P-ISSN 2586-9000
E-ISSN 2586-9027

Homepage : *https://tci-thaijo.org/index.php/SciTechAsia* **Science & Technology Asia**

Vol. 25 No.2 April - June 2020 Page: [1-14]

Original research article

Wald Confidence Intervals for the Parameter in a Bernoulli Component of Zero-Inflated Poisson and Zero-Altered Poisson Models with Different Link Functions

Patchanok Srisuradetchai*, Sunisa Junnumtuam

Department of Mathematics and Statistics, Faculty of Science and Technology,
Thammasat University, Pathum Thani 12120, Thailand

Received 5 March 2019; Received in revised form 3 July 2019
Accepted 24 July 2019; Available online 29 June 2020

ABSTRACT

This paper aims to study the Wald confidence intervals for the parameters in a Bernoulli component of the zero-inflated Poisson (ZIP) and zero-altered Poisson (ZAP) models. The effects of the model choices between ZIP and ZAP with three different link functions: logit, probit, and complementary log-log, are investigated. Akaike's information criterion (AIC) is normally used for comparing models with different links in the literature. However, use of AIC is not advisable for the model comparison of non-nested models. To study the performance of the confidence intervals with different links, the coverage probability (CP) should be used because the AIC criterion can be misleading. The effects of the parameters in ZIP and ZAP distributions were also studied. The CPs are estimated from Monte-Carlo simulations, where the data are generated from both ZIP and ZAP distributions. The results show that when the employed model corresponds to the distribution of data, the link function and parameters of the distributions do not have much impact on the CPs. Conversely, if the wrong model is used, the components of the Bernoulli and the mean of count data are essential to determine the CPs of the intervals. Overall, the ZIP models tend to outperform the ZAP models, and the Wald confidence intervals with different links have approximately the same performance, regardless of any model used for fitting the data. If the mean of positive counts is large, both the ZAP and ZIP models tend to produce the same CPs or have the same performance.

Keywords: Count data; Hurdle model; Link functions; Wald Confidence Interval

***Corresponding author:** spatchan@tu.ac.th

doi: 10.14456/scitechasia.2020.16

1. Introduction

Count data are often applied in daily life, for example, the number of people who went to a certain hospital last year or the number of accidents that occurred at one intersection. These numbers are considered non-negative integer-valued random variables, and count models are typically employed for fitting such data. In recent years, many disciplines, including actuarial science, engineering, economics, and political science, have extensively used advanced statistical modelling methods, as a number of statistical programs are available, many of them at no cost. The most commonly used count models are Poisson and Negative binomial; however, when count data include excess zeros, these models may not be appropriate. Many models such as the zero-inflated Poisson model (ZIP) and the hurdle model (also called the zero-altered Poisson model, or ZAP) are more appropriate. The general form of hurdle regression models, consisting of two components: zero and non-zero parts, was proposed by Mullahy [1]. This assumes that all zero data are from one source and positive data follow either truncated Poisson distribution or truncated negative binomial distribution. In particular, when the "hurdle" is not crossed with probability π , we observe a count of zero, and if the hurdle is crossed, positive counts are observed with the count density $f(y)$. In some situations, the ZAP predicts too few zeros [2]. The ZIP model, proposed by Lambert [3], is considered a finite mixture model that can be used for zero-inflated data.

Choosing an appropriate model for a particular set of data is a challenging problem. This has led many researchers to compare and attempt to find an appropriate model for count data with excess zeros, such as the negative binomial model, the ZIP model, the zero-inflated negative binomial (ZINB) model, and the ZAP model. Ridout *et al.* [4] applied the ZIP and ZINB models to real data having an over-dispersion problem and used two criteria, namely AIC and BIC, for comparing the models; the results showed that the ZINB is the most appropriate model. Miller [5] employed the Poisson model, the ZAP model, and the ZIP model to fit count data with excess zeros generated from negatively skewed, positively skewed, and normal distributions and compared them by using the deviance statistic and Akaike's Information Criterion (AIC) as the criteria. The results showed that the ZAP model tends to have the best performance among all skewed data; however, in some cases, other models are better. In addition, Yang [6] used the ordinary least-squares regression model, the Poisson regression model, the negative binomial regression model, ZIP, ZINB, ZAP models, and the zero-altered negative binomial regression model (ZANB) to fit simulated data and real data. The AICs were compared in order to show the effectiveness of the models under different conditions regarding the proportion of zeros and the over-dispersion of the data.

In general, the ZIP and ZAP models have structures similar to GLMs, in which a link function connects a random component with linear predictors. The link function g that transforms the mean λ to the natural parameter η is called the canonical link, where $\eta = \sum_{j=1}^{p} \beta_j x_j$ and p is the number of regressors. The canonical link function can be obtained from the exponential of the response's density function [7]. For the Poisson distribution, the canonical link is a log link, and it is the logit link for the parameter π. However, the canonical link cannot guarantee a good fit for all data, as there are statistical tests to protect against link misspecification, and usually these are large sample tests such as the likelihood ratio test [8].

In the literature, all research studies involving comparisons of link functions have been conducted for binary data. Koenker and

Yoon [9] compared the Gosset link with the Pregibon link for binary data and explored the Bayesian and maximum likelihood methods for estimation and inference, and it was found that the misspecification of the link function can create serious bias. Li [10] compared three link functions: the logit, probit, and cloglog, and each link was used with different models; the information indices, including BIC and AIC, the posterior predictive distribution in a Bayesian approach, and the receiver operating characteristic (ROC), were considered as the criteria for the comparisons. The results showed that the probit and logit link functions are appropriate for symmetric data, while the cloglog link performs well for asymmetric data. Gunduz and Fokoue [11] proposed the definition of both structural and predictive equivalence of link functions-based binary regression models and explored the various ways in which they are either similar or dissimilar. From the predictive view, it was shown that not only are the probit and logit links perfectly equivalent, but the other link functions such as Cauchit and cloglog also have a high percentage of predictive equivalence. Damisa *et al.* [12] compared the logit, probit, and cloglog link functions for binary data where the sample size is small (< 1000) under symmetric and asymmetric assumptions. When using the AIC as the criterion, the probit link is preferred under the symmetric assumption, while the cloglog should be used under the asymmetric assumption. Wu and Lord [13] examined the influence of link function misspecification in the regression models for real data and found that the misuse of the link function for one or more variables can result in biased estimates.

