


## Article

# EM Estimation for Zero- and $k$ -Inflated Poisson Regression Model

Monika Arora <sup>1</sup>  and N. Rao Chaganty <sup>2,\*</sup><sup>1</sup> Department of Mathematics, Indraprastha Institute of Information Technology, Delhi 110020, India; monika@iiitd.ac.in<sup>2</sup> Department of Mathematics and Statistics, Old Dominion University, Norfolk, VA 23529-0077, USA

\* Correspondence: rchagant@odu.edu

**Abstract:** Count data with excessive zeros are ubiquitous in healthcare, medical, and scientific studies. There are numerous articles that show how to fit Poisson and other models which account for the excessive zeros. However, in many situations, besides zero, the frequency of another count  $k$  tends to be higher in the data. The zero- and  $k$ -inflated Poisson distribution model (ZkIP) is appropriate in such situations. The ZkIP distribution essentially is a mixture distribution of Poisson and degenerate distributions at points zero and  $k$ . In this article, we study the fundamental properties of this mixture distribution. Using stochastic representation, we provide details for obtaining parameter estimates of the ZkIP regression model using the Expectation–Maximization (EM) algorithm for a given data. We derive the standard errors of the EM estimates by computing the complete, missing, and observed data information matrices. We present the analysis of two real-life data using the methods outlined in the paper.

**Keywords:** poisson regression; zero-inflated data; zero- and  $k$ -inflated data; EM algorithm



**Citation:** Arora, M.; Chaganty, N.R. EM Estimation for Zero- and  $k$ -Inflated Poisson Regression Model. *Computation* **2021**, *9*, 94. <https://doi.org/10.3390/computation9090094>

Academic Editor: Demos T. Tsahalidis

Received: 28 June 2021

Accepted: 21 August 2021

Published: 26 August 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Data that count the number of occurrences of certain events or the number of subjects or items that fall into certain categories arise in many scientific investigations, medical, and social science research. The most commonly used models to analyze such data are developed using the Poisson probability distribution. The Poisson distribution possesses the equi-dispersion property because its mean and variance are equal. However, in real-life examples, most often the data are over-dispersed or under-dispersed. The occurrence of over-dispersion is more common than under-dispersion. In the absence of equi-dispersion the most commonly used alternative to the Poisson distribution is the negative Binomial distribution.

There could be several reasons that lead to over-dispersion in the data. A primary cause of over-dispersion in the counts is an inflated number of zeros in excess of the number expected under the Poisson distribution. In such cases, an appropriate model is the zero-inflated Poisson (ZIP). The ZIP models are extensively studied in the literature. The earliest paper on the ZIP model was by Cohen [5]. In a seminal paper, Lambert [3] introduced and studied the ZIP regression model using the Expectation–Maximization (EM) approach [6]. Lambert [3] applied the ZIP model to count data where the response variable was the number of defects in a manufacturing process along with covariates masking, soldering, etc. The ZIP model with random effects has been studied by Min and Agresti [7] and Yau and Lee [8]. Ghosh et al. [9] explored the Bayesian approach for small to moderate sample sizes. The ZIP models using the Bayesian approach for spatial data were studied by Agarwal et al. [10]. Furthermore, ZIP models for censored data were studied by Saffari and Adnan [11], Yang and Simpson [12], and Nguyen and Dupuy [13]. Altun [14] introduced a zero-inflated Poisson–Lindley regression model, while, recently,

Bakouch et al. [15] introduced COS-Poisson distribution and the corresponding regression model for zero-inflated count data.

In health science research, zero-inflated count models have been shown to perform better than traditional count models [16,17]. The ZIP models have been applied across a wide spectrum of academic disciplines, including biology [18], ecology [19], psychology [20,21], and education [22]. The ZIP models have also been studied in economics [23–25]. In industry, the ZIP models have been applied in manufacturing [3,9], transportation [26,27], and insurance [28]. Recently, Motalebi et al. [29] applied the ZIP models for monitoring social networks. A good review and applications of ZIP models is given in Bohning and Seidel [30] and Ridout et al. [18]. The other ZIP-like models are zero-inflated negative binomial (ZINB), zero-inflated geometric (ZIG), and zero-inflated Binomial (ZIB). For example, Hall [31] illustrated the use of ZIP and ZIB in horticulture.

The zero-inflated models can be fitted easily using available packages in SAS and R software. There are two procedures in SAS that deal with zero-inflated models. In SAS, the finite mixture model (FMM) and count regression (COUNTREG) procedures can be used to study zero-inflated models. They provide estimates, standard errors, and AIC values similar to glm procedure. The high-dimensional count regression procedure (HPCOUNTREG) in SAS can handle big data. In R, the package ‘pscl’ includes functions for handling zero-inflated discrete distributions with various link options. The inflated count models are also available in the ‘VGAM’ package.

In addition to zero, some data sets may have an inflated count of additional value  $k > 0$  due to multiple effects, including the design of the study. Research questionnaire studies are examples with zero- and  $k$ -inflated count data sets typically resulting either in the way the questions were asked or the way the responses were provided. For example, one study investigated the frequency of pap smear tests in women for the last six years. The survey had a large number of women who never had a pap smear and many who had pap smears on an annual basis. Thus, the survey resulted in large frequencies of zero and six. The other source of inflation is the nature of the response. For example, Arora et al. [32] considered a study that counts the number of days a subject exercised per week. The reply of the non-exercising subjects was zero, and the reply of regularly exercising subjects was 5. Hence, the data have 0 and 5 counts inflated. Lin and Tsai [1] describe a survey where adults were asked about the number of cigarettes they consume daily. The responses tend to be none or a pack. Since a pack consists of 20 cigarettes, the data result in inflated frequencies for 0 and 20. Lin and Tsai [1] proposed a zero- and  $k$ -inflated Poisson regression model (ZkIP) to analyze such data. Sheth et al. [2] also introduced two forms of ZkIP models, known as doubly inflated Poisson (DIP) models. In this article, we study the ZkIP form given by Lin and Tsai [1]. It is the same as the second DIP model proposed by Sheth et al. [2].

The ZkIP is a finite mixture model. It has three components. The first is degenerate at zero with probability  $\pi_1$ . The second distribution is degenerate at  $k$  with probability  $\pi_2$ , and the third distribution is Poisson with mean  $\lambda$  with probability  $\pi_3 = (1 - \pi_1 - \pi_2)$ . The mixture leads to heterogeneity in the data, which is not captured by the Poisson model. These components can also be interpreted as three groups of the population. A special case of the ZkIP model is the zero- and one-inflated Poisson model (ZOIP). Zhang et al. [33] studied the properties and inference on the parameters of the ZOIP distribution without covariates. The inference of ZOIP without covariates was described by Alshkaki [34]. A Bayesian approach for the ZOIP model was examined by Tang et al. [35]. The ZOIP regression model using maximum likelihood and the Bayesian approach was also studied by [36]. The zero- and one-inflated count data using truncated Poisson was studied by [37]. Lin and Tsai [1] introduced the ZkIP regression model and used the nonlinear optimization method to obtain the maximum likelihood (ML) estimates and standard errors. The ZkIP has also been studied by Finkelman et al. [38] for grouped psychological data. In this article, we study the ZkIP model using the Expectation–Maximization (EM) approach.

Furthermore, we pursue the method outlined by Louis [4] to obtain the standard errors for the EM parameter estimates.

The outline of the article is as follows: we present the derivation of the zero- and  $k$ -inflated Poisson (ZkIP) distribution in Section 2. Section 3 contains the corresponding ZkIP regression model that incorporates observed covariates on each subject. We describe the EM algorithm steps to estimate the regression and mixing parameters in Section 3.1. Computational details of the standard errors for the regression estimates using the method described by Louis [4] are presented in Section 3.2. The criteria for model selection and goodness of fit are described in Section 4. We illustrate our methods on two real-life data sets in Section 5, including identification of significant covariates.

