test p-value$<$ 5% for the wrong model over repeated samples.

4.4 Results

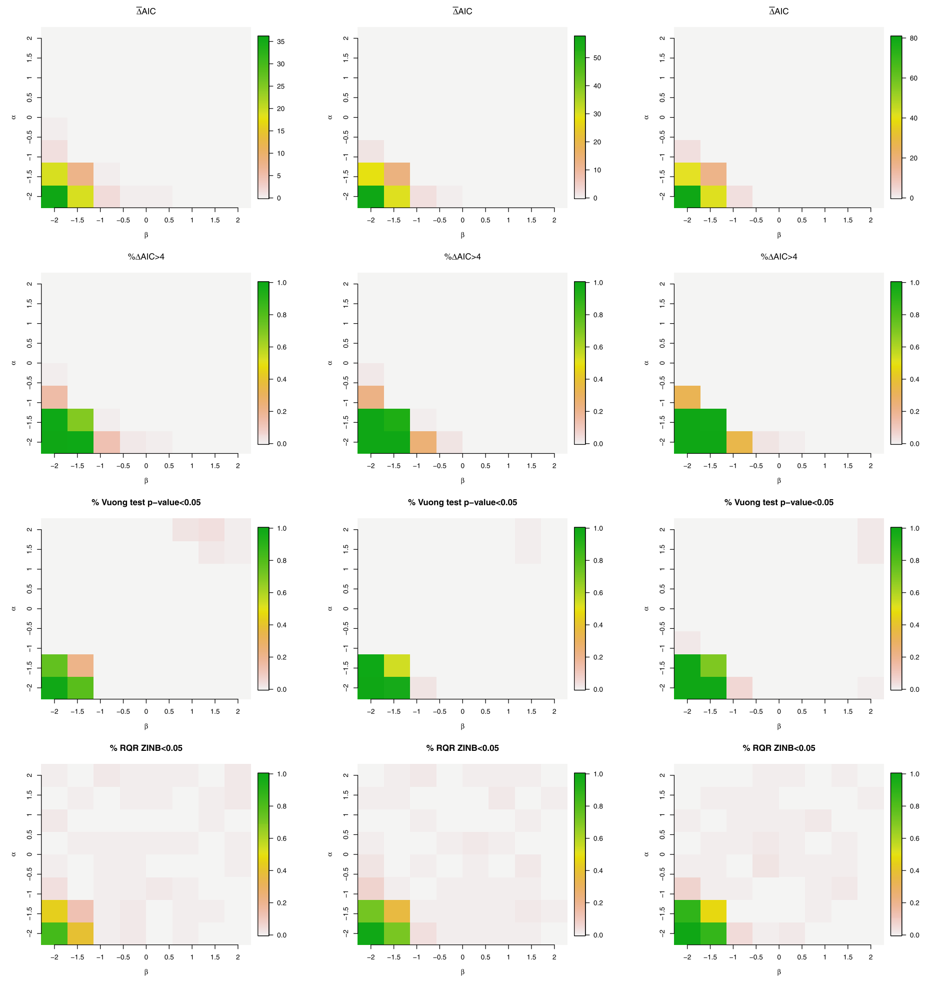
4.4.1 Simulation setting #1 (True model: HNB)

First, we evaluate the performance of ZINB and HNB models when the data are simulated from a HNB model. Figure 5 plots the relative fit measures and absolute fit measures when the data are simulated from a HNB with a single binary covariate generated from a Bernoulli distribution with probability parameter 0.5. In both the logistic and log-linear components, the regression coefficients vary between -2 to 2 at sample size $n = 300, 500$ and 700. Our simulation results demonstrate that when the data contains zero-deflated data points as depicted in the left panel of Fig. 1, the ZINB model performs poorly as compared with the counterpart HNB model, yielding a higher AIC and significant difference in model fits according to the Vuong's test (Fig. 5). RQRs are also not normally distributed under the ZINB model as the percentage of zero-deflated data points increases. As expected, the evidence of rejecting the ZINB model becomes stronger as the sample size increases. On the other hand, when the percentage of zero deflation in the data approaches zero, hurdle and ZINB models yield equivalent fits.