However, there has been no research studying the proper link functions of π (the probability of occurring zeros) for count data with excess zeros. It has also been found in the literature that AIC is generally used for model comparisons. Unfortunately, standard criteria, such as AIC or the likelihood ratio test, are not appropriate for model comparison of non-nested models [14]. Therefore, the AIC may not be sufficient to determine the accuracy of the estimates of π, which have never been studied in the literature. In this paper, the Wald confidence intervals of parameter π will be constructed and the corresponding CPs will be investigated when only count responses are available. In addition, the ZIP and ZAP models will be compared with three different link functions, namely the logit, probit, and cloglog, by using the simulated data from ZIP and ZAP distributions. In the perspective of statistical theory, using the proper link function is equivalent to improving the quadratic approximation of the normalized log-likelihood around $\hat{\pi}$:

$$\log\frac{L(\hat{\pi})}{L(\pi)} \approx -\frac{1}{2}I(\hat{\pi})(\pi-\hat{\pi})^2 \qquad (1)$$

by choosing the appropriate transformation $\psi = g(\pi)$, e.g., the logit link that corresponds to $g(\pi)=\log(\pi/(1-\pi))$. The "quality" of the Wald confidence interval is based on two levels of approximation; the log-likelihood is approximated by a quadratic function and the confidence level is approximate [15]. The Wald interval can be written in the form:

$$\left\{\pi \left| -\frac{1}{2}I(\hat{\pi})(\pi-\hat{\pi})^2 \geq -\frac{1}{2}\chi^2_{1,(1-\alpha)}\right.\right\}$$
$$=\left\{\pi \left| \sqrt{I(\hat{\pi})}\left|\pi-\hat{\pi}\right| \leq z_{1-\alpha/2}\right.\right\}, \qquad (2)$$

where $I(\hat{\pi})$ is the observed Fisher information. The familiar form of the Wald interval for π is $\hat{\pi} \pm z_{1-\alpha/2}\sqrt{I^{-1}(\hat{\pi})} = \hat{\pi} \pm z_{1-\alpha/2}s.e.(\hat{\pi})$. If the transformation $\psi = g(\pi)$ is used, the improved confidence interval for $g(\pi)$ will be $g(\hat{\pi}) \pm z_{1-\alpha/2} \times$

$s.e.\left(g\left(\hat{\pi}\right)\right)$. To illustrate that the AIC and CP are not necessarily concordant, the CPs of the Wald confidence interval for π and the AICs are estimated by Monte-Carlo simulations and then investigated.

2. Materials and Methods

A brief description of the hurdle model and the zero-inflated Poisson models when there is no independent variable will be presented. Thus, the models in this paper are in the forms of $g_1\left(\pi\right)=\beta_1$ and $g_2\left(\lambda\right)=\beta_2$ where π and λ are the parameters in the Bernoulli and Poisson components, respectively.

2.1 Zero-altered poisson models

The ZAPs (hurdle models) are composed of two parts. The first component is responsible for generating all zeros and the second component generates positive counts. The hurdle model uses a binomial logistic regression model to assign a probability π that will determine whether a count will be zero or positive. If the positive count is recognized, then the 'hurdle' is crossed, and the counts are modelled by a truncated-at-zero count model. The zero-truncated density is $f(y)/(1-f(0))$, and the unconditional probability mass function for Y is

$$\mathrm{P}\left(Y=y\right)=\begin{cases} \pi & \text{if } y=0 \\ \dfrac{1-\pi}{1-f(0)}f(y) & \text{if } y>0 \end{cases}.$$

If the zero-truncated Poisson distribution is chosen, the unconditional probability mass function for Y is

$$\mathrm{P}\left(Y=y\right)=\begin{cases} \pi & ,\text{ if } y=0 \\ \left(\dfrac{1-\pi}{1-e^{-\lambda}}\right)\dfrac{e^{-y}\lambda^{y}}{y!} & ,\text{ if } y\geq 1 \end{cases} \quad (3)$$

If the logit link is used for π and the log link is used for λ, without independent variables, π will be $e^{\beta_1}\big/\left(1+e^{\beta_1}\right)$ and λ will be e^{β_2}. Define a binary indicator as follows:

$$d_i=\begin{cases} 0, & \text{if } y_i=0 \\ 1, & \text{if } y_i\geq 1 \end{cases}.$$

In general, $\pi=g_1^{-1}(\beta_1)$ and $\lambda=g_2^{-1}(\beta_2)$. Thus, the log-likelihood function is

$$\begin{aligned}
&\log L\left(\beta_1,\beta_2|y_1,\ldots,y_n\right)\\
&=\log\left\{\begin{matrix}\displaystyle\prod_{i=1,2,\ldots,n}\left[g_1^{-1}(\beta_1)\right]^{1-d_i}\\ \left[\left(\dfrac{1-g_1^{-1}(\beta_1)}{1-e^{-g_2^{-1}(\beta_2)}}\right)\dfrac{e^{-g_2^{-1}(\beta_2)}\left(g_2^{-1}(\beta_2)\right)^{y_i}}{y_i!}\right]^{d_i}\end{matrix}\right\}\\
&=\left\{\begin{matrix}\left(n-\displaystyle\sum_{i=1}^{n}d_i\right)\log\left(g_1^{-1}(\beta_1)\right)+\\ \left(\displaystyle\sum_{i=1}^{n}d_i\right)\log\left(1-g_1^{-1}(\beta_1)\right)\end{matrix}\right\}+\\
&\left\{\sum_{i=1}^{n}\left[d_i\log\left(\frac{e^{-g_2^{-1}(\beta_2)}\left(g_1^{-1}(\beta_2)\right)^{y_i}}{\left(1-e^{-g_2^{-1}(\beta_2)}\right)y_i!}\right)\right]\right\}\\
&=L_1(\beta_1|y_1,\ldots,y_n)+L_2(\beta_2|y_1,\ldots,y_n) \qquad (4)
\end{aligned}$$

Thus, the joint likelihood can be maximized by separately maximizing each component [2]. In this paper, the function 'hurdle' in the package 'pscl' in R [16] is used to find the maximum likelihood estimates of β_1 and β_2. This package uses the method of Nelder and Mead [17] as a default method for optimization.