## 2. Zero- and $k$ -Inflated Poisson Distribution

The Poisson distribution is widely used to model nonnegative integer count data. The zero-inflated Poisson (ZIP) distribution is a popular model for count data containing excessive zeros. The ZIP distribution is a mixture of degenerate distribution at zero with probability  $\pi_1$  and Poisson distribution with probability  $1 - \pi_1$ . Additionally, if another count value  $k > 0$  in the data are also inflated, a suitable model is the Poisson distribution mixed with two-point masses  $\pi_1$  and  $\pi_2$  at 0 and  $k$ , respectively. The probability mass function of a random variable  $Y$  with this mixture distribution is given by

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

where  $0 < \pi_1 < \pi_1 + \pi_2 < 1$ ,  $\pi_3 = (1 - \pi_1 - \pi_2)$ ,  $y \in \mathbb{N}$ , and  $\lambda > 0$ . The distribution (1) is known as the zero- and  $k$ -inflated Poisson (ZkIP) distribution [1]. The moment generating function of the ZkIP distribution is  $M_Y(t) = E(e^{tY}) = \pi_1 + \pi_2 e^{tk} + \pi_3 e^{\lambda(e^t - 1)}$  and the probability generating function is  $G_Y(z) = E(z^Y) = \pi_1 + \pi_2 z^k + \pi_3 e^{\lambda(z - 1)}$ . Using  $M_Y(t)$ , it is easy to show that the mean and variance of the ZkIP distribution are

$$\begin{aligned} E(Y) &= k\pi_2 + \pi_3\lambda \\ \text{Var}(Y) &= k^2\pi_2(1 - \pi_2) + \pi_3\lambda(1 + \pi_1\lambda + \pi_2\lambda - 2k\pi_2). \end{aligned}$$

Since the ZkIP distribution is essentially a mixture of Poisson and two degenerate distributions at zero and  $k$  with probabilities  $\pi_1$  and  $\pi_2$ , respectively, it reduces to ZIP when  $\pi_2 = 0$  and becomes the Poisson distribution if  $\pi_1 = \pi_2 = 0$ . The following stochastic representation is instrumental in elucidating the properties of the ZkIP distribution. Consider a latent variable  $\mathbf{z} = (z_1, z_2, z_3)$  distributed as multinomial with parameters  $(1, \pi_1, \pi_2, \pi_3)$ . Note that  $\mathbf{z}$  takes values  $(1, 0, 0)$  with probability  $\pi_1$ ,  $(0, 1, 0)$  with probability  $\pi_2$ , and  $(0, 0, 1)$  with probability  $\pi_3$ . That is,

$$P(\mathbf{z} = (z_1, z_2, z_3)) = \begin{cases} \pi_1 & \text{if } z_1 = 1, z_2 = 0, z_3 = 0 \\ \pi_2 & \text{if } z_1 = 0, z_2 = 1, z_3 = 0 \\ \pi_3 & \text{if } z_1 = 0, z_2 = 0, z_3 = 1. \end{cases} \quad (2)$$

Furthermore, let us assume the conditional distribution of  $Y|\mathbf{z}$  is

$$P(Y = y|\mathbf{z} = (z_1, z_2, z_3)) = \begin{cases} 1 & \text{for } z_1 = 1, y = 0 \\ 1 & \text{for } z_2 = 1, y = k \\ \frac{\lambda^y e^{-\lambda}}{y!} & \text{for } z_3 = 1, y = 0, 1, \dots \end{cases} \quad (3)$$

Thus, the joint distribution of  $(Y, \mathbf{z})$  obtained by multiplying (2) and (3) is

$$P(Y = y, \mathbf{z} = (z_1, z_2, z_3)) = \begin{cases} \pi_1 & \text{for } z_1 = 1, y = 0 \\ \pi_2 & \text{for } z_2 = 1, y = k \\ \pi_3 \frac{\lambda^y e^{-\lambda}}{y!} & \text{for } z_3 = 1, y = 0, 1, \dots \end{cases} \quad (4)$$

The marginal of  $Y$  can be obtained from (4) by summing over the three possible values of  $\mathbf{z}$ . Thus, we get

$$\begin{aligned} P(Y = 0) &= P(Y = 0, z_1 = 1) + P(Y = 0, z_2 = 1) + P(Y = 0, z_3 = 1) \\ &= \pi_1 + \pi_3 e^{-\lambda}, \\ P(Y = k) &= P(Y = k, z_1 = 1) + P(Y = k, z_2 = 1) + P(Y = k, z_3 = 1) \\ &= \pi_2 + \pi_3 \frac{\lambda^k e^{-\lambda}}{k!}, \end{aligned}$$

and

$$\begin{aligned} P(Y = y) &= P(Y = y, z_1 = 1) + P(Y = y, z_2 = 1) + P(Y = y, z_3 = 1) \\ &= \pi_3 \frac{\lambda^y e^{-\lambda}}{y!}, \quad \text{for } y \geq 1, y \neq k, \end{aligned}$$

which is equivalent to the ZkIP distribution defined by (1). Furthermore, the posterior distribution  $P(\mathbf{z}|Y) = P(\mathbf{z})P(Y|\mathbf{z})/P(Y)$  can be summarized as in Table 1.

**Table 1.**  $P(\mathbf{z} = (z_1, z_2, z_3)|Y = y)$  of ZkIP.

$\mathbf{z} = (z_1, z_2, z_3)$	$y = 0$	$y = k$	$y \neq 0, k$
$z_1 = 1$	$\frac{\pi_1}{\pi_1 + \pi_3 p_0}$	0	0
$z_2 = 1$	0	$\frac{\pi_2}{\pi_2 + \pi_3 p_k}$	0
$z_3 = 1$	$\frac{\pi_3 p_0}{\pi_1 + \pi_3 p_0}$	$\frac{\pi_3 p_k}{\pi_2 + \pi_3 p_k}$	1

In Section 3, we build a ZkIP regression model using (1), and, in Section 3.1, we use the conditional probabilities in Table 1 to develop the EM algorithm for estimation of the ZkIP parameters from the data.

### 3. Zero- and k-Inflated Poisson Regression Model

Let  $\mathbf{y} = (y_1, y_2, \dots, y_n)$  be a vector of  $n$  independent count responses. We assume that the number of  $y_i$ 's that are equal to 0 (or  $k$ ) is high and corresponding to each  $y_i$ , a vector  $\mathbf{x}_i^T = (1, x_{i1}, \dots, x_{ip})$  of covariates has been observed. A reasonable model for the

distribution of each  $y_i$  is given by (1) with different parameters  $\lambda_i$  but the same mixing parameters  $\pi_1$  and  $\pi_2$ . In this case, the likelihood function of the observed data is

$$L_{obs}(\pi_1, \pi_2, \lambda|\mathbf{y}) = \prod_{i:y_i=0} (\pi_1 + \pi_3 p_{0i}(\lambda_i)) \prod_{i:y_i=k} (\pi_2 + \pi_3 p_{ki}(\lambda_i)) \prod_{i:y_i \neq 0,k} (\pi_3 p_{yi}(\lambda_i)), \quad (5)$$

where  $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_n)$  and  $p_{yi}(\lambda_i) = e^{-\lambda_i} \lambda_i^{y_i} / y_i!$  for  $y_i = 0, 1, 2, \dots$ . To incorporate the covariates into the model, we follow the standard generalized linear model (GLM) framework for the multinomial distribution. The three mixing distributions can be viewed as three nominal categories (degenerate(0), degenerate( $k$ ), and Poisson) with probabilities  $\pi_1$ ,  $\pi_2$ , and  $\pi_3$ , respectively. Following the GLM baseline category logit models for the multinomial, we re-parametrize and set

$$\log\left(\frac{\pi_1}{\pi_3}\right) = \gamma \quad \text{and} \quad \log\left(\frac{\pi_2}{\pi_3}\right) = \delta. \quad (6)$$

We treat the Poisson distribution as the baseline category, leading to two equations for the other two categories. As in loglinear models, the ZkIP regression model assumes the Poisson parameter  $\lambda_i$  is a loglinear function of the covariates, and it is given by

$$\log(\lambda_i) = \mathbf{x}_i^T \boldsymbol{\beta}.$$

where  $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \dots, \beta_p)^T$  is a  $p + 1$  dimensional unknown regression parameter vector. For simplicity, we assume that the parameters  $\gamma$  and  $\delta$  are constants. The generalization of the case where  $\gamma$  and  $\delta$  are functions of the covariates is straightforward. Thus, the parameters of our ZkIP regression model are  $\boldsymbol{\beta}$ ,  $\gamma$ , and  $\delta$ . In the next section, we consider estimating the parameters  $\boldsymbol{\beta}$ ,  $\gamma$ , and  $\delta$  using the observed data.

