Current Research in Biostatistics

Review

Review of Zero-Inflated Models with Missing Data

[1]T. Martin Lukusa, [2]Shen-Ming Lee and [3]Chin-Shang Li

[1]Institute of Statistical Science, Academia Sinica, Taiwan, R.O.C., Taiwan
[2]Department of Statistics, Feng Chia University, Taiwan, R.O.C., Taiwan
[3]Division of Biostatistics, Department of Public Health Sciences, University of California, Davis, USA

Corresponding Author:
Chin-Shang Li
Division of Biostatistics, Department of Public Health Sciences, University of California, Davis, USA
Email: csli2003@gmail.com

Abstract: The literature of count regression models covers a large scope of studies and applications that implemented simple and standard models for count response variables by using Poisson regression models, binomial regression models, negative binomial regression models, geometric regression models, or generalized Poisson regression models. These regression models have received considerable attention in various situations. Nevertheless in many fields, the distribution of the count response variable may display a feature of excess zeros for which the aforementioned regression models may fail to provide an adequate fit. To remedy this handicap, a class of distributions known as zero-inflated models is considered as the most appropriate approach for dealing properly with this issue of excess zeros. In addition to the zero-inflated problem, it happens quite often that the sample data sets under investigation are not completely observed. This refers to the missing data problem. In this study, our primary interest is in reviewing studies that considered simultaneously the missing data problem and the zero-inflated feature in modeling zero-inflated data. Moreover, we discuss their methodologies and results and some potential directions of the future research.

Keywords: Count Data, Estimating Methods, Missing at Random, Missing Completely at Random, Missing Data, Missing Not at Random, Zero-Inflated Models

Introduction to Zero-Inflated Data

A regression model fit is generally a statistical methodology that helps estimate the strength and direction of the relationship between two or more variables. It is considered as one of the most powerful and popular tools used for making important decisions or investigating some assertions in many statistical studies and across various domains of science. Similarly, a regression approach to count data is one of the most important statistical techniques, which plays a big role in decision making and investigation. This simple but powerful tool may become frustrating even misleading if less sufficient attention is paid to some important aspects of statistical modeling, such as the assumptions of models, the specific features or patterns displayed by the data set, or the presence of missing data. Many computer software programs have made the implementations of estimation of regression models, e.g., for count data, easier than before, but there is still a high chance to obtain a bad fit, especially when fewer attention is paid to the underlining assumptions of models or the complexity of the data. For instance, the presence of excess zeros in the response count variable requires some precautions prior to proceeding with a model fit.

In this review work two main issues are of great concern, including the presence of a Zero-Inflated (ZI) feature in response data and the presence of the missing response or covariate data. Note that count data cover a considerable portion of data in statistical inference and they arise from various fields to include the social sciences, medicine and industry among others. For instance, for the number of new friends added on a user's facebook account a week, the number of customers' mails a day that a business company receives regarding goods lost or damaged, or the number of doctor and hospital visits occurring throughout the weekday or weekend; regular Poisson regression models, negative binomial regression models or generalized Poisson regression models may be appropriate to fit this kind of data. Among these standard regression models, Poisson regression models are the most popular tool used to fit

count data (Cameron and Trivedi, 2013) because of their simplicity in application and interpretation of the results. Notwithstanding those advantages, a regular Poisson regression model cannot capture the ZI feature because it has only one parameter that is its mean. In the presence of the ZI feature, fitting a regular Poisson regression model has a tendency to overstate the significance level or underestimate the standard errors of the estimators of the model parameters. Consequently, inference based on the regular Poisson regression model fit is misleading and not credible in this situation. On the other hand, the negative binomial regression models (Cameron and Trivedi, 2013) and the generalized Poisson regression models (Consul and Famoye, 1992) are mostly seen as the backup solutions in case the regular Poisson regression model fit is not adequate. Contrary to the regular Poisson regression model, the negative binomial regression models and the generalized Poisson regression models and the generalized Poisson regression models have an extra parameter that can capture an additional effect, such as the ZI feature. Nevertheless, in many analyses of count data with excess zeros, these two regressions models may fail to adequately fit the data under study. In this case, ZI regression models or other mixture regression models (Mullahy, 1986) are better options (Allison, 2012). Ridout *et al*. (1998) and Ismail and Jemain (2007) provided a comprehensive introduction to the class of the ZI regression models. ZI models provide a wide and intensive area of research (Tu and Liu, 2016). Interestingly, the Scopus search engine developed by Elsevier reveals that in the last ten years. ZI models have been mentioned over 1,410 times as titles, abstracts or keywords among all articles. Compared to standard regression models, ZI models are considered to be more advanced methods and are required in order to account properly for the feature of excess zeros. For instance, there are Zero-Inflated Poisson (ZIP) models, Zero-Inflated Negative Binomial (ZINB) models, Zero-Inflated Binomial (ZIB) models and Zero-Inflated Generalized Poisson (ZIGP) models. Other models closely related to ZI models are hurdle models (Mullahy, 1986) and two-part models (Heilbron, 1994).

In general, a ZI model can be thought of as a mixture distribution of two components, including a count distribution, such as Poisson, binomial, negative binomial, or geometric and the degenerated distribution at zero. These ZI regression models differ from others in terms of the nature of the count distribution used for the probability mass function as given in expression (1). The ZI feature is generated by both sources (processes), including the count distribution component (random zeros) and the component of excess zeros (structural zeros). To the best of our knowledge, among the most used ZI models, the ZIP regression models proposed by Lambert (1992) are the most used in many applications. Besides that, the ZINB regression models (Ridout *et al*., 2001), ZIB regression models (Hall, 2000), Zero-Inflated Geometric (ZIG) regression models (Nagesh *et al*., 2015) and ZIGP regression models (Famoye and Singh, 2006) have been proposed in some situations to account for the feature of excess zeros, where a ZIP regression model could not fit the data well. Note that the ZIB regression models and the ZIG regression models have received very little attention compared to the most used ZI models. Besides using expression (1) as a generic form, the zero-inflated power series regression models as given in Gupta *et al*. (1995) can be seen as another form of presenting the count data with excess zeros (see Section 3.4). Up to this day, different orientations have been taken under the ZI models and many interesting results are found in the literature. But most of these works have left aside the potential question of missing data.

Besides the issue of excess zeros in count data, another important issue that has been addressed in the ZI data analysis literature is the missing data problem. ZI data are very active in many statistical studies or applications in practice. Therefore, the response count variable or some covariates involved in a regression model are likely to have missing data. There are many reasons behind the missing data appearance. Some missings are intentionally created for technical or confidential reasons, while others are due to happenstance. In these past decades, many researchers have proved that missing data were not avoidable in statistical studies; see, e.g., Little and Rubin (2002) and Schafer and Graham (2002). Consequently, the problem of missing response or covariate data attracts great attention. Any failures in addressing properly the presence of missing data while analyzing a ZI data set possibly yield inaccurate estimates. Little (1992) pointed out that the missing process and the missing pattern needed to be well understood in order to apply appropriate methods in response to missing data. Therefore, methods summarized in Table 1 are very useful in dealing with missing data. Due to the importance of these matters, we review only those works that simultaneously studied the ZI feature and the missing data problem. We introduce briefly the ZI model framework and some important concepts related to missing data in Section 2. Section 3 presents only the most popular ZI models and their related missing data treatments. A conclusion is given in Section