2.2 Zero-inflated poisson models

Unlike hurdle models, the zero-inflated Poisson models, or ZIPs, are finite mixture models with two components. The mixture weights for the two components are

π and $1-\pi$. The probability mass function of the ZIP model can be expressed as

$$Y \sim \begin{cases} \delta_0 & \text{, with Prob} = \pi \\ \text{Poisson}(\lambda) & \text{, with Prob} = 1-\pi, \end{cases}$$

where δ_0 is a degenerate distribution at 0, or

$$\mathrm{P}(Y=y)=\begin{cases} \pi+(1-\pi)e^{-\lambda} & \text{if } y=0 \\ (1-\pi)\dfrac{e^{-\lambda}\lambda^{y}}{y!} & \text{if } y\geq 1 \end{cases} \qquad (5)$$

The joint log-likelihood function for the two parts of the ZIP model is

$$\log L\left(\beta_1,\beta_2 | y_1,\ldots,y_n\right)=$$

$$\log\left\{\prod_{i=1,..,n} \left[g_1^{-1}(\beta_1)+(1-g_1^{-1}(\beta_1))e^{-g_2^{-1}(\beta_2)}\right]^{1-d_i} \left[(1-g_1^{-1}(\beta_1))\frac{e^{-g_2^{-1}(\beta_2)}\left(g_2^{-1}(\beta_2)\right)^{y_i}}{y_i!}\right]^{d_i}\right\}$$

$$=\log\left\{L\left(\beta_1,\beta_2 | y_1,\ldots,y_n\right)\right\} \qquad (6)$$

It can be seen that the log-likelihood function cannot be divided into two components as the ZAPs can. Thus, theoretically, both parameters are correlated. We use the function ‘zeroinfl’ in the package ‘pscl’ [16] in R to find the maximum likelihood estimates of β_1 and β_1. Similar to the function ‘hurdle’, the method of Nelder and Mead [17] is used.

2.3 Link functions

A link function relates the expected value of the response to the linear predictors in generalized linear models. If there is no independent variable, the link function can be considered as the transformation $g\left(\pi\right)$. In this study, the link functions are logit, probit, and cloglog links, as the following:

logit link: $g\left(\pi\right)=\log\left(\pi/(1-\pi)\right)$

probit link: $g\left(\pi\right)=\Phi^{-1}(\pi)$

cloglog link: $g\left(\pi\right)=\log(-\log(1-\pi))$

2.4 Wald confidence intervals

To construct the Wald confidence intervals, the Fisher information matrix, $I\left((\hat{\beta}_1,\hat{\beta}_2)\right)=\begin{bmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{bmatrix}$, is needed, where

$$I_{ij}=-\frac{\partial^2}{\partial\beta_i\partial\beta_j}L\left(\beta_1,\beta_2\right)\Bigg|_{\beta_i=\hat{\beta}_i,\beta_j=\hat{\beta}_j}$$

and $L\left(\beta_1,\beta_2\right)$ can be either Eq. (4) or Eq. (6), and the inverse of the Fisher information matrix is

$$I^{-1}\left((\hat{\beta}_1,\hat{\beta}_2)\right)=\begin{bmatrix} I^{11} & I^{12} \\ I^{21} & I^{22} \end{bmatrix}. \qquad (7)$$

Because $\beta_1=g_1\left(\pi\right)$ is of interest, only I_{11} is required. Thus, the Wald $(1-\alpha)100\%$ CI for β_1 is (L,U), where $L=\hat{\beta}_1-z_{1-\alpha/2}\sqrt{I^{11}}$ and $U=\hat{\beta}_1+z_{1-\alpha/2}\sqrt{I^{11}}$ [15]. If transforming β_1 to the original scale of π, the $(1-\alpha)100\%$ CI for π will be

$$\left(g^{-1}(L), g^{-1}(U)\right). \qquad (8)$$

3. Research Methodology

In this section, the method of simulating the data and constructing the Wald confidence intervals will be described.

3.1 Simulated data and models

Only the response values are generated. The π value of the Bernoulli distribution and the mean λ of the Poisson distribution are varied. Lambert [3] used the sample sizes of 25, 50, and 100, but the sample sizes equal to 25 and 50 resulted in the occurrence of singularities of the Fisher information matrix. Thus, the sample sizes in this study were set at 100. The estimated intercepts in the ZIP and ZAP models were estimated from the function ‘zeroinfl’ and the function ‘hurdle’, respectively. This will give the estimate of Eq. (7), so that the standard errors of $\hat{\beta}_1$ and $\hat{\beta}_2$ can be calculated from the square root of I^{11} and I^{22}, respectively.

The logit, probit, and cloglog link functions were applied in order to observe the effects of the link specification. The conditions of this study were as follows: (1) the parameter π was set as 0.1, 0.2, …, 0.9; (2) the parameter λ was set as 1, 2, 5, and 10; and (3) the number of simulated data was set at 3,000 for each combination of the π, λ, link function, and distribution (ZIP or ZAP).