For the scenario when the data are simulated from a HNB model with a continuous covariate generated from a standard normal distribution, Fig. 6 again confirms that the comparison of the model fits between the HNB and the ZINB model closely align with the percentage of zero-deflated data across all the data points as depicted in the right panel of Fig. 1. More specifically, when there is no zero-deflated data, both HNB and ZINB fit the data equivalently well; whereas, as the percentage of zero deflation increases, the difference in the model fits between the HNB and ZINB models increases. We also remark that the all model comparison measures between the HNB and ZINB models increase as the sample size increases, which suggests that the relative predictive gain by the HNB model increases with increasing sample size. This is not surprising because as the sample size increases, the statistical power of identifying model misspecification increases.

4.4.2 Simulation setting #2 (True model: ZINB)

In the second simulation setting, we compare the overall goodness of fit between the ZINB and HNB model when the data are simulated from a ZINB model. Figure 7 plots the relative and absolute fit measures when the data are simulated from a ZINB model containing a single binary covariate generated from a Bernoulli distribution with probability parameter 0.5. In both the logistic and log-linear components, the regression coefficients of the covariate vary between -2 to 2 at sample size $n = 300, 500$ and 700. Our simulation results demonstrate HNB model can govern the prediction equivalently well as the ZINB model in all scenarios. Recall as shown in the left panel of Fig. 4, even when the structural zeros and sampling zeros are simulated from two largely different processes,

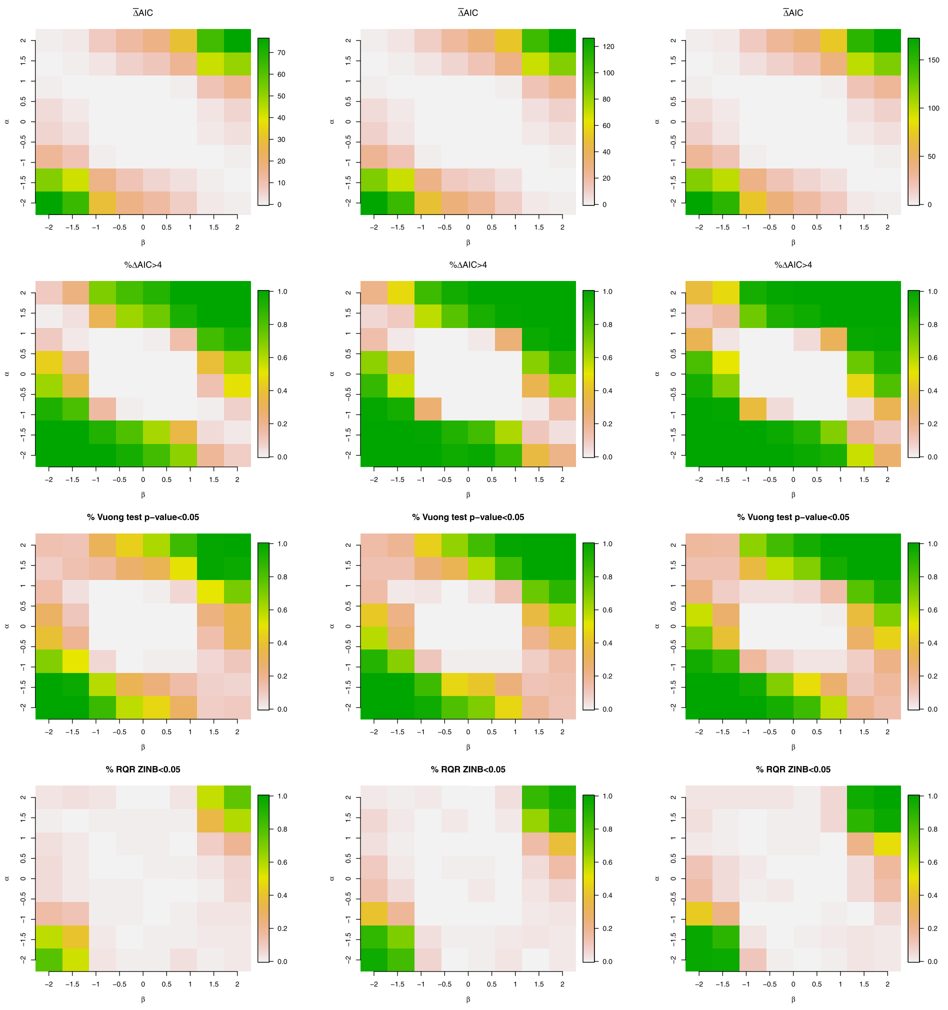


(a) n=300 (b) n=500 (c) n=700

Fig. 5 Simulation results for the simulation setting #1 (true model: HNB model with a single binary covariate generated from a Bernoulli distribution with probability parameter 0.5). Evaluation criteria include $\bar{\Delta}$AIC (mean difference in AICs of the ZINB and HNB models); %ΔAIC >4 (percentage of the differences in AICs between the ZINB and HNB models that are above 4; percentage of Vuong's test p-value< 5% and percentage of the SW normality test of the RQRs for the ZINB model < 5%

the percentage of the large discrepancies between excessive and sampling zeros is close to zero, which provide strong justification to use the discrepancy measure between excessive and sampling zeros developed in Section 2.3.2 to characterize the feature of a ZI model as compared to a hurdle model.

