Analysis of Extra Zero Counts using Zero-inflated Poisson Models

Saranya Numna

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science in Mathematics and Statistics
Prince of Songkla University
2009

Thesis Title	Analysis of extra zero counts using Zero-inflated Poisson Models
Author	Miss. Saranya Numna
Major Program	Mathematics and Statistics

Major Advisor	**Examining Committee:**
....	....Chairperson
(Asst. Prof. Dr. Naratip Jansakul)	(Dr. Thawatchai Sirikantisophon)
....	....
(Dr. Natthada Jibenja)	(Asst. Prof. Dr. Saengla Chaimongkol)
	....
	(Asst. Prof. Dr. Naratip Jansakul)
	....
	(Dr. Natthada Jibenja)

The Graduate School, Prince of Songkla University, has approved this thesis as partial fulfillment of the requirements for the Master of Science Degree in Mathematics and Statistics

....
(Assoc. Prof. Dr. Krerkchai Thongnoo)
Dean of Graduate School

Thesis Title	Analysis of Extra Zero Counts using Zero-inflated Poisson Models
Author	Miss. Saranya Numna
Major Program	Mathematics and Statistics
Academic Year	2009

ABSTRACT

Poisson regression models are basically modelling for counts. There are two strong assumptions for Poisson model to be checked: one is that events occur independently over of time or exposure period, the other is that the conditional mean and variance are equal. In practice, counts have greater variance than the mean are described as overdispersion. This indicates that Poisson regression is not adequate. There are two common causes that can lead to overdispersion are additional variation to the mean or heterogeneity, an Negative Binomial model is often used and other cause counts with excess zeros or zero-inflated Poisson counts, since the excess zeros will give smaller conditional mean than the true value, this can be modeled by using zero-inflated Poisson (ZIP).

This thesis concentrates on the use of ZIP model for analysis counts data including maximum likelihood estimation for regression coefficients using Fisher scoring method, compare between Poisson and ZIP models by various tests: likelihood ratio test, score test, chi - square test, test based on a confidence interval test and Cochran test presented in literature. Model selection using Akaike information criteria (AIC) and checking adequacy of the model using half-normal plots with a simulated envelope.

We developed a Wald test (W_ω) for ZIP model in a single sample case for detecting zero-inflation in Poisson model and conduct a small simulation study in order to investigate sampling distribution of W_ω and power of W_ω. From our study we found that its distribution is an equal mixture of a χ_0^2 (a constant of zero) and a χ_1^2 distribution and can be used to detect the zero-inflation in counts.

We applied presented procedure and our Wald test by using three sets of data: the set of AIDS-related data for ZIP regression analysis with mean counts display on covariate, the foetal lamb movement data and the death notice data of women 80 years of age and over, on each day for three consecutive years appearing in the London "Times", as a single sample case. It is showed that ZIP model is appropriate with the foetal lamb movement data and the death notice data of London times but the set of AIDS-related data, ZIP model is not suitable.

ACKNOWLEDGEMENTS

My deeply debt of gratitude is owned to my advisors, Asst. Prof. Dr. Naratip Jansakul and Dr. Natthada Jibenja, for their initial idea, guidance, constant instruction, kindness and encouragement which enable me to carry out my study successfully.

My special appreciation is expressed to Dr. Thawatchai Sirikantisophon and Asst. Prof. Dr. Saengla Chaimongkol for many valuable comments and helpful suggestions.

Thanks are also due to all other lecturer staffs of the department of mathematics, Prince of Songkla University for their patience, encouragement and impressive teaching.

I gratefully appreciate my beloved parents for their love, sacrifices and encouragement. I would like to express my special thanks to my friends for their suggestions and encouragement.

Finally, I would like to thanks every one, who supported me, but I did not mention above.

Saranya Numna

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AIC	Akaike information criteria
BIC	Baysians information criteria
d.f	Degree of freedom
glms	Generlized linear models
MLE	Maximum likelihood estimate
NB	Negative Binomial
NB1	Linear Mean-Variance Negative Binonial
p.d.f	Probability density function
p.m.f	Probability mass function
s.e.	Standard error
ZINB	Zero-inflated Negative Binomial
ZIP	Zero-inflated Poisson

CHAPTER 1

Introduction

1.1 Background and Motivation

Linear regression analysis is a statistical method for investigating the relationship between the variable to be predicted, called the response or dependent variable and variables expected to be related to the response variable. The related variables are called explanatory or independent variables. The response variable must be continuous random variable which is assumed to have normal distribution with constant variance. The explanatory variables can be either quantitative or qualitative variables.

Gardner et al. (1995) characterized method of regression analysis in terms of two ideas: (a) a model for the mean which says how the expected value of the response variable depends on a set of explanatory variables and (b) a model for the variation of the response variable scores around the expected value.

Counts of events occuring in a given time or exposure period are discrete random variable having Poisson distribution. Even though, the Poisson distribution can be approximated by the normal distribution when the mean count is sufficient large, applying ordinary linear regression to count data is problematic on those two ideas. That is the linear model relating the expected counts to the predictors is likely to produce negative predicted values or the validity of hypothesis tests in linear regression depending on constant variance assumption of response variable is unlikely to be met in count data (Gardner et al., 1995). Based on the basic concept of generalized linear models (*glms*), the relationship between explanatory variables and response variable of Poisson counts can be described by Poisson regression or log linear models. The Poisson model is formed under two principal assumptions: one is that events occur independently over given time or exposure period and the other is that the conditional mean and variance are equal. However, in practice, the equality of the mean and variance rarely occurs; the variance may be either greater or less than the mean. If the variance is greater than the mean, it means that counts are more variable than specified by the Poisson events and are describe as overdispersion. If the variance is less than the mean, it means that counts are less variable than specified by the Poisson events and are describe as underdispersion. However, in practice, underdispersion is less common (McCullagh and Nelder, 1989).

One general cause of overdispersion is cluster sampling, such as litters, families, households, etc. where there is additional variability between clusters that produces response variance larger than the nominal value (McCullagh and Nelder, 1989). Besides clustering, other possible causes of over dispersion, for example, correlation between individual response that breaks the individual independent assumption and aggregate level data which leads to compound distribution (Hinde and Demétrio, 1998). Moreover, overdispersion can be caused by excess number of observed zero counts, since the excess zeros will give smaller conditional mean than the true value. The count data with excess zeros, is known as zero-inflated Poisson counts. Of course it is possible to have fewer zero counts than expected, but this is less common in practice (Ritout et al., 1998).

When there is overdispersion or zero-inflation and we fail to take it into account, it can lead to misinterpretion of the fitted model (Hinde and Demet́rio, 1998), since the overdispersion or zero-inflation produces:

1) smaller standard errors of the parameter estimates than the true values. Therefore we may incorrectly choose explanatory variables for the model that are not required;

2) too large a reduction of deviance associated with model selection tests. This leads to selecting more complicated models.

In the literature of statistical modelling for counts there are number of models proposed to handle zero-inflated Poisson counts, for example, Hurdle model (Germu et al., 1996), Two-part model (Heilbron, 1994), Zero-modified distributions (Dietz and Böhning, 2000), and Zero-inflated Poisson (ZIP) models (Lambert, 1992). ZIP models are more widely used as all important statistical inferences can be carried out more easily and conveniently than the others. Applications of ZIP models can be found in many areas, such as, agriculture (Ridout et al., 1998), epidemiology (Böhning et al., 1999), biostatistics (Van den Broek, 1995) and industry (Lambert, 1992).

This thesis will be focussed only on ZIP models and will proposed a Wald test for comparing between the standard Poisson and ZIP models.

1.2 Objectives

1. To explore some aspects of the analysis of counts with excess zeros;
2. To develop a Wald test for zero-inflation parameter used to compare between Poisson and ZIP models;
3. To investigate the distribution and properties of the Wald test;
4. To investigate power of the Wald test and compare that with other tests for zero-inflation parameter, presented in the literature.

1.3 Scope and Methodology

This thesis focuses on

1. studying characteristics, theories and properties of Generalize linear models (glms) and ZIP regression models;
2. analysing counts with many zeros;
3. exploring various tests for zero-inflation parameter used to compare between Poisson and ZIP models;

4. develop a Wald test for zero-inflation parameter in the case of single homogenous sample, investigate its properties and compare with the others explored in 3. by conducting simulation study;

5. example used in this thesis for illustrating the use of ZIP and the proposed Wald test include the set of AIDS-related for ZIP regression analysis with mean counts depending on covariate, the foetal lamb movement data and the death notice data of London times, as single sample cases.

1.4 Advantages

1. Propose the Wald test for ZIP models in the literature related to statistical modelling for zero-inflated counts;

2. We have a new alternative statistic test for comparing between Poisson and ZIP models;

3. We have an idea to develop the Wald test for comparing between a Negative Binomial model and Zero-inflated Negative Binomial model.

1.5 Period and Plan of Study:

October 2008 - September 2009

Table 1.1: Plan of Study

Task	2008			2009								
	10	11	12	1	2	3	4	5	6	7	8	9
Select topic	*											
Study literature	*	*	*									
Write thesis proposal		*	*	*								
Present thesis proposal				*								
Study and research				*	*	*	*					
Create research draft				*	*	*	*	*				
Adjust draft report				*	*	*	*	*				
Write final report									*	*	*	
Present thesis												*

1.6 Place of Study

Department of Mathematics, Faculty of Science,
Prince of Songkla University, Hat Yai, Songkhla 90112, Thailand.

CHAPTER 2

Review of Literature

Here we reviewed literature associated with generalized linear models, abbreviated by *glms*, Poisson regression and zero-inflated Poisson (ZIP) models followed by test statistics for zero-inflation proposed in ZIP literatures.

2.1 Exponential Family Distributions

A random variable Y has a distribution within the exponential family if the probability (mass) density function can be written in the canonical form

$$f(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\} \tag{2.1}$$

for some specific functions $a(.), b(.)$ and $c(.)$. θ is the *natural* or *canonical* parameter. $a(\phi)$ is called the *scale* or *dispersion* parameter, such as σ^2 for normal distribution. For known ϕ, this equation is the *linear exponential* family. The canonical parameter θ is a function of the mean and so in turn can be related to the linear predictor. A natural choice of link function is to take $\theta = \eta = g(\mu)$, which is known as the *canonical link*.

Given a vector of n observations $\mathbf{y} = (y_1, \ldots, y_n)^T$ of Y from the exponential family distribution, the likelihood function, $L = L(\theta, \phi; \mathbf{y})$ is

$$L(\theta,\phi;\mathbf{y}) = \prod_{i=1}^{n} f(y_i;\theta,\phi) = \exp\left[\sum_{i=1}^{n}\left\{\frac{y_i\theta - b(\theta)}{a(\phi)} + c(y_i,\phi)\right\}\right], \tag{2.2}$$

and the corresponding log-likelihood function, $\ell = \ell(\theta, \phi; \mathbf{y}) = \ln L$ is

$$\ell = \ell(\theta,\phi;\mathbf{y}) = \sum_{i=1}^{n}\left\{\frac{y_i\theta - b(\theta)}{a(\phi)} + c(y_i,\phi)\right\}. \tag{2.3}$$

The log-likelihood function (2.3) has some elementary properties that play an important role in the context of statistical modelling. These are

$$\begin{aligned} \mathrm{E}\left(\frac{\partial \ell}{\partial \theta}\right) &= 0 \\ \mathrm{E}\left(\frac{\partial^2 \ell}{\partial \theta^2}\right) + \mathrm{Var}\left(\frac{\partial \ell}{\partial \theta}\right) &= 0. \end{aligned} \tag{2.4}$$

Simple calculation shows that the mean and variance of Y_i are

$$\mathrm{E}(Y_i) = b^{'}(\theta) = \mu$$

and

$$\mathrm{Var}(Y_i) = a(\phi)b^{''}(\theta) = a(\phi)V(\mu).$$

Here $b^{'}(\theta)$ and $b^{''}(\theta)$ denote the first and second derivative of $b(\theta)$ with respect to θ, respectively, and $b^{''}(\theta)$ can be defined as $V(\mu)$, because it depends on μ through $b^{'}(\theta)$ $V(\mu)$ is commonly known as the *variance function* of the model.

2.2 Generalized linear models

Generalized linear models abbreviated by *glms*, originally proposed by Nelder and Wedderburn (1972) are statistical models, defined by following three components:

1) The random part: responses $Y_i, i = 1, \ldots, n$ is a random variable that its distribution is within the linear exponential family distribution with mean μ_i and the constant dispersion parameter $a(\phi)$;

2) The systematic part: the associated explanatory variables $\mathbf{x_i} = (x_{0i}, x_{1i}, \ldots, x_{pi})^T$, $i = 1, \ldots, n$, give a linear predictors

$$\eta_i = \mathbf{x_i}^T \boldsymbol{\beta}, \tag{2.5}$$

where $\boldsymbol{\beta}$ is a vector of $p + 1$ unknown parameters, $x_{0i} = 1$, for all i to include the constants or intercept term in model. In the case of defining the systematic part in term of the vector of linear predictor, (2.5) is then

$$\boldsymbol{\eta} = X\boldsymbol{\beta}, \tag{2.6}$$

where $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^T$ is an $n \times (p+1)$ design, or covariate matrix;

3) The link function: a function g that links the mean of Y in 1) and the linear predictor in 2) that is

$$g(\mu_i) = \eta_i = \mathbf{x_i}^T\boldsymbol{\beta}. \tag{2.7}$$

The choice of link function depends upon the particular distribution.

Well-known distributions belong to the linear exponential family, such as the following examples:

- If Y is a normal distributed random variable with mean μ and variance σ^2, which is commonly denoted by $Y \sim \mathrm{N}(\mu, \sigma^2)$, its probability density function (p.d.f) can be defined as

$$f(y;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{1}{2\sigma^2}(y-\mu)^2\right\}, \quad -\infty < y < \infty$$

This can be rewritten as

$$f(y;\mu,\sigma^2) = \exp\left\{\frac{y\mu - \frac{1}{2}\mu^2}{\sigma^2} + \left(\frac{-y^2}{2\sigma^2} - \frac{1}{2}\mathrm{ln}(2\pi\sigma^2)\right)\right\}. \tag{2.8}$$

Here the distribution of Y is in the canonical exponential form with canonical parameter $\theta = \mu$, $b(\theta) = \frac{1}{2}\mu^2$, $c(y,\phi) = \frac{-y^2}{2\sigma^2} - \frac{1}{2}\mathrm{ln}(2\pi\sigma^2)$, scale parameter $a(\phi) = \sigma^2, \mathrm{E}(Y) = \mu$ and the variance function ,$V(\mu) = 1$. The canonical link functionis the *identity*, then the i^{th} linear predictor, $g(\mu_i) = \mu_i = \mathbf{x_i}^T\boldsymbol{\beta}$. That is the classical linear model or linear regression.

- If Y is a Poisson distributed random variable with mean μ which is normally denoted by $Y \sim \mathrm{Pois}(\mu)$, its probability mass function (p.m.f) is

$$f(y;\mu) = \frac{\mathrm{e}^{-\mu}\mu^y}{y!}, \quad y = 0, 1, 2, \ldots$$

The distribution of Y can be written in the form

$$f(y;\mu) = \exp\left\{\frac{y\mathrm{ln}\mu - \mu}{1} - \mathrm{ln}y!\right\}, \tag{2.9}$$

which is in the linear exponential family with canonical parameter $\theta = \ln\mu$, $b(\theta) = \mu = \mathrm{e}^{\theta}, c(y, \phi) = \ln y!$ and $a(\phi) = 1$, *i.e.* no unknown scale parameter, $\mathrm{E}(Y) = \mu$, and the variance function, $V(\mu) = \mu$. The canonical link function, $g(\mu_i) = \ln\mu_i = \mathbf{x_i}^T\boldsymbol{\beta}$, leads to a *log-linear model* for Poisson counts.