3.2 Comparisons of the models

The coverage probabilities of π from the 95% Wald confidence intervals were estimated in order to study the effects of the links, models, parameters π and λ. The general form of the confidence intervals is as seen in Eq. (8). If $g(\pi)=\log(\pi/(1-\pi))$ $=\beta_1$, the $(1-\alpha)$ 100% Wald's confidence interval for π is

$$\frac{\exp\left(\hat{\beta}_1 \pm z_{1-\alpha/2}\sqrt{I^{11}}\right)}{1+\exp\left(\hat{\beta}_1 \pm z_{1-\alpha/2}\sqrt{I^{11}}\right)}, \qquad (9)$$

where I^{11} is in Eq. (7). In the same manner, the $(1-\alpha)$100% Wald's confidence interval for π constructed using the probit link is

$$\Phi\left[\hat{\beta}_1 \pm z_{1-\frac{\alpha}{2}}\sqrt{I^{11}}\right], \qquad (10)$$

where Φ is the cumulative distribution function of the standard normal distribution. Also, the $(1-\alpha)$100% Wald's confidence interval for π constructed using the cloglog link is

$$1-\exp\left[-\exp\left(\hat{\beta}_1 \pm z_{1-\alpha/2}\sqrt{I^{11}}\right)\right] \quad (11).$$

And, the coverage probability was estimated from $CP=\sum_{i=1}^{3000} I_{[L_i,U_i]}(\pi) \Big/ 3000$, where (L_i,U_i) is a CI in the *i*th iteration. To show that AIC and CP are not necessarily in agreement, it was necessary to calculate

$$\mathrm{AIC}=-2\log(\text{likelihood})+2k,$$

where k is the number of parameters to be estimated. The model that gives the lowest AIC value will be considered the best model. Note that the use of this criterion is not advisable for comparisons of non-nested models [14].

4. Results and Discussion

The CPs of π's in the ZAP models are presented in Tables 1 and 2, and the corresponding AICs are shown respectively in Tables 5 and 6, whereas the CPs and AICs of the ZIP models are shown in Tables 3, 4, 7, and 8. It can be seen that the different link functions give nearly the same estimated CPs, especially when the distribution generating the data and the model used to fit are the same. For example, in Table 1 when π equals 0.8 and λ equals 2, the CPs are approximately 0.95, and in Table 4 when π equals 0.1 and λ equals 1, the CPs are all near 0.91. The differences among the links will be noticeable if the model is wrong or does not correspond to the distribution generating the data. For example, in Table 2 when π equals 0.8 and λ equals 1, both the logit and probit links have a CP close to 0.49, while the cloglog gives 0.64. These are much lower than the desired confidence level, 0.95. In many cases, the CP values are not very different, especially when λ is increasingly higher.

Table 1. The estimated coverage probabilities of π_i in ZAP models when the data are generated from ZAP distribution.

π	$\lambda = 1$			$\lambda = 2$			$\lambda = 5$			$\lambda = 10$		
	logit	**probit**	**cloglog**	**logit**	**probit**	**cloglog**	**logit**	**probit**	**cloglog**	**logit**	**probit**	**cloglog**
0.1	0.953	0.958	0.951	0.957	0.958	0.957	0.953	0.953	0.959	0.955	0.959	0.962
0.2	0.938	0.958	0.964	0.948	0.954	0.955	0.943	0.946	0.951	0.939	0.950	0.953
0.3	0.958	0.947	0.951	0.960	0.949	0.950	0.969	0.953	0.949	0.960	0.947	0.950
0.4	0.951	0.943	0.960	0.947	0.950	0.961	0.943	0.948	0.959	0.945	0.939	0.961
0.5	0.939	0.945	0.956	0.938	0.937	0.954	0.947	0.939	0.953	0.942	0.942	0.950
0.6	0.945	0.953	0.952	0.947	0.947	0.959	0.951	0.948	0.960	0.952	0.948	0.957
0.7	0.957	0.954	0.953	0.961	0.954	0.953	0.965	0.953	0.954	0.961	0.952	0.956
0.8	0.944	0.954	0.953	0.940	0.952	0.955	0.932	0.964	0.952	0.946	0.956	0.953
0.9	0.953	0.958	0.957	0.954	0.961	0.947	0.954	0.953	0.954	0.960	0.954	0.954

Table 2. The estimated coverage probabilities of π_i in ZAP models when the data are generated from ZIP distribution.

π	$\lambda = 1$			$\lambda = 2$			$\lambda = 5$			$\lambda = 10$		
	logit	**probit**	**cloglog**	**logit**	**probit**	**cloglog**	**logit**	**probit**	**cloglog**	**logit**	**probit**	**cloglog**
0.1	0.000	0.000	0.000	0.048	0.082	0.081	0.934	0.956	0.941	0.956	0.953	0.953
0.2	0.000	0.000	0.000	0.247	0.325	0.300	0.938	0.952	0.951	0.941	0.952	0.953
0.3	0.001	0.000	0.002	0.507	0.500	0.510	0.962	0.953	0.950	0.966	0.943	0.950
0.4	0.004	0.006	0.009	0.596	0.614	0.686	0.943	0.942	0.962	0.951	0.947	0.960
0.5	0.031	0.031	0.041	0.708	0.697	0.774	0.944	0.944	0.961	0.942	0.945	0.950
0.6	0.121	0.116	0.170	0.802	0.805	0.863	0.948	0.948	0.957	0.948	0.950	0.959
0.7	0.335	0.248	0.349	0.893	0.854	0.882	0.960	0.950	0.959	0.959	0.950	0.957
0.8	0.498	0.491	0.643	0.893	0.901	0.936	0.944	0.954	0.952	0.937	0.956	0.951
0.9	0.885	0.762	0.875	0.973	0.929	0.957	0.947	0.956	0.959	0.954	0.954	0.948

Table 3. The estimated coverage probabilities of π_i in ZIP models when the data are generated from ZAP distribution.