Zero-Inflated Models and Missing Data Concepts

Zero-Inflated Distributions

Prior to describing some popular ZI models and their applications, we first define a generic form for all ZI models. Let Y be a count response variable. The

probability mass function of a ZI distribution can then be expressed as follows:

$$P(Y=y)=\begin{cases} p+(1-p)f(y;\eta,d), & y=0, \\ (1-p)f(y;\eta,d), & y>0. \end{cases} \quad (1)$$

Here $p \in [0, 1]$ is a mixing weight for the accommodation of extra zeros. $f(y; \eta, d)$ represents a regular count distribution; for instance, Poisson distribution, binomial distribution, geometric distribution and negative binomial distribution. In general, $f(y;\eta, d)$ possesses two parameters η and d, where η and d represent its expected value and dispersion parameter, respectively. In practice p is linked to a set of covariates (χ_1) via a logit-linear predictor such that $p = H(u) = H(\beta^T\chi_1)$, where $H(u) = [1+\exp(-u)]^{-1}$, whereas η is linked to another set of covariates (χ_2) via a log-linear predictor $\eta = \exp(\gamma^T\chi_2)$ for unbounded count data. In many applications, the parameter d is neither modeled as a function of χ_1 nor χ_2. Naturally, χ_1 and χ_2 do not have to be identical. For instance, Lambert (1992) assumed that $\chi_1 \neq \chi_2$, whereas Lukusa *et al*. (2016) assumed that $\chi_1 = \chi_2 = \chi$, where $\chi = (1, X^T, Z^T)^T$ for X and Z being vectors of categorical or continuous covariates. A special case is when p is a constant not depending on covariates (Li, 2011). To have a comprehensive review, we define $p = H(\beta^T\chi_1)$ and $\eta = \exp(\gamma^T\chi_2)$. Other appropriate linear predictors can be used to model p; for instance, the probit-linear predictor given by $p = \Phi(\beta^T\chi_1)$ can be used instead of $p = H(\beta^T\chi_1)$, where Φ is the cumulative distribution function of the standard normal distribution.

One of the most interesting features about the ZI models is that they are related to each other based on the behaviors of parameters p, d and η in expression (1). For instance, when $d \to \infty$, the zero-inflated negative binomial distribution reduces to a zero-inflated Poisson distribution. When $d = 0$, the zero-inflated generalized Poisson distribution reduces to a zero-inflated Poisson distribution. But when $p = 0$, the zero-inflated negative binomial distribution, the zero-inflated generalized Poisson distribution and the zero-inflated Poisson distribution reduce to the negative binomial distribution, the generalized Poisson distribution and the Poisson distribution, respectively. Various relations can be established for the entire family of ZI distributions. The ZI regression models aim at estimating the unknown parameter vector $\theta = (\beta^T, \gamma^T, d)^T$ by means of different optimization techniques, such as Newton-Raphson method and expectation-maximization algorithm (Dempster *et al*., 1977).

Some Important Concepts of Missing Data

Missing data are described as various codes indicating lack of response (Schafer and Graham, 2002). Missing data are generally caused by technical problems or designs. But in some specific cases, e.g., privacy, missing data are deliberately created. The missing data should not be overlooked without a specific reason. Before applying any appropriate methods to deal with missing data, as listed in the taxonomy (Little, 1992), a data set needs to be described by means of descriptive statistics in order to obtain the information related to the missing data. If it is revealed that there are missing data, then the first important step should be to understand the missing patterns and the missing mechanisms. Let n be the sample size, Y the non-negative count outcome variable and X and Z covariate vectors, where Z is always observed. Assume that X is partially observed and W is a surrogate variable able to provide enough information about the missing variable. To account for missingness, an indicator variable, $\delta = 1$ if X is observed and $\delta = 0$ otherwise, is included. Similarly, the idea of X having missing can be extended to a situation where the response variable Y is incomplete. For the sake of illustration, when X is missing at random, the basic data structure is as follows:

$$\begin{bmatrix} X_1 \\ Y_1 \\ Z_1 \\ W_1 \end{bmatrix}, \begin{bmatrix} X_2 \\ Y_2 \\ Z_2 \\ W_2 \end{bmatrix}, \cdots, \begin{bmatrix} X_{n_v} \\ Y_{n_v} \\ Z_{n_v} \\ W_{n_v} \end{bmatrix}, \begin{bmatrix} Y_{n_v+1} \\ Z_{n_v+1} \\ W_{n_v+1} \end{bmatrix}, \cdots, \begin{bmatrix} Y_{n_v+m} \\ Z_{n_v+m} \\ W_{n_v+m} \end{bmatrix}, \cdots, \begin{bmatrix} Y_n \\ Z_n \\ W_n \end{bmatrix},$$

where n_v denotes the number of validation data.

The data set structure is often arranged in arrays, which is allowed to visualize clearly the different patterns of missing values. There are three main missing patterns (Rubin, 1976), including (i) the univariate pattern where missing data occur only on a single item (single variable) or group of variables of the same nature, while others are completely observed, (ii) the monotone pattern where missing values on items can be arranged in an increasing proportion from items with least missing values to items with more missing values and (iii) the general pattern where missing values scattered everywhere. Compared with the general pattern, the univariate and monotone patterns are not hard to handle in practice.

Let $V = (Z,W)$ and the data set $D = D_0 \cup D_1$, where $D_0 = \{(Y_i, V_i): \delta_i = 0, i = 1, 2,..., n\}$ and $D_1 = \{(Y_i, X_i, V_i): \delta_i = 1, i = 1, 2,..., n\}$. The missing mechanism plays an important role in dealing with missing data problems. Rubin (1976) distinguished among three processes of missing mechanisms, including Missing Completely At Random (MCAR), Missing At Random (MAR) and Missing Not At Random (MNAR). Under the MCAR, the selection probability is expressed as

$P(\delta = 1|Y,X, V) = \pi$. When missing is MAR, the selection probability is expressed as $P(\delta = 1|Y,X, V) = \pi(Y, V)$. Note that the MCAR and the MAR mechanisms are ignorable missing mechanisms. Under the MNAR, the selection probability is given by $P(\delta = 1|Y,X, V) = \pi(Y,X, V)$. The MNAR mechanism is a nonignorable missing mechanism. Note that in survey studies, clinical studies or other statistical techniques for data collection, it is difficult to distinguish between the MAR and MNAR, even if the MAR is from the MCAR unless additional information is available. Therefore, it is important to understand clearly the whole process and circumstances during the data collection stage. Due to the importance of the missing mechanisms, the estimation of the selection probability has been of great concern. For instance, Rosenbaum and Rubin (1983) and Robins *et al*. (1994) proposed a parametric estimation method, whereas Wang *et al*. (1997) and Wang and Wang (2001) proposed a nonparametric estimation method. Many techniques can be applied to estimate the selection probability provided that the condition that the estimate of selection probability $\pi \in [0, 1]$ holds.

Taxonomy of Techniques for Handling Missing

Although some methods for dealing with missing data are seen as more powerful than others, they all have some limitations. This happens when the model assumptions in the presence of missing data are not well understood or when the proportion of missing increases considerably. Based on Little (1992), Pigot (2001) and Ibrahim *et al*. (2012), the most popular methods for handling the missing data are summarized in Table 1. Note that contents in Table 1 are more technical and general than specific.

The introduction of ZI models and the missing data problem help understand various orientations authors have taken regarding the zero excess and the missing data issues. Table 1 will serve as a guidance of methods potential to be applied under the ZI regression models. Next, we review the most popular ZI regression models and the missing data problem.