- If Y is a binomial distributed random variable with probability of success, π and number of trial, m which is generally denoted by $Y \sim \mathrm{Bin}(m, \pi)$, its p.m.f is

$$f(y; m, \pi) = \binom{m}{y}\pi^{y}(1-\pi)^{(m-y)}, \quad y = 0, 1, 2, \ldots, m$$

This can be written as

$$f(y; \pi) = \exp\left\{y\ln\left(\frac{\pi}{1-\pi}\right) + m\ln(1-\pi) + \ln\binom{m}{y}\right\}, \tag{2.10}$$

which is in the linear exponential family with canonical parameter $\theta = \ln\left(\frac{\pi}{1-\pi}\right)$, $b(\theta) = -m\ln(1-\pi) = -m\ln\left(\frac{1}{1+\mathrm{e}^{\theta}}\right)$, $c(y, \phi) = \ln\binom{m}{y}$ and $a(\phi) = 1$, $\mathrm{E}(Y) = m\pi$ and the variance function, $V(\pi) = \pi(1-\pi)$. The canonical link function is logit, defined by $g(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{x_i}^T\boldsymbol{\beta}$. This transforms the range for π from $[0, 1]$ to $[-\infty, \infty]$ giving a sensible scale for modelling; giving standard *logistic* regression.

To obtain the maximum likelihood estimate, $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ in (2.7), we need to find the values of $\boldsymbol{\beta}$ that maximize $\ell(\boldsymbol{\beta}, \phi; \mathbf{y})$, the log-likelihood function of $\boldsymbol{\beta}$ and ϕ conditional on respones $\mathbf{y}$ via $\ell(\boldsymbol{\theta}, \phi; \mathbf{y})$, using the chain rule. Differentiating ℓ with respect to each β_j in turn, we have the likelihood estimation equation

$$\frac{\partial\ell}{\partial\beta_j} = \frac{\partial\ell}{\partial\theta_i}\frac{\partial\theta_i}{\partial\mu_i}\frac{\partial\mu_i}{\partial\eta_i}\frac{\partial\eta_i}{\partial\beta_j}$$

giving

$$\frac{\partial\ell}{\partial\beta_j} = \sum_{i=1}^{n}\frac{(y_i - \mu_i)x_{ij}}{\mathrm{Var}(\mathbf{Y}_i)\acute{g}(\mu_i)} = 0, \quad j = 0, 1, \ldots, p. \tag{2.11}$$

In general, equations (2.11) are not linear in $\boldsymbol{\beta}$, so that they need to be solved by using an iterative scheme. The iterative schemes commonly used are either

Newton-Raphson or Fisher scoring. Using these methods, current parameter estimates are obtained by correcting the estimates of previous step using the first and second derivatives of the log-likelihood function with respect to the parameters of interest. Based on the log-likelihood (2.3) and the model (2.7), a Newton-Raphson for $\hat{\boldsymbol{\beta}}$ at the $(m+1)^{th}$ iteration is

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + [I^{(m)}]^{-1}\mathbf{s}^{(m)}, \tag{2.12}$$

where

$$\mathbf{s}^{(m)} = \begin{bmatrix} \frac{\partial \ell}{\partial \beta_0} \\ \frac{\partial \ell}{\partial \beta_1} \\ \vdots \\ \frac{\partial \ell}{\partial \beta_p} \end{bmatrix}_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(m)}} \quad \text{and}$$

$$I^{(m)} = \begin{bmatrix} -\frac{\partial^2 \ell}{\partial \beta_0^2} & -\frac{\partial^2 \ell}{\partial \beta_0 \partial \beta_1} & \cdots & -\frac{\partial^2 \ell}{\partial \beta_0 \partial \beta_p} \\ -\frac{\partial^2 \ell}{\partial \beta_1 \partial \beta_0} & -\frac{\partial^2 \ell}{\partial \beta_1^2} & \cdots & -\frac{\partial^2 \ell}{\partial \beta_1 \partial \beta_p} \\ \vdots & \vdots & & \vdots \\ -\frac{\partial^2 \ell}{\partial \beta_p \partial \beta_0} & -\frac{\partial^2 \ell}{\partial \beta_p \partial \beta_1} & \cdots & -\frac{\partial^2 \ell}{\partial \beta_p^2} \end{bmatrix}_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(m)}}$$

$\mathbf{s}^{(m)}$ and $I^{(m)}$ are the score vector and observed information matrix evaluated at $\boldsymbol{\beta} = \boldsymbol{\beta}^{(m)}$, respectively. The method of Fisher scoring has the same idea as the Newton-Raphson method except that I is replaced by $\mathcal{I} = -\mathrm{E}(I)$, the expected information matrix. That is

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + [\mathcal{I}^{(m)}]^{-1}\mathbf{s}^{(m)}. \tag{2.13}$$

With good starting values $\boldsymbol{\beta}^{(0)}$ the iterative scheme converges in a few step, convergence is obtained with a stopping rule, such as $|\ell^{(m+1)} - \ell^{(m)}| \leq \epsilon$, where $\ell^{(m)}$ and $\ell^{(m+1)}$ are the log-likelihood, ℓ evaluated using the estimates of $\boldsymbol{\beta}$ from the (m) and $(m+1)$ iterations, respectively. The asymtotic variance-covariance matrix for $\hat{\boldsymbol{\beta}}$ is automatically provided in final iteration. The method of Newton-Raphson and Fisher scoring are implemented in software package for fitting *glms*, such as S-Plus, R etc. which include automatic calculation of starting values.

2.3 Poisson regression model

Poisson regression models provide a standard framework for the analysis of count data. Let $Y_i, i = 1, \ldots, n$ represent counts of events occuring in a given time or exposure periods with rate μ_i. Y_i are Poisson random variables which the p.m.f. is characterized by

$$f(y_i; \mu_i) = \frac{\mathrm{e}^{-\mu_i} {\mu_i}^{y_i}}{y_i!}, \qquad y_i = 0, 1, 2, \ldots$$

with

$$\mathrm{E}(Y_i) = \mathrm{Var}(Y_i) = \mu_i.$$

The log-likelihood function is

$$\ell(\boldsymbol{\mu}) = \ell(\boldsymbol{\mu}; \mathbf{y}) = \sum_{i=1}^{n} \{y_i \ln(\mu_i) - \mu_i - \ln(y_i!)\}. \qquad (2.14)$$

Let X be an $n \times (p+1)$ matrix of explanatory variables. The relationship between Y_i and i^{th} row vector of X, $\mathbf{x}_i$ linked by $g(\mu_i)$ is

$$\ln(\mu_i) = \eta_i = \mathbf{x}_i^T \boldsymbol{\beta}. \qquad (2.15)$$

This model is known as the Poisson regression or log-linear model. The maximum likelihood estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ can be obtained via the method of Fisher scoring.

There are two principal assumptions in the Poisson model we need to regard: one is that events occur independently over of time or exposure period, the other is that the conditional mean and variance are equal (Cameron and Trivedi, 1986). The latter assumption is quite important. If it fails, the fitted model should be reconsidered.

There are two basic criteria commonly used to check the presence of overdispersion : the deviance, $D(\mathbf{y}; \hat{\boldsymbol{\mu}})$ or the Pearson (χ^2) statistic be greater than its degrees of freedom (Lindsey, 1999 and Hilbe, 2007). For the Poisson regression, $D(\mathbf{y}; \hat{\boldsymbol{\mu}})$ and χ^2 are respectively defined in expression (2.16) and (2.17)

$$D(\mathbf{y}; \hat{\boldsymbol{\mu}}) = 2 \times \sum_{i}^{n} \left\{ y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i) \right\}; \qquad (2.16)$$

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu_i})^2}{\hat{\mu_i}}. \tag{2.17}$$

However, these two rules of thumb can yield misleading inference from a direct likelihood point of view. Therefore, selecting between Poisson regression and an overdispersed Poisson model should be performed using some appropriate modelling procedure.

2.4 Zero-inflated poisson model: ZIP

Zero-inflated poisson (ZIP) model, well described by Lambert (1992) is a simple mixture model for count data with excess zeros. The model is a combination of a Poisson distribution and a degenerate distribution at zero. Specifically if Y_i are independent random variables having a zero-inflated Poisson distribution, the zeros are assumed to arise in to ways corresponding to distinct underlying states. The first state occurs with probability ω_i and produces only zeros, while the other state occurs with probability $1-\omega_i$ and leads to a standard Poisson count with mean λ_i and hence a chance of further zeros. In general, the zeros from the first state are called *structural zeros* and those from the Poisson distribution are called *sampling zeros* (Jansakul and Hinde, 2002). This two-state process gives a simple two-component mixture distribution with p.m.f

$$\Pr(Y_i = y_i) \quad = \quad \begin{cases} \omega_i + (1-\omega_i)\mathrm{e}^{-\lambda_i}, & y_i = 0 \\ \\ (1-\omega_i)\dfrac{e^{-\lambda_i}\lambda^{y_i}}{y_i!}, & y_i = 1, 2, \ldots, \quad 0 \leq \omega_i \leq 1 \end{cases} \tag{2.18}$$

which we denote by $Y_i \sim \mathrm{ZIP}(\lambda_i, \omega_i)$. The mean and variance of Y_i are

$$\mathrm{E}(Y_i) = (1-\omega_i)\lambda_i = \mu_i$$

and

$$\mathrm{Var}(Y_i) = \mu_i + \left(\frac{\omega_i}{1-\omega_i}\right)\mu_i^2, \tag{2.19}$$

indicating that the marginal distribution of Y_i exhibits overdispersion, if $\omega_i > 0$. It is clear that this reduces to the standard Poisson model when $\omega_i = 0$.

For positive values of ω_i we have zero-inflation, however, it is poissible for $\omega_i < 0$ and to still obtain a valid probability distribution (this corresponds to a deficit of zeros – zero-deflation) (Jansakul and Hinde, 2002).

For a random sample of observations $y_1, \ldots, y_n$, the log-likelihood function is given by

$$\begin{aligned} \ell = \ell(\lambda, \omega; \mathbf{y}) &= \sum_i \{ I_{(y_i=0)} \ln[\omega_i + (1-\omega_i)e^{-\lambda_i}] \\ &+ I_{(y_i>0)}[\ln(1-\omega_i) - \lambda_i + y_i \ln\lambda_i - \ln(y_i!)]\}, \end{aligned} \tag{2.20}$$

where $I(.)$ is the indicator function for the specified event, *i.e.* equal to 1 if the event is true and 0 otherwise. To apply the zero-inflated Poisson model in practical modelling situations, Lambert (1992) suggested the following joint models for $\boldsymbol{\lambda}$ and $\boldsymbol{\omega}$

$$\ln(\boldsymbol{\lambda}) = X\boldsymbol{\beta} \quad \text{and} \quad \ln\left(\frac{\boldsymbol{\omega}}{1-\boldsymbol{\omega}}\right) = G\boldsymbol{\gamma}, \tag{2.21}$$

where X and G are covariate matrices and $\boldsymbol{\beta}, \boldsymbol{\gamma}$ are $(p+1) \times 1$ and $(q+1) \times 1$ vectors of unknown parameters respectively.

In the case of single homogeneous sample (λ and ω are constant or do not depend on X and G), the log-likelihood function (2.20) can be written as

$$\ell(\lambda, \omega) = n_0 \ln\left[\omega + (1-\omega)\mathrm{e}^{-\lambda}\right] + \sum_{j=1}^{J} n_j \ln\left[(1-\omega)\frac{\mathrm{e}^{-\lambda}\lambda^j}{j!}\right], \tag{2.22}$$

where J is the largest observed count value, n_j is the frequency of each possible count value, $j = y = 0, 1, 2, \ldots, J$ then n_0 is the number of observed zeros and $\sum_{j=0}^{J} n_j = n$, the total number of observations or the sample size.

2.5 Test statistics for zero-inflation proposed in ZIP literature

Within the family of ZIP models, testing if a Poisson model is adequate corresponding to testing

$$\mathrm{H}_0 : \omega = 0; \quad \mathrm{H}_1 : \omega > 0, \tag{2.23}$$

where ω here is taken a constant.

There are a number of test statistics proposed for testing the hypothesis (2.23) including score test, likelihood ratio test, chi-square test, test based on a confidence interval of probability zero-inflated counts and Cochran test. Those are collected and investigated the properties by Xie et al. (2001). Most of the tests mentioned were derived based on a single homogeneous sample, *i.e.* λ and ω are not depend upon covariate or are constant. The expressions of the tests and their sampling distributions are summarized in following subsections.

2.5.1 Likelihood ratio test

The Likelihood Ratio test is a test of a null hypothesis H_0 against an alternative H_1 based on the ratio of two log-likelihood functions. The likelihood ratio test for hypothesis (2.23) can be computed from the following formula:

$$R_\omega = -2 \times [\ell(\hat{\lambda}) - \ell(\hat{\lambda}, \hat{\omega})] \tag{2.24}$$

where $\ell(\hat{\lambda})$ and $\ell(\hat{\lambda}, \hat{\omega})$ are the maximized log-likelihood under the Poisson and the ZIP regression model, respectively. From (2.14), (2.22) and (2.24) the likelihood ratio test can be written as

$$R_\omega = 2\left\{ n_0 \ln\left(\frac{n_0}{n}\right) + (n - n_0)\left(\ln\left(\frac{\bar{y}}{\hat{\lambda}}\right) - \hat{\lambda}\right) + n\bar{y}(\ln\hat{\lambda} + 1 - \ln\bar{y}) \right\}, \tag{2.25}$$

where $\bar{y}$ is the mean of the observations under H_0 and $\hat{\lambda}$ is the estimated positive mean counts under H_1. This test statistic R_ω approximately follows chi-square distribution on 1 degree of freedom (d.f) under the null hypothesis.

2.5.2 Score test

A score test for the hypothesis (2.23) is proposed by Van den broek (1995). The test is derived based on the log-likelihood (2.22) to obtain the ratio of the score vector, $\begin{bmatrix} \dfrac{\partial \ell(\lambda, \omega)}{\partial \lambda} \\ \\ \dfrac{\partial \ell(\lambda, \omega)}{\partial \omega} \end{bmatrix}$ and minus expected information matrix,

$$\begin{bmatrix} -\mathrm{E}\left(\dfrac{\partial^2\ell(\lambda,\omega)}{\partial\lambda^2}\right) & -\mathrm{E}\left(\dfrac{\partial^2\ell(\lambda,\omega)}{\partial\lambda\partial\omega}\right) \\ -\mathrm{E}\left(\dfrac{\partial^2\ell(\lambda,\omega)}{\partial\omega\partial\lambda}\right) & -\mathrm{E}\left(\dfrac{\partial^2\ell(\lambda,\omega)}{\partial\omega^2}\right) \end{bmatrix}, \text{ evaluated at } \omega = 0$$

or under H_0 true. Using mathematical algebra, the score statistic is defined by

$$S_\omega = \frac{(n_0 - np_0)^2}{np_0(1-p_0) - n\bar{y}p_0^2}, \tag{2.26}$$

where $p_0 = e^{-\hat{\lambda}_0}$, in which $\hat{\lambda}_0$ is the estimate of the Poisson parameter under the null hypothesis or $\bar{y}$. This statistic will have an asymptotic chi-square distribution on 1 d.f under the null hypothesis.

2.5.3 Chi-square test

The chi-square statistic χ^2 is used to test if a sample of data came from a population with a specific distribution. The χ^2 is commonly defined by

$$\chi^2_\omega = \sum_{k=1}^{c} \frac{(O_k - E_k)^2}{E_k} \tag{2.27}$$

where c denotes the number of classes(categories) decided for a given data set, O_k and E_k are observed frequencies and expected frequencies under the null hypothesis of the k^{th} class, respectively. When the null hypothesis is valid, χ^2_ω follows an asymtotic chi-square distribution on $c-1$ d.f.