### 3.1. Estimation of the Regression Parameters

In this section, we study methods for estimating the parameters of the ZkIP regression model. The two popular methods are the maximum likelihood (ML) and Expectation–Maximization (EM) approach. The ML technique involves optimizing the likelihood or the log-likelihood function with respect to the unknown parameters  $\boldsymbol{\beta}$ ,  $\gamma$ , and  $\delta$ . Substituting the reparametrizations (6) in the likelihood function (5), we get

$$\begin{aligned} \ell_{obs}(\boldsymbol{\beta}, \gamma, \delta) &= \log L_{obs}(\boldsymbol{\beta}, \gamma, \delta|\mathbf{y}) \\ &= \sum_{i:y_i=0} \log(e^\gamma + p_{0i}(\lambda_i)) + \sum_{i:y_i=k} \log(e^\delta + p_{ki}(\lambda_i)) \\ &\quad + \sum_{i:y_i \neq 0,k} \log(p_{yi}(\lambda_i)) - n \log(1 + e^\gamma + e^\delta), \end{aligned} \quad (7)$$

where  $\log \lambda_i = \mathbf{x}_i^T \boldsymbol{\beta}$ . The ML estimates can be obtained by maximizing the log-likelihood (7) directly with respect to the parameters or taking the partial derivatives and solving the three score equations:

$$\begin{aligned} \sum_{i:y_i=0} \frac{e^\gamma}{e^\gamma + p_{0i}(\lambda_i)} &= \frac{ne^\gamma}{(1 + e^\gamma + e^\delta)} \\ \sum_{i:y_i=k} \frac{e^\delta}{e^\delta + p_{ki}(\lambda_i)} &= \frac{ne^\delta}{(1 + e^\gamma + e^\delta)} \\ \sum_{i:y_i \neq 0,k} (y_i - \lambda_i) \mathbf{x}_i &= \sum_{i:y_i=0} \frac{\lambda_i p_{0i}(\lambda_i)}{e^\gamma + p_{0i}} \mathbf{x}_i - \sum_{i:y_i=k} \frac{(k - \lambda_i) p_{ki}(\lambda_i)}{e^\delta + p_{ki}(\lambda_i)} \mathbf{x}_i. \end{aligned} \quad (8)$$

Equation (8) can be solved iteratively using the Newton–Raphson method. In theory, this seems fine, but, in practice, there are convergence issues with ML estimation. An

alternative to ML and a popular method for parameter estimation is the Expectation–Maximization (EM) approach, used by Lambert [3] for the ZIP model. Here, we extend her ideas for the ZkIP model.

The EM approach treats the observed data  $\mathbf{y} = (y_1, y_2, \dots, y_n)$  as the part of a complete data that includes a latent vector  $\mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_n)$ , which is regarded as missing. Here, each  $\mathbf{z}_i = (z_{i1}, z_{i2}, z_{i3})$  is a three-component vector with a probability distribution given by (2), and the conditional distribution of  $y_i$  given  $\mathbf{z}_i$  is given by (3). Then, the joint distribution of the observed and missing data are given by

$$P(y_i, \mathbf{z}_i) = \begin{cases} \pi_1 & \text{for } y_i = 0, z_{1i} = 1 \\ \pi_2 & \text{for } y_i = k, z_{2i} = 1 \\ \pi_3 p_{yi}(\lambda_i) & \text{for } y_i = 0, 1, \dots, z_{3i} = 1, \end{cases}$$

where  $p_{yi}(\lambda_i)$  is the Poisson probability mass function with mean  $\lambda_i$ . Therefore, the complete data likelihood function of the ZkIP model is given by

$$L_{comp}(\pi_1, \pi_2, \lambda | \mathbf{y}, \mathbf{z}) = \prod_{i: y_i=0} (\pi_1)^{z_{1i}} \prod_{i: y_i=k} (\pi_2)^{z_{2i}} \prod_{i=1}^n (\pi_3 p_{yi}(\lambda_i))^{z_{3i}},$$

and the log-likelihood of the complete data,  $(\mathbf{y}, \mathbf{z})$  for the ZkIP model is

$$\begin{aligned} \ell_{comp}(\pi_1, \pi_2, \lambda | \mathbf{y}, \mathbf{z}) &= \sum_{i: y_i=0} (z_{1i} \pi_1 + z_{3i} (\log \pi_3 + \log p_{0i}(\lambda_i))) \\ &+ \sum_{i: y_i=k} (z_{2i} \log \pi_2 + z_{3i} (\log \pi_3 + \log p_{ki}(\lambda_i))) \\ &+ \sum_{i=1}^n (z_{3i} \log \pi_3 + \log p_{yi}(\lambda_i)). \end{aligned} \quad (9)$$

Using the reparametrization given in (6), we can write the log-likelihood of the complete data as

$$\ell_{comp}(\gamma, \delta, \lambda | \mathbf{y}, \mathbf{z}) = \sum_{i=1}^n (z_{1i} \gamma + z_{2i} \delta - \log(1 + e^\gamma + e^\delta)) + \sum_{i=1}^n z_{3i} \log p_{yi}(\lambda_i). \quad (10)$$

Note that, when  $\pi_2 = 0$ , the ZkIP reduces to the ZIP model. Thus, from (9), the log-likelihood of the ZIP for the complete data is

$$\begin{aligned} \ell_{comp}(\pi_1, \lambda | \mathbf{y}, \mathbf{z}_1) &= \sum_{i: y_i=0} (z_{1i} \pi_1 + (1 - z_{1i}) (\log(1 - \pi_1) + \log p_{0i}(\lambda_i))) \\ &+ \sum_{i: y_i>0} ((1 - z_{1i}) \log(1 - \pi_1) + \log p_{yi}(\lambda_i)). \end{aligned}$$

From (10), the log-likelihood of the ZIP for the complete data can be written as

$$\ell_{comp}(\gamma, \lambda | \mathbf{y}, \mathbf{z}_1) = \sum_{i=1}^n (z_{1i} \gamma - \log(1 + e^\gamma) + \sum_{i=1}^n (1 - z_{1i}) \log p_{yi}(\lambda_i)). \quad (11)$$

Lambert [3] used Equation (11) as the complete data log-likelihood for the ZIP model to get the EM estimates.

We now proceed to describe in detail the EM algorithm for the ZkIP model. The first step in the EM algorithm involves selecting some initial values for the unknown parameters.

The choice of the initial values is important for the convergence of the algorithm. An incorrect choice of the initial values could result in slow convergence or breakdown of the algorithm. We recommend using the proportions of zeros and  $k$ 's from the observed data as initial values for the parameters  $\pi_1$  and  $\pi_2$ , respectively. Then, we use the relations (6) to get initial values  $\gamma_0$  and  $\delta_0$  for the parameters  $\gamma$  and  $\delta$ , respectively. The initial values of  $\beta$  can be obtained by fitting a Poisson model on the data. The initial values  $\beta_0$  can be used as the coefficients of the covariates.

The next step involves filling the latent values  $\mathbf{z}_i$  by its expectations, which is the E-step. We use the conditional expected values of  $E(\mathbf{z}|\mathbf{y})$  given in Table 2 to generate  $\mathbf{z}_i$ 's. Note that Table 2 is a reparametrized version of Table 1.