Popular Zero-Inflated Regression Models

Zero-Inflated Negative Binomial Models

A ZINB distribution can be seen as a mixture of two distributions, including a Negative Binomial (NB) distribution and a degenerated distribution at zero (Ridout *et al*., 1998; 2001). Therefore, the ZINB distribution can be derived from expression (1) such that the function f(y; η, d) is a NB distribution, expressed as follows:

$$f(y;\mu,d) = \frac{\Gamma(y+d)}{\Gamma(y+1)\Gamma(d)}\left(\frac{\mu}{\mu+d}\right)^{y}\left(\frac{d}{\mu+d}\right)^{d}, \tag{2}$$

where $\eta = \mu$ and p and d are identical to those in expression (1). Then, the probability mass function of the ZINB distribution is expressed as follows:

$$P(Y=y) = \begin{cases} p+(1-p)\left(\frac{d}{\mu+d}\right)^{d}, & y=0, \\ (1-p)\frac{\Gamma(y+d)}{\Gamma(y+1)\Gamma(d)}\left(\frac{\mu}{\mu+d}\right)^{y}\left(\frac{d}{\mu+d}\right)^{d}, & y>0, \end{cases} \tag{3}$$

where $\Gamma(\cdot)$ is the gamma function. Note that when $p = 0$, the ZINB distribution reduces to a NB distribution and when $p = 0$ and $d \to \infty$, the ZINB distribution reduces to a regular Poisson distribution.

Table 1. Taxonomy of popular methods for missing data

Approach	MCAR	MAR	MNAR	References
Case deletion methods	Consistent	Not consistent	Not consistent	Little (1992); Graham (2012)
Stochastic regression imputation methods	Consistent but inefficient	Consistent but inefficient	Not consistent and inefficient	Graham (2012); Enders (2010)
Multiple imputation methods	Consistent	Consistent under mild conditions	Consistent under strong conditions	Rubin (1987); Graham (2012)
Maximum Likelihood with EM algorithm	Consistent	Consistent under mild conditions	Consistent under correct missing model specifications	Horton and Liard (1999); Ibrahim *et al*. (2005)
Bayesian method	Consistent	Consistent	Consistent under some conditions	Mason *et al*. (2012)
Weighted methods	Consistent	Mostly consistent under some conditions	Mostly consistent under some conditions	Zhao and Lipsitz (1992); Robins *et al*. (1994)

Consistent, consistent estimates; inefficient, inefficient estimates
Case deletion methods: Complete case method and available case method
Single imputations: Regression imputation, mean or median imputation, etc.
Weighted methods: Inverse probability weighted method and augmented inverse probability weighting method

Let $\{(y_i, X_i, Z_i): i = 1,..., n\}$ be the data set and $\theta = (\beta^T, \gamma^T, d)^T$. The likelihood function of the ZINB distribution can then be expressed as follows:

$$L(\beta,\gamma,d)=\prod_{i=1}^{n}\left[p_i+(1-p_i)\left(\frac{d}{\mu_i+d}\right)^d\right]^{I(y_i=0)} \prod_{i=1}^{n}\left\{(1-p_i)\frac{\Gamma(y_i+d)}{\Gamma(y_i+1)\Gamma(d)}\left(\frac{\mu_i}{\mu_i+d}\right)^{y_i}\left(\frac{d}{\mu_i+d}\right)^d\right\}^{I(y_i>0)}, \quad (4)$$

where $I(\cdot)$ is an indicator function, $p_i = H(\beta^T\chi_i)$, $\mu_i = \exp(\gamma^T\chi_i)$ and $\chi_i = \left(1, X_i^T, Z_i^T\right)^T = 1,..., n$. The log-likelihood function of the ZINB distribution is then $\ell(\beta, \gamma, d) = \log L(\beta, \gamma, d)$. The likelihood-based method can be used to obtain estimates of β, γ and d via the Expectation Maximization (EM) algorithm (Dempster *et al*., 1977). The ZINB distribution has been used in some interesting studies; for instance, Preisser *et al*. (2012) provided a review of some ZI models for data of dental caries indices in epidemiology. There are other interesting works; nevertheless, they do not tackle the missing data issue. We now turn our attention to the missing data problem under the ZINB regression model framework. To the best of our knowledge, neither the missing response nor the missing covariates has been fully explored under the ZINB regression model framework except for Chen and Fu (2011) who conducted a model selection where the ZINB model was a model candidate and Samani *et al*. (2012) considered the ZINB model as a candidate model for model selection under the zero-inflated power series (ZIPS) model framework.

Zero-Inflated Generalized Poisson Models

Extended from the Generalized Poisson (GP) regression model developed by Consul and Famoye (1992), the ZIGP regression model (Famoye and Singh, 2006) is a competitor of the ZINB regression model. It has the flexibility to handle any inflation or deflation in count data. Alike to the ZINB distribution, the ZIGP distribution is a mixture of two distributions, including the GP distribution that can be represented as $f(y, \eta, d)$ in expression (1), where $\eta = \mu$ and the degenerated distribution at zero. In the past decades, the ZIGP regression models received significant interest and attention due to its flexibility to handle some unusual features of count data. Consequently, different variants of the ZIGP regression models have been developed and applied. The probability mass function of the ZIGP distribution is expressed as follows:

$$P(Y=y)=\begin{cases}p+(1-p)f(y;\mu,d), & y=0,\\ (1-p)f(y;\mu,d), & y>0,\end{cases} \quad (5)$$

where

$$f(y;\mu,d)=\left(\frac{\mu}{1+d\mu}\right)^y\frac{(1+dy)^{y-1}}{y!}\exp\left[-\frac{\mu(1+dy)}{1+d\mu}\right]. \quad (6)$$

When $d = 0$, the ZIGP distribution reduces to the ZIP distribution. When $p = 0$, the ZIGP distribution reduces to the GP distribution. For more details about the likelihood function for the ZIGP model and its optimization, see, e.g., Famoye and Singh (2006) and Ismail and Zamani (2013). Among the ZI models, the ZIGP model is becoming more attractive to researchers and its scope takes various directions. For example, Gupta *et al*. (1995) developed the zero-inflated modified power series distributions that cover many distributions, e.g., the ZIGP regression model. Famoye and Singh (2006) applied the ZIGP regression model under the frequentist context to fit domestic violence data with excess zeros. They found that it converged in all situations when fitting the ZIGP regression model, whereas it converged only in some cases when fitting the ZINB regression model. That supported the view that for this kind of data, the ZINB and ZIP models could not provide an adequate fit. Angers and Biswas (2003) investigated the fit of ZIGP regression model under a Bayesian framework where they discussed the use of noninformative priors to obtain the posteriors and to compare the performance of the ZIGP model with that of the Poisson and ZIP models used for the fetal movement data. Regarding the missing data problem, researchers have not yet shown much interest in studying the missing data problem under the ZIGP model framework, despite the indication that there is a growing tendency of work related to ZIGP models. Thus, there are no ZIGP models with missing data work to study, unlike the ZIP model with missing data or the ZINB model with missing data that are both illustrated under the ZIPS model (Samani *et al*., 2012).