2.5.4 Test based on a confidence interval of probability zero-inflated counts

It is possible to derive a test based on asymtotic normality of the estimate of the parameters. Following the statistical properties of ZIP,

$$\mathrm{E}(\bar{Y}) = \mathrm{E}(Y) = (1-\omega)\lambda = \mu$$

and

$$\mathrm{Var}(\bar{Y}) = \frac{1}{n}\mathrm{Var}(Y) = \frac{1}{n}\left\{\frac{(1-\omega)\mu + \omega\mu^2}{1-\omega}\right\}.$$

From the central limit theorem, the confidence interval can be written as

$$1 - \frac{\bar{y} - Z_{\alpha/2}\sqrt{\{\bar{y} + \bar{y}[\hat{\lambda} - \bar{y}]\}/n}}{\hat{\lambda}} \leq \omega \leq 1 - \frac{\bar{y} + Z_{\alpha/2}\sqrt{\{\bar{y} + \bar{y}[\hat{\lambda} - \bar{y}]\}/n}}{\hat{\lambda}}.$$

Hence, a test based on a positive one sided confidence interval of probability zero-inflated counts can be obtained as

$$CI_\omega = 1 - \frac{\bar{y} + Z_\alpha\sqrt{\{\bar{y} + \bar{y}[\hat{\lambda} - \bar{y}]\}/n}}{\hat{\lambda}}, \tag{2.28}$$

where $\hat{\lambda}$ is the estimated positive mean counts under H_1. The critical region of this test method is simply $CI_\omega > 0$. That is,when $CI_\omega > 0$, we reject the null hypothesis at α level of significance and the ZIP model should be used instead of a Poisson model.

2.5.5 The Cochran test

An early test is proposed in Cochran (1954), and is commonly called the C test. The C test is used to test the assumption of constant variance of the residuals in the analysis of variance. This test is a ratio that relates the largest empirical variance of a particular treatment to the sum of the variances of the remaining treatments. The C test statistic for ZIP model was developed by Xie et al. (2001) can be written as follows:

$$C_\omega = \frac{(n_0 - n\mathrm{e}^{-\bar{y}})}{[n\mathrm{e}^{-\bar{y}}(1 - \mathrm{e}^{-\bar{y}} - \bar{y}\mathrm{e}^{-\bar{y}})]^{1/2}}. \tag{2.29}$$

Under the null hypothesis, the test statistic C_ω is approximately normally distributed with zero mean and unit variance.

Following the relationship between $\mathrm{N} \sim (0, 1)$ and χ^2_1, we found that C^2_ω, can be obtained as

$$C^2_\omega = \frac{(n_0 - n\mathrm{e}^{-\bar{y}})^2}{n\mathrm{e}^{-\bar{y}}(1 - \mathrm{e}^{-\bar{y}} - \bar{y}\mathrm{e}^{-\bar{y}})}. \tag{2.30}$$

C^2_ω has the form exactly the same as Score test in (2.26), where the detail of this is shown in Appendix 1.

2.6 Basic idea of the Wald test

The Wald test is a statistical test, typically used to test whether an effect exists or not. A Wald test can be used in a great variety of different models including models for dichotomous or binary variables and models for continuous variables. Under the aspect of the Wald statistical test, named after Abraham Wald, the maximum likeilhood estimate $\hat{\theta}$ of parameter (s) of interest θ of the random variable Y, with p.d.f or p.m.f $f(y;\theta)$ is compared with the proposed value θ_0 ($\mathrm{H}_0 : \theta = \theta_0$), under the assumption that $\hat{\theta} - \theta_0$ will be approximately normal. Typically the square of the difference is compared to a chi-squared distribution. In the univariate case, the Wald statistic is

$$W = \frac{(\hat{\theta} - \theta_0)^2}{\mathrm{Var}(\hat{\theta})}. \tag{2.31}$$

Under H_0 true, W is a chi-square distribution on 1 d.f. Alternatively, the difference can be compared to normal distribution. In this case the test statistic (2.31) is rewritten as

$$\frac{\hat{\theta} - \theta_0}{\mathrm{s.e.}(\hat{\theta})}$$

where s.e.$(\hat{\theta})$ is the standard error of the maximum likelihood estimate, $\hat{\theta}$. A reasonable estimate of the standard error for the MLE is obtained by $\dfrac{1}{\sqrt{I_n(\hat{\theta})}}$, where $I_n \;=\; -\dfrac{\partial^2 \ell}{\partial \theta^2}$, evaluated at $\theta = \hat{\theta}$, is the Fisher information of the parameter (Wikipedia, the free encyclopedia). Using the elementary properties of the log-likelihood displayed in (2.3), I_n can be replace by the expected information $\mathcal{I} = -\mathrm{E}\left(\dfrac{\partial^2 \ell}{\partial \theta^2}\right)$.

In this thesis, we will use this basic idea to develop a Wald test for ZIP model with constant λ and ω or for testing the hypothesis (2.23), where its sampling distribution and power of the test are also investigated.

CHAPTER 3

Zero-inflated Poisson Models

In this chapter we focuss on a zero-inflated (ZIP) model to take account of zero-inflation in Poisson counts. We consider the use of Fisher scoring method to obtain maximum likelihood estimates for the model. Moreover, we develop the Wald test for single homogeneous ZIP sample with constant λ and ω and conduct a small simulation study to investigate the sampling distribution and power of the Wald test.

3.1 Maximum Likelihood Estimation for ZIP Regression Models

Based on the ZIP model (2.18), the log-likelihood function (2.20) and the model for $\boldsymbol{\lambda}$ and $\boldsymbol{\omega}$, displayed in (2.21), it is obvious that being a finite mixture the ZIP distribution is not a member of the exponential family distribution and so standard glm fitting procedures will not be adequate. To obtain the parameter estimates of ZIP regression models, $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$, the Newton-Raphson method or the method of Fisher scoring can be used. However, the method of scoring is more appropriate for ZIP regression because the second derivative of $\ell(\boldsymbol{\lambda}, \boldsymbol{\omega}; \mathbf{y})$ can be simplified by taking expectations.

3.1.1 The Method of Fisher Scoring

Assuming that $\boldsymbol{\lambda}$ and ω in (2.21) are not functionally related. The first and second derivatives of ℓ with respect to $\boldsymbol{\beta}$ and ω are

$$\begin{aligned}\frac{\partial \ell}{\partial \beta_j} &= \frac{\partial \ell}{\partial \lambda_i}\frac{\partial \lambda_i}{\partial \beta_j} \qquad (3.1)\\ &= \sum_{i=1}^{n}\left\{ I_{(y_i=0)}\left[\frac{-(1-\omega_i)\mathrm{e}^{-\lambda_i}}{\omega_i+(1-\omega_i)\mathrm{e}^{-\lambda_i}}\right]\lambda_i + I_{(y_i>0)}(y_i-\lambda_i)\right\} x_{ij},\end{aligned}$$

$j = 0, 1, 2, \ldots, p;$

$$\frac{\partial \ell}{\partial \omega_i} = \sum_{i=1}^{n}\left\{ I_{(y_i=0)}\left[\frac{(1-\mathrm{e}^{-\lambda_i})}{\omega_i+(1-\omega_i)\mathrm{e}^{-\lambda_i}}\right] + I_{(y_i>0)}\left[\frac{-1}{1-\omega_i}\right]\right\}; \qquad (3.2)$$

and

$$\begin{aligned}\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_k} &= \sum_{i=1}^{n}\left\{ I_{(y_i=0)}\left[\frac{-\mathrm{e}^{-\lambda_i}[(1-\lambda_i)\omega_i+(1-\omega_i)\mathrm{e}^{-\lambda_i}](1-\omega_i)\lambda_i}{[\omega_i+(1-\omega_i)\mathrm{e}^{-\lambda_i}]^2}\right]\right.\\ &+ \left. I_{(y_i>0)}(-\lambda_i)\right\} x_{ij}x_{ik}, \quad j,k = 0,1,2,\ldots,p; \qquad (3.3)\end{aligned}$$

$$\frac{\partial^2 \ell}{\partial \omega_i^2} = \sum_{i=1}^{n}\left\{ I_{(y_i=0)}\left[\frac{-(1-\mathrm{e}^{-\lambda_i})^2}{[\omega_i+(1-\omega_i)\mathrm{e}^{-\lambda_i}]^2}\right] + I_{(y_i>0)}\left[\frac{-1}{(1-\omega_i)^2}\right]\right\}; \quad (3.4)$$

$$\frac{\partial^2 \ell}{\partial \beta_j \partial \omega_i} = \frac{\partial^2 \ell}{\partial \omega_i \partial \beta_j} = \sum_{i=1}^{n}\left\{ I_{(y_i=0)}\left[\frac{\lambda_i \mathrm{e}^{-\lambda_i}}{[\omega_i+(1-\omega_i)\mathrm{e}^{-\lambda_i}]^2}\right]\right\} x_{ij}. \qquad (3.5)$$

Using the fact that

$$\begin{aligned}\mathrm{E}[I_{(y_i=0)}] &= \Pr(Y_i=0) = \omega_i + (1-\omega_i)\mathrm{e}^{-\lambda_i} \quad \text{and}\\ \mathrm{E}[I_{(y_i>0)}] &= \Pr(Y_i>0) = (1-\omega_i)(1-e^{-\lambda_i})\end{aligned}$$

we have

$$\begin{aligned}-\mathrm{E}\left[\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_k}\right] &= \sum_{i=1}^{n}\left\{ \frac{\mathrm{e}^{-\lambda_i}[(1-\lambda_i)\omega_i+(1-\omega_i)\mathrm{e}^{-\lambda_i}](1-\omega_i)\lambda_i}{\omega_i+(1-\omega_i)\mathrm{e}^{-\lambda_i}}\right.\\ &+ \left. \lambda_i(1-\omega_i)(1-e^{-\lambda_i})\right\} x_{ij}x_{ik}; \qquad (3.6)\end{aligned}$$

$$-\mathrm{E}\left[\frac{\partial^2 \ell}{\partial \omega_i^2}\right] = \sum_{i=1}^{n}\left\{ \frac{(1-e^{-\lambda_i})^2}{\omega_i+(1-\omega_i)\mathrm{e}^{-\lambda_i}} + \frac{(1-e^{-\lambda_i})}{(1-\omega_i)}\right\}; \qquad (3.7)$$

$$-\mathrm{E}\left[\frac{\partial^2 \ell}{\partial \beta_j \partial \omega_i}\right] = \sum_{i=1}^{n}\left\{ \frac{-\lambda_i \mathrm{e}^{-\lambda_i}}{\omega_i+(1-\omega_i)\mathrm{e}^{-\lambda_i}}\right\} x_{ij}. \qquad (3.8)$$

Hence the estimates of $\boldsymbol{\beta}$ and ω at the $(m+1)^{th}$ iteration, denoted by $\boldsymbol{\beta}^{(m+1)}$ and $\omega^{(m+1)}$, are given by

$$\begin{bmatrix} \boldsymbol{\beta}^{(m+1)} \\ \omega^{(m+1)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\beta}^{(m)} \\ \omega^{(m)} \end{bmatrix} + [\mathcal{I}^{(m)}(\boldsymbol{\beta}, \omega)]^{-1} \mathbf{s}^{(m)}(\boldsymbol{\beta}, \omega), \tag{3.9}$$

where the score vector and the expected information matrix respectively, evaluated at $\boldsymbol{\beta} = \boldsymbol{\beta}^{(m)}$ and $\omega = \omega^{(m)}$ are as follows.

$$\mathbf{s}(\boldsymbol{\beta}, \omega) = \begin{bmatrix} \mathbf{s}_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \omega) \\ \mathbf{s}_{\omega}(\boldsymbol{\beta}, \omega) \end{bmatrix} = \begin{bmatrix} \dfrac{\partial \ell}{\partial \boldsymbol{\beta}} \\ \dfrac{\partial \ell}{\partial \omega} \end{bmatrix}, \tag{3.10}$$

and

$$\mathcal{I}(\boldsymbol{\beta}, \omega) = \begin{bmatrix} \mathcal{I}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}, \omega) & \mathcal{I}_{\boldsymbol{\beta}\omega}(\boldsymbol{\beta}, \omega) \\ \mathcal{I}_{\omega\boldsymbol{\beta}}(\boldsymbol{\beta}, \omega) & \mathcal{I}_{\omega\omega}(\boldsymbol{\beta}, \omega) \end{bmatrix}, \tag{3.11}$$

where the elements $\mathcal{I}_{\boldsymbol{\beta}\boldsymbol{\beta}}, \mathcal{I}_{\boldsymbol{\beta}\omega} = \mathcal{I}_{\omega\boldsymbol{\beta}}$ and $\mathcal{I}_{\omega\omega}$ are, respectively,

$$-\mathrm{E}\left[\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}\right], \quad -\mathrm{E}\left[\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \omega}\right] \quad \text{and} \quad -\mathrm{E}\left[\frac{\partial^2 \ell}{\partial \omega^2}\right].$$

With good starting values $\boldsymbol{\beta}^{(0)}, \omega^{(0)}$ and hence $\boldsymbol{\lambda}^{(0)}, \omega^{(0)}$ the iterative scheme converges in a few step, convergence is obtained with a stopping rule, such as $|\ell^{(m+1)} - \ell^{(m)}| \leq \epsilon$, where $\ell^{(m)}$ and $\ell^{(m+1)}$ are the log-likelihood, $\ell(\boldsymbol{\lambda}, \omega; y)$ evaluated using the estimates of $\boldsymbol{\lambda}$ and ω from the (m) and $(m+1)$ iterations, respectively. The asymtotic variance-covariance matrix for $(\hat{\boldsymbol{\beta}}, \hat{\omega})$ is automatically provided in final iteration.

3.1.2 Maximum Likelihood Estimation for ZIP Models with no Covariates

Based on the log-likelihood function (2.22), the respectively maximum likelihood estimates, $\hat{\lambda}$ and $\hat{\omega}$ for λ and ω are the roots of the equations $\dfrac{\partial \ell(\lambda, \omega)}{\partial \lambda} = 0$ and $\dfrac{\partial \ell(\lambda, \omega)}{\partial \omega} = 0$. Here we have

$$\frac{\partial \ell(\lambda, \omega)}{\partial \lambda} = \frac{-n_0(1-\omega)e^{-\lambda})}{[\omega + (1-\omega)e^{-\lambda}]} - \sum_{j=1}^{J} n_j + \frac{\sum_{j=1}^{J}(n_j \times j)}{\lambda}, \tag{3.12}$$

and

$$\frac{\partial \ell(\lambda,\omega)}{\partial\omega} = \frac{n_0(1-e^{-\lambda})}{[\omega+(1-\omega)e^{-\lambda}]} - \frac{\sum_{j=1}^{J} n_j}{1-\omega}. \tag{3.13}$$

Setting each of these equal to zero gives

$$\frac{n_0(1-\hat{\omega})e^{-\hat{\lambda}}}{[\hat{\omega}+(1-\hat{\omega})e^{-\hat{\lambda}}]} + \sum_{j=1}^{J} n_j = \frac{\sum_{j=1}^{J}(n_j \times j)}{\hat{\lambda}}, \tag{3.14}$$

and

$$\frac{n_0(1-\hat{\omega})}{\hat{\omega}+(1-\hat{\omega})e^{-\hat{\lambda}}} = \frac{\sum_{j=1}^{J} n_j}{(1-e^{\hat{\lambda}})} \tag{3.15}$$

Substituting (3.15) into (3.14) gives

$$\begin{aligned}
\frac{\mathrm{e}^{-\hat{\lambda}}\sum_{j=1}^{n} n_j}{(1-\mathrm{e}^{-\hat{\lambda}})} + \sum_{j=1}^{n} n_j &= \frac{\sum_{j=1}^{n}(n_j \times j)}{\hat{\lambda}} \\
\sum_{j=1}^{n} n_j \left\{ \frac{\mathrm{e}^{-\hat{\lambda}} + 1 - \mathrm{e}^{-\hat{\lambda}}}{(1-\mathrm{e}^{-\hat{\lambda}})} \right\} &= \frac{\sum_{j=1}^{n}(n_j \times j)}{\hat{\lambda}} \\
\frac{\hat{\lambda}}{(1-\mathrm{e}^{-\hat{\lambda}})} &= \frac{\sum_{j=1}^{n}(n_j \times j)}{\sum_{j=1}^{n} n_j}
\end{aligned}$$

$$\hat{\lambda} = \frac{(1-\mathrm{e}^{-\hat{\lambda}})\sum_{j=1}^{n}(n_j \times j)}{\sum_{j=1}^{n} n_j}. \tag{3.16}$$

Note that this does not depend on ω or n_0. From (3.15),

$$\begin{aligned}
n_0(1-\mathrm{e}^{-\hat{\lambda}})(1-\hat{\omega}) &= [\hat{\omega}+(1-\hat{\omega})\mathrm{e}^{-\hat{\lambda}}]\sum_{j=1}^{n} n_j \\
n_0(1-\mathrm{e}^{-\hat{\lambda}}) - \hat{\omega}n_0(1-\mathrm{e}^{-\hat{\lambda}}) &= \hat{\omega}\sum_{j=1}^{n} n_j + (1-\hat{\omega})\mathrm{e}^{-\hat{\lambda}}\sum_{j=1}^{n} n_j \\
\hat{\omega}\sum_{j=1}^{n} n_j - \hat{\omega}\mathrm{e}^{-\hat{\lambda}}\sum_{j=1}^{n} n_j + \hat{\omega}n_0(1-\mathrm{e}^{-\hat{\lambda}}) &= n_0(1-\mathrm{e}^{-\hat{\lambda}}) - \mathrm{e}^{-\hat{\lambda}}\sum_{j=1}^{n} n_j
\end{aligned}$$

giving

$$\begin{aligned}
\hat{\omega} &= \frac{n_0 - (n_0 + \sum_{j=1}^{n} n_j)\mathrm{e}^{-\hat{\lambda}}}{(\sum_{j=1}^{n} n_j + n_0) - (\sum_{j=1}^{n} n_j + n_0)\mathrm{e}^{-\hat{\lambda}}} \\
&= \frac{n_0 - n\mathrm{e}^{-\hat{\lambda}}}{n(1-\mathrm{e}^{-\hat{\lambda}})}. \qquad (3.17)
\end{aligned}$$

These lead to a simple scheme for obtaining the maximum likelihood estimates of λ and ω :

- Step 1: fit a standard Poisson model to obtain an initial value, $\lambda^{(0)}$ for λ ;
- Step 2: use an iterative scheme for $\hat{\lambda}$

  $$\hat{\lambda}^{(m+1)} = \frac{(1 - \mathrm{e}^{-\hat{\lambda}^{(m)}}) \sum_{j=1}^{n} (n_j \times j)}{\sum_{j=1}^{n} n_j}. \tag{3.18}$$

  The iterations are repeated until converge, using a stopping rule, $|\hat{\lambda}^{(m+1)} - \hat{\lambda}^{(m)}| < \epsilon$, where$\hat{\lambda}^{(m)}$ and $\hat{\lambda}^{(m+1)}$ are estimates of λ at the $(m)^{th}$ and $(m+1)^{th}$ iteration, respectively.