**Table 2.**  $E(\mathbf{z}|\mathbf{y})$  for the ZkIP regression model.

$\mathbf{z} = (z_1, z_2, z_3)$	$\mathbf{y} = 0$	$\mathbf{y} = k$	$\mathbf{y} \neq 0, k$
$z_1$	$\frac{e^\gamma}{e^\gamma + p_{0i}(\lambda_i)}$	0	0
$z_2$	0	$\frac{e^\delta}{e^\delta + p_{ki}(\lambda_i)}$	0
$z_3$	$\frac{p_{0i}(\lambda_i)}{e^\gamma + p_{0i}(\lambda_i)}$	$\frac{p_{ki}(\lambda_i)}{e^\delta + p_{ki}(\lambda_i)}$	1

We use Table 2 to estimate the missing values in the expectation step of the EM algorithm as follows:

$$\begin{aligned}\widehat{z}_{1i} &= E(z_{1i}|y_i = 0) = \frac{e^\gamma}{e^\gamma + p_{0i}(\lambda_i)} \quad \text{and} \quad \widehat{z}_{1i} = E(z_{1i}|y_i = k) = 0, \\ \widehat{z}_{2i} &= E(z_{2i}|y_i = k) = \frac{e^\delta}{e^\delta + p_{ki}(\lambda_i)} \quad \text{and} \quad \widehat{z}_{2i} = E(z_{2i}|y_i \neq k) = 0.\end{aligned}\quad (12)$$

For the maximization step in the EM algorithm, instead of maximizing the complete likelihood directly, we solve the score equations

$$\begin{aligned}\frac{\partial \ell_{comp}}{\partial \beta} &= \sum_{i=1}^n \widehat{z}_{3i} (y_i - e^{\mathbf{x}_i^T \beta}) \mathbf{x}_i = 0 \\ \frac{\partial \ell_{comp}}{\partial \gamma} &= \sum_{i=1}^n \widehat{z}_{1i} - \frac{ne^\gamma}{(1 + e^\gamma + e^\delta)} = 0 \\ \frac{\partial \ell_{comp}}{\partial \delta} &= \sum_{i=1}^n \widehat{z}_{2i} - \frac{ne^\delta}{(1 + e^\gamma + e^\delta)} = 0,\end{aligned}\quad (13)$$

where  $\widehat{z}_{3i} = (1 - \widehat{z}_{1i} - \widehat{z}_{2i})$  and  $\ell_{comp}$  is defined in (10). In summary, the EM algorithm to estimate the parameters  $\gamma$ ,  $\delta$ , and the regression parameter  $\beta$  for the ZkIP regression model can be summarized as follows.

1. Select initial values  $\beta_0$ ,  $\gamma_0$ ,  $\delta_0$  for the parameters  $\beta$ ,  $\gamma$ , and  $\delta$  respectively.
2. E-step: Estimate  $\widehat{z}_{1i}$ ,  $\widehat{z}_{2i}$  using Equation (12).
3. M-step: Solve the score Equation (13) and obtain an updated estimates  $\beta_1$ ,  $\gamma_1$ ,  $\delta_1$ .
4. Repeat the E-step and the M-step until the parameter estimates converge.

In the next section, we will discuss how to obtain standard errors of the EM estimates.

### 3.2. Standard Errors for EM Estimates

The most commonly used method to get the standard errors in the mixture models is to compute the matrix of partial derivatives of the log-likelihood for the observed data, that



is, to calculate the information matrix from the observed data. The optimization algorithms routinely output a numerically computed Hessian matrix for the functions that are being optimized. Lambert [3] used this method for computing the standard errors for the ZIP regression model. Lin and Tsai [1] used the Hessian matrix to get the standard errors for the ZkIP model without actually computing second-order partial derivatives of the log-likelihood.

However, for the EM framework, an appropriate and easier approach for obtaining the standard errors is the method outlined by Louis [4]. The method is based on the complete and missing data log-likelihoods. The relation between the likelihood of the complete, observed and missing data is given by

$$L_{comp}(\theta | \mathbf{y}, \mathbf{z}) = L_{obs}(\theta | \mathbf{y}) L_{miss}(\theta | (\mathbf{z} | \mathbf{y})), \quad (14)$$

where  $\mathbf{y}$  and  $\mathbf{z}$  stand for the observed and missing data, respectively. Taking log on both sides of (14), we get

$$\ell_{comp}(\theta | \mathbf{y}, \mathbf{z}) = \ell_{obs}(\theta | \mathbf{y}) + \ell_{miss}(\theta | (\mathbf{z} | \mathbf{y})) \quad (15)$$

We can see from Equation (15) that the information matrices for the complete, observed, and missing data satisfy the following equation:

$$\mathcal{I}_{comp} = \mathcal{I}_{obs} + \mathcal{I}_{missing}, \quad (16)$$

where  $\mathcal{I}_{comp}$  is obtained using (15) as

$$\mathcal{I}_{comp} = \begin{bmatrix} -\frac{\partial^2 \ell_{comp}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} & -\frac{\partial^2 \ell_{comp}}{\partial \boldsymbol{\beta} \partial \gamma} & -\frac{\partial^2 \ell_{comp}}{\partial \boldsymbol{\beta} \partial \delta} \\ -\frac{\partial^2 \ell_{comp}}{\partial \gamma \partial \boldsymbol{\beta}} & -\frac{\partial^2 \ell_{comp}}{\partial \gamma^2} & -\frac{\partial^2 \ell_{comp}}{\partial \gamma \partial \delta} \\ -\frac{\partial^2 \ell_{comp}}{\partial \delta \partial \boldsymbol{\beta}} & -\frac{\partial^2 \ell_{comp}}{\partial \delta \partial \gamma} & -\frac{\partial^2 \ell_{comp}}{\partial \delta^2} \end{bmatrix}. \quad (17)$$

Equation (16) can be re-written as

$$\mathcal{I}_{obs} = \mathcal{I}_{comp} - \mathcal{I}_{miss}. \quad (18)$$

Since the right-hand side of Equation (18) depends on the missing data, Louis [4] has recommended taking the conditional expected value of the missing data given the observed data. Therefore, we have

$$\mathcal{I}_{obs} = E(\mathcal{I}_{obs} | \mathbf{y}) = E(\mathcal{I}_{comp} | \mathbf{y}) - E(\mathcal{I}_{miss} | \mathbf{y}). \quad (19)$$

Thus, the estimate of the observed information matrix is given by

$$\widehat{\mathcal{I}}_{obs} = E(\mathcal{I}_{comp} | \mathbf{y}) - E(\mathcal{I}_{miss} | \mathbf{y}). \quad (20)$$



The elements of the expected information matrix  $E(\mathcal{I}_{comp}|\mathbf{y})$  are given in Appendix A.1. Using (7), (10), and (15), the log-likelihood of the missing data for the ZkIP regression model is given by

$$\begin{aligned} \ell_{miss}(\boldsymbol{\beta}, \gamma, \delta) = & \sum_{i=1}^n (z_{1i}\gamma + z_{2i}\delta + z_{3i} \log p_{yi}(\lambda_i)) - \sum_{i:y_i=0} \log(e^\gamma + p_{0i}(\lambda_i)) \\ & - \sum_{i:y_i=k} \log(e^\delta + p_{ki}(\lambda_i)) - \sum_{i:y_i \neq 0,k} \log p_{yi}(\lambda_i). \end{aligned} \quad (21)$$

The elements of the matrix  $E(\mathcal{I}_{miss}|\text{observed})$  are the negative of the expected value of second-order derivatives of (21), and these are given in Appendix A.2. Using these second-order derivatives, we can compute  $\widehat{\mathcal{I}}_{obs}$  given in Equation (20). The square root of diagonal elements of  $(\widehat{\mathcal{I}}_{obs})^{-1}$  gives the standard errors of the EM estimates.