In the setting when the data are simulated from a ZINB model with a continuous covariate generated from a standard normal distribution, the differences between HNB and ZINB model are observed in Fig. 8. The simulation scenarios where the differences occur are consistent with the settings identified for the large differences between the excessive zeros and sampling zeros as shown in the right panel of Fig. 4. More specifically, the ZINB model has a better fit to the data than the HNB model according to the relative fit measures; whereas, RQRs did not significantly identify inadequacy of the HNB model. This indicates that RQRs are not sensitive to detect small differences between the two models.



(a) n=300 (b) n=500 (c) n=700

Fig. 6 Simulation results for the simulation setting #1 (true model: HNB model with a single continuous covariate generated from a standard normal distribution). Evaluation criteria include $\bar{\Delta}$AIC (mean difference in AICs of the ZINB and HNB models); %ΔAIC >4 (percentage of the differences in AICs between the ZINB and HNB models that are above 4; the percentage of Vuong's test p-value< 5% and the percentage of the SW normality test of the RQRs for the ZINB model < 5%

5 Conclusion and future work

This study reviewed ZI and hurdle models, which are commonly used for modeling zero-inflated count data. This study provided a better understanding of the differences between these two types of models regarding their characteristics and overall model fits.

Our simulation study showed that, with zero-inflated data, zero deflation could occur at certain levels of the covariate, in which case, the hurdle model tends to outperform the ZI model, since only the hurdle model can handle zero-deflated data. Such evidence becomes stronger as the proportion of data points that are zero-deflated increases. Therefore, if there exist a group of subjects in the data with fewer zeros than the sampling zeros from a conventional counts regression model, the hurdle model may be more appropriate than a ZI model. Hurdle and ZI models perform almost equivalently in the overall model fit