- Step 3: obtain $\hat{\omega}$ by substituting $\hat{\lambda}$ given by the final iteration of step 2 in equation (3.17).

  Note that this gives a closed form expression for $\hat{\omega}$ and hence no iteration is required. Also (3.17) can be rewritten as $n[\hat{\omega} + (1 - \hat{\omega})\mathrm{e}^{-\hat{\lambda}}] = n_0$ which shows that the observed and fitted zero frequencies are identical.

Estimated covariance matrix of $\hat{\lambda}$ and $\hat{\omega}$, denoted by $\mathrm{Cov}(\hat{\lambda}, \hat{\omega})$, can be simply obtained using the expected information matrix, that is $\mathrm{Cov}(\hat{\lambda}, \hat{\omega}) = \mathcal{I}^{-1}(\hat{\lambda}, \hat{\omega})$, where

$$\mathcal{I}(\hat{\lambda}, \hat{\omega}) = \begin{bmatrix} \mathcal{I}_{\lambda\lambda}(\lambda, \omega) & \mathcal{I}_{\lambda\omega}(\lambda, \omega) \\ \mathcal{I}_{\omega\lambda}(\lambda, \omega) & \mathcal{I}_{\omega\omega}(\lambda, \omega) \end{bmatrix}_{\lambda=\hat{\lambda}, \omega=\hat{\omega}} . \tag{3.19}$$

The elements $\mathcal{I}_{\lambda\lambda}, \mathcal{I}_{\lambda\omega} = \mathcal{I}_{\omega\lambda}$ and $\mathcal{I}_{\omega\omega}$ are, respectively,

$$-\mathrm{E}\left[\frac{\partial^2 \ell(\lambda, \omega)}{\partial \lambda^2}\right], \quad -\mathrm{E}\left[\frac{\partial^2 \ell(\lambda, \omega)}{\partial \lambda \partial \omega}\right], \quad \text{and} \quad -\mathrm{E}\left[\frac{\partial^2 \ell(\lambda, \omega)}{\partial \omega^2}\right].$$

with

$$\begin{aligned} \frac{\partial^2 \ell(\lambda, \omega)}{\partial \lambda^2} &= \frac{n_0 \omega (1 - \omega) \mathrm{e}^{-\lambda}}{[\omega + (1 - \omega)\mathrm{e}^{-\lambda}]^2} - \frac{\sum_j (n_j \times j)}{\lambda^2}; \\ \frac{\partial^2 \ell(\lambda, \omega)}{\partial \lambda \partial \omega} &= \frac{n_0 \mathrm{e}^{-\lambda}}{[\omega + (1 - \omega)\mathrm{e}^{-\lambda}]^2}; \\ \frac{\partial^2 \ell(\lambda, \omega)}{\partial \omega^2} &= \frac{-n_0 (1 - \mathrm{e}^{-\lambda})^2}{[\omega + (1 - \omega)\mathrm{e}^{-\lambda}]^2} - \frac{\sum_j n_j}{(1 - \omega)^2}. \end{aligned} \tag{3.20}$$

Using the fact that

$$\begin{aligned} \mathrm{E}[I_{(y_i=0)}] &= \Pr(Y_i = 0) = \omega_i + (1 - \omega_i)\mathrm{e}^{-\lambda_i} \\ \text{and} \qquad \mathrm{E}[I_{(y_i>0)}] &= \Pr(Y_i > 0) = (1 - \omega_i)(1 - e^{-\lambda_i}) \end{aligned}$$

we have

$$\begin{aligned}
\mathcal{I}_{\lambda\lambda} = -\mathrm{E}\left[\frac{\partial^2\ell(\lambda,\omega)}{\partial\lambda^2}\right] &= n\left[\frac{1-\hat{\omega}}{\hat{\lambda}} - \frac{\hat{\omega}(1-\hat{\omega})\mathrm{e}^{-\hat{\lambda}}}{\hat{\omega}+(1-\hat{\omega})\mathrm{e}^{-\hat{\lambda}}}\right];\\
\mathcal{I}_{\omega\lambda} = -\mathrm{E}\left[\frac{\partial^2\ell(\lambda,\omega)}{\partial\lambda\partial\omega}\right] &= \frac{-n\mathrm{e}^{-\hat{\lambda}}}{\hat{\omega}+(1-\hat{\omega})\mathrm{e}^{-\hat{\lambda}}};\\
\mathcal{I}_{\omega\omega} = -\mathrm{E}\left[\frac{\partial^2\ell(\lambda,\omega)}{\partial\omega^2}\right] &= \frac{n(1-\mathrm{e}^{-\hat{\lambda}})}{(1-\hat{\omega})[\hat{\omega}+(1-\hat{\omega})\mathrm{e}^{-\hat{\lambda}}]}.
\end{aligned} \tag{3.21}$$

$\mathrm{Var}(\hat{\lambda})$ and $\mathrm{Var}(\hat{\omega})$ are obtained from $\mathcal{I}^{-1}(\lambda,\omega)$, in (3.19) using inverse of partitioned matrix (Searle, 1966) as follow:

$$\mathrm{Var}(\hat{\lambda}) = \left(\mathcal{I}_{\lambda\lambda} - \mathcal{I}_{\lambda\omega}\mathcal{I}_{\omega\omega}^{-1}\mathcal{I}_{\omega\lambda}\right)^{-1} \tag{3.22}$$

and

$$\mathrm{Var}(\hat{\omega}) = \left(\mathcal{I}_{\omega\omega} - \mathcal{I}_{\omega\lambda}\mathcal{I}_{\lambda\lambda}^{-1}\mathcal{I}_{\lambda\omega}\right)^{-1}. \tag{3.23}$$

3.2 Model Selection

Selecting an appropriate model can be used a standard likelihood information criteria, for example, Akaike information criteria (Akaike, 1973) or Baysians information criteria (Raftery, 1986) abbreviated by AIC and BIC , respectively, where

$$\mathrm{AIC} = -2\times\ell(\hat{\lambda},\hat{\omega}) + 2(\text{No.fitted parameters}); \tag{3.24}$$

$$\mathrm{BIC} = -2\times\ell(\hat{\lambda},\hat{\omega}) + \ln(n)\times(\text{No.fitted parameters}). \tag{3.25}$$

The model with smallest value of AIC or of BIC is preferable.

3.3 The Wald test for ZIP Models

In this thesis, we will develop a Wald test for ZIP model with constant λ and ω or for testing the hypothesis

$$\mathrm{H}_0:\omega=0;\quad \mathrm{H}_1:\omega>0. \tag{3.26}$$

Based on the basic idea of obtaining the Wald test, it is

$$W_\omega = \frac{\hat{\omega}^2}{\mathrm{Var}(\hat{\omega})} \tag{3.27}$$

where $\hat{\omega}$ is the maximum likelihood estimate of ω under the ZIP given in (3.17). That is

$$\hat{\omega} = \frac{n_0 - n\mathrm{e}^{-\hat{\lambda}}}{n(1-\mathrm{e}^{-\hat{\lambda}})}$$

and $\mathrm{Var}(\hat{\omega})$ given in (3.23) is

$$\mathrm{Var}(\hat{\omega}) = \left(\mathcal{I}_{\omega\omega} - \mathcal{I}_{\omega\lambda}\mathcal{I}_{\lambda\lambda}^{-1}\mathcal{I}_{\lambda\omega}\right)^{-1}.$$

Then

$$\begin{aligned}
\mathcal{I}_{\omega\omega} - \mathcal{I}_{\omega\lambda}\mathcal{I}_{\lambda\lambda}^{-1}\mathcal{I}_{\lambda\omega} &= n\left\{\frac{(1-\mathrm{e}^{-\hat{\lambda}})}{(1-\hat{\omega})[\hat{\omega}+(1-\hat{\omega})\mathrm{e}^{-\hat{\lambda}}]}\right. \\
&- \left.\frac{\hat{\lambda}\mathrm{e}^{-2\hat{\lambda}}}{(1-\hat{\omega})[\hat{\omega}+(1-\hat{\omega})\mathrm{e}^{-\hat{\lambda}}]\left\{[\hat{\omega}+(1-\hat{\omega})\mathrm{e}^{-\hat{\lambda}}]-\hat{\omega}\hat{\lambda}\mathrm{e}^{-\hat{\lambda}}\right\}}\right\} \\
&= \frac{\left\{[\hat{\omega}+(1-\hat{\omega})\mathrm{e}^{-\hat{\lambda}}]-\hat{\omega}\hat{\lambda}\mathrm{e}^{-\hat{\lambda}}\right\}-\hat{\lambda}\mathrm{e}^{-2\hat{\lambda}}}{(1-\hat{\omega})[\hat{\omega}+(1-\hat{\omega})\mathrm{e}^{-\hat{\lambda}}]\left\{[\hat{\omega}+(1-\hat{\omega})\mathrm{e}^{-\hat{\lambda}}]-\hat{\omega}\hat{\lambda}\mathrm{e}^{-\hat{\lambda}}\right\}}
\end{aligned} \tag{3.28}$$

Since $\hat{\omega}+(1-\hat{\omega})\mathrm{e}^{-\hat{\lambda}} = \dfrac{n_0}{n}$ and $(1-\omega) = \dfrac{\bar{y}}{\hat{\lambda}}$, (3.28) is simply defined as

$$\mathcal{I}_{\omega\omega} - \mathcal{I}_{\omega\lambda}\mathcal{I}_{\lambda\lambda}^{-1}\mathcal{I}_{\lambda\omega} = \frac{n^2\hat{\lambda}\left\{(1-\mathrm{e}^{-\hat{\lambda}})[n_0-(\hat{\lambda}-\bar{y})n\mathrm{e}^{-\hat{\lambda}}]-n\hat{\lambda}\mathrm{e}^{-2\hat{\lambda}}\right\}}{n_0\bar{y}(n_0-(\hat{\lambda}-\bar{y})n\mathrm{e}^{-\hat{\lambda}})}. \tag{3.29}$$

Hence,

$$\begin{aligned}
\mathrm{Var}(\hat{\omega}) &= (\mathcal{I}_{\omega\omega} - \mathcal{I}_{\omega\lambda}\mathcal{I}_{\lambda\lambda}^{-1}\mathcal{I}_{\lambda\omega})^{-1} \\
&= \frac{n_0\bar{y}[n_0-n\mathrm{e}^{-\hat{\lambda}}(\hat{\lambda}-\bar{y})]}{n^2\hat{\lambda}\left\{(1-\mathrm{e}^{-\hat{\lambda}})[n_0-n\mathrm{e}^{-\hat{\lambda}}(\hat{\lambda}-\bar{y})]-n\hat{\lambda}\mathrm{e}^{-2\hat{\lambda}}\right\}}.
\end{aligned} \tag{3.30}$$

The Wald test for the ZIP is then

$$W_\omega = \frac{\left(\dfrac{n_0 - n\mathrm{e}^{-\hat{\lambda}}}{n(1-\mathrm{e}^{-\hat{\lambda}})}\right)^2}{\dfrac{n_0\bar{y}[n_0 - n\mathrm{e}^{-\hat{\lambda}}(\hat{\lambda}-\bar{y})]}{n^2\hat{\lambda}\left\{(1-\mathrm{e}^{-\hat{\lambda}})[n_0 - n\mathrm{e}^{-\hat{\lambda}}(\hat{\lambda}-\bar{y})] - n\hat{\lambda}\mathrm{e}^{-2\hat{\lambda}}\right\}}}$$

$$= \frac{(n_0 - n\mathrm{e}^{-\hat{\lambda}})^2\hat{\lambda}\left\{(1-\mathrm{e}^{-\hat{\lambda}})[n_0 - n\mathrm{e}^{-\hat{\lambda}}(\hat{\lambda}-\bar{y})] - n\hat{\lambda}\mathrm{e}^{-2\hat{\lambda}}\right\}}{n_0\bar{y}(1-\mathrm{e}^{-\hat{\lambda}})^2[n_0 - n\mathrm{e}^{-\hat{\lambda}}(\hat{\lambda}-\bar{y})]}. \quad (3.31)$$

According to standard asymptotic theory, the sampling distribution of W_ω should be χ^2_1 distribution under H_0. However, for the ZIP model and the hypothesis (2.23) the null hypothesis corresponds to ω being on the boundary of the parameter space and the appropriate reference distribution is a mixture of chi-square distribution, see Self and Kung-Yee Liang (1987). For the simple constant ω the appropriate reference distribution is an equal mixture of a χ^2_0 (a constant of zero) and a χ^2_1 distribution, with p-value given by $\frac{1}{2}[\Pr\{\chi^2_1 \geq W_\omega\}]$.

3.4 Simulation study of the Wald test

In this section we conduct a small simulation study using R (R Development Core Team 2008) in order to investigate sampling distribution of W_ω and power of W_ω, including comparing its power with other tests described in section 2.5

3.4.1 Sampling distribution of W_ω

In order to investigate distributional approximation of W_ω, a small simulation study was carried out using R (R Development Core Team 2008). We explored single homogeneous samples and simulating as follow:

1. In the simulation, we explored for sample of size $n = 25, 50, 100$ and 200. For each sample size n we simulated 3000 set of responses are generated under

Poisson model with $\lambda = 1.25, 1.50, 2.00$ and 2.25, and fit a ZIP regression model to calculate W_ω.

2. In order to check whether the distribution of W_ω is the expected asymptotic χ_1^2 distribution, we computed the half of proportion of number of W_ω greater than or equal the critical value, $\chi_{1,\alpha}^2$. That is

$$\frac{\sum_{i=1}^{3000} I_{(W_\omega \geq \chi_{1,\alpha}^2)}}{3000 \times 2}, \tag{3.32}$$

α is the nominal size of the test. Here, we consider $\alpha = 0.01, 0.05$. Results from the study are presented in Table 3.1.