Zero-Inflated Poisson Models

Among the ZI models, the ZIP regression model (Lambert, 1992) is the most popular. The ZIP distribution can be thought of as a population that includes two latent groups of subjects: The non-susceptible group consisting of those who are not at risk of an event of interest and the susceptible group consisting of those who are at risk of the event and may have experienced the event several times during a specific time period (Dietz and Böhning, 1997). From expression (1), the ZIP distribution is a mixture distribution that includes the Poisson distribution denoted as $f(y; \lambda)$, where $\lambda = \eta$ and the degenerated function at zero (Singh, 1963; Johnson *et al*., 2005). Alternatively, if $d \to \infty$, then $f(y; \lambda, d) \to f(y; \lambda)$, where $f(y; \lambda) = e^{-\lambda}\lambda^y/y!$ and $f(y; \lambda, d)$ is a NB

distribution function as given in expression (2). The probability mass function of the ZIP distribution is then expressed as follows:

$$P(Y=y)=\begin{cases}p+(1-p)e^{-\lambda}, & y=0,\\ (1-p)\dfrac{e^{-\lambda}\lambda^{y}}{y!}, & y>0\end{cases}$$
$$=pI(y=0)+(1-p)\frac{e^{-\lambda}\lambda^{y}}{y!}, y=0,1,2,... \quad (7)$$

where λ is the Poisson mean. The ZIP distribution reduces to a regular Poisson distribution when $p = 0$. The likelihood of the ZIP model can be expressed as follows:

$$L(\theta)=\prod_{i=1}^{n}L_i(\theta)$$
$$=\prod_{i=1}^{n}\left\{p_i+(1-p_i)e^{-\lambda_i}\right\}^{I(y_i=0)}\left\{(1-p_i)\frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}\right\}^{I(y_i>0)}. \quad (8)$$

Here $p_i = H(\beta^T\chi_i)$ and $\lambda_i = \exp(\gamma^T\chi_i)$, $i = 1,..., n$, so that the vector of parameters of interest is $\theta = (\beta^T, \gamma^T)^T$. The estimate $\hat{\theta}$ can be obtained by maximizing the log-likelihood $\ell(\theta) = \log L(\theta) = \sum_{i=1}^{n}\log L_i(\theta) = \sum_{i=1}^{n}\ell_i(\theta)$ via the EM algorithm as done by Lambert (1992). The ZIP regression models have been further disseminated and used successfully by some authors, e.g., Böhning *et al*. (1999), Yau and Lee (2001), Cheung (2002), Lu *et al*. (2004). Hall and Shen (2010) proposed a robust expectation-solution estimation method for ZIP regression models to overcome the case where the Maximum Likelihood Estimator (MLE) is highly sensitive to the presence of outliers. In addition, Li (2011) proposed a semiparametric ZIP regression model that can be used to assess the lack of fit of a postulated parametric ZIP model. Jansakul and Hinde (2002) proposed a score test for a ZIP model against a Poisson model. Li (2012) proposed a score test for a semiparametric ZIP regression model versus a semiparametric Poisson regression model. Similar to the ZINB and ZIGP models, researchers have not yet shown enough interest in exploring the missing data problem under the missing data framework. To the best of our knowledge, little has been done so far. At this level, we present the work of Lukusa *et al*. (2016).

By assuming that the missing covariates were MAR (Rubin, 1976), Lukusa *et al*. (2016) proposed a semiparametric Inverse Probability Weighting (IPW) estimator of a ZIP regression model in the spirit of Zhao and Lipsitz (1992) and Flander and Greenland (1991). The proposed estimating method was a Horvitz and Thompson (1952)-type weighted estimating method where the selection probability was $\pi(Y, V) = P(\delta = 1|Y,X, V)$. Following Wang *et al*. (1997) and Reilly and Pepe (1995), Lukusa *et al*. (2016) expressed the nonparametric selection probability estimator of $\pi(y, v)$ as $\hat{\pi}(y,v)=\frac{\sum_{k=1}^{n}\delta_k I(Y_k=y,V_k=v)}{\sum_{i=1}^{n}I(Y_i=y,V_i=v)}$, where $y = 0, 1, 2,...$ and $v\in\{v_1, v_2,..., v_m\}$ for $v_1, v_2,..., v_m$ being the distinct values of the V_is. In order to improve the precision of $\hat{\pi}(y,v)$, an auxiliary variable is included. They proposed the semiparametric IPW estimating function expressed as follows:

$$U(\theta,\hat{\pi})=\sum_{i=1}^{n}\frac{\delta_i}{\hat{\pi}(Y_i,V_i)}S_i(\theta), \quad (9)$$

where $S_i(\theta) = \partial\ell_i(\theta)/\partial\theta$ and $\hat{\pi}=(\hat{\pi}_1,...,\hat{\pi}_n)$ for $\hat{\pi}_i=\hat{\pi}(Y_i,V_i)$, $i = 1,..., n$. By solving $U(\theta,\hat{\pi})=0$, they obtained $\hat{\theta}$, an estimator of θ. Here $\hat{\pi}(y,v)$ plays a crucial role in obtaining the estimate of θ because observed data are inversely weighted by $\hat{\pi}(y,v)$. Moreover, they studied the limiting behavior of $\hat{\theta}$ and showed that $\hat{\theta}\xrightarrow{p}\theta$ and $\sqrt{n}(\hat{\theta}-\theta)\xrightarrow{d}N(0,\Delta_{ws})$ as $n\to\infty$, where $\Delta_{ws} = G_F^{-1}(\theta)\left[J(\theta,\pi)-\left(J^*(\theta,\pi)-C^*(\theta,\pi)\right)\right]$ $G_F^{-1}(\theta)$, $G_F(\theta)=E\left(-\frac{\partial S_1(\theta)}{\partial\theta^T}\right)$, $J(\theta,\pi)=E\left(-\frac{S_1(\theta)S_1^T(\theta)}{\pi(Y_1,V_1)}\right)$ $, J^*(\theta,\pi)=E\left(-\frac{S_1^*(\theta)S_1^{*T}(\theta)}{\pi(Y_1,V_1)}\right)$, $C^*(\theta,\pi)=E\left(S_1^*(\theta)S_1^{*T}(\theta)\right)$ and $S_1^*(\theta)=E(S_1(\theta)|Y_1, V_1)$.

A simulation study was conducted to compare the semiparametric IPW estimator, true weight IPW estimator, the CC estimator and the MLE that was considered as the benchmark. Comparisons were made based on the average bias, standard deviation, standard error and the 95% coverage probability. Overall, the semiparametric IPW estimator was found to be asymptotically unbiased and more efficient than the CC estimator that was seriously biased and the true weight estimator $\pi(Y, V)$ that had a bigger standard error. It means that even if $\pi(Y, V)$ is known, which is not always the case, it should be substituted by $\hat{\pi}(Y,V)$ in the estimating function using the true weight. Moreover, they illustrated the practical use of the proposed methodology with a data set from a survey study conducted in Taiwan in 2007 that consists of 7,386 respondents. The response count variable was the number of speed regulations that a motorcycle rider violated in a year (about 90% of motorcycle riders not

violating speed regulations). Only the covariate related to the distance covered in kilometers had 15% of data missing while data of other covariates were fully observed. The analysis results overall showed that the performance of the semiparametric IPW estimator was very close to that of the CC estimator.