#### 4. Model Selection and Model Fit

In statistical inference, estimation of the parameters is usually followed by testing the significance of the parameters and selecting the best model for the data. Hence, in this section, we discuss the hypothesis testing to the significance of inflation at zero and  $k$ —in other words, whether ZkIP significantly fits the data better than the ZIP or the Poisson model. There are various criteria to select the best model. We use the Akaike Information Criterion (AIC) and the likelihood-ratio method to arrive at the best model that fits the data. These details will be illustrated with a couple of real-life data analyses in Section 5.

##### 4.1. Hypothesis Testing

Here, we discuss hypothesis testing to determine significant parameters and covariates. In the ZkIP model, the parameters  $\pi_1$  and  $\pi_2$  represent the proportion of observations that come from degenerate distributions, and the parameter  $\boldsymbol{\beta}$  determines the effects of the covariates in the model. Let  $\hat{\boldsymbol{\theta}} = (\hat{\pi}_1, \hat{\pi}_2, \hat{\boldsymbol{\beta}})$  denote the EM estimates of these parameters. Assume that the true value  $\boldsymbol{\theta}^0 = (\pi_1^0, \pi_2^0, \boldsymbol{\beta}^0)$  is in the interior of the parameter space, that is,  $0 < \pi_1^0 + \pi_2^0 < 1$  and  $-\infty < \boldsymbol{\beta}^0 < \infty$ . Under usual regularity conditions,  $\hat{\boldsymbol{\theta}}$  is asymptotically normal with mean  $\boldsymbol{\theta}^0$  and covariance matrix is given by  $(\widehat{\mathcal{I}}_{obs})^{-1}$ . We can use this result to construct a Wald's test for testing the hypotheses that a specified proportion  $0 < \pi_2^0 < 1$  of observations come from a degenerate distribution at  $k$  or a specified proportion  $0 < \pi_1^0 < 1$  come from the degenerate distribution at zero. Similarly, the hypothesis  $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}^0$  could be tested for significance using Wald's test.

The FMM and Countreg procedures in SAS use the parameters  $\gamma = \log(\pi_1/\pi_3)$  and  $\delta = \log(\pi_2/\pi_3)$  and test for the hypothesis  $H_0 : (\gamma, \delta) = (0, 0)$ . This hypothesis is equivalent to testing  $H_0 : (\pi_1, \pi_2) = (\pi_1^0 = 1/3, \pi_2^0 = 1/3)$ , which can be done using Wald's test because  $\pi_1^0 = 1/3$  and  $\pi_2^0 = 1/3$  are values in the interior of the parameter space.

As we discussed in Section 2, the ZkIP, ZIP, and the Poisson model form a group of three nested models in the sense that Poisson is a special case of ZIP which is a special case of ZkIP. Thus, one could use the likelihood ratio test (LRT) to test the significance of the nested models, that is, whether the ZkIP model could be replaced by the ZIP model or whether the ZIP model could be replaced by the Poisson model. We need to test the null hypothesis  $H_0 : \pi_2 = 0$  to see whether there is a significant or insignificant inflated frequency at count  $k$ . The acceptance of the null hypothesis implies that we can replace the ZkIP model with the ZIP model. Similarly, the acceptance of  $H_0 : \pi_1 = 0$  implies that inflation at zero is insignificant, and the ZIP model can be replaced by the Poisson model. Since  $0 \leq \pi_i \leq 1, (i = 1, 2)$ , the null hypothesis  $H_0 : \pi_i = 0$  corresponds to testing a parameter value on the boundary. Therefore, the standard asymptotic theory for the likelihood ratio statistic is not applicable. The asymptotic distribution of the likelihood

ratio statistic is not a  $\chi^2$  distribution, but is a mixture of  $\chi^2$  distributions [39,40]. In fact, the test statistic  $-2 \log L$  approaches a 50:50 mixture of  $\chi_0^2$  and  $\chi_1^2$ .

#### 4.2. Model Selection

We can use several criteria for selecting the appropriate model between the three competing models, Poisson, ZIP, and ZkIP. A popular criterion is the Akaike Information Criteria (AIC). The AIC was introduced by [41], and it is calculated as  $-2\ell + 2m$ , where  $\ell$  is the maximum value of the log-likelihood and  $m$  is the number of parameters for the model under consideration. The log-likelihood tends to increase as we move from a simpler model to a complex one. The constant  $2m$  penalizes the complex model since it will have more parameters than the simple model. This avoids overfitting the model for the data. To select the best model, we use minimum AIC criteria and apply Burnham and Anderson's approach [42]. The interpretation of AIC is weighty when different values are compared. Thus, it is a relative term and not an absolute term that is of importance. The approach given in [42] is based on AIC differences  $\Delta_i = AIC_i - AIC_{min}$ , where  $AIC_i$  is the AIC of the  $i$ -th model and  $AIC_{min}$  is the minimum AIC of the models in the study. The lower values of  $\Delta_i$  imply that there is not much difference between model  $i$  and model with minimum AIC—while, from higher values of  $\Delta_i$ , we can infer that the model with minimum AIC is better than model  $i$  (Table 3).

**Table 3.** Rough rules of thumb for model selection based on AIC differences.

$\Delta_i$	Level of Empirical Support of Model $i$
$0 \leq \Delta_i \leq 2$	Substantial
$4 \leq \Delta_i \leq 7$	Considerably less
$\Delta_i > 10$	Essentially none

#### 4.3. Goodness of Fit

For count data, the most commonly used statistic for testing the goodness-of-fit test is the Pearson chi-square statistic  $\chi^2 = \sum_{i=1}^c (o_i - e_i)^2 / e_i$ , where  $o_i$  is observed frequency and  $e_i$  is the expected frequency of the  $i$ -th category, and  $c$  is the total number of categories. Asymptotically, the  $\chi^2$  statistic follows a chi-square distribution with  $(c - 1)$  degrees of freedom. The test is not the best when there are inflated frequencies. An alternate and simple measure for checking the goodness-of-fit among competing models is the sum of Absolute Error (ABE), which is defined as

$$\text{Sum of ABE} = \sum_{i=1}^c |o_i - e_i|.$$

The model that has a minimum sum of ABE has the least deviation between the observed and expected frequencies. Hence, the model with a minimum error fits data the best.

### 5. Applications

In this section, we illustrate the results presented in Sections 3.1 and 3.2 on two real-life data sets. These data sets were obtained from the National Health Interview Survey (NHIS) conducted by the National Center for Health Sciences (NCHS). Since 1957, NCHS has been collecting and archiving data on US residents. The data are collected annually on various health topics, including immunizations, depression, hepatitis, cancer, tobacco use, and other variables related to health. For our illustrations, we took a subset of data that was collected in the year 2015. We fit the zero- and  $k$ -inflated Poisson (ZkIP) model for both the data sets and compare them to the zero-inflated Poisson (ZIP) and Poisson models. The first example illustrates a ZkIP model with inflations at 0 and  $k = 6$ , while the second

example demonstrates a zero- and one-inflated Poisson (ZOIP) model with inflations at 0 and  $k = 1$ .

### 5.1. Pap Smear Data

Cervical cancer is a major concern for the female population. A common preventive and early detection screening procedure for cervical cancer is the pap smear test. In this example, the data consist of the number of pap smear tests a female took in the last six years for females aged more than 18 years. The count variable represents responses to two questions in the survey: (1) *Have you ever had a Pap smear or Pap test?* and (2) *How many Pap tests have you had in the last 6 years?*. If the reply to the first question is a 'No', then the number of tests done is reported zero, while, if the reply is a 'Yes', then the number of tests done is the same as the reply to the second question. The data also consist of the age of the female respondent and her answer to the question, *"Do you ever received HPV shot or vaccine?"*. Here, age is a continuous variable, whereas the response to HPV shot/vaccine is a dichotomous variable. Both of these variables could be treated as covariates in the model.