Table 3.1: Estimated upper tail probabilities for the Wald test at $\chi_{1,\alpha}^2$ based on 3000 samples

n		$\lambda = 1.25$		$\lambda = 1.50$		$\lambda = 2.00$		$\lambda = 2.25$	
	α	0.01	0.05	0.01	0.05	0.01	0.05	0.01	0.05
25		0.004	0.002	0.003	0.018	0.009	0.024	0.009	0.035
50		0.004	0.017	0.004	0.024	0.005	0.024	0.007	0.033
100		0.006	0.023	0.004	0.028	0.005	0.024	0.009	0.025
200		0.004	0.025	0.007	0.027	0.005	0.026	0.006	0.028

From the results on estimated upper tail probabilities for the Wald test we can see that the χ_1^2 distribution can be used as a reference sampling distribution in all situation studied here. For a fixed value of λ, estimate upper tail probabilities for W_ω at χ_1^2 close to $\frac{\alpha}{2}$ when n increases. Similarly pattern is also found for increasing value of λ and fixed n.

Noth here that $\frac{\sum_{i=1}^{3000} I_{(W_\omega \geq \chi_{1,\alpha}^2)}}{3000 \times 2}$ is approximately equal to $\frac{\sum_{i=1}^{3000} I_{(W_\omega \geq \chi_{1,\alpha/2}^2)}}{3000}$.

3.4.2 Power of W_ω and the compared test for ZIP

In order to investigate power of W_ω and compare that with the other tests reviewed in section 2.5, a small simulation study was carried out using R (R Development Core Team 2008). For each sample size $n = 25, 50, 100$ and 200 we simulate 3000 set of responses under ZIP model with $\lambda = 1.25, 1.50, 2.00, 2.25$ and $\omega = 0.25, 0.35, 0.45, 0.55, 0.65, 0.75$. For each set of generated data, a ZIP model is fitted for calculating the W_ω and the other mentioned tests in section 2.5 followed by the powers of the tests. Results from the simulation study are presented in Table 3.2-3.9.

From the results in Table 3.2-3.9, we can see that for a fixed value of ω, the power of the those tests increases when sample size n increases. Similarly pattern is also found for increasing value of ω and fixed n. It is interesting to note that when the value of ω increases, number of excess zeros is large. The power of those tests are also good for this the situation. For example, when sample size is 200, and parameters are 2.00 and 0.65 for λ and ω, respectively, the powers are higher than 0.997.

Moreover, these tests are all good for comparing between Poisson and ZIP models. It can be seen that the Wald test is as good as the cochran test, but it is better than likelihood ratio test, score test and chi - square test. Thus, the Wald test can be an alternative test for comparing between Poisson and ZIP models.

Table 3.2: The power of the six tests for $\lambda = 1.25(\alpha = 0.01)$

	ω	0.25	0.35	0.45	0.55	0.65	0.75
$n = 25$	Wald test	0.137	0.271	0.438	0.579	0.700	-
	Likelihood ratio test	0.041	0.090	0.162	0.233	0.287	-
	Score test	0.041	0.090	0.163	0.234	0.294	-
	Chi-square test	0.093	0.142	0.203	0.273	0.316	-
	Confidence interval test	0.127	0.277	0.478	0.637	0.769	-
	Cochran test	0.063	0.126	0.219	0.314	0.412	-
$n = 50$	Wald test	0.260	0.494	0.686	0.826	0.879	0.927
	Likelihood ratio test	0.110	0.251	0.411	0.544	0.603	0.628
	Score test	0.117	0.263	0.426	0.593	0.656	0.719
	Chi-square test	0.163	0.248	0.383	0.504	0.589	0.638
	Confidence interval test	0.285	0.569	0.792	0.922	0.952	0.972
	Cochran test	0.161	0.327	0.493	0.651	0.716	0.768
$n = 100$	Wald test	0.481	0.783	0.934	0.978	0.991	0.991
	Likelihood ratio test	0.326	0.604	0.798	0.907	0.943	0.939
	Score test	0.328	0.607	0.800	0.914	0.955	0.964
	Chi-square test	0.292	0.486	0.685	0.812	0.899	0.912
	Confidence interval test	0.563	0.859	0.973	0.995	0.998	0.999
	Cochran test	0.416	0.694	0.857	0.937	0.969	0.977
$n = 200$	Wald test	0.816	0.975	0.999	1.000	1.000	1.000
	Likelihood ratio test	0.712	0.945	0.993	0.999	0.999	1.000
	Score test	0.710	0.945	0.993	0.999	0.999	1.000
	Chi-square test	0.552	0.845	0.964	0.993	0.998	0.999
	Confidence interval test	0.883	0.993	1.000	1.000	1.000	1.000
	Cochran test	0.786	0.964	0.998	1.000	1.000	1.000

- The procedures gives missing values, because the probability of ω is closed to one.

Table 3.3: The power of the six tests for $\lambda = 1.25(\alpha = 0.05)$

	ω	0.25	0.35	0.45	0.55	0.65	0.75
$n = 25$	Wald test	0.211	0.370	0.544	0.667	0.762	-
	Likelihood ratio test	0.107	0.197	0.301	0.383	0.444	-
	Score test	0.107	0.198	0.303	0.388	0.461	-
	Chi-square test	0.136	0.191	0.266	0.343	0.398	-
	Confidence interval test	0.262	0.456	0.653	0.781	0.864	-
	Cochran test	0.186	0.307	0.427	0.526	0.595	-
$n = 50$	Wald test	0.385	0.634	0.803	0.905	0.920	0.942
	Likelihood ratio test	0.272	0.462	0.615	0.742	0.778	0.790
	Score test	0.271	0.462	0.615	0.750	0.800	0.837
	Chi-square test	0.233	0.356	0.502	0.635	0.687	0.728
	Confidence interval test	0.472	0.739	0.890	0.960	0.971	0.974
	Cochran test	0.360	0.573	0.712	0.826	0.864	0.897
$n = 100$	Wald test	0.651	0.875	0.967	0.991	0.995	0.995
	Likelihood ratio test	0.541	0.801	0.922	0.964	0.981	0.976
	Score test	0.541	0.805	0.925	0.970	0.984	0.984
	Chi-square test	0.410	0.631	0.804	0.897	0.943	0.952
	Confidence interval test	0.753	0.945	0.991	0.998	0.999	0.999
	Cochran test	0.650	0.867	0.961	0.984	0.991	0.990
$n = 200$	Wald test	0.910	0.992	1.000	1.000	1.000	1.000
	Likelihood ratio test	0.872	0.984	0.999	1.000	1.000	1.000
	Score test	0.868	0.983	0.999	1.000	1.000	1.000
	Chi-square test	0.714	0.932	0.989	0.999	0.999	0.999
	Confidence interval test	0.955	0.998	1.000	1.000	1.000	1.000
	Cochran test	0.918	0.992	1.000	1.000	1.000	1.000

- The procedures gives missing values, because the probability of ω is closed to one.

Table 3.4: The power of the six tests for $\lambda = 1.50(\alpha = 0.01)$

	ω	0.25	0.35	0.45	0.55	0.65	0.75
$n = 25$	Wald test	0.140	0.277	0.434	0.577	0.700	-
	Likelihood ratio test	0.041	0.093	0.154	0.234	0.287	-
	Score test	0.041	0.093	0.155	0.236	0.294	-
	Chi-square test	0.097	0.142	0.203	0.277	0.316	-
	Confidence interval test	0.129	0.282	0.468	0.644	0.769	-
	Cochran test	0.062	0.135	0.217	0.319	0.412	-
$n = 50$	Wald test	0.240	0.476	0.681	0.821	0.893	0.928
	Likelihood ratio test	0.109	0.245	0.399	0.538	0.622	0.637
	Score test	0.115	0.252	0.417	0.579	0.685	0.735
	Chi-square test	0.150	0.266	0.375	0.503	0.611	0.653
	Confidence interval test	0.273	0.552	0.773	0.909	0.964	0.977
	Cochran test	0.156	0.318	0.492	0.643	0.742	0.781
$n = 100$	Wald test	0.481	0.789	0.935	0.981	0.989	0.993
	Likelihood ratio test	0.327	0.617	0.822	0.908	0.939	0.935
	Score test	0.330	0.622	0.824	0.915	0.949	0.961
	Chi-square test	0.267	0.482	0.689	0.811	0.893	0.914
	Confidence interval test	0.568	0.870	0.974	0.995	0.998	0.998
	Cochran test	0.410	0.702	0.870	0.941	0.969	0.972
$n = 200$	Wald test	0.801	0.978	0.999	1.000	1.000	1.000
	Likelihood ratio test	0.777	0.945	0.992	0.999	1.000	1.000
	Score test	0.699	0.945	0.992	0.999	1.000	1.000
	Chi-square test	0.545	0.839	0.966	0.993	0.997	0.998
	Confidence interval test	0.883	0.994	1.000	1.000	1.000	1.000
	Cochran test	0.774	0.969	0.997	1.000	1.000	1.000

- The procedures gives missing values, because the probability of ω is closed to one.

Table 3.5: The power of the six tests for $\lambda = 1.50(\alpha = 0.05)$

	ω	0.25	0.35	0.45	0.55	0.65	0.75
$n = 25$	Wald test	0.226	0.380	0.542	0.677	0.762	-
	Likelihood ratio test	0.112	0.205	0.306	0.385	0.444	-
	Score test	0.115	0.206	0.306	0.389	0.461	-
	Chi-square test	0.139	0.207	0.271	0.349	0.398	-
	Confidence interval test	0.282	0.464	0.648	0.791	0.864	-
	Cochran test	0.203	0.324	0.432	0.531	0.595	-
$n = 50$	Wald test	0.385	0.616	0.786	0.891	0.933	0.945
	Likelihood ratio test	0.258	0.447	0.617	0.736	0.793	0.801
	Score test	0.258	0.447	0.617	0.740	0.812	0.852
	Chi-square test	0.234	0.364	0.493	0.630	0.713	0.739
	Confidence interval test	0.471	0.719	0.880	0.949	0.977	0.979
	Cochran test	0.345	0.552	0.712	0.820	0.879	0.903
$n = 100$	Wald test	0.653	0.877	0.967	0.991	0.994	0.996
	Likelihood ratio test	0.546	0.804	0.926	0.969	0.978	0.973
	Score test	0.546	0.806	0.930	0.973	0.982	0.985
	Chi-square test	0.405	0.641	0.809	0.893	0.942	0.951
	Confidence interval test	0.754	0.944	0.989	0.998	0.999	0.999
	Cochran test	0.651	0.867	0.963	0.985	0.989	0.991
$n = 200$	Wald test	0.906	0.993	1.000	1.000	1.000	1.000
	Likelihood ratio test	0.861	0.986	0.999	1.000	1.000	1.000
	Score test	0.858	0.985	0.999	1.000	1.000	1.000
	Chi-square test	0.699	0.926	0.989	0.999	1.000	1.000
	Confidence interval test	0.956	0.999	1.000	1.000	1.000	1.000
	Cochran test	0.918	0.995	1.000	1.000	1.000	1.000

- The procedures gives missing values, because the probability of ω is closed to one.

Table 3.6: The power of the six tests for $\lambda = 2.00(\alpha = 0.01)$

	ω	0.25	0.35	0.45	0.55	0.65	0.75
$n = 25$	Wald test	0.156	0.292	0.439	0.598	0.700	-
	Likelihood ratio test	0.049	0.098	0.157	0.236	0.287	-
	Score test	0.049	0.098	0.158	0.238	0.294	-
	Chi-square test	0.100	0.152	0.206	0.269	0.316	-
	Confidence interval test	0.148	0.298	0.474	0.658	0.769	-
	Cochran test	0.070	0.140	0.213	0.332	0.412	-
$n = 50$	Wald test	0.250	0.477	0.670	0.808	0.884	0.925
	Likelihood ratio test	0.111	0.240	0.392	0.533	0.611	0.648
	Score test	0.117	0.248	0.413	0.574	0.671	0.746
	Chi-square test	0.157	0.260	0.369	0.487	0.583	0.642
	Confidence interval test	0.276	0.555	0.778	0.910	0.963	0.971
	Cochran test	0.152	0.320	0.473	0.632	0.733	0.787
$n = 100$	Wald test	0.460	0.783	0.938	0.983	0.992	0.995
	Likelihood ratio test	0.319	0.605	0.818	0.907	0.945	0.937
	Score test	0.322	0.612	0.822	0.912	0.960	0.970
	Chi-square test	0.271	0.492	0.687	0.827	0.898	0.917
	Confidence interval test	0.555	0.865	0.974	0.996	1.000	0.999
	Cochran test	0.329	0.689	0.872	0.942	0.973	0.977
$n = 200$	Wald test	0.796	0.979	1.000	1.000	1.000	1.000
	Likelihood ratio test	0.694	0.945	0.993	0.999	1.000	1.000
	Score test	0.691	0.945	0.993	0.999	1.000	1.000
	Chi-square test	0.539	0.840	0.963	0.990	0.997	0.997
	Confidence interval test	0.877	0.993	1.000	1.000	1.000	1.000
	Cochran test	0.767	0.968	0.998	1.000	1.000	1.000

- The procedures gives missing values, because the probability of ω is closed to one.

Table 3.7: The power of the six tests for $\lambda = 2.00(\alpha = 0.05)$

	ω	0.25	0.35	0.45	0.55	0.65	0.75
$n = 25$	Wald test	0.234	0.389	0.546	0.683	0.762	-
	Likelihood ratio test	0.121	0.208	0.301	0.401	0.444	-
	Score test	0.122	0.209	0.302	0.405	0.461	-
	Chi-square test	0.139	0.209	0.271	0.341	0.398	-
	Confidence interval test	0.284	0.470	0.650	0.801	0.864	-
	Cochran test	0.204	0.324	0.437	0.545	0.595	-
$n = 50$	Wald test	0.388	0.617	0.786	0.894	0.929	0.940
	Likelihood ratio test	0.258	0.447	0.598	0.724	0.785	0.806
	Score test	0.258	0.447	0.599	0.731	0.807	0.851
	Chi-square test	0.232	0.366	0.480	0.611	0.693	0.730
	Confidence interval test	0.469	0.727	0.887	0.949	0.975	0.973
	Cochran test	0.351	0.550	0.705	0.811	0.868	0.900
$n = 100$	Wald test	0.640	0.880	0.967	0.991	0.998	0.997
	Likelihood ratio test	0.519	0.798	0.930	0.969	0.982	0.979
	Score test	0.520	0.800	0.934	0.973	0.986	0.986
	Chi-square test	0.404	0.645	0.812	0.902	0.948	0.955
	Confidence interval test	0.747	0.942	0.990	0.999	1.000	1.000
	Cochran test	0.639	0.871	0.960	0.986	0.993	0.994
$n = 200$	Wald test	0.909	0.992	1.000	1.000	1.000	1.000
	Likelihood ratio test	0.863	0.985	1.000	1.000	1.000	1.000
	Score test	0.858	0.985	1.000	1.000	1.000	1.000
	Chi-square test	0.694	0.925	0.987	0.998	0.999	0.999
	Confidence interval test	0.955	0.999	1.000	1.000	1.000	1.000
	Cochran test	0.918	0.994	1.000	1.000	1.000	1.000

- The procedures gives missing values, because the probability of ω is closed to one.