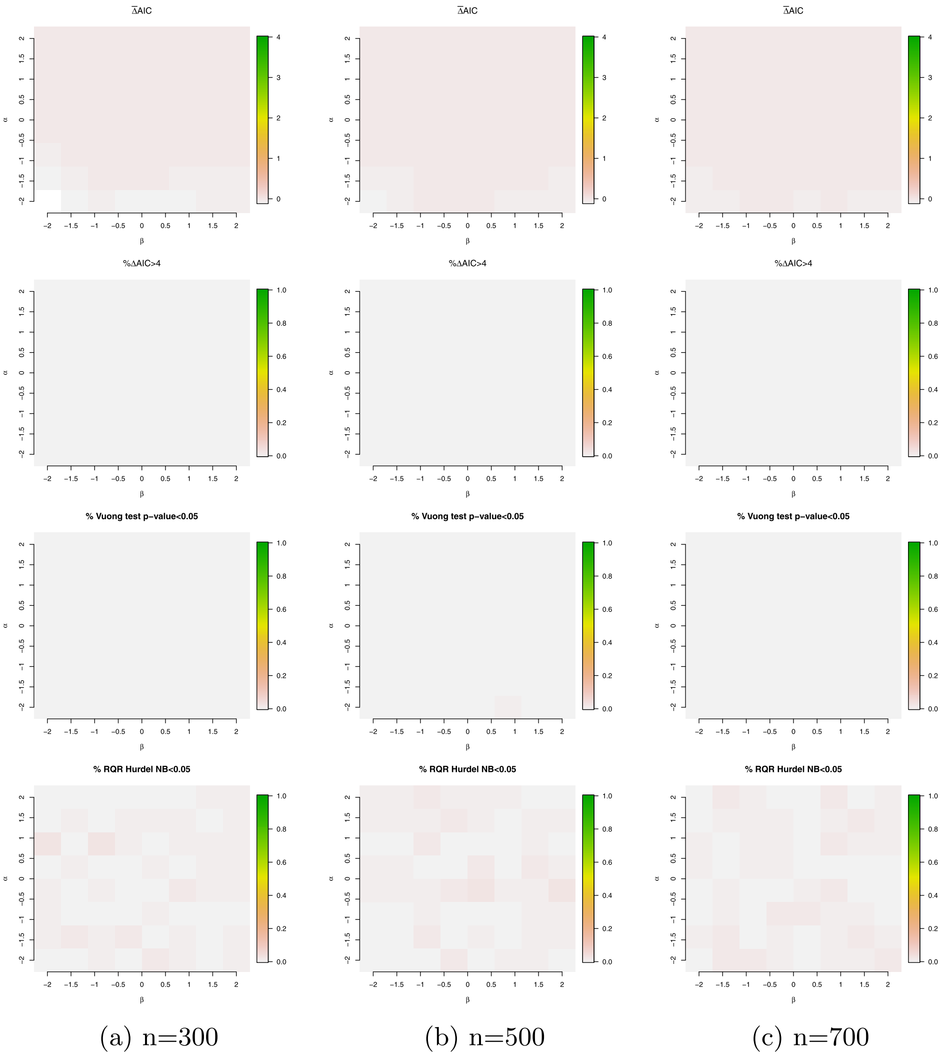
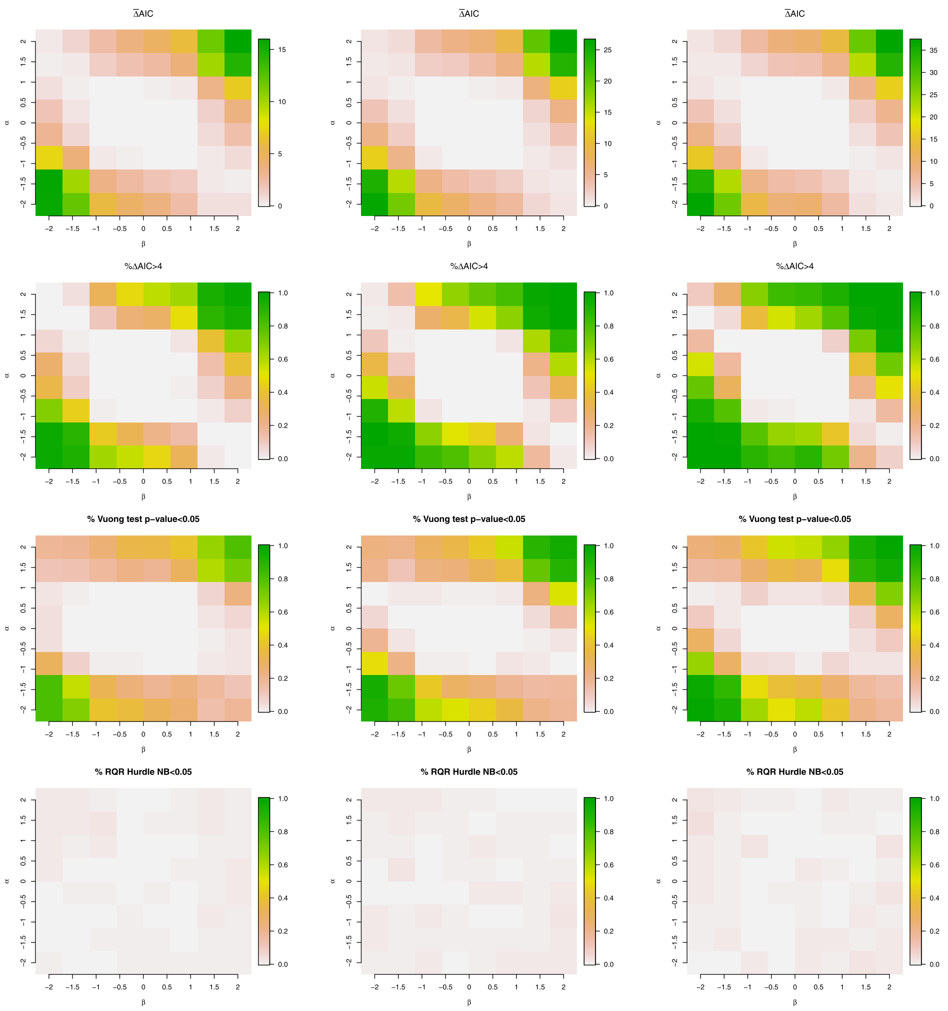


Fig. 7 Simulation results for the simulation setting #2 (true model: ZINB model with a single binary covariate generated from a Bernoulli distribution with probability parameter 0.5). Evaluation criteria include $\bar{\Delta}$AIC (mean difference in AICs of the HNB and ZINB models); %ΔAIC >4 (percentage of the differences in AICs between the HNB and ZINB models that are above 4; the percentage of Vuong's test p-value< 5% and the percentage of the SW normality test of the RQRs for the HNB model $<$ 5%

when there are no or few zero deflations across all the data points when the data are simulated from a hurdle model.