There were 33,672 females interviewed in the survey, out of which about 3.5% choose not to answer, or their response was not recorded. We performed a list-wise deletion to clean the data and ended up with a data set consisting of 12,014 independent observations. The mean number of the pap smear tests for the data thus obtained is 3.40 and the variance is 5.25. The percentage (count) of females who never took a pap smear test was 15.68% (1884), and the percentage (count) of females who had one pap smear each year for a total of six in the last six years was 29.17% (3504). The proportions of zero and six in the data set are inflated, and both of these proportions are more than what we would expect under a Poisson model. Therefore, an appropriate model for these data is the zero and six inflated Poisson model or the ZkIP model with  $k = 6$ .

Using the methods described in Sections 3.1 and 3.2, we fitted ZkIP, ZIP, and the Poisson models for this pap smear data. We tested the significance of age and HPV shot covariates in the models using Wald's test. The variable age was not significant in the ZkIP and ZIP models. Thus, age was removed in subsequent analysis, and we reran the models with only HPV shot as the covariate.

The regression parameter is significant for all the models at  $\alpha = 0.05$ . The estimates obtained by the EM algorithm and the corresponding standard errors for the EM estimates described in Section 3.2 are presented in Table 4. For the ZkIP model, the mixing parameter estimates were  $\hat{\pi}_1 = 0.126$  and  $\hat{\pi}_2 = 0.26$ , meaning about 12.6% of the zeros were from the degenerate distribution and 26% of the observed frequencies of six pap smear count were from a degenerate distribution at six. The table also has the AIC value and the maximum value of the log-likelihood function for different models. The AIC values of the ZkIP, ZIP, and Poisson models are 46,523.89, 52,205.70, and 56,061.88, respectively. The ZkIP model has minimum AIC, and  $\Delta_{ZIP}$  is greater than 5000 and  $\Delta_{Poisson} > 9000$ . Thus, according to Table 3, the empirical support for both the Poisson and ZIP model is "essentially none". Thus, for these data, adding one more distribution, which is degenerate at six to the model or the ZkIP with  $k = 6$ , is a better model than the ZIP or Poisson model.

Recall that the three models Poisson, ZIP, and ZkIP, are nested models, and we could use the likelihood ratio criterion described in Section 4 to decide whether the complex model could be reduced to the simpler model. The LRT statistic, which compares the Poisson model with the ZIP, is  $-2 \log \Lambda = 3860.18$  and the  $p$ -value is computed using the limiting distribution, which is a mixture of two  $\chi^2$ 's with equal weights, is less than 0.0001. This implies that the inflation at zero is significant, and the ZIP model is significantly better than the Poisson model. Similarly, we use LRT to compare ZkIP with the ZIP model. The value of the test statistic is  $-2 \log \Lambda = 1469.85$ , which is again highly significant with a  $p$ -value less than 0.0001. Hence, ZkIP is significantly better than the ZIP model.

**Table 4.** Estimates and standard errors for pap smear data. The significant regression coefficients are marked with an asterisk.

Parameter	ZkIP	ZIP	Poisson
Intercept	1.0837 * (0.0086)	1.3696 * (0.0054)	1.2192* (0.0053)
HPV shot	0.0727 * (0.0235)	0.0333 * (0.0154)	0.0276 * (0.0152)
$\hat{\gamma}$	−1.5844 (0.0331)	−1.8132 (0.0184)	—
$\hat{\delta}$	−0.8526 (0.0235)	—	—
$\hat{\pi}_1$	0.1257 (0.0026)	0.1402 (0.0022)	—
$\hat{\pi}_2$	0.2613 (0.0018)	—	—
$\log L_{obs}$	−25,363.93	−26,098.85	−28,028.94
AIC	46,523.89	52,205.70	56,061.88

Furthermore, we check the goodness-of-fit of the models by comparing the observed frequencies and the expected frequencies. The observed and predicted frequencies are in Table 5. Table 5 shows that the Poisson model has the highest sum of absolute error (ABE) and does not provide a good fit to the data. The error 5685.69 of the ZIP model is lower than that of the Poisson model 8086.16, while the sum of the absolute difference between the observed and expected frequency is minimum (1130.93) for the ZkIP model. Therefore, the ZkIP, which is able to capture inflated frequencies at both zero and 6, is a superior model for these data compared with ZIP and the Poisson model.

**Table 5.** Frequency comparisons for pap smear data.

Count	Observed	ZkIP	ZIP	Poisson
0	1884	1884.24	1883.47	402.81
1	1417	1112.73	785.54	1366.85
2	1362	1661.47	1553.79	2323.12
3	1536	1648.59	2043.78	2627.90
4	1115	1228.13	2016.95	2230.12
5	905	732.45	1592.78	1514.29
6	3504	3504.14	1048.87	857.41
>6	291	162.46	600.28	421.80
Sum of ABE		1130.93	5685.69	8086.16
$\chi^2$		297.64	7263.93	15,312.36

## 5.2. Emergency Room Data

The data for this example were taken from the NHIS 2015 database on children aged less than 18 years. The count variable is the number of visits to an emergency room (ER) of children in a year. We choose age (0–17) and gender (Male/Female) as the covariates. We remove the cases where the response or the covariates are missing and end up with a clean data set of  $n = 12,223$  children. The average number of visits to the ER in our sample is 0.26, and the variance is 0.45. In the data, the count values 0 and 1 have frequencies 10,046 and 1466, respectively. These frequencies are high because they account for 82.19 and 11.99 percentages of the total sample.

We fit zero- and one-inflated Poisson (ZOIP), zero-inflated Poisson (ZIP), and the Poisson model for these data. The significance of the regression variables is tested using Wald Test. In the first iteration, the gender variable was insignificant in all three models,

so it was removed from the models. The analysis is again performed with only age as the covariate. The model estimates and standard errors are presented in Table 6.

**Table 6.** Estimates and SE for ER data. The significant regression coefficients are marked with an asterisk.

Parameter	ZkIP	ZIP	Poisson
Intercept	0.1173 * (0.0610)	-0.0314 * (0.0395)	-1.0512 * (0.0312)
Age	-0.0217 * (0.0044)	-0.0252 * (0.0039)	-0.0358 * (0.0033)
$\hat{\gamma}$	1.0959 (0.1210)	0.7098 (0.0427)	—
$\hat{\delta}$	-2.0450 (0.3853)	—	—
$\hat{\pi}_1$	0.7260 (0.0213)	0.6704 (0.0094)	—
$\hat{\pi}_2$	0.0314 (0.0679)	—	—
$\log L_{obs}$	-7736.62	-7741.19	-8295.23
AIC	15,481.24	15,488.39	16,594.00

The AIC value of the ZOIP, ZIP, and Poisson models are 15,481.24, 15,488.39, and 16,594.00, respectively. On calculating the AIC differences, we obtain  $\Delta_{ZIP} = 7.15$ ,  $\Delta_{Poisson} = 1112.76$ . The AIC difference between ZIP and Poisson models gives  $\Delta = 1105.61$ . Clearly, the ZIP model performs better than the Poisson model. Furthermore, the ZIP model has “considerably less” support than the ZOIP model (Table 3). We also performed the likelihood ratio test for model selection. The LRT statistic for testing the Poisson model over ZIP is given by  $-2 \log \Lambda = 1108.08$ , which is highly significant. The LRT statistic  $-2 \log \Lambda = 9.15$  shows that the ZOIP model is significantly better than the ZIP. Hence, both the AIC and LRT criteria show that ZOIP fits best for these data.

The observed and expected frequencies of the ZOIP, ZIP, and Poisson models are in Table 7. The ZIP model is able to capture the inflation at count zero. However, the ZOIP model is able to capture the inflation at count zero and one as well. The conclusions are also supported by the sum of ABE measure. In conclusion, the ZOIP model gives a good fit to the observed data.