Besides the work of Lukusa *et al*. (2016), Pahel *et al*. (2011) used expression (1) to predict the missing dental caries data, particularly under the ZI regression model framework. Similar to Lambert (1992) and Lukusa *et al*. (2016), they used the ZIP regression model where process 1 and process 2 were generated by $p = H(\beta^T \chi_1)$ and $\lambda = \exp(\gamma^T \chi_2)$, respectively, for χ_1 and χ_2 as sets of covariates. Similar to Lukusa *et al*. (2016), Pahel *et al*. (2011) assumed $\chi_1 = \chi_2 = \chi$. In order to impute the missing dental caries data, they considered the complete case and some additional information. In Step 1, estimate ZIP model with non-missing caries data. In Step 2, generate predictions for levels of caries based on estimated coefficients. In addition, if the predicted probability in process 1 is less than say t, a uniform (0, 1) random variate, then the missing is filled in with 0 (meaning no dental caries); otherwise process 2 is used to fill in the missing values. The final result is a summary of all imputations based on the formula of Rubin (1987). To illustrate the performance of the proposed multiple imputation techniques for missing dental caries data, they used a real example where they compared the three model imputations: The Poisson model, the ZIP model and the ZIP model with random effects, respectively. Under the missing not MCAR for the response variable, they imputed the three models and computed their Akaike Information Criterion (AIC) value (Akaike, 1974). Although Pahel *et al*. (2011) and Böhning *et al*. (1999) studied the ZI data for the dental caries, the main difference is that Pahel *et al*. (2011) considered the missing data problem, whereas Böhning *et al*. (1999) focused on the problem of missing teeth.

Zero-Inflated Power Series Distributions

The ZIPS model is a two-component mixture model that consists of a Power Series (PS) distribution, such as Poisson, binomial, negative binomial and geometric distributions and a degenerated distribution at zero. The probability mass function of the PS distribution is expressed as follows:

$$f(y;\lambda) = \frac{b(y)\lambda^y}{h(\lambda)}, \tag{10}$$

where $h(\lambda) = \sum_{y=0}^{\infty} b(y)\lambda^y$, $y = 1, 2,..., b(y) > 0$ and λ is the PS model parameter to be estimated. The ZIPS distribution is given by:

$$P(Y = y) = \begin{cases} p + (1-p) f(y;\lambda), & y = 0, \\ (1-p) f(y;\lambda), & y > 0. \end{cases} \tag{11}$$

More interestingly, the ZIPS models include most of the ZI models except for the ZIGP models. Let y_i be a realization of a random variable Y_i that has a ZIPS distribution. Let χ_{1i} and χ_{2i} be covariate sets for the ith subject and define $\theta = (\beta^T, \gamma^T)^T$ a vector of parameters to be estimated. The likelihood function of the ZIPS model is then expressed as follows:

$$L(\theta) = \prod_{i=1}^{n} \left\{ p_i + (1-p_i)\frac{b(y_i)}{h(\lambda_i)} \right\}^{I(y_i=0)} \left\{ (1-p_i)\frac{b(y_i)\lambda_i^{y_i}}{h(\lambda_i)} \right\}^{I(y_i>0)}, \tag{12}$$

where $p_i = H(\beta^T \chi_{1i})$ and $\lambda_i = \exp(\gamma^T \chi_{2i})$, $i = 1,..., n$. The MLE of θ is obtained by optimizing the log-likelihood function $\ell(\theta) = \log L(\theta)$. The ZIPS model framework seems to become more attractive for many researchers. For example, Bhattacharya *et al*. (2008) provided a general Bayesian setup to test for the ZI feature in a ZIPS distribution. Samani *et al*. (2012) used a likelihood-based approach. Under the ZIPS regression model framework, Samani *et al*. (2012) proposed the mixed Stochastic EM (SEM) and EM algorithms (M-SEM-EM algorithm) for parameter estimation in the likelihood-based approach. Unfortunately, so far, that was the unique work that addressed the missing data problem. We briefly present their approach. To capture the ZI feature, Samani *et al*. (2012) extended the idea of Samani (2011) known as the missing inflated power series distribution model.

Assuming the response Y to be MNAR (Rubin, 1976), Samani *et al*. (2012) expressed the joint incomplete data model as follows:

$$\begin{cases} Y_i \sim ZIPS(p_i, \lambda_i) \\ \text{logit}\left[P(\delta_i = 1 \mid Y_i) \right] = \alpha_1^T \chi_{3i} + \alpha_2 Y_{i1}^* + \alpha_3 Y_{i2}^*, \; i = 1,..., n. \end{cases} \tag{13}$$

Here $\text{logit}(p_i) = \beta^T \chi_{1i}$. $\log(\lambda_i) = \gamma^T \chi_{2i}$ for $\chi_{1i} \neq \chi_{2i}$. δ_i is a binary missing indicator variable defined as $\delta_i = 1$ if y_i is observed and $\delta_i = 0$ otherwise. χ_{3i} is another covariate vector. Y_{i1}^* and Y_{i2}^* are defined, respectively, as $Y_{i1}^* = 1$ if $Y_i = 0$ and $Y_{i1}^* = Y_i$ otherwise, whereas $Y_{i2}^* = Y_i$ if Y_i is from the PS family and $Y_{i2}^* = 0$ otherwise. In expression (13), α_1, α_2 and α_3 are parameters of the MNAR selection probability model. The log-likelihood for the ZIPS joint model under the response MNAR is then given as follows:

$$\ell(\theta, y) = \sum_{i=1}^{n} \left\{ \delta_i \log P\left(\delta_i = 1 \mid y_i, \alpha_1, \alpha_2, \alpha_3\right) + \left(1-\delta_i\right) \log P\left(\delta_i = 1 \mid y_i, \alpha_1, \alpha_2, \alpha_3\right) + \log P\left(Y_i = y_i \mid \beta, \gamma\right) \right\},$$

where $\theta = (\beta^T, \gamma^T, \alpha^T)^T$ and $\alpha = \left(\alpha_1^T, \alpha_2, \alpha_3\right)^T$.

To estimate θ, they maximized $\ell(\theta, y)$ by a variant of the EM algorithm, known as a MSEM-EM algorithm. Furthermore, they computed the AIC value (Akaike, 1974) in order to compare different regression models under the ZIPS model framework. Overall, their simulation study results showed that the larger the sample, the better the estimate of θ. They applied the proposed methodology to the data set from the British Household Panel Survey regarding the number of visits to a hospital during the year by using the AIC to compare the Poisson, NB, ZIP and ZINB regression models. They found that the ZIP regression model was the best model because it had the smallest AIC value. They proceeded to fit the ZIP regression model and the results confirmed that it was MNAR because the number of visits showed a significant Poisson status on the probability of the nonresponse; Samani *et al*. (2012) for more details. Besides Samani *et al*. (2012), Chen and Fu (2011) proposed a parametric model selection when some covariates were MAR (Rubin, 1976). Considering the ZIPS distribution, they proposed a new model selection criterion in place of the classical AIC (Akaike, 1974) in order to account for the missing covariates that were assumed to be MAR. In effect, in the presence of missing data, the classical AIC misleads the conclusion of model selection. Interestingly, the proposed method can be implemented in the presence of missing data or without missing data. Chen and Fu (2011) showed that their developed method is a modified version of Monte Carlo EM (MCEM) algorithm that is based on the data augmentation scheme. We briefly present their idea. Let $x = (x_{obs}, x_{mis})$ be the vector of covariates partially observed. In order to reduce the number of nuisance parameters that need to be estimated via the MCEM algorithm and allow a more convenient model specification for the distribution of covariates, following Ibrahim *et al*. (2005), Chen and Fu (2011) modeled the vector of covariates $x_i = (x_{obs,i}, x_{mis,i})$ by developing the following probability model:

$$\begin{aligned} &P\left(x_{i1}, \ldots, x_{iq} \mid \alpha\right) \\ &= P\left(x_{iq} \mid x_{i1}, \ldots, x_{iq-1}, \alpha_q\right) \cdots P\left(x_{i2} \mid x_{i1}, \alpha_2\right) P\left(x_{i1} \mid \alpha_1\right), \end{aligned} \tag{14}$$