Alternatively, when the data are simulated from a ZI model, the ZI model is more favorable than the hurdle model when there are substantial differences between the probability of structural zeros and sampling zeros. This phenomenon is more evident when the model contains a continuous covariate from a standard normal distribution as compared to the model containing only a single binary covariate. If the processes of generating sampling zeros and structural zeros are not substantially different, the two models yield almost identical model fits.

Overall, our simulation studies indicate the inappropriate application of the ZI and hurdle models could have an undesirable impact on overall model fit. The performances



(a) n=300 (b) n=500 (c) n=700

Fig. 8 Simulation results for the simulation setting #2 (true model: ZINB model with a single continuous covariate generated from a standard normal distribution). Evaluation criteria include $\bar{\Delta}$AIC (mean difference in AICs of the HNB and ZINB models); %ΔAIC >4 (percentage of the differences in AICs between the HNB and ZINB models that are above 4; the percentage of Vuong's test *p*-value< 5% and the percentage of the SW normality test of the RQRs for the HNB model < 5%

of the two types of models depend on the percentage of the zero-deflated data points in the data and the discrepancy in the data generating processes between the structural zeros and sampling zeros. It is therefore important to recognize the distinct features of these two types of models.

In the current research, we only considered a single covariate to illustrate the model performance depends on the type of covariates included in the model. Additional research needs to be conducted to expand these results to models with multiple covariates. Further, our simulation study only considered independent data. It would be an interesting research topic to consider various correlation structures in the data to assess if the strength of the correlation and correlation structure play a role in choosing between ZI and hurdle model.

Abbreviations

ZI: Zero-inflated; ZIP: Zero-inflated poisson; NB: Negative binomial; ZINB: Zero-inflated negative binomial; HNB: Hurdle negative binomial; PMF: Probability mass function; CDF: Cumulative distributions function; RQR: Randomized quantile residuals; SW: Shaprio-Wilk; AIC: Akaike information criterion

Acknowledgements

The author is grateful to the reviewers and editor for their careful reading and very helpful suggestions that helped improve the paper.

Authors' contributions

CF conceptualized the study, conducted simulation studies, and drafted the manuscript. The author(s) read and approved the final manuscript.

Funding

This research is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant. The funder played no role in any of the design of the study, analysis, interpretation of data, and writing the manuscript.

Availability of data and materials

The R codes for the simulation study are available from the corresponding author.

Declarations

Competing interests

The author declare that they have no competing interests.

Received: 2 February 2021 Accepted: 19 May 2021
Published online: 24 June 2021

References

Agarwal, D. K., Gelfand, A. E., Citron-Pousty, S.: Zero-inflated models with application to spatial count data. Environ. Ecol. Stat. **9**, 341–355 (2002)

Akaike, H.: Akaike's Information Criterion(Lovric, M., ed.) Springer, Berlin (2011)

Akaike, H., Petrov, B. N., Csaki, F.: Second international symposium on information theory. Akadémiai Kiadó, Budapest (1973)

Atkins, D., Gallop, R.: Rethinking how family researchers model infrequent outcomes: A tutorial on count regression and zero-inflated models. J. Fam. Psychol. **21**(4), 726–735 (2007)

Austin, P. C.: Using the standardized difference to compare the prevalence of a binary variable between two groups in observational research. Commun. Stat. Simul. Comput. **38**(6), 1228–1234 (2009)

Bohning, D., Dietz, E., Schlattmann, P., Mendonca, L., Kirchner, U.: The zero-inflated poisson model and the decayed, missing and filled teeth index in dental epidemiology. J. R. Stat. Soc. Ser. A. **162**(2), 195–209 (1999)

Burnham, K. P., Anderson, D. R.: Multimodel inference: Understanding AIC and BIC in model selection. Sociol. Methods Res. **33**(2), 261–304 (2004)

Buu, A., Li, R., Tan, X., Zucker, R. A.: Statistical models for longitudinal zero-inflated count data with applications to the substance abuse field. Stat. Med. **31**, 4074–4086 (2012)