Table 3.8: The power of the six tests for $\lambda = 2.25(\alpha = 0.01)$

	ω	0.25	0.35	0.45	0.55	0.65	0.75
$n = 25$	Wald test	0.156	0.292	0.439	0.598	0.700	-
	Likelihood ratio test	0.049	0.098	0.157	0.236	0.287	-
	Score test	0.049	0.098	0.158	0.238	0.294	-
	Chi-square test	0.100	0.152	0.206	0.269	0.316	-
	Confidence interval test	0.148	0.298	0.474	0.658	0.769	-
	Cochran test	0.070	0.140	0.213	0.332	0.412	-
$n = 50$	Wald test	0.248	0.488	0.670	0.818	0.894	0.926
	Likelihood ratio test	0.117	0.236	0.397	0.536	0.616	0.637
	Score test	0.125	0.245	0.411	0.575	0.675	0.732
	Chi-square test	0.154	0.256	0.378	0.498	0.599	0.640
	Confidence interval test	0.280	0.561	0.773	0.913	0.961	0.974
	Cochran test	0.157	0.314	0.489	0.645	0.739	0.780
$n = 100$	Wald test	0.458	0.770	0.940	0.979	0.990	0.994
	Likelihood ratio test	0.297	0.589	0.813	0.907	0.944	0.934
	Score test	0.298	0.596	0.816	0.917	0.956	0.969
	Chi-square test	0.259	0.467	0.685	0.819	0.889	0.916
	Confidence interval test	0.544	0.851	0.978	0.992	0.999	0.999
	Cochran test	0.388	0.670	0.872	0.944	0.970	0.979
$n = 200$	Wald test	0.804	0.974	0.998	1.000	1.000	1.000
	Likelihood ratio test	0.703	0.940	0.992	0.998	0.999	0.999
	Score test	0.702	0.940	0.992	0.999	1.000	1.000
	Chi-square test	0.546	0.847	0.963	0.991	0.998	0.999
	Confidence interval test	0.875	0.990	0.999	1.000	1.000	1.000
	Cochran test	0.775	0.961	0.997	1.000	1.000	1.000

- The procedures gives missing values, because the probability of ω is closed to one.

Table 3.9: The power of the six tests for $\lambda = 2.25 (\alpha = 0.05)$

	ω	0.25	0.35	0.45	0.55	0.65	0.75
$n = 25$	Wald test	0.234	0.389	0.546	0.683	0.762	-
	Likelihood ratio test	0.121	0.208	0.301	0.401	0.444	-
	Score test	0.122	0.209	0.302	0.405	0.461	-
	Chi-square test	0.139	0.209	0.271	0.341	0.393	-
	Confidence interval test	0.284	0.470	0.650	0.801	0.864	-
	Cochran test	0.204	0.324	0.437	0.545	0.595	-
$n = 50$	Wald test	0.388	0.627	0.791	0.895	0.934	0.942
	Likelihood ratio test	0.261	0.463	0.610	0.731	0.797	799
	Score test	0.261	0.463	0.611	0.741	0.820	0.848
	Chi-square test	0.241	0.360	0.486	0.623	0.702	0.734
	Confidence interval test	0.471	0.744	0.883	0.950	0.973	0.974
	Cochran test	0.352	0.568	0.705	0.819	0.879	0.898
$n = 100$	Wald test	0.633	0.870	0.969	0.987	0.994	0.996
	Likelihood ratio test	0.514	0.782	0.929	0.969	0.978	0.978
	Score test	0.515	0.785	0.931	0.971	0.983	0.987
	Chi-square test	0.384	0.628	0.806	0.900	0.936	0.951
	Confidence interval test	0.738	0.943	0.992	0.998	0.999	1.000
	Cochran test	0.638	0.858	0.964	0.983	0.990	0.993
$n = 200$	Wald test	0.902	0.991	0.999	1.000	1.000	1.000
	Likelihood ratio test	0.863	0.982	0.998	1.000	1.000	1.000
	Score test	0.860	0.982	0.998	1.000	1.000	1.000
	Chi-square test	0.702	0.928	0.987	0.997	1.000	1.000
	Confidence interval test	0.949	0.998	0.999	1.000	1.000	1.000
	Cochran test	0.913	0.992	1.000	1.000	1.000	1.000

- The procedures gives missing values, because the probability of ω is closed to one.

3.5 Model Diagnostics

Firstly, Half-normal plots proposed by Atkinson (1985) are very useful graphical method for verifying those model diagnostic using residuals. The plot developed from the Q-Q plot by using the absolute values of the normal linear model residuals. Superimposed with simulated envelope the plot can be applied to detect both the systematic departure from the model and influential values.

Such a plot is particular valuable when trying to decide whether an unacceptably large deviance is due to a small number of outlying observations (with only the largest residuals lying above the upper envelope) or a more general lack of fit or overdispersion (with both large and small residual lying above the upper envenlope). The graphic is obtained by drawing some ordered absolute values of a suitable diagnostic measure, $d_{(i)}$ against the half-normal scores or the expected ordered statistics.

Half-normal plots with a simulated envelope can be used to check the adequacy of the selected ZIP regression model, especially, the model for $\boldsymbol{\lambda}$. For the model for $\boldsymbol{\omega}$, it is known that overdispersion cannot occour in the binary data (Jansakul, 2005). Here, diagnostic quantities, the standardized Pearson residual will be used as. For given $\hat{\omega}_i$ the i^{th} term of this residual is

- Standardized Pearson residual can be written as

$$\dot{r}_{zip,i} = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i \left(1 + \frac{\hat{\omega}_i}{1 - \hat{\omega}_i}\hat{\mu}_i\right)}}. \tag{3.33}$$

Using this residual for construction of the half-normal plot with simulated envelope are given below.

1. Calculate the ordered absolute value of a diagnostic measurement denoted by $d_{(i)}$ from a selected ZIP regression model with $\hat{\boldsymbol{\lambda}}$ and $\hat{\boldsymbol{\omega}}$.

2. Simulate 19 sample for the response variable, $Y_{ji} \sim \text{ZIP}(\hat{\lambda}_i, \hat{\omega}_i)$, $j = 1, 2, \ldots, 19; i = 1, 2, \ldots, n$ by simulating pairs of random variables

$Y_{pj,i} \sim \text{Pois}(\hat{\lambda}_i)$ and $Z_{ji} \sim \text{Bin}(1, \hat{\omega}_i)$ and then forming
$Y_{ji} = 0 \times Z_{ji} + (1 - Z_{ji}) \times Y_{pj,i}$.

3. Refit the model using the same explanatory variables to each sample and calculate the ordered absolute value of a diagnostic of interest, $d^*_{j(i)}$.

4. For each i calculate the minimum, mean and maximum of $d^*_{j(i)}$.

5. Plot these values and the observed $d_{(i)}$ against the half-normal order statistics.

If the selected model is appropriate, the observed $d_{(i)}$ should lie within the simulated envelope.

CHAPTER 4

Applications

This chapter illustrates the use of ZIP and the proposed Wald test using three sets of data: the set of AIDS-related data discussed in Heilbron (1994) for ZIP regression analysis with mean counts depending on covariate, the foetal lamb movement data and the death notice data of London times (Gupta et al 1996), as single sample cases. In these three example, the authors showed that a standard Poisson model is not adequate because the data have many zeros and have gone on to consider the use of ZIP.

4.1 AIDS-related Data

The set of AIDS-related data discussed in Heilbron (1994). The study involved 1115 respondents aged 18-49. The response variable was the self-reported number of times that the respondents had anal intercourse with opposite sex partners during the study period classified by to dichotomous explanatory variables; sex (male, female) and having a risky (yes, no) main sexual partner. The data are presented in Table 4.1 with sex and risk are dummy variable:

$$\text{sex} = \begin{cases} 0 & \text{male} \\ 1 & \text{female} \end{cases}$$

and

$$\text{risk} = \begin{cases} 0 & \text{no} \\ 1 & \text{yes} \end{cases}$$

Table 4.1: AIDS-related Data

		Gender			
		Male		Female	
	Risky Partner	No	Yes	No	Yes
Y					
0		541	102	238	103
1		19	5	8	6
2		17	8	.	4
3		16	2	2	2
4		3	1	1	.
5		6	4	1	1
6		5	1	1	.
7		2	.	1	.
10		6	.	.	.
12		1	.	.	.
15		.	1	1	.
20		3	.	.	.
30		1	.	.	.
37		.	1	.	.
50		.	.	.	1
n		620	125	253	117
Mean		0.56	0.87	0.20	0.64
Variance		5.43	13.76	1.46	21.72

Fitting all possible Poisson models, the model with smallest value of AIC is the full interaction model, see Table 4.2.

Table 4.2: AIDS-related Data: all possible Poisson models

Models ($\ln(\boldsymbol{\lambda})$)	Residual deviance	d.f	-2ℓ	AIC
$\hat{\beta}_0$	3182.0	1114	3564.11	3566.1
$\hat{\beta}_0 + \hat{\beta}_1 sex$	3143.8	1113	3525.86	3529.9
$\hat{\beta}_0 + \hat{\beta}_1 risk$	3151.9	1113	3533.96	3538.0
$\hat{\beta}_0 + \hat{\beta}_1 sex + \hat{\beta}_2 risk$	3099.5	1112	3481.57	3487.6
$\hat{\beta}_0 + \hat{\beta}_1 sex + \hat{\beta}_2 risk + \hat{\beta}_3 sex : risk$	3087.0	1111	3469.78	3477.8

Where the maximum likelihood estimates of the linear predictor coefficients and their standard error are displayed in equation (4.1).

$$\begin{array}{l} \ln(\boldsymbol{\lambda}) = -0.575 - 1.027(sex) + 0.438(risk) + 0.719(sex : risk) \\ \qquad\quad (0.054) \qquad (0.150) \qquad\quad (0.110) \qquad\qquad (0.212) \end{array} \tag{4.1}$$

However, the half-normal plots with a simulated envelope, Figure 4.1 indicates that the Poisson model (4.1) is not appropriate. It shows overdispersion.

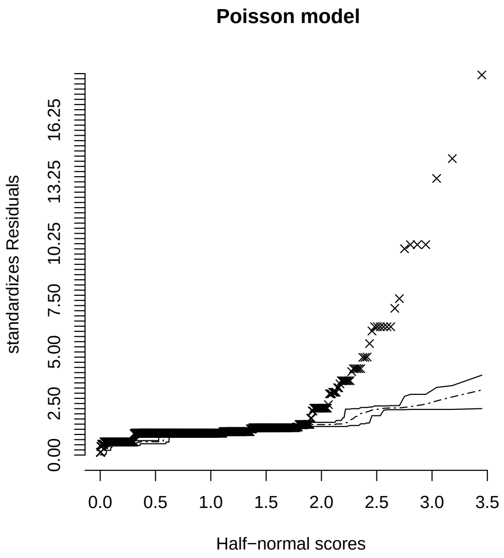

Figure 4.1: Half normal plot of AIDS-related Data

We tried fitting all possible ZIP regression with constant ω, we found that the interaction model for $\boldsymbol{\lambda}$ gives smallest value of AIC, and the value of -2ℓ is reduced nearly to a half of one given by the model (4.1), see Table 4.3.

Table 4.3: AIDS-related Data: all possible ZIP models

Model				
$\ln(\boldsymbol{\lambda})$	ω	d.f	-2ℓ	AIC
$\hat{\beta}_0$	0.865	1113	1869.3	1873.3
$\hat{\beta}_0 + \hat{\beta}_1 sex$	0.865	1112	1869.1	1875.1
$\hat{\beta}_0 + \hat{\beta}_2 risk$	0.865	1112	1866.0	1872.0
$\hat{\beta}_0 + \hat{\beta}_1 sex + \hat{\beta}_2 risk$	0.865	1111	1865.2	1873.2
$\hat{\beta}_0 + \hat{\beta}_1 sex + \hat{\beta}_2 risk + \hat{\beta}_3 sex : risk$	0.865	1110	1860.9	1870.9

However, the half-normal plots, Figure 4.2 indicates that the ZIP model is not consistent with the data.

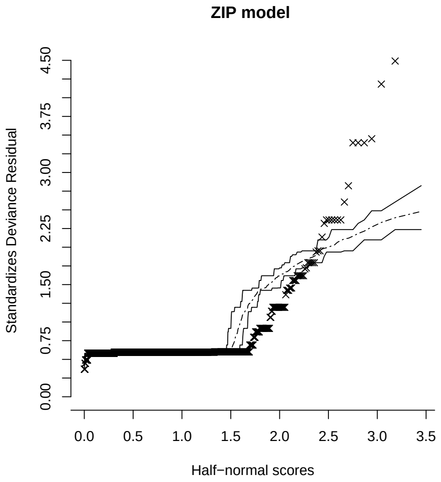

Figure 4.2: Half normal plot of AIDS-related Data

From above result show that the observed counts include many zeros, but the ZIP model is not still suitable. It indicates that the observed zeros are not structural zeros. They occur under Poisson state and have additional variation to the mean or heterogeneity. An appropriate model for this data might be an Negative Binomial model (NB).

Following Natchadamon (2006), we fitted linear mean variance negative binomial (NB1) model to this set data and present all possible NB1 model in Table 4.4.

Table 4.4: AIDS-related Data: all possible NB1 models

Model			
$\ln(\boldsymbol{\lambda})$	d.f	-2ℓ	AIC
$\hat{\beta}_0$	1113	1407.7	1411.7
$\hat{\beta}_0 + \hat{\beta}_1 sex$	1112	1403.9	1409.9
$\hat{\beta}_0 + \hat{\beta}_2 risk$	1112	1404.3	1410.3
$\hat{\beta}_0 + \hat{\beta}_1 sex + \hat{\beta}_2 risk$	1111	1392.8	1400.8
$\hat{\beta}_0 + \hat{\beta}_1 sex + \hat{\beta}_2 risk + \hat{\beta}_3 sex : risk$	1110	1392.3	1402.3

The procedure worked well and selected model indicated by AIC of 1400.8 is

$$\ln(\boldsymbol{\lambda}) = \underset{(0.152)}{-0.584} - \underset{(0.213)}{0.685}(sex) + \underset{(0.196)}{0.489}(risk) \qquad \underset{(1.765)}{\hat{\alpha} = 9.934} \tag{4.2}$$

The half-normal plots of NB1 model (4.2) displayed in Figure 4.3 indicates that the model is consistent with the data.

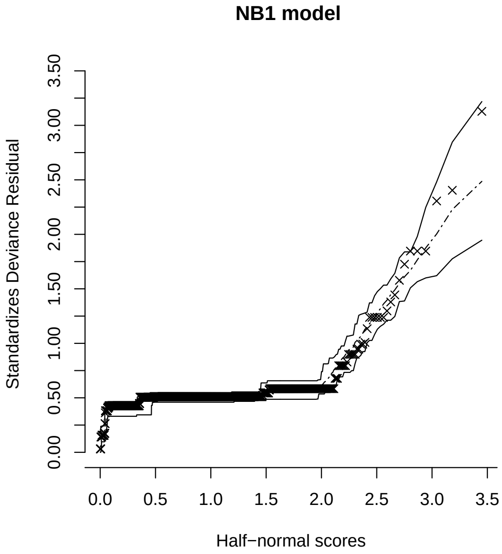

Figure 4.3: Half normal plot of AIDS-related Data

4.2 foetal lamb movement data

Foetal lamb movement data are number of movements made by a foetal lamb in each of 240 consecutive 5-second intervals which are summarized in Table 4.5.

Table 4.5: movements made by a foetal lamb in 240 consecutive 5-second intervals.

Number of movements	0	1	2	3	4	5	6	7
Number of intervals	182	41	12	2	0	2	0	1

The mean and variance under the constant Poisson model are 0.3667 and 0.7269, respectively. It shows overdispersion. To check whether the overdispersion made by excess zeros or zero-inflation, we calculate the Wald test and the others mentioned here with and the MLE forλ and ω under ZIP model: $\hat{\lambda} = 0.8999$ and $\hat{\omega} = 0.5927$, respectively. The computed test statistics are presented in Table 4.6

Table 4.6: Test statistics for the foetal lamb movement data at $\alpha = 0.05$

Test	Test statistics	P-value	Reject/not reject H_0
Likelihood ratio test	24.7867	0.000001	Reject
Score test	27.9372	0.000000	Reject
Chi-square test*	17.6341	0.000523	Reject
Confidence interval test	0.4674	-	Reject
C test	5.2856	0.000000	Reject
Wald test	86.9585	0.000000	Reject

* We defined the classes for this data are {{0}, {1}, {2}, {3 and above}}.

The results from Table 4.6 show that the ZIP model with $\hat{\lambda} = 0.8999$ (0.155) and $\hat{\omega} = 0.5927$ (0.0636) is consistent with the data. The quantities in parentheses are estimated standard errors. The half-normal plots for the Poisson fit and the ZIP model are displayed in Figure 4.4 and 4.5 respectively, Show that the ZIP model is a better fit than the Poisson model, although there are some outliers.