where α_j is a vector of indexing parameters for the jth conditional distribution, $\alpha = \left(\alpha_1^T, \ldots, \alpha_q^T\right)^T$ and the α_j's are distinct. They defined the missing value indicator vector $r_i = (r_{i1}, \ldots, r_{iq})$ of the covariate vector x_i as $r_{ij} = 0$ if x_{ij} is observed and $r_{ij} = 1$ if x_{ij} is missing. Under the assumption and excluding the missing data indicator from the model, the complete data probability function of subject i from the ZIPS regression model is given by $P(y_i, x_i, r_i|\theta) \propto P(y_i, x_i|\theta) = P(y_i|x_i,\beta,\gamma)P(x_i|\alpha)$ that leads to the complete data log-likelihood:

$$\ell_c(\theta) = \sum_{i=1}^{n} \log\left[P\left(y_i \mid x_i, \beta, \gamma\right)\right] + \log\left[P\left(x_i \mid \alpha\right)\right], \tag{15}$$

where $\theta = (\beta^T, \gamma^T, \alpha^T)^T$. In order to obtain the estimate of θ, Chen and Fu (2011) used the data augmentation techniques (Ghosh *et al*., 2006) and a modified version of the MCEM algorithm to maximize the log-likelihood $\ell_{ad}(\theta)$ that was obtained by including the latent variable into $\ell_c(\theta)$. Following Claeskens and Consentino (2008) who derived a version of AIC (Akaike, 1974) that is suitable for the situation of missing covariates, they proposed the new criterion for ZIPS regression models with missing covariates. See Chen and Fu (2011) for more details regarding their methodology. They conducted a simulation study to illustrate the application of the proposed method in selecting the best model among the four candidate models: Poisson, NB, ZIP and ZINB regression models. They illustrated the practical use of the proposed methodologies by using a real data set from the Female Consumer Lifestyle Study in which the whole data set was collected in six cities of China in 2003 on broad topics, such as lifestyle, the frequency of buying goods for slim and the average amount of purchases.

Hurdle Models

Closely related to the ZI models, the hurdle models were developed by Mullahy (1986) and were popularized by Cameron and Trivedi (2013) in order to deal with count data sets having more zero counts than allowed for by the Poisson and NB models. The difference between the hurdle and ZI count models is that the later can separately model the zero and non-zero counts. The hurdle models are two-component models: A hurdle component for zeros versus non-zeros and a truncated count component for positive counts. For the hurdle component, either a binomial model or a censored count distribution, such as a censored Poisson, geometric, or NB distribution, can be used to model zeros versus non-zeros. For a truncated count component, a Poisson, geometric or NB model can be used for positive counts. More specifically, the hurdle model combines a zero hurdle model $P_{\text{zero}}(Y = y)$ (right-censored at 1) and a count data model $P_{\text{count}}(Y = y)$ (left-truncated at 1), expressed as follows:

$$P_{hurdle}\left(Y = y\right) = \begin{cases} P_{zero}\left(Y = y\right), & y = 0, \\ \left[1 - P_{zero}\left(Y = 0\right)\right] \dfrac{P_{count}\left(Y = y\right)}{1 - P_{count}\left(Y = 0\right)}, & y > 0. \end{cases} \tag{16}$$

For example, Mullahy (1986) used a Poisson distribution to model the zeros versus non-zeros and a zero-truncated Poisson distribution to model the positive counts. The hurdle models have been intensively applied in many studies. In the last decade, the hurdle models have been mentioned for about 2,000 times as titles or keywords; see the Scopus engine from Elsevier. Nevertheless, none of them addressed the missing data problem.

Multivariate Zero-Inflated Models

Besides the univariate ZI count model frame, it is possible to have several outcomes measured on each individual. For instance, Li *et al*. (1999) studied multivariate zero-inflated Poisson models in order to model outcomes of manufacturing processes producing numerous defect-free products while Wang (2003) studied the bivariate zero-inflated negative binomial models for bivariate count data with excess zeros and applied the proposed model to analyze the data of health-care utilization with a sample of 5,190 single-person households from the 1977-1978 Australian Health Survey.

Yang *et al*. (2016) proposed a flexible MCEM algorithm for estimation of the Bivariate Zero-Inflated Poisson (BZIP) regression model when the response count is MAR. They applied the proposed methodology to a bivariate data set regarding the demand for health care in Australia. More details can be found in Yang *et al*. (2016).

Discussion

Table 2 (column 2) provides the number of appearances in terms of titles, keywords and abstracts for the most used ZI models. For instance, ZINB is mentioned about 620 times overall and 39 times as a title in the literature according to the Scopus engine from Elsevier. Similarly, the Scopus search engine reveals that multivariate ZI models are mentioned about 40 times only and there are about 12 titles only pointing out bivariate ZI data. Moreover, Table 2 provides the existing references for ZIP models with missing data, the purpose of study and the methodology used to deal with missing data.

Based on the missing mechanisms, Pahel *et al*. (2011) considered the missing in the count response variable as MAR. Samani *et al*. (2012) assumed the missing in the count response variable was MNAR. Chen and Fu (2011) and Lukusa *et al*. (2016) considered the missing in covariates as MAR. Yang *et al*. (2016) proposed a joint MAR mechanism for the bivariate count response variable. Regarding the purpose and methodology, Chen and Fu (2011) and Samani *et al*. (2012) considered the ZIPS regression model framework and implemented different variants of the EM algorithm to estimate the model parameters and to compute the AIC values for model selection in the presence of missing data. Lukusa *et al*. (2016) developed the semiparametric inverse probability weighting method for estimation of the ZIP regression model that used a nonparametric selection probability and showed that the proposed estimator had good asymptotic properties. In addition, they showed that their estimator was more efficient than the estimator that uses the true weight and the CC estimator that was seriously biased. Pahel *et al*. (2011) developed a multiple imputation method for missing dental caries data under the ZIP regression model.

Table 2. Summary of popular zero-inflated models

Popular model	Number of appearances	Reference with missing	Missing mechanism	Proposed method
ZINB	620 (39)	not yet addressed	−−	−−
ZIGP	132 (9)	not yet addressed	−−	−−
ZIP	731 (110)	Lukusa *et al*. (2016)	Covariate MAR	Semiparametric inverse probability weighting
		Pahel *et al*. (2011)	Response not MCAR	Multiple imputation of missing dental caries
ZIPS	16 (9)	Chen and Fu (2011)	Covariate MAR	Modified MCEM under likelihood-based approach
		Samani *et al*. (2012)	Response MAR	M-SEM-EM under likelihood based approach
Hurdle	2108 (79)	not yet addressed	−−	−−
BZIP	40 (12)	Yang *et al*. (2016)	Response MAR	MCEM under likelihood-based approach

ZINB, zero-inflated negative binomial; ZIGP, zero-inflated generalized Poisson; ZIP, zero-inflated Poisson; ZIPS, zero-inflated Poisson series; BZIP, bivariate zero-inflated Poisson

Column 2, the number represents the frequency of citations as title, keywords or abstracts. The number in bracket represents the frequency of the article titles related to the corresponding zero inflated regression model.

−− means there is not yet an article of the corresponding zero-inflated model with missing data.

MCEM, Monte Carlo expectation-maximization; M-SEM-EM, mixed stochastic expectation maximization and expectation-maximization.