Cox, D. R.: Some remarks on overdispersion. Biometrika. **70**(1), 269–274 (1983)

Dean, C. B.: Testing for overdispersion in poisson and binomial regression models. J. Am. Stat. Assoc. **87**(418), 451–457 (1992)

Dean, C. B., Lundy, E. R.: Overdispersion. Wiley, New Jersey (2016)

DeSantis, S. M., Bandyopadhyay, D.: Hidden Markov models for zero-inflated Poisson counts with an application to substance use. Stat. Med. **30**(14), 1678–94 (2011)

Dunn, P. K., Smyth, G. K.: Randomized quantile residuals. J. Comput. Graph. Stat. **5**(3), 236–244 (1996)

Feng, C.: Zero-inflated models for adjusting varying exposures: a cautionary note on the pitfalls of using offset. J. Appl. Stat. **0**(0), 1–23 (2020)

Feng, C. X., Dean, C. B.: Joint analysis of multivariate spatial count and zero-heavy count outcomes using common spatial factor models. Environmetrics. **23**(6), 493–508 (2012)

Feng, C. X., Li, L., Sadeghpour, A.: A comparison of residual diagnosis tools for diagnosing regression models for count data. BMC Med. Res. Methodol. **20**(175), 1–21 (2020)

Hedges, I. L. V., Olkin: Statistical Methods for Meta-Analysis. Academic Press, San Diego (1985)

Heilbron, D. C.: Zero-altered and other regression models for count data with added zeros. Biom. J. **36**, 531–547 (1994)

Hu, M. C., Pavlicova, M., Nunes, E. V.: Zero-inflated and hurdle models of count data with extra zeros: examples from an HIV-risk reduction intervention trial. Am. J. Drug Alcohol Abuse. **37**(5), 367–375 (2011)

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

Min, Y., Agresti, A.: Random effect models for repeated measures of zero-inflated count data. Stat. Model. **5**, 1–19 (2005)

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

Neelon, B. H., Ghosh, P., Loebs, P. F.: A spatial Poisson hurdle model for exploring geographic variation in emergency department visits. J. R. Stat. Soc. Ser. A. **176**, 389–413 (2013)

Neelon, B. H., O'Malley, A. J., Normand, S. L.: A Bayesian model for repeated measures zero-inflated count data with application to outpatient psychiatric service use. Stat. Modell. **10**, 421–439 (2010)

Neelon, B., O'Malley, A. J., Smith, V. A.: Modeling zero-modified count and semicontinuous data in health services research Part 1: background and overview. Stat. Med. **35**(27), 5070–5093 (2016)

Payne, E. H., Hardin, J. W., Egede, L. E., Ramakrishnan, V., Selassie, A., Gebregziabher, M.: Approaches for dealing with various sources of overdispersion in modeling count data: Scale adjustment versus modeling. Stat. Methods Med. Res. **26**(4), 1802–1823 (2017)

Perumean-Chaney, S. E., Morgan, C., McDowall, D., Aban, I.: Zero-inflated and overdispersed: what's one to do? J. Stat. Comput. Simul. **83**(9), 1671–1683 (2013)

Rathbun, S., Fei, S. L.: A spatial zero-inflated Poisson regression model for oak regeneration. Environ. Ecol. Stat. **13**, 409–426 (2006)

Rose, C., Martin, S., Wannemuehler, K., Plikaytis, B.: On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data. J. Biopharm. Stat. **16**(4), 463–481 (2006)

Sharker, S., Balbuena, L., Marcoux, G., Feng, C. X.: Modeling socio-demographic and clinical factors influencing psychiatric inpatient service use: a comparison of models for zero-inflated and overdispersed count data. BMC Med. Res. Methodol. **20**(232), 1–10 (2020)

Tüzen, F., Erbaş, S., Olmuş, H.: A simulation study for count data models under varying degrees of outliers and zeros. Commun. Stat. Simul. Comput. **0**(0), 1–11 (2018)

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

Xu, L., Paterson, A. D., Turpin, W., Xu, W.: Assessment and selection of competing models for zero-inflated microbiome data. PLOS ONE. **10**, 1–30 (2015)

Yau, K. K., Lee, A. H.: Zero-inflated poisson regression with random effects to evaluate an occupational injury prevention programme. Stat. Med. **20**, 2907–2920 (2001)

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

YB, C.: Zero-inflated models for regression analysis of count data: a study of growth and development. Stat. Med. **21**, 1461–1469 (2002)

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.