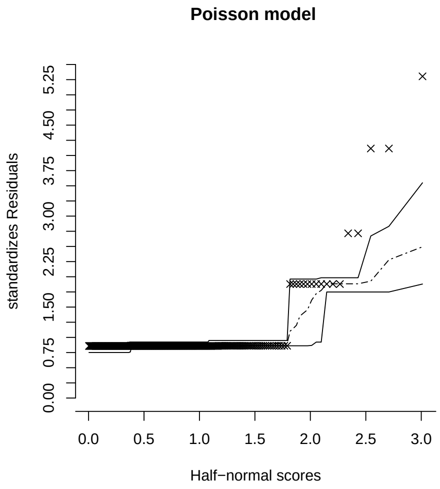

Figure 4.4: Half normal plot of foetal lamb movement data

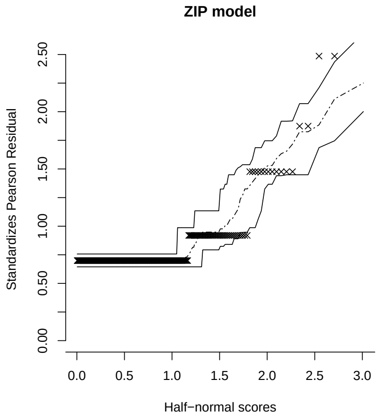

Figure 4.5: Half normal plot of foetal lamb movement data

4.3 Death Notice Data of London Times

The set of Death Notice Data of London Times (Gupta et al, 1996) shown in Table 4.7 presents the number of death notice, y_k of women 80 years of age and over, appearing in the London "Times"on each day for three consecutive years with the frequency, f_k for each $k = 0, 1, 2, \ldots$.

Table 4.7: Death Notice Data of London Times

Number of Death notices, y_k	Frequency (f_k)
0	162
1	267
2	271
3	185
4	111
5	61
6	27
7	8
8	3
9	1

The mean and variance under the constant Poisson model are 2.1569 and 2.6073, respectively. It shows overdispersion. To check whether the overdispersion made by excess zeros or zero-inflation, we calculate the Wald test and the others mentioned here with and the MLE forλ and ω under ZIP model: $\hat{\lambda} = 2.2693$ and $\hat{\omega} = 0.0495$, respectively. The computed test statistics are presented in Table 4.8.

Table 4.8: Test statistics for Death Notice Data of London Times at $\alpha = 0.05$

Test	Test statistics	Critical region	Reject/not reject H_0
Likelihood ratio test	14.7832	0.000121	Reject
Score test	15.4085	0.000087	Reject
Chi-square test*	30.8484	0.000066	Reject
Confidence interval test	0.0156	-	Reject
C test	3.9254	0.000008	Reject
Wald test	13.7581	0.000208	Reject

* We defined the classes for this data are {{0}, {1}, {2},{3},{4}, {5}, {6}, {7 and above}}.

The results from Table 4.8 show that the ZIP model with $\hat{\lambda} = 2.2693$ (0.0543) and $\hat{\omega} = 0.0495$ (0.0134) is consistent with the data. The quantities in parentheses are estimated standard errors.

The half-normal plots for the Poisson fit and the ZIP model are displayed in Figure 4.6 and 4.7 respectively, show that the ZIP model is a better fit than the Poisson model.

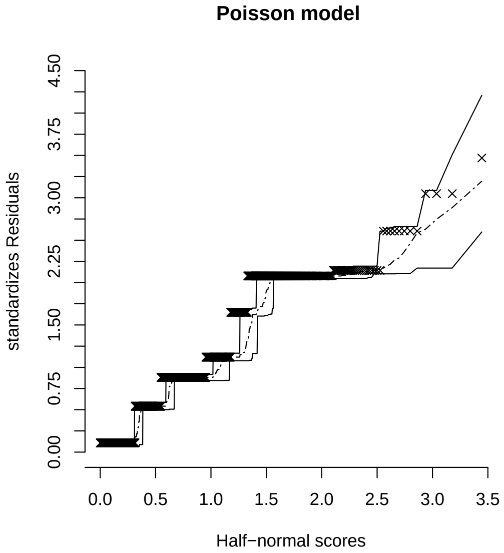

Figure 4.6: Half normal plot of Death Notice Data of London Times

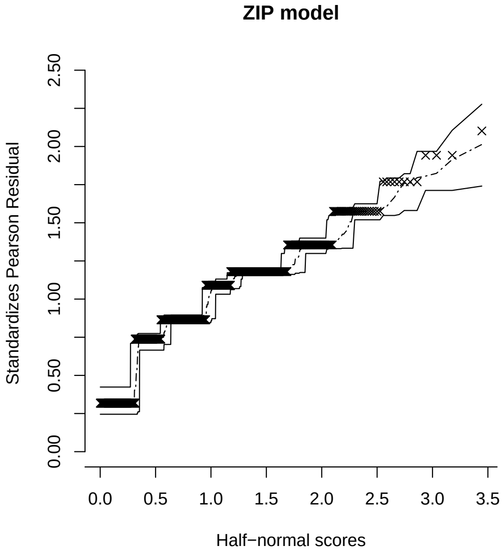

Figure 4.7: Half normal plot of Death Notice Data of London Times

CHAPTER 5

Conclusion and Discussion

In this thesis, we studied characteristics, theories and properties of glms and ZIP regression models and explore various tests for zero-inflation parameter used to compare between Poisson and ZIP models proposed in literature. Moreover, we developed a Wald test for a ZIP model in the case of single homogenous sample for testing the hypothesis (2.23). In order to investigated its sampling distribution and compared power of the test between W_ω and the others a small simulation study is conducted using R (R Development Core Team 2008).

5.1 Conclusion

Even though Poisson regression or log linear model is a basic model for count data analysis, in practice, it is commonly replaced by more complicated models. This because of the equality of mean and variance assumption under Poisson distribution fails. Counts that have greater variance than the mean are described as overdispersed Poisson counts (McCullagh and Nelder 1989, Hilbe 2007). Counts with less variance than the mean are termed underdispersed Poisson counts but this phenomenon is rarely occurred in real data (McCullagh and Nelder 1989). There are various causes that can lead to overdispersion, such as litters, families, households, etc. Furthermore, overdispersion can be caused by excess number of observed zero counts, since the excess zeros will give smaller conditional mean than the true value. When the data present too many zeros, Poisson model might not be appropriate and a ZIP model is commonly used as an alternative. In this thesis we reviewed the procedure of fitting ZIP regression and developed a Wald test for detecting zero-inflation in Poisson model. From our investigation by conducting simulation study, we found that its distribution follows mixture of a χ_0^2 (a constant of zero) and a χ_1^2 distribution and it can be used to detect the zero-inflation in counts.

5.2 Discussion

1. Our developed Wald test for ZIP model with constant λ and ω for compared between Poisson and ZIP models can be extended to a general situation on where the λ and ω is allowed to depend on covariates.

2 In some set of count data with excess zero, the ZIP model might not be appropriate because of showing overdispersion, then other model, such as the zero-inflated Negative Binomial model (ZINB) (Ridout et al., 2001) or zero-inflated random effect model can be used (Jansakul and Hinde, 2009).

3 A further extensions of the ideas in this thesis is to develop the Wald test for comparing between a Negative Binomial model and Zero-inflated Negative Binomial model.

BIBLIOGRAPHY

Akaike, H. 1973. Information theory and extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki (Eds.), Proceedings 2^{nd} International Symposium on Inference Theory, Budapest: Akadmiai kiado. 267-281.

Atkinson, A. 1985. Plots, Transformations and Regression. An introduction to Graphical Methods of Diagnostic Regression Analysis, Oxford: Clarendon Press.

Böhning, D., Dietz, E. and Schlattmann, P. 1999. The zero-inflated poisson model and decayed, missing and filled teeth index in dental epidemiology. Journal of the Royal Statistical Society, Series A. 162(2): 195-209.

Cochran, W. G. 1954. Some Methods for strengthening the common χ^2 tests. Biometrics. 10(4): 417-451.

Dempster, A., Laird, N. M. and Rubin, D. B. 1997. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B. 39: 1-38.

Dietz, E. and Böhning, D. 2000. On estimation of the Poisson parameter in zero-modifined Poisson models. Computational Statistics and Data Analysis. 34(4): 441-459.

Gardner, W., Mulvey, E. P. and Shaw, E. C. 1995. Regression analyses of counts and rates: Poisson, Overdispersed Poisson, and Negative binomial models. Psychological Bulletin. 11: 392-404.

Germu, S. and Trivedi, P. K. 1996. Excess zeros in count models for Recreational Trips. Journal of Business and Economic Statistics. 14: 469-477.

Gupta, P.L., Gupta, R. C. and Tripathi, R. C. 1996. Analysis of zero-adjusted count data. Computational Statistics and Data Analysis. 23: 207-218.

Heilbron, D. 1994. Zero-altered and other regression models for count data with added zeros. Biometrical Journal. 36: 531-547.

Hilbe, J. M. 2007. Negative Binomial Regression. Cambridge University Press.

Hinde, J. P. and Demet́rio, C. G. B. 1998. Overdispersion: models and estimation. Computational Statistics and Data Analysis. 27: 151-170.

Jansakul, N. 2005. Fitting a zero-inflated negative binomial model via R. Proceedings of the 20^{th} International Workshop on Statistical Modelling. Sydney, July 2005. 277-284.

Jansakul, N. and Hinde, J. P. 2002. Score tests for zero-inflated Poisson models. Computational Statistics and Data Analysis. 40: 75-96.

Jansakul, N. and Hinde, J. P. 2009. Score tests for extra-zero models in zero-inflated Negative Binomial models. Communications in statistics-simulation and computation. 38: 92-108.

Lambert, D. 1992. Zero-inflated Poisson regression with application to defects in manufacturing. Technometrics. 41(1): 29-38.

Lindsey, J. K. 1999. On the use of corrections for overdispersion. Applied Statistics. 48: 553-561.

McCullagh, P. and Nelder, J. A. 1989. Generalized Linear Models. second edition. Chapman and Hall: London.

Natchadamon Watcharachutimanon. 2006. Performance of Linear Mean-Variance Negative Binomial Models. Master of Science Thesis, Prince of Songkla University, Songkhla, Thailand.

Nelder, J. A. and Wedderburn, R. W. M. 1972. Generalized Linear Models. Journal of the Royal Statistical Society,Series A. 135(3): 370-384.

R Development Core Team. 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing , Vienna, Austria. ISBN 3-900051-07-0. URL http://www.R-project.org.

Raftery, A. E. 1986. A note on Bayes factors for log-linear contigency table model with vague prior Information, Journal of Rayal Statistical Society. 48: 249-250.

Ridout, M. S., Demet́rio, C. G. B. and Hinde, J. P. 1998. Model for count data with many zeros. International Biometric Conference. 179-190.

Ridout, M. S., Demt́rio, C. G. B., Hinde, J. P. 2001. A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. Biometrics 57: 219223.

Searle, S. R. 1966. Matrix Algebra for Biological Sciences: Including Applications in Statistics. John Wiley and Sons, Inc, New York, NY.

Self, S. G. and Kung-Yee Liang. 1987. Asymptotic properties of the maximum likelihood estimators and likelihood ratio test under non standard conditions. Journal of the American Statistical Association. 82: 605-610.

van den Broek, J. 1995. A score test for zero inflation in a Poisson distribution. Biometrics. 51: 738-743.

Wikipedia, the free encyclopedia. 2008. Wald test. http: //en.wikipedia.org/wiki/ Wald-test. 25/12/2551.

Xie, M. , He, B. and Goh, T.N. 2001. Zero-inflated Poisson model in statistical process control. Computational Statistics and Data Analysis. 38: 191-201.

APPENDIX A

Test statistics for zero-inflation

In the case of single homogeneous sample, the log-likelihood function (2.20) can be written as

$$\ell(\lambda, \omega) = n_0 \ln \left[\omega + (1-\omega)\mathrm{e}^{-\lambda}\right] + \sum_{j=1}^{J} n_j \ln \left[(1-\omega)\frac{\mathrm{e}^{-\lambda}\lambda^j}{j!}\right],$$

where J is the largest observed count value, n_j is the frequency of each possible count value, $j = y = 0, 1, 2, \ldots, J$ then n_0 is the number of zeros in the observed and $\sum_{j=0}^{J} n_j = n$, the total number of observations or the sample size.

1. Likelihood ratio test for ZIP model

The Likelihood ratio test based on the ratio of two log-likelihood functions can be written as.

$$R_\omega = -2 \times [\ell(\hat{\lambda}) - \ell(\hat{\lambda}, \widehat{\omega})] \tag{1}$$

where $\ell(\hat{\lambda})$ and $\ell(\hat{\lambda}, \hat{\omega})$ are the maximized log-likelihood under the Poisson and the ZIP regression model, respectively. This can be computed from the following formula:

$$
\begin{aligned}
R_\omega &= -2\left\{\sum_{i}^{n}\{y_i\text{ln}(\mu_i)-\mu_i-\ln(y_i!)\}-n_0\ln\left[\omega+(1-\omega)\mathrm{e}^{-\hat{\lambda}}\right]\right.\\
&- \left.\sum_{y=1}^{J}n_y\ln\left[(1-\omega)\frac{\mathrm{e}^{-\hat{\lambda}}\hat{\lambda}^y}{y!}\right]\right\}\\
&= -2\left\{\sum_{i}\text{ln}\mu_i-\sum_{i}\mu_i-\sum_{i}\text{ln}y_i!-n_0\text{ln}\left[\omega+(1-\omega)\mathrm{e}^{-\hat{\lambda}}\right]\right.\\
&- \left.\sum_{y=1}^{J}n_y\ln\left[(1-\omega)\frac{\mathrm{e}^{-\hat{\lambda}}\hat{\lambda}^y}{y!}\right]\right\}\\
&= -2\left\{n\bar{y}\text{ln}\bar{y}-n\bar{y}-\sum_{i}\text{ln}y_i!-n_0\ln\left[\omega+(1-\omega)\mathrm{e}^{-\hat{\lambda}}\right]\right.\\
&- \left.\sum_{y=1}^{J}n_y\ln\left[(1-\omega)\frac{\mathrm{e}^{-\hat{\lambda}}\hat{\lambda}^y}{y!}\right]\right\}\\
&= -2\left\{n\bar{y}\text{ln}\bar{y}-n\bar{y}-n_0\text{ln}\left(\frac{n_0}{n}\right)-\sum_{y}n_y[\ln(1-\omega)-\hat{\lambda}]-n\bar{y}\text{ln}\hat{\lambda}\right\}\\
&= -2\left\{n\bar{y}[\text{ln}\bar{y}-1-\text{ln}\hat{\lambda}]-n_0\text{ln}\left(\frac{n_0}{n}\right)-(n-n_0)\left(\ln\left(\frac{\bar{y}}{\hat{\lambda}}\right)-\hat{\lambda}\right)\right\}\\
&= \quad 2\left\{n_0\text{ln}\left(\frac{n_0}{n}\right)+(n-n_0)\left(\ln\left(\frac{\bar{y}}{\hat{\lambda}}\right)-\hat{\lambda}\right)+n\bar{y}(\text{ln}\hat{\lambda}+1-\text{ln}\bar{y})\right\}
\end{aligned}
$$

where $\bar{y}$ is the mean of the observations under H_0 and $\hat{\lambda}$ is the estimated positive mean counts under H_1. This test statistic R_ω approximately follows chi-square distribution on 1 degree of freedom (d.f).