**Table 7.** Frequency comparisons for ER data.

Count	Observed	ZOIP	ZIP	Poisson
0	10,046	10,047.60	10,049.72	9450.22
1	1466	1465.98	1436.96	2499.80
2–3	548	523.82	588.65	356.52
4–5	92	170.26	162.21	34.31
6–7	37	41.01	33.08	2.43
8–9	12	6.92	4.65	0.12
10–12	13	0.86	0.47	0.00
>12	9	0.21	0.10	0.00
Sum of ABE		134.07	176.32	1947.20
$\chi^2$		586.40	1150.84	304,381.20

## 6. Discussion

In this article, we developed the Expectation–Maximization (EM) algorithm for the ZkIP model generalizing the results of seminal work by Lambert [3]. The EM algorithm is a

computationally simpler approach to get the estimates of the unknown model parameters. However, unlike Lambert [3], we obtain the standard errors of the parameters using the approach developed specifically for the EM algorithm by Louis [4], which we believe is the right approach. We demonstrate our methods on two real-life data, showing that, in count data, if there is inflation at two points zero and  $k$ , then ZkIP outperforms the simpler Poisson models, ZIP, and Poisson according to AIC and LRT criteria.

In our regression model, for simplicity, we assumed the Poisson parameter depends on the covariates, and the obvious extension is linking the covariates to the inflation parameters  $\pi_1$  and  $\pi_2$  as well. In that case, Equation (6) becomes  $\log(\pi_1/\pi_3) = u_i^T \gamma$  and  $\log(\pi_2/\pi_3) = v_i^T \delta$ . The covariate vectors  $u_i$  and  $v_i$  may or may not be the same. This will result in higher dimensionality of the information matrix. The variable selection methodologies in this article could be used to obtain a simpler model. Other possible extensions are obtained replacing the Poisson distribution with generalized Poisson or Conway–Maxwell Poisson (CMP). In particular, we could implement the EM algorithm for the ZkICMP model studied by [32]. These extensions are currently our work in progress.

**Author Contributions:** Conceptualization, N.R.C.; data curation, M.A. and N.R.C.; writing—original draft, M.A. and N.R.C. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable

**Informed Consent Statement:** Not applicable

**Data Availability Statement:** The data that support the findings of this study are openly available in “Data Files” at [https://www.cdc.gov/nchs/nhis/nhis\\_2015\\_data\\_release.htm](https://www.cdc.gov/nchs/nhis/nhis_2015_data_release.htm) (accessed on 21 August 2021).

**Conflicts of Interest:** The authors declare no conflict of interest.

## Appendix A

As discussed in Section 3.2, the information matrix of the complete data can be used to get the standard errors of the EM estimates. Here, we provide the elements of the information matrix of the complete and missing data.

### Appendix A.1. Information Matrix of the Complete Data

The elements of the matrix  $E(\mathcal{I}_{comp}|\mathbf{y})$  are the expected values of the negative of second-order partial derivatives of the complete data log-likelihood (10), and they are given by

$$\begin{aligned} E \left[ \frac{-\partial^2 \ell_{comp}}{\partial \beta \partial \beta^T} \right] &= \sum_{i=1}^n \frac{[p_{0i}(\lambda_i) p_{ki}(\lambda_i) - e^{\gamma+\delta}] \lambda_i}{[e^{\gamma} + p_{0i}(\lambda_i)] [e^{\delta} + p_{ki}(\lambda_i)]} (\mathbf{x}_i \mathbf{x}_i^T) \\ E \left[ \frac{-\partial^2 \ell_{comp}}{\partial \gamma^2} \right] &= \frac{ne^{\gamma}(1 + e^{\delta})}{(1 + e^{\gamma} + e^{\delta})^2} \\ E \left[ \frac{-\partial^2 \ell_{comp}}{\partial \gamma \partial \delta} \right] &= \frac{-ne^{\gamma+\delta}}{(1 + e^{\gamma} + e^{\delta})^2} \\ E \left[ \frac{-\partial^2 \ell_{comp}}{\partial \delta^2} \right] &= \frac{ne^{\delta}(1 + e^{\gamma})}{(1 + e^{\gamma} + e^{\delta})^2}. \end{aligned}$$

The two elements  $-\partial^2 \ell_{comp} / \partial \beta \partial \gamma$  and  $-\partial^2 \ell_{comp} / \partial \beta \partial \delta$  are equal to zero and the other elements are obtained by symmetry.



### Appendix A.2. Information Matrix of the Missing Data

The elements of the matrix  $E(\mathcal{I}_{miss}|_{observed})$  are the negative of the expected value of second-order derivatives of (21). These are given by the following equations:

$$\begin{aligned} E\left[\frac{-\partial^2 \ell_{miss}}{\partial \beta \partial \beta^T}\right] &= -E\left[\left(\sum_{i=1}^n z_{3i} \lambda_i \mathbf{x}_i \mathbf{x}_i^T - \sum_{y_i \neq 0, k} \lambda_i \mathbf{x}_i \mathbf{x}_i^T \right. \right. \\ &\quad \left. \left. - \sum_{i: y_i=0} \frac{e^\gamma p_{0i}(1-\lambda_i) + p_{0i}^2}{(e^\gamma + p_{0i})^2} \lambda_i \mathbf{x}_i \mathbf{x}_i^T \right. \right. \\ &\quad \left. \left. - \sum_{i: y_i=k} \frac{e^\delta p_{ki}(\lambda_i - (k-\lambda_i)^2) + p_{ki}^2 \lambda_i}{(e^\delta + p_{ki})^2} \mathbf{x}_i \mathbf{x}_i^T \right) | \mathbf{y}\right] \\ &= \sum_{i=1}^n \frac{p_{0i} p_{ki} - e^{\gamma+\delta}}{(e^\gamma + p_{0i})(e^\delta + p_{ki})} \lambda_i \mathbf{x}_i \mathbf{x}_i^T - \sum_{i: y_i=0} \frac{e^\gamma p_{0i}(1-\lambda_i) + p_{0i}^2}{(e^\gamma + p_{0i})^2} \lambda_i \mathbf{x}_i \mathbf{x}_i^T \\ &\quad - \sum_{i: y_i=k} \frac{e^\delta p_{ki}(\lambda_i - (k-\lambda_i)^2) + p_{ki}^2 \lambda_i}{(e^\delta + p_{ki})^2} \mathbf{x}_i \mathbf{x}_i^T \end{aligned}$$

and

$$\begin{aligned} E\left[\frac{-\partial^2 \ell_{miss}}{\partial \beta \partial \gamma}\right] &= \sum_{i: y_i=0} \frac{e^\gamma p_{0i} \lambda_i \mathbf{x}_i}{(e^\gamma + p_{0i})^2} \\ E\left[\frac{-\partial^2 \ell_{miss}}{\partial \beta \partial \delta}\right] &= - \sum_{i: y_i=k} \frac{e^\delta p_{ki} (k - \lambda_i) \mathbf{x}_i}{(e^\delta + p_{ki})^2} \\ E\left[\frac{-\partial^2 \ell_{miss}}{\partial \gamma^2}\right] &= \sum_{i: y_i=0} \frac{e^\gamma p_{0i}}{(e^\gamma + p_{0i})^2} \\ E\left[\frac{-\partial^2 \ell_{miss}}{\partial \gamma \partial \delta}\right] &= 0, \quad E\left[\frac{-\partial^2 \ell_{miss}}{\partial \delta^2}\right] = \sum_{i: y_i=k} \frac{e^\delta p_{ki}}{(e^\delta + p_{ki})^2}. \end{aligned}$$

$E(\mathcal{I}_{miss}|_{observed})$  is a symmetric matrix, so the remaining off-diagonal elements can be easily obtained.