π	$\lambda = 1$			$\lambda = 2$			$\lambda = 5$			$\lambda = 10$		
	logit	**probit**	**cloglog**	**logit**	**probit**	**cloglog**	**logit**	**probit**	**cloglog**	**logit**	**probit**	**cloglog**
0.1	1.000	1.000	1.000	1.000	1.000	1.000	0.973	0.966	0.975	0.947	0.956	0.955
0.2	1.000	1.000	1.000	1.000	0.999	1.000	0.959	0.957	0.948	0.935	0.959	0.953
0.3	1.000	1.000	1.000	0.795	0.707	0.876	0.951	0.949	0.951	0.964	0.951	0.940
0.4	0.989	0.965	0.998	0.750	0.708	0.782	0.956	0.955	0.952	0.947	0.949	0.959
0.5	0.732	0.638	0.893	0.784	0.770	0.814	0.947	0.953	0.946	0.940	0.946	0.953
0.6	0.521	0.475	0.701	0.816	0.811	0.855	0.949	0.950	0.952	0.950	0.955	0.966
0.7	0.540	0.556	0.698	0.844	0.856	0.895	0.950	0.953	0.954	0.959	0.951	0.958
0.8	0.666	0.692	0.782	0.888	0.896	0.913	0.959	0.937	0.954	0.937	0.954	0.956
0.9	0.824	0.847	0.901	0.918	0.920	0.928	0.955	0.958	0.959	0.953	0.950	0.962

Table 4. The estimated coverage probabilities of π_i in ZIP models when the data are generated from ZIP distribution.

π	$\lambda = 1$			$\lambda = 2$			$\lambda = 5$			$\lambda = 10$		
	logit	**probit**	**cloglog**	**logit**	**probit**	**cloglog**	**logit**	**probit**	**cloglog**	**logit**	**probit**	**cloglog**
0.1	0.913	0.918	0.902	0.953	0.954	0.940	0.965	0.951	0.966	0.946	0.954	0.956
0.2	0.943	0.951	0.931	0.963	0.967	0.960	0.954	0.957	0.953	0.934	0.950	0.948
0.3	0.958	0.959	0.943	0.958	0.965	0.964	0.947	0.948	0.953	0.961	0.951	0.954
0.4	0.957	0.962	0.956	0.962	0.958	0.961	0.953	0.955	0.947	0.949	0.942	0.961
0.5	0.965	0.964	0.950	0.960	0.946	0.946	0.958	0.953	0.943	0.944	0.941	0.955
0.6	0.971	0.965	0.956	0.948	0.953	0.953	0.956	0.944	0.950	0.948	0.952	0.961
0.7	0.970	0.968	0.955	0.959	0.954	0.953	0.944	0.957	0.943	0.959	0.956	0.946
0.8	0.973	0.974	0.962	0.956	0.959	0.949	0.955	0.949	0.955	0.942	0.959	0.954
0.9	0.978	0.971	0.973	0.960	0.960	0.966	0.954	0.957	0.959	0.951	0.958	0.959

Table 5. The average AICs of ZAP models when the data are generated from ZAP distribution.

π	$\lambda = 1$			$\lambda = 2$			$\lambda = 5$			$\lambda = 10$		
	logit	**probit**	**cloglog**	**logit**	**probit**	**cloglog**	**logit**	**probit**	**cloglog**	**logit**	**probit**	**cloglog**
0.1	251.24	251.06	251.49	339.45	339.29	339.05	459.32	459.25	458.89	527.84	528.20	528.17
0.2	265.52	265.78	265.80	344.35	344.28	344.36	450.73	450.69	450.83	511.87	511.92	511.89
0.3	267.25	267.29	267.28	336.25	335.82	335.56	429.54	428.92	429.39	482.85	482.75	482.99
0.4	259.55	259.46	259.21	318.20	317.99	318.12	397.69	397.73	398.19	443.33	443.73	444.19
0.5	243.21	242.73	243.08	292.39	291.69	291.71	359.45	358.18	358.51	396.79	396.94	396.99
0.6	218.27	218.83	217.99	257.12	257.03	257.50	310.76	311.60	310.81	341.18	341.81	341.15
0.7	185.11	185.47	185.90	215.34	215.08	214.93	255.29	255.63	254.69	278.07	277.80	278.25
0.8	143.13	143.01	142.59	162.42	162.25	162.90	189.01	189.21	188.52	205.47	203.57	204.96
0.9	87.20	87.45	87.23	96.81	97.49	97.03	110.65	111.31	110.81	118.23	117.81	118.33

Table 6. The average AICs of ZAP models when the data are generated from ZIP distribution.

π	$\lambda = 1$			$\lambda = 2$			$\lambda = 5$			$\lambda = 10$		
	logit	**probit**	**cloglog**	**logit**	**probit**	**cloglog**	**logit**	**probit**	**cloglog**	**logit**	**probit**	**cloglog**
0.1	255.23	255.00	255.90	343.59	343.37	343.36	458.86	458.90	459.00	527.82	527.95	527.93
0.2	244.56	243.74	243.81	335.41	335.15	334.75	449.27	450.02	449.80	511.91	511.84	512.04
0.3	230.20	229.99	229.68	319.25	319.49	319.33	428.00	427.89	427.77	482.80	482.10	482.65
0.4	212.44	212.77	212.27	296.90	297.65	297.28	396.65	395.88	395.99	444.31	443.40	444.46
0.5	191.32	191.62	191.09	269.48	269.19	269.68	357.68	356.54	357.42	396.56	395.64	396.30
0.6	167.02	167.00	166.78	235.52	235.91	235.79	309.02	310.25	309.73	341.39	341.15	340.91
0.7	137.69	137.34	138.37	194.63	195.14	194.09	253.36	254.11	253.37	277.79	277.03	277.67
0.8	103.37	103.10	104.10	146.59	146.78	146.57	188.91	187.85	188.73	204.09	204.47	204.32
0.9	62.68	62.10	61.25	87.29	86.03	86.57	110.01	110.05	109.73	119.29	118.09	118.39

Table 7. The average AICs of ZIP models when the data are generated from ZAP distribution.