Based on the non-missing caries data, they imputed missing caries data by using the Poisson model, ZIP model and ZIP model with random effects. Indeed, the ZIP model with random effects yielded the best result because this model accounted for the cluster effects. Yang *et al*. (2016) applied a straightforward likelihood approach to estimate parameters of a BZIP model when the bivariate count variable is missing at random. The EM algorithm and MCEM algorithm are the most used estimation methods in the literature of ZI regressions models with missing data. Moreover, in simulation studies from Chen and Fu (2011), the proportion of zeros generated is not clearly presented. Therefore, the ZI feature may not be clearly perceived. Chen and Fu (2011) and Samani *et al*. (2012) also generated the missing data where the missing rate was less than 25%. Regarding the work of Lukusa *et al*. (2016), their simulation study used only moderate and large samples. It can be interesting to see how the proposed method performed under small or moderate samples. Pahel *et al*. (2011) assumed the missing mechanism to be not MCAR. Nevertheless, it is not clear whether it was a MAR or MNAR mechanism. Yang *et al*. (2016), who fit a BZIP regression model, pointed out that in some cases, the estimator based on the CC method was closer to the MLE that was obtained by using the MCEM method. However, often under the MAR the CC estimate is expected to be biased. Further investigations are needed. In general, most of the methods developed in the literature of ZI models with missing data agree with the summary of the most used methods given in Table 1.

Conclusion

The missing data problem has been intensively studied from various angles in the regression model literature. Some studies investigated the missing data under specific distribution models or vice versa; particularly, we have reviewed the literature of ZI models with missing data. It is crystal clear that fewer works related to the ZI models dealt with the missing data problem and the ZI feature simultaneously. Chen and Fu (2011), Pahel *et al*. (2011), Samani *et al*. (2012) Lukusa *et al*. (2016) and Yang *et al*. (2016) seem to be the only appealing works; see Table 2. Surprisingly, the ZINB, ZIGP and hurdle regressions models are among the most used models for ZI count data. However, these three regression models with missing data, exclusively, have not yet been investigated. On the other hand, the ZIP, ZIPS and BZIP regression models have less than three works each on the missing data problems. Table 2 gives the whole picture of the ZI data literature in terms of the regression model appearance, the missing data mechanisms considered, the references and the methodology used to handle the missing data. We wish to inspire researchers to discover the research regarding ZI models with missing data. There are many extensions or future studies to be carried on. For instance, Chen and Fu (2011) and Samani *et al*. (2012) could include the asymptotic behavior of the proposed AIC. Lukusa *et al*. (2016) could assume a MNAR mechanism. Yang *et al*. (2016) might also consider the case where covariates in the BZIP regression model are MNAR. Finally, ZI data with missing values still have plenty of orientations yet to be investigated. ZI data are important in many studies and sectors of life. A relationship between Table 1 and 2 shows many potential studies that could be done. Thus with the information from Table 1 and 2, researchers are invited to come out with some comprehensive and intensive studies of ZI data with missing values.

Acknowledgment

We would like to thank the anonymous reviewers and editors for their reviews. The research was supported by Ministry of Science and Technology (MOST) grant of Taiwan, ROC, MOST-105-2118-M-035-005-MY2 (S.M. Lee).

Author's Contributions

T. Martin Lukusa: Concept, design, drafting the manuscript, critical review of manuscript, approval of final version.

Shen-Ming Lee: Concept, design, critical review/revision of manuscript, approval of final version.

Chin-Shang Li: Concept, design, critical review/revision of manuscript, approval of final version.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the authors have read and approved the manuscript and no ethical issues involved.

References

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Automatic Control, 19: 716-723. DOI: 10.1109/TAC.1974.1100705

Allison, P.D., 2012. Do we really need zero-inflated models?

Angers, J.F. and A. Biswas, 2003. A Bayesian analysis of zero-inflated generalized Poisson model. Comput. Stat. Data Anal., 42: 37-46. DOI: 10.1016/S0167-9473(02)00154-8

Bhattacharya, A., B.S. Clarke and G.S. Datta, 2008. A Bayesian test for excess zeros in a zero-inflated power series distribution. Inst. Math. Stat., 1: 89-104. DOI: 10.1214/193940307000000068

Böhning, D., E. Dietz, P. Schlattmann, L. Mendonca and U. Kirchner, 1999. The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. J. Royal Stat. Society, 162: 195-209. DOI: 10.1111/1467-985X.00130

Cameron, A.C. and P.K. Trivedi, 2013. Regression Analysis of Count Data. 2nd Edn., Cambridge University Press, New York, ISBN-10: 1107667275, pp: 596.

Chen, X.D. and Y.Z. Fu, 2011. Model selection for zero-inflated regression with missing covariates. Comput. Stat. Data Anal., 55: 765-773. DOI: 10.1016/j.csda.2010.06.023

Cheung, Y.B., 2002. Zero-inflated models for regression analysis of count data: A study of growth and development. Stat. Med., 21: 1461-1469. DOI: 10.1002/sim.1088

Claeskens, G. and F. Consentino, 2008. Variable selection with incomplete covariate data. Biometrics, 64: 1062-1096. DOI: 10.1111/j.1541-0420.2008.01003.x

Consul, P.C. and F. Famoye, 1992. Generalized Poisson regression model. Commun. Stat. - Theory Meth., 21: 89-109. DOI: 10.1080/03610929208830766

Dempster, A.P., N.M. Laird and D.B. Rubin, 1977. Maximum likelihood from incomplete data via the EM algorithm. J. Royal Stat. Society, 39: 1-38.

Dietz, E. and D. Böhning, 1997. The use of two-component mixture models with one completely or partly known component. Comput. Stat., 12: 219-234.

Enders, C.K., 2010. Applied Missing Data Analysis. 1st Edn., The Guilford Press, New York, ISBN-10: 1606236393, pp: 377.

Famoye, F. and K.P. Singh, 2006. Zero-inflated generalized Poisson model with an application to domestic violence data. J. Data Sci., 4: 117-130.

Flander, W.D. and S. Greenland, 1991. Analytic methods for two-stage case-control studies and other stratified designs. Stat. Med., 10: 739-747. DOI: 10.1002/sim.4780100509

Ghosh, S.K., P. Mukhopadhyay and J.C. Lu, 2006. Bayesian analysis of zero-inflated regression models. J. Stat. Plann. Inference, 136: 1360-1375. DOI: 10.1016/j.jspi.2004.10.008

Graham, J.W., 2012. Missing Data: Analysis and Design. 1st Edn., Springer, New York, ISBN-10: 1461440181, pp: 324.

Gupta, P.L., R.C. Gupta and R.C. Tripathi, 1995. Inflated modified power series distributions with applications. Commun. Stat. - Theory Meth., 24: 2355-2374. DOI: 10.1080/03610929508831621

Hall, D.B., 2000. Zero-inflated Poisson and binomial regression with random effects: A case study. Biometrics, 56: 1030-1039. DOI: 10.1111/j.0006-341X.2000.01030.x

Hall, D.B. and J. Shen, 2010. Robust estimation for zero-inflated Poisson regression. Scandinavian J. Stat., 37: 237-252. DOI: 10.1111/j.1467-9469.2009.00657.x

Heilbron, D.C., 1994. Zero-altered and other regression models for count data with added zeros. Biometrical J., 36: 531-547. DOI: 10.1002/bimj.4710360505

Horton, N.J. and N.M. Laird, 1999. Maximum likelihood analysis of generalized linear models with missing covariates. Stat. Meth. Med. Res., 8: 37-50. DOI: 10.1177/096228029900800104

Horvitz, D.G. and D.J. Thompson, 1952. A generalization of sampling without replacement from a finite universe. J. Am. Stat. Assoc., 47: 663-685. DOI: 10.1080/01621459.1952.10483446

Ibrahim, J.G., M.H. Chen, S.R. Lipsitz and A.H. Herring, 2005. Missing-data methods for generalized linear models: A comparative review. J. Am. Stat. Assoc., 100: 332-346. DOI: 10.1198/016214504000001844

Ibrahim, J.G., H. Chu and M.H. Chen, 2012. Missing data in clinical studies: Issues and methods. J. Clin. Oncol., 30: 3297-3303. DOI: 10.1200/JCO.2011.38.7589

Ismail, N. and A.A. Jemain, 2007. Handling overdispersion with negative binomial and generalized Poisson regression models. Casualty Actuarial Society Forum, Winter.