### References

1. Lin, T.H.; Tsai, M.H. Modeling health survey data with excessive zero and  $k$  responses. *Stat. Med.* **2012**, *32*, 1572–1583. [\[CrossRef\]](#)
2. Sheth-Chandra, M.; Chaganty, N.R.; Sabo, R.T. *A Doubly Inflated Poisson Distribution and Regression Model*; Springer International Publishing: Berlin/Heidelberg, Germany, 2019; pp. 131–145.
3. Lambert, D. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* **1992**, *34*, 1–14. [\[CrossRef\]](#)
4. Louis, T.A. Finding the observed information matrix when using the EM algorithm. *J. R. Stat. Soc. Ser. B (Methodol.)* **1982**, *44*, 226–233.
5. Cohen, A.C. Estimating the parameters of a modified Poisson distribution. *J. Am. Stat. Assoc.* **1960**, *55*, 139–143. [\[CrossRef\]](#)
6. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B (Methodol.)* **1977**, *39*, 1–22.
7. Min, Y.; Agresti, A. Random effect models for repeated measures of zero-inflated count data. *Stat. Model.* **2005**, *5*, 1–19. [\[CrossRef\]](#)
8. Yau, K.; Lee, A. Zero-inflated Poisson regression with random effects to evaluate an occupational injury prevention programme. *Stat. Med.* **2001**, *20*, 2907–2920. [\[CrossRef\]](#)
9. Ghosh, S.K.; Mukhopadhyay, P.; Lu, J.C. Bayesian analysis of zero-inflated regression models. *J. Stat. Plan. Inference* **2006**, *136*, 1360–1375. [\[CrossRef\]](#)
10. Agarwal, D.K.; Gelfand, A.E.; Citron-Pousty, S. Zero-inflated models with application to spatial count data. *Environ. Ecol. Stat.* **2002**, *9*, 341–355. [\[CrossRef\]](#)
11. Saffari, S.E.; Adnan, R. Zero-inflated Poisson regression models with right censored count data. *Matematika* **2011**, *27*, 21–29.



12. Yang, Y.; Simpson, D.G. Conditional decomposition diagnostics for regression analysis of zero-inflated and left-censored data. *Stat. Methods Med. Res.* **2012**, *21*, 393–408. [[CrossRef](#)] [[PubMed](#)]
13. Nguyen, V.T.; Dupuy, J.F. Asymptotic results in censored zero-inflated Poisson regression. *Commun. Stat.-Theory Methods* **2021**, *50*, 2759–2779. [[CrossRef](#)]
14. Altun, E. A new zero-inflated regression model with application. *J. Stat. Stat. Actuar. Sci.* **2018**, *2*, 73–80.
15. Bakouch, H.; Chesneau, C.; Karakaya, K.; Kuş, C. The Cos-Poisson model with a novel count regression analysis. *Hacet. J. Math. Stat.* **2021**, *50*, 559–578. [[CrossRef](#)]
16. Gupta, P.L.; Gupta, R.C.; Tripathi, R.C. Analysis of zero-adjusted count data. *Comput. Stat. Data Anal.* **1996**, *23*, 207–218. doi: 10.1016/S0167-9473(96)00032-1. [[CrossRef](#)]
17. Umbach, D. On inference for a mixture of a Poisson and a degenerate distribution. *Commun. Stat.-Theory Methods* **1981**, *10*, 299–306. [[CrossRef](#)]
18. Ridout, M.; Demetrio, C.; Hinde, J. Models for count data with many zeros. In Proceedings of the International Biometric Conference, Cape Town, South Africa, 14–18 December 1998.
19. Welsh, A.; Cunningham, R.; Donnelly, C.; Lindenmayer, D. Modelling the abundance of rare species: Statistical models for counts with extra zeros. *Ecol. Model.* **1996**, *88*, 297–308. [[CrossRef](#)]
20. Atkins, D.; Gallop, R. Rethinking how family researchers model infrequent outcomes: A tutorial on count regression and zero-inflated models. *J. Fam. Psychol.* **2007**, *21*, 726–735. [[CrossRef](#)] [[PubMed](#)]
21. Loeys, T.; Moerkerke, B.; De Smet, O.; Buysse, A. The analysis of zero-inflated count data: Beyond zero-inflated Poisson regression. *Br. J. Math. Stat. Psychol.* **2012**, *65*, 163–180. [[CrossRef](#)] [[PubMed](#)]
22. Salehi, M.; Roudbari, M. Zero-inflated Poisson and negative binomial regression models: Application in education. *Med. J. Islam. Repub. Iran* **2015**, *29*, 297.
23. Cameron, A.C.; Trivedi, P.K. *Regression Analysis of Count Data*; Cambridge Press: London, UK, 2013.
24. Greene, W. *Accounting for Excess Zeros and Sample Selection in Poisson And Negative Binomial Regression Models*; Department of Economics, New York University: New York, NY, USA, 1994.
25. Gurmu, S.; Trivedi, P. Excess zeros in count models for recreational trips. *J. Bus. Econ. Stat.* **1996**, *14*, 469–477.
26. Lord, D.; Washington, S.P.; Ivan, J.N. Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: Balancing statistical fit and theory. *Accid. Anal. Prev.* **2005**, *37*, 35–46. [[CrossRef](#)] [[PubMed](#)]
27. Qin, X.; Ivan, J.N.; Ravishanker, N. Selecting exposure measures in crash rate prediction for two-lane highway segments. *Accid. Anal. Prev.* **2004**, *36*, 183–191. [[CrossRef](#)]
28. Chen, K.; Huang, R.; Chan, N.H.; Yau, C.Y. Subgroup analysis of zero-inflated Poisson regression model with applications to insurance data. *Insur. Math. Econ.* **2019**, *86*, 8–18. [[CrossRef](#)]
29. Motalebi, N.; Owlia, M.S.; Amiri, A.; Fallahnezhad, M.S. Monitoring social networks based on zero-inflated Poisson regression model. *Commun. Stat.-Theory Methods* **2021**. [[CrossRef](#)]
30. Bohning, D.; Seidel, W. Editorial: Recent developments in mixture models. *Comput. Stat. Data Anal.* **2003**, *41*, 349–357. [[CrossRef](#)]
31. Hall, D.B. Zero-inflated Poisson and binomial regression with random effects: A case study. *Biometrics* **2000**, *56*, 1030–1039. [[CrossRef](#)]
32. Arora, M.; Chaganty, N.R.; Sellers, K.F. A flexible regression model for zero- and k-inflated count data. *J. Stat. Comput. Simul.* **2021**, *91*, 1815–1845. [[CrossRef](#)]
33. Zhang, C.; Tian, G.L.; Ng, K. Properties of the zero- and one-inflated Poisson distribution and likelihood based inference methods. *Stat. Its Interface* **2016**, *9*, 11–32. [[CrossRef](#)]
34. Alshkaki, R.S.A. On the zero-one inflated Poisson distribution. *Int. J. Stat. Distrib. Appl.* **2016**, *2*, 42–48.
35. Tang, Y.; Liu, W.; Xu, A. Statistical inference for zero- and one-inflated Poisson models. *Stat. Theory Relat. Fields* **2017**, *1*, 216–226. [[CrossRef](#)]
36. Liu, W.; Tang, Y.; Xu, A. Zero- and one-inflated Poisson regression model. *Stat. Pap.* **2021**, *62*, 915–934. [[CrossRef](#)]
37. Tsai, M.H.; Lin, T.H. Modeling data with a truncated and inflated Poisson distribution. *Stat. Methods Appl.* **2017**, *26*, 383–401. [[CrossRef](#)]
38. Finkelstein, M.D.; Green, J.G.; Gruber, M.J.; Zaslavsky, A.M. A zero- and k-inflated mixture model for health questionnaire data. *Stat. Med.* **2011**, *30*, 1028–1043. [[CrossRef](#)] [[PubMed](#)]
39. Chant, D. On asymptotic tests of composite hypotheses in nonstandard conditions. *Biometrika* **1974**, *61*, 291–298. [[CrossRef](#)]
40. Shapiro, A. Asymptotic distribution of test statistics in the analysis of moment structures under inequality constraints. *Biometrika* **1985**, *72*, 133–144. [[CrossRef](#)]
41. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Autom. Control* **1974**, *19*, 716–723. [[CrossRef](#)]
42. Burnham, K.P.; Anderson, D. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*; Springer: New York, NY, USA, 2002.