π	$\lambda = 1$			$\lambda = 2$			$\lambda = 5$			$\lambda = 10$		
	logit	**probit**	**cloglog**	**logit**	**probit**	**cloglog**	**logit**	**probit**	**cloglog**	**logit**	**probit**	**cloglog**
0.1	274.07	274.55	274.46	341.56	341.26	341.17	459.45	459.00	459.10	527.77	527.77	528.52
0.2	274.28	273.81	273.65	344.48	344.06	343.88	450.15	450.68	450.28	511.64	511.36	511.89
0.3	269.36	269.82	269.26	335.92	335.82	336.04	429.16	429.02	428.96	483.11	482.73	482.80
0.4	259.78	259.88	259.56	318.11	317.82	318.49	397.88	398.73	398.04	443.80	443.09	443.97
0.5	242.81	242.75	242.82	291.83	292.00	291.62	358.96	358.89	359.21	396.25	397.08	397.46
0.6	217.84	218.54	218.74	257.49	257.74	257.73	310.68	311.21	311.64	341.31	341.00	341.51
0.7	185.82	185.82	185.18	215.49	215.05	214.76	255.03	254.38	254.98	278.71	277.66	277.04
0.8	143.99	142.94	142.74	161.46	162.20	162.71	189.09	190.74	189.69	205.18	204.12	204.33
0.9	87.61	87.68	87.60	97.32	97.98	97.67	110.33	111.08	110.51	119.37	117.74	117.66

Table 8. The average AICs of ZIP models when the data are generated from ZIP distribution.

π	$\lambda = 1$			$\lambda = 2$			$\lambda = 5$			$\lambda = 10$		
	logit	**probit**	**cloglog**	**logit**	**probit**	**cloglog**	**logit**	**probit**	**cloglog**	**logit**	**probit**	**cloglog**
0.1	255.76	255.65	255.04	343.63	342.89	343.52	459.20	458.86	459.38	527.96	528.11	527.93
0.2	244.38	244.30	244.36	335.03	334.46	334.78	450.22	449.52	449.72	511.94	511.53	511.54
0.3	230.31	230.29	229.35	319.30	319.42	319.60	428.31	427.56	427.78	482.88	482.60	481.88
0.4	212.22	212.01	211.97	297.92	297.42	297.27	396.53	395.79	396.40	443.60	443.50	444.20
0.5	191.43	191.78	191.84	269.46	269.53	268.71	357.18	357.33	356.77	397.41	396.66	397.11
0.6	166.86	167.46	166.28	235.80	235.38	235.59	309.54	310.32	309.45	341.40	341.74	340.64
0.7	138.09	137.93	137.53	194.95	195.46	195.40	253.19	254.28	253.27	278.26	277.13	278.53
0.8	103.74	104.28	103.67	146.01	146.30	147.22	187.97	188.52	188.12	203.93	204.81	204.25
0.9	62.24	62.00	62.19	86.33	86.47	87.06	110.15	109.66	109.50	119.47	118.42	118.71

Table 9. Proportions of zeros in the data generated from the ZAP and ZIP models.

π	**ZAP**				**ZIP**			
	$\lambda = 1$	$\lambda = 2$	$\lambda = 5$	$\lambda = 10$	$\lambda = 1$	$\lambda = 2$	$\lambda = 5$	$\lambda = 10$
0.1	0.10	0.10	0.10	0.10	0.43	0.22	0.11	0.10
0.2	0.20	0.20	0.20	0.20	0.50	0.31	0.20	0.20
0.3	0.30	0.30	0.30	0.30	0.57	0.40	0.30	0.30
0.4	0.40	0.40	0.40	0.40	0.63	0.49	0.41	0.40
0.5	0.50	0.50	0.50	0.50	0.69	0.58	0.51	0.51
0.6	0.60	0.60	0.60	0.60	0.76	0.66	0.61	0.61
0.7	0.70	0.70	0.70	0.70	0.82	0.75	0.71	0.71
0.8	0.80	0.80	0.80	0.80	0.88	0.83	0.81	0.81
0.9	0.90	0.90	0.90	0.90	0.94	0.92	0.91	0.91

Considering the AIC values, they may not be in agreement with the CPs. Having a low value of AIC does not mean that there is a high value in the CP. The differences in CPs can be significant, but corresponding AIC values are only slightly different. For example, in Table 2, when π equals 0.8 and λ equals 1, the CPs of the logit, probit, and cloglog links are 0.50, 0.49, and 0.64, respectively, but the corresponding AICs (seen in Table 6) are 103.37, 103.10, and 104.10, respectively. In Table 3, when π equals 0.9 and λ equals 1, the CPs of the logit, probit, and cloglog links are 0.82, 0.85, and 0.90, respectively, but the corresponding AIC values (shown in Table 7) are nearly the same, which is approximately 87.6. Therefore, the AIC does not represent the performance of the intervals or coverage probabilities, and this suggests that the AIC should not be used to evaluate the fitted models with different links.

When the model corresponds to the distribution, the CP values (as shown in Tables 1 and 4) are approximately 0.95 regardless of the λ values, π values, and link functions, except in the case where λ equals 1 and π equals 0.1, as seen in Table 4. In this case, the CPs are approximately 0.91, less than 0.95. In conclusion, if the model is correct (or agrees with the distribution), the link functions and values of λ and π do not significantly affect the coverage probabilities.

When the model used to fit the data is incorrect, the value of λ has a greatly significant effect on both the CP and AIC values. As seen in Table 2, when λ equals 1, the Wald confidence interval cannot contain the true π, especially when π is of a low value, i.e., $\pi = 0.1$, 0.2, or 0.3. This is because the average length of the intervals of ZAP is very short for all values of π and λ. Note that the value of parameter π is the same as the proportion of zeros in the population, as shown Table 9. This characteristic likely makes the ZAP model have a low value of $\sqrt{I^{11}}$, which appears in Eq. (9), Eq. (10), and Eq. (11). The ZIP distribution has two sources of zeros, so the ZIP tends to have a higher percentage of zeros than the ZAP, especially in cases with low values of λ. The proportion of zeros in the ZAP model is represented by π, but this is not true for ZIP distribution. For example, in Table 9, the ZIP distribution in which $\lambda = 1$ and $\pi = 0.1$ has zeros of about 43% while the ZAP has only 10%, which corresponds to $\pi = 0.1$. Moreover, it is also

Table 10. The average length of interval of π_i in ZAP models when the data are obtained from ZIP distribution.