Ismail, N. and H. Zamani, 2013. Estimation of claim count data using negative binomial, generalized Poisson, zero-inflated negative binomial and zero-inflated generalized Poisson regression models. Casualty Actuarial Society E-Forum, 41: 1-28.

Jansakul, N. and J.P. Hinde, 2002. Score tests for zero-inflated Poisson models. Comput. Stat. Data Anal., 40: 75-96. DOI: 10.1016/S0167-9473(01)00104-9

Johnson, N.L., S. Kotz and A.W. Kemp, 2005. Univariate Discrete Distributions. 3rd Edn, John Wiley and Sons, New York.

Lambert, D., 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics, 34: 1-14. DOI: 10.2307/1269547

Li, C.S., 2011. A Lack-of-fit test for parametric zero-inflated Poisson models. J. Stat. Comput. Simulat., 81: 1081-1098. DOI: 10.1080/00949651003677410

Li, C.S., 2012. Score tests for semiparametric zero-inflated Poisson models. Int. J. Stat. Probability, 1: 1-7. DOI: 10.5539/ijsp.v1n2p1

Li, C.S., J.C. Lu, J. Park, K.M. Kim and P.A. Brinkley *et al*., 1999. Multivariate zero-inflated Poisson models and their applications. Technometrics, 41: 29-38. DOI: 10.1080/00401706.1999.10485593

Little, R.J.A., 1992. Regression with missing X's: A review. J. Am. Stat. Assoc., 87: 1227-1237. DOI: 10.2307/2290664

Little, R.J.A. and D.B. Rubin, 2002. Statistical Analysis with Missing Data. 2nd Edn., John Wiley and Sons, New York.

Lu, S.E., Y. Lin and W.C.J. Shih, 2004. Analyzing excessive no changes in clinical trials with clustered data. Biometrics, 60: 257-267. DOI: 10.1111/j.0006-341X.2004.00155.x

Lukusa, T.M., S.M. Lee and C.S. Li, 2016. Semiparametric estimation of a zero-inflated Poisson regression model with missing covariates. Metrika, 79: 457-483. DOI: 10.1007/s00184-015-0563-7

Mason, A., S. Richardson, I. Plewis and N. Best, 2012. Strategy for modelling nonrandom missing data mechanisms in observational studies using Bayesian methods. J. Official Stat., 28: 279-302.

Mullahy, J., 1986. Specification and testing of some modified of some count data models. J. Econometr., 33: 341-365. DOI: 10.1016/0304-4076(86)90002-3

Nagesh, S., G. Nanjundan, R. Suresh and S. Pasha, 2015. A characterization of zero-inflated geometric model. Int. J. Math. Trends Technol., 23: 71-73. DOI: 10.14445/22315373/IJMTT-V23P510

Pahel, B.T., J.S. Preisser, S.C. Stearns and R.G. Rozier, 2011. Multiple imputation of dental caries data using a zero-inflated Poisson regression model. J. Public Health Dentistry, 71: 71-78. DOI: 10.1111/j.1752-7325.2010.00197.x

Pigot, T.D., 2001. A review of methods for missing data. Educ. Res. Evaluat., 4: 353-383. DOI: 10.1076/edre.7.4.353.8937

Preisser, J.S., J.W. Stamm, D.L. Long and M.E. Kincade, 2012. Review and recommendations for zero-inflated count regression modeling of dental caries indices in epidemiological Studies. Caries Res., 46: 413-23. DOI: 10.1159/000338992

Reilly, M. and M.S. Pepe, 1995. A mean score method for missing and auxiliary covariate data in regression models. Biometrika, 82: 299-314. DOI: 10.1093/biomet/82.2.299

Ridout, M., C.G.B. Demetrio and J. Hinde, 1998. Models for count data with many zeros. Proceedings of the 19th International Biometric Conference, (IBC' 98), Cape Town, South Africa, pp: 179-190.

Ridout, M., J. Hinde and C.G.B. Dem'eAtrio, 2001. A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. Biometrics, 57: 219-23. DOI: 10.1111/j.0006-341X.2001.00219.x

Robins, J.M., A. Rotnitzky and L.P. Zhao, 1994. Estimation of regression coefficients when some regressors are not always observed. J. Am. Stat. Assoc., 89: 846-866. DOI: 10.1080/01621459.1994.10476818

Rosenbaum, P.R. and D.B. Rubin, 1983. The central role of the propensity score in observational studies for causal effects. Biometrika, 70: 41-55. DOI: 10.1093/biomet/70.1.41

Rubin, D.B., 1976. Inference and missing data. Biometrika, 63: 581-592. DOI: 10.1093/biomet/63.3.581

Rubin, D.B., 1987. Multiple Imputation for Nonresponse in Surveys. 1st Edn., John Wiley and Sons, New York.

Samani, E.B., 2011. A missing inflated power series model for regression analysis of the British Household Panel Survey (BHPS) data. Austr. J. BASIC Applied Sci., 5: 325-331.

Samani, E.B., M. Ganjali and Y. Amirian, 2012. Zero-inflated power series joint model to analyze count data with missing responses. J. Stat. Theory Pract., 6: 334-343. DOI: 10.1080/15598608.2012.673892

Schafer, J.L. and J.W. Graham, 2002. Missing data: Our view of the state of the art. Psychol. Meth., 7: 147-177. DOI: 10.1037/1082-989X.7.2.147

Singh, S., 1963. A note on inflated Poisson distribution. J. Ind. Stat. Assoc., 1: 140-144.

Tu, W. and H. Liu, 2016. Zero-inflated data. Wiley StatsRef: Statistics Reference Online.

Wang, P., 2003. A bivariate zero-inflated negative binomial regression model for count data with excess zeros. Econom. Lett., 78: 373-378. DOI: 10.1016/S0165-1765(02)00262-8

Wang, S. and C.Y. Wang, 2001. A note on kernel assisted estimators in missing covariate regression. Stat. Probability Lett., 55: 439-449. DOI: 10.1016/S0167-7152(01)00167-5

Wang, C.Y., S. Wang, L.P. Zhao and S.T. Ou, 1997. Weighted semiparametric estimation in regression analysis with missing covariate data. J. Am. Stat. Assoc., 92: 512-525. DOI: 10.1080/01621459.1997.10474004

Yang, M., K. Das and A. Majumdar, 2016. Analysis of bivariate zero inflated count data with missing responses. J. Multivariate Anal., 148: 73-82. DOI: 10.1016/j.jmva.2016.02.010

Yau, K.K.W. and A.H. Lee, 2001. Zero-inflated Poisson regression with random effects to evaluate an occupational injury prevention programme. Stat. Med., 20: 2907-2920. DOI: 10.1002/sim.860

Zhao, L.P. and S. Lipsitz, 1992. Designs and analysis of two-stage studies. Stat. Med., 11: 769-782. DOI: 10.1002/sim.4780110608