π	$\lambda = 1$			$\lambda = 2$			$\lambda = 5$			$\lambda = 10$		
	logit	**probit**	**cloglog**	**logit**	**probit**	**cloglog**	**logit**	**probit**	**cloglog**	**logit**	**probit**	**cloglog**
0.1	0.191	0.191	0.191	0.162	0.161	0.160	0.123	0.120	0.118	0.120	0.117	0.115
0.2	0.193	0.193	0.193	0.178	0.178	0.178	0.157	0.156	0.155	0.156	0.156	0.154
0.3	0.191	0.192	0.193	0.189	0.189	0.189	0.178	0.178	0.178	0.177	0.177	0.177
0.4	0.187	0.187	0.188	0.192	0.193	0.193	0.189	0.190	0.189	0.189	0.189	0.189
0.5	0.180	0.180	0.182	0.191	0.191	0.192	0.193	0.193	0.193	0.193	0.193	0.193
0.6	0.168	0.168	0.170	0.183	0.184	0.185	0.189	0.189	0.190	0.189	0.189	0.190
0.7	0.153	0.152	0.154	0.170	0.170	0.171	0.177	0.177	0.178	0.177	0.177	0.178
0.8	0.131	0.129	0.134	0.148	0.147	0.150	0.155	0.155	0.158	0.156	0.155	0.158
0.9	0.101	0.097	0.102	0.114	0.111	0.115	0.120	0.117	0.121	0.120	0.117	0.121

Table 11. The average length of interval of π_i in ZIP models when the data are obtained from ZAP distribution.

π	$\lambda = 1$			$\lambda = 2$			$\lambda = 5$			$\lambda = 10$		
	logit	**probit**	**cloglog**	**logit**	**probit**	**cloglog**	**logit**	**probit**	**cloglog**	**logit**	**probit**	**cloglog**
0.1	1.000	1.000	1.000	0.939	0.926	0.956	0.122	0.120	0.124	0.119	0.117	0.121
0.2	0.999	0.999	1.000	0.402	0.361	0.429	0.157	0.157	0.160	0.156	0.155	0.158
0.3	0.970	0.952	0.977	0.238	0.231	0.249	0.178	0.179	0.181	0.177	0.177	0.179
0.4	0.757	0.723	0.787	0.238	0.237	0.241	0.191	0.191	0.192	0.189	0.189	0.190
0.5	0.533	0.520	0.582	0.237	0.238	0.240	0.194	0.195	0.195	0.193	0.193	0.194
0.6	0.432	0.421	0.455	0.231	0.232	0.233	0.190	0.191	0.190	0.189	0.189	0.189
0.7	0.382	0.388	0.399	0.217	0.217	0.216	0.178	0.179	0.178	0.178	0.178	0.176
0.8	0.343	0.352	0.341	0.191	0.190	0.190	0.157	0.156	0.155	0.156	0.155	0.153
0.9	0.294	0.298	0.296	0.155	0.151	0.148	0.121	0.119	0.117	0.120	0.117	0.116

observed that when λ equals 5, as seen in Table 2, the CPs are not very far from 0.95 for all π's. A plausible explanation is that the proportions of zeros from the ZIP and ZAP distributions are nearly the same. When this occurs, the ZAP model can be used for the data generated from the ZIP distribution.

Conversely, if the ZIP model is used when the data comes from the ZAP distribution, the CPs tend to be higher than 0.95 in many situations. In Table 3, when λ equals 1 and π equals 0.1, 0.2, 0.3, or 0.4, the CPs are approximately 1 for all link functions, which is much higher than 0.95. When π is greater than 0.4, the CP starts decreasing until π is about 0.6 and begins increasing again. Thus, it can be concluded that the ZIP model tends to give excessively conservative confidence intervals when π and λ are of low values. Such problems may result from the differences in the proportions of zeros, as the proportions of zeros in the data generated from the ZAP distribution are much lower than those from the ZIP distribution, especially when λ equals 1 and π is at a low value. Again, when the proportions of zeros from ZIP and ZAP distributions are nearly the same, e.g., ZAP and ZIP with $\lambda = 5,10$, the ZIP model can be used for data generated from the ZAP distribution. For example, with $\lambda = 5$ and $\pi = .3$, the percentage of zeros in the ZIP and ZAP distributions are approximately 30%, as seen in Table 9, and the corresponding CPs are approximately 0.95 for all links.

When examining further, the average length of intervals from the ZAP models tends be much smaller than that from the ZIP models, as shown in Tables 10 and 11. However, when λ is large, i.e., λ equals 5 or 10, the difference between the interval length of the ZIP and ZAP models is very small, regardless of the π values. In addition, there is less variability of the average length from the ZAP than that from the ZIP.

5. Real Data

To illustrate the applications of the Wald confidence intervals, the beetle egg-laying data is used as the example to calculate the CIs. Tauber *et al.* [18] observed a particular response called diapause in the females. They expected that the females might enter diapause and produce no eggs with a higher probability for some conditions than for others. One of the responses in the experiment was the number of eggs laid by the females not in diapause. The data were provided by Bilder and Loughin [19] and are presented in Fig. 1.

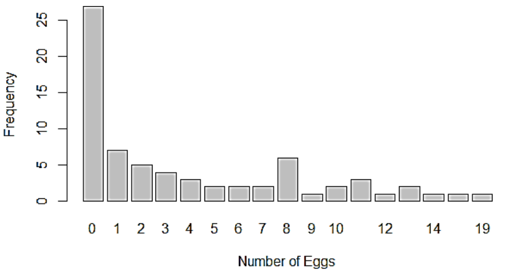

Fig. 1. The number of eggs laid by the females not in diapause.

The number of zeros in the data is too high to be a regular Poisson distribution. The proportion of zeros is approximately 40%, and the mean of positive counts is 6.4. From the simulation results in Section 4, when the value of λ is high, both the ZIP and ZAP models will have nearly the same performance. This can help predict the result of the real data. Here, if the ZIP is used, the CIs using logit, probit, and cloglog links will be (0.2784, 0.5032), (0.2769, 0.5023), (0.2821, 0.5092), respectively, and if the ZAP is applied, the CIs using the logit, probit, and cloglog links will produce (0.2796, 0.5040), (0.2780, 0.5030), and (0.2727, 0.4973), respectively. Both the ZIP and ZAP models result in similar CIs, and all three link functions do not seem to affect the CIs.

6. Conclusion

If the model used to fit the data is correct or corresponds to the distribution of the data, the link functions, proportion of zeros, probability π, and mean λ do not have a very significant effect on the coverage probabilities. However, when the model is incorrect, the values of λ and π play an important role in determining the coverage probabilities. If any combination of λ and π values results in the ZAP and ZIP distributions having nearly the same proportions of zeros, both the ZAP and ZIP models can be used for fitting the data. This especially occurs when λ equals 5 or 10. For example, in Fig. 2 and Fig. 3, when λ equals 5, the CPs can achieve a confidence level of 0.95 for all π values, even though the models are wrong. In Table 9, it can be seen that the proportions (for both ZIP and ZAP) of zeros in each π are approximately the same. However, when λ equals 1 or 2, the proportions of zeros are different. In such situations, using the wrong model will give a CP below the desirable confidence level, 0.95. Overall, the Wald-type confidence intervals with different link functions have nearly the same coverage probabilities whether the model used for fitting the data is correct or not. In addition, the AIC and CP do not necessarily agree, so the AIC should not be used as the criterion for selecting a model to construct the Wald confidence interval.

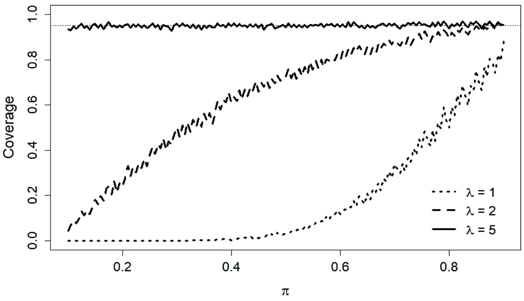

Fig. 2. The estimated coverage probabilities of π_i in ZAP models using the logit link when the data are obtained from ZIP distribution.

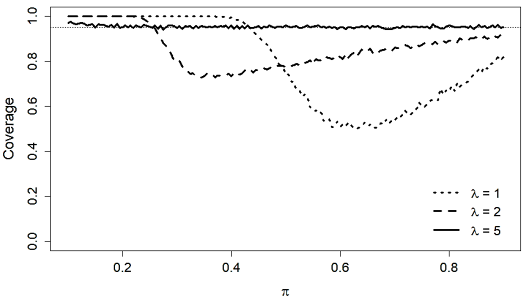

Fig. 3. The estimated coverage probabilities of π_i in ZIP models using the logit link when the data are obtained from ZAP distribution.

In practice, analysts do not know the distribution of data in advance, so in this study, the ZIP would be recommended, as it tends to give higher CPs than the ZAP. Also, the descriptive statistics such the proportion of zeros, $\hat{\pi}$, and the mean of positive counts, $\hat{\lambda}$, should be calculated to help select between the ZIP and ZAP models. When $\hat{\lambda}$ is large, both models will be usable, and if $\hat{\lambda}$ is small, the ZIP would be recommended as it is conservative.

References

[1] Mullahy J. Specification and testing of some modified count data models. J Econom 1986;33(3): 341-65.

[2] Cameron AC, Trivedi PK. Regression analysis of count data. 2nd ed. New York: Cambridge University Press; 2013.

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

[4] Ridout M, Demétrio C, Hinde J. Models for count data with many zeros. The XIXth International Biometric Conference 1998; 179–92.

[5] Miller JM. Comparing Poisson, Hurdle and ZIP model fit under varying degrees of skew and Zero-inflation [Ph.D. dissertation]. Gainesville: University of Florida; 2007

[6] Yang S, Harlow L, Puggioni G, Redding C. A comparison of different methods of zero-inflated data analysis and an application in health surveys. J Mod Appl Stat Methods 2017;16(1):518-43.

[7] Agresti A. Foundations of linear and generalized linear models. New Jersey: John Wiley & Sons; 2015. Czado C, Munk A. Noncanonical links in generalized linear models – when is the effort justified?. J Stat Plan Inference 2000;87(2):317-34.

[8] Czado C, Munk A. Noncanonical links in generalized linear models – when is the effort justified?. J Stat Plan Infer 2000;87(2):317-34.

[9] Koenker R, Yoon J. Parametric links for binary choice models: A Fisherian–Bayesian colloquy. J Econom 2009;152(2):120-30.

[10] Li J. Choosing the proper link function for binary data [M.S. thesis]. Austin: University of Texas; 2014.

[11] Gunduz N, Fokoue E. On the predictive analytics of the probit and logit link functions. Available from: https://scholar works.rit.edu/article/1235.

[12] Damisa SA, Tasi'u M, Bello SY, Musa FN, Ajadi NA, Agboola S. On the comparison of some link functions of binary response analysis under symmetric and asymmetric assumptions. BSI 2017;2(5):145–9.

[13] Wu L, Lord D. Examining the influence of link function misspecification in conventional regression models for developing crash modification factors. Accid Anal Prev 2017; 102:123-35.

[14] Czado C, Pfettner J, Gschlößl S, Schiller F. Nonnested model comparison of GLM and GAM count regression models for life insurance data; 2009.

[15] Pawitan Y. In all likelihood: Statistical Modelling and Inference Using Likelihood. Oxford: Clarendon Press; 2013.

[16] Zeileis A, Kleiber C, Jackman S. Regression models for count data in R. J Stat Softw 2008;27(8):246-56.

[17] Nelder JA, Mead R. A simplex method for function minimization. Comput J 1965;7(4):308-13.

[18] Tauber M, Tauber C, Nechols J. Life history of Galerucella nymphaeae and implications of reproductive diapause for

rearing univoltine chrysomelids. Physiol Entomol 1996;21:317–24.

[19] Bilder CR, Loughin TM. Analysis of Categorical Data with R. NY: CRC Press; 2015.