2 Score test for ZIP models

Based on the log-likelihood function given in (2.22) the score vector is

$$
S(\lambda,\omega)=\begin{bmatrix} S_\lambda(\lambda,\omega)\\ S_\omega(\lambda,\omega)\end{bmatrix}=\begin{bmatrix}\dfrac{\partial\ell(\lambda,\omega)}{\partial\lambda}\\ \dfrac{\partial\ell(\lambda,\omega)}{\partial\omega}\end{bmatrix}
$$

where

$$
\frac{\partial\ell(\lambda,\omega)}{\partial\lambda}=\frac{-n_0(1-\omega)e^{-\lambda})}{[\omega+(1-\omega)e^{-\lambda}]}-\sum_{j=1}^{J}n_j+\frac{\sum_{j=1}^{J}(n_j\times j)}{\lambda}, \tag{2}
$$

and

$$\frac{\partial \ell(\lambda,\omega)}{\partial \omega} = \frac{n_0(1-e^{-\lambda})}{[\omega+(1-\omega)e^{-\lambda}]} - \frac{\sum_{j=1}^{J} n_j}{1-\omega}. \tag{3}$$

The expected information matrix $\mathcal{I}(\boldsymbol{\beta},\boldsymbol{\gamma})$ can be partitioned as

$$\mathcal{I}(\lambda,\omega) = \left[\begin{array}{cc} \mathcal{I}_{\lambda\lambda}(\lambda,\omega) & \mathcal{I}_{\lambda\omega}(\lambda,\omega) \\ \mathcal{I}_{\omega\lambda}(\lambda,\omega) & \mathcal{I}_{\omega\omega}(\lambda,\omega) \end{array}\right]$$

where the elements $\mathcal{I}_{\lambda,\lambda}, \mathcal{I}_{\lambda\omega} = \mathcal{I}_{\omega\lambda}$ and $\mathcal{I}_{\omega\omega}$ are, respectively,

$$-\mathrm{E}\left[\frac{\partial^2 \ell(\lambda,\omega)}{\partial \lambda^2}\right], \quad -\mathrm{E}\left[\frac{\partial^2 \ell(\lambda,\omega)}{\partial \lambda \partial \omega}\right], \quad \text{and} \quad -\mathrm{E}\left[\frac{\partial^2 \ell(\lambda,\omega)}{\partial \omega^2}\right]$$

with

$$\begin{aligned} \frac{\partial^2 \ell(\lambda,\omega)}{\partial \lambda^2} &= \frac{n_0\omega(1-\omega)\mathrm{e}^{-\lambda}}{[\omega+(1-\omega)\mathrm{e}^{-\lambda}]^2} - \frac{\sum_j (n_j \times j)}{\lambda^2}; \\ \frac{\partial^2 \ell(\lambda,\omega)}{\partial \lambda \partial \omega} &= \frac{n_0\mathrm{e}^{-\lambda}}{[\omega+(1-\omega)\mathrm{e}^{-\lambda}]^2}; \\ \frac{\partial^2 \ell(\lambda,\omega)}{\partial \omega^2} &= \frac{-n_0(1-\mathrm{e}^{-\lambda})^2}{[\omega+(1-\omega)\mathrm{e}^{-\lambda}]^2} - \frac{\sum_j n_j}{(1-\omega)^2}. \end{aligned}$$

Using the fact that

$$\begin{aligned} & \mathrm{E}[I_{(y_i=0)}] &=& \Pr(Y_i=0) = \omega_i + (1-\omega_i)\mathrm{e}^{-\lambda_i} \\ \text{and} \quad & \mathrm{E}[I_{(y_i>0)}] &=& \Pr(Y_i>0) = (1-\omega_i)(1-e^{-\lambda_i}) \end{aligned}$$

we have

$$\begin{aligned} \mathcal{I}_{\lambda\lambda} = -\mathrm{E}\left[\frac{\partial^2 \ell(\lambda,\omega)}{\partial \lambda^2}\right] &= n\left[\frac{1-\hat{\omega}}{\hat{\lambda}} - \frac{\hat{\omega}(1-\hat{\omega})\mathrm{e}^{-\hat{\lambda}}}{\hat{\omega}+(1-\hat{\omega})\mathrm{e}^{-\hat{\lambda}}}\right]; \\ \mathcal{I}_{\omega\lambda} = -\mathrm{E}\left[\frac{\partial^2 \ell(\lambda,\omega)}{\partial \lambda \partial \omega}\right] &= \frac{-n\mathrm{e}^{-\hat{\lambda}}}{\hat{\omega}+(1-\hat{\omega})\mathrm{e}^{-\hat{\lambda}}}; \\ \mathcal{I}_{\omega\omega} = -\mathrm{E}\left[\frac{\partial^2 \ell(\lambda,\omega)}{\partial \omega^2}\right] &= \frac{n(1-\mathrm{e}^{-\hat{\lambda}})}{(1-\hat{\omega})[\hat{\omega}+(1-\hat{\omega})\mathrm{e}^{-\hat{\lambda}}]}. \end{aligned} \tag{4}$$

Under the null hypothesis, $\omega = 0$, the general score test is then

$$S_{\boldsymbol{\omega}} = S_{\omega}^{T}(\hat{\lambda},0)C^{-1}S_{\omega}(\hat{\lambda},0), \tag{5}$$

where $\hat{\lambda}$ is the maximum likelihood estimate under the Poisson model and

$$S_\omega(\hat{\lambda},0) = \left[\frac{n_0 - n\mathrm{e}^{-\hat{\lambda}}}{\mathrm{e}^{-\hat{\lambda}}}\right], \tag{6}$$

$$C = \mathcal{I}_\omega(\hat{\lambda},0) - \mathcal{I}_{\lambda\omega}(\hat{\lambda},0)\mathcal{I}_{\lambda\lambda}^{-1}(\hat{\lambda},0)\mathcal{I}_{\omega\lambda}(\hat{\lambda},0)$$

with

$$\begin{aligned}
\mathcal{I}_{\lambda\lambda}(\hat{\lambda},0) &= \frac{n}{\hat{\lambda}}, \\
\mathcal{I}_{\lambda\omega}(\hat{\lambda},0) &= -n \\
\mathcal{I}_{\omega\omega}(\hat{\lambda},0) &= \frac{n(1-\mathrm{e}^{-\hat{\lambda}})}{\mathrm{e}^{-\hat{\lambda}}}.
\end{aligned} \tag{7}$$

Score test for ZIP model is

$$S_\omega = \frac{(n_0 - n\mathrm{e}^{-\hat{\lambda}})^2}{n\mathrm{e}^{-\hat{\lambda}}(1-\mathrm{e}^{-\hat{\lambda}}) - n\hat{\lambda}\mathrm{e}^{-2\hat{\lambda}}}. \tag{8}$$

Under the null hypothesis, from (2) we have

$$\begin{aligned}
-n_0 - (n - n_0) + \frac{\sum_{j=1}^{J}(n_j \times j)}{\hat{\lambda}} &= 0; \\
-n + \frac{\sum_i^n y_i}{\hat{\lambda}} &= 0; \\
\hat{\lambda} &= \bar{y}.
\end{aligned} \tag{9}$$

Hence, S_ω in (8) we can rewritten as:

$$\begin{aligned}
S_\omega &= \frac{(n_0 - n\mathrm{e}^{-\bar{y}})^2}{n\mathrm{e}^{-\bar{y}}(1-\mathrm{e}^{-\bar{y}}) - n\bar{y}\mathrm{e}^{-2\bar{y}}} \\
&= \frac{(n_0 - np_0)^2}{np_0(1-p_0) - n\bar{y}p_0^2} \quad \text{where} \quad p_0 = e^{-\bar{y}}.
\end{aligned}$$

Under the null hypothesis this statistic will have asymptotic chi-squared distribution with 1 degree of freedom (Van den broek, 1995).

3 Confidence interval test for ZIP models

It is possible to derive a test based on asymtotic normality of the estimate of the parameters. Following the statistical properties of ZIP model we have

$$\mathrm{E}(\bar{Y}) = \mathrm{E}(Y) = (1-\omega)\lambda\hat{\lambda} = \mu$$

and

$$\begin{aligned}
\mathrm{Var}(\bar{Y}) &= \frac{1}{n}\mathrm{Var}(Y) \\
&= \frac{1}{n}\left\{\frac{(1-\omega)\mu+\omega\mu^2}{1-\omega}\right\} \\
&= \frac{1}{n}\left\{(1-\omega)\lambda+\omega(1-\omega)\lambda^2\right\} \\
&= \frac{1}{n}\left\{(1-\omega)\lambda+(1-\omega)\lambda\omega\lambda\right\} \\
&= \frac{1}{n}\left\{(1-\omega)\lambda+(1-\omega)\lambda(\lambda-(1-\omega)\lambda)\right\}.
\end{aligned}$$

From the central limit theorem, the confidence interval can be written as

$$\begin{aligned}
-Z_{\alpha/2} &\le \frac{\bar{Y}-(1-\omega)\lambda}{\sqrt{\mathrm{Var}(\bar{Y})}} \le Z_{\alpha/2} \\
-Z_{\alpha/2}\sqrt{\mathrm{Var}(\bar{Y})} &\le \bar{Y}-(1-\omega)\lambda \quad \le Z_{\alpha/2}\sqrt{\mathrm{Var}(\bar{Y})} \\
\frac{-\bar{Y}-Z_{\alpha/2}\sqrt{\mathrm{Var}(\bar{Y})}}{\lambda} &\le \quad -(1-\omega) \quad \le \frac{-\bar{Y}+Z_{\alpha/2}\sqrt{\mathrm{Var}(\bar{Y})}}{\mu} \\
\frac{-\bar{Y}-Z_{\alpha/2}\sqrt{\mathrm{Var}(\bar{Y})}}{\lambda} &\le \quad \omega-1 \quad \le \frac{-\bar{Y}+Z_{\alpha/2}\sqrt{\mathrm{Var}(\bar{Y})}}{\lambda} \\
1-\frac{\bar{Y}-Z_{\alpha/2}\sqrt{\frac{1}{n}\left\{\frac{(1-\omega)\mu+\omega\mu^2}{1-\omega}\right\}}}{\lambda} &\le \quad \omega \quad \le 1-\frac{\bar{Y}+Z_{\alpha/2}\sqrt{\frac{1}{n}\left\{\frac{(1-\omega)\mu+\omega\mu^2}{1-\omega}\right\}}}{\lambda}
\end{aligned}$$

can be obtained as

$$\begin{aligned}
&1-\frac{\bar{Y}-Z_{\alpha/2}\sqrt{\{E(Y)+E(Y)[\lambda-E(Y)]\}/n}}{\lambda} \\
&\le \omega \le 1-\frac{\bar{Y}+Z_{\alpha/2}\sqrt{\{E(Y)+E(Y)[\lambda-E(Y)]\}/n}}{\lambda}.
\end{aligned}$$

In practice, a set of data, can compute the above confidence interval by substituting λ and $E(Y)$ by the maximum likelihood estimate of λ, say $\hat{\lambda}$ and the sample mean $\bar{y}$, respectively. Thus, the confidence interval can be written as

$$1-\frac{\bar{y}-Z_{\alpha/2}\sqrt{\{\bar{y}+\bar{y}[\hat{\lambda}-\bar{y}]\}/n}}{\hat{\lambda}} \le \omega \le 1-\frac{\bar{y}+Z_{\alpha/2}\sqrt{\{\bar{y}+\bar{y}[\hat{\lambda}-\bar{y}]\}/n}}{\hat{\lambda}}.$$

Hence, a test based on a positive one sided confidence interval of probability zero-inflated counts can be obtained as

$$CI_{\omega} = 1-\frac{\bar{y}+Z_{\alpha}\sqrt{\{\bar{y}+\bar{y}[\hat{\lambda}-\bar{y}]\}/n}}{\hat{\lambda}}, \tag{10}$$

where $\hat{\lambda}$ is the estimated positive mean counts under H_1. The critical region of this test method is simply $CI_\omega > 0$.

4 Cochran test for ZIP models

First, it is well known that if one random variable $X \sim \chi^2_1$ then $\sqrt{X} \sim \mathrm{N}(0,1)$. In fact, the Cocharn test is transformed their original their original Chi-square from into its corresponding standard normal form. Cochran(1954) proposed to test any single deviation $(f_i - m_i)$ when m ia estimated from the data,

$$\begin{array}{c} \text{from} \quad \chi^2 = \dfrac{L^2}{\mathrm{Var}(L)} \quad \Longrightarrow \quad C_\omega = \dfrac{L}{\sqrt{\mathrm{Var}(L)}} \sim \mathrm{N}(0,1) \\ L = (f_i - m_i) : \mathrm{Var}(L) = m_i - \dfrac{m_i^2}{N}\left\{1 + \dfrac{(i-m)^2}{m}\right\} \\ \text{set} \quad i = 0 \quad \text{let} \quad f_0 = n_0, m_0 = ne^{-\bar{y}}, m = \bar{y}, N = n \end{array}$$

$$\begin{aligned} L &= n_0 - ne^{-\bar{y}} \\ \mathrm{Var}(L) &= \mathrm{Var}(n_0 - ne^{-\bar{y}}) \\ &= ne^{-\bar{y}} - \frac{(ne^{-\bar{y}})^2}{n}\{1+\bar{y}\} \\ &= ne^{-\bar{y}} - ne^{-2\bar{y}}\{1+\bar{y}\} \\ &= ne^{-\bar{y}} - ne^{-2\bar{y}} - n\bar{y}e^{-2\bar{y}} \\ &= ne^{-\bar{y}}(1 - e^{-\bar{y}} - \bar{y}e^{-\bar{y}}) \end{aligned}$$

The C test statistic written by the symbols used here is as follows:

$$\begin{aligned} C_\omega &= \frac{L}{\sqrt{\mathrm{Var}(L)}} \\ &= \frac{n_0 - ne^{-\bar{y}}}{\sqrt{\mathrm{Var}(n_0 - ne^{-\bar{y}})}} \\ &= \frac{(n_0 - \mathrm{n}e^{-\bar{y}})}{[n\mathrm{e}^{-\bar{y}}(1 - \mathrm{e}^{-\bar{y}} - \bar{y}\mathrm{e}^{-\bar{y}})]^{1/2}} \sim \mathrm{N}(0,1) \end{aligned}$$

Under the null hypothesis, the test statistic C_ω is approximately normally distributed with zero mean and unit variance.

Following the relationship between N $\sim (0,1)$ and χ^2_1, we found that C^2_ω, can be obtained as

$$C^2_\omega = \frac{(n_0 - n\mathrm{e}^{-\bar{y}})^2}{n\mathrm{e}^{-\bar{y}}(1-\mathrm{e}^{-\bar{y}}-\bar{y}\mathrm{e}^{-\bar{y}})}. \tag{11}$$

Under null hypothesis C^2_ω approximately follows chi-square distribution on one d.f. Moreover, C^2_ω has the form exactly the same as Score test as follow:

$$\begin{aligned} C^2_\omega &= \frac{(n_0 - n\mathrm{e}^{-\bar{y}})^2}{n\mathrm{e}^{-\bar{y}}(1-\mathrm{e}^{-\bar{y}}-\bar{y}\mathrm{e}^{-\bar{y}})} \\ &= \frac{(n_0 - n\mathrm{e}^{-\bar{y}})^2}{n\mathrm{e}^{-\bar{y}}(1-\mathrm{e}^{-\bar{y}}) - n\bar{y}\mathrm{e}^{-2\bar{y}}} \\ &= \frac{(n_0 - np_0)^2}{np_0(1-p_0) - n\bar{y}p_0^2} = S_\omega \quad \text{where} \quad p_0 = e^{-\bar{y}}. \end{aligned}$$

APPENDIX B

R functions

```
# A function for creating a simulated envelope and drawing the
graph.

# use standardized deviance residual

hnp <- function(object,models=c("Normal","Poisson","Binomial"))
 {
      rmax <- 0;    drmax <- 0
      rme <- 0;     drme <- 0
      minr <- 0;    dminr <- 0
            models<-match.arg(models)
            if(models=="Poisson"){
            resids <- hnp.pois(object)
            }
            else if (models=="Binomial"){
            resids <- hnp.logit(object)
            }

            else {
            resids <- hnp.norm(object)
            }

            r.sim <- resids
            sres <- r.sim[[1]]


  #++++++++++++++++ simulated envelope +++++++++++++

    for(i in 1:nrow(r.sim[[2]])) {

          rmax[i] <- max(r.sim[[2]][i,] )
            rme[i]  <- mean(r.sim[[2]][i,] )
            minr[i] <- min(r.sim[[2]][i,] )
                                 }

            rmin <- minr;     rmax <- rmax;  rmean <- rme
            sres <- sres

      n <- length(sres)
      i <- seq(1, n, by = 1 )
      nd <- qnorm((i + n - 0.125)/(2*n + 0.5))
# A function for creating a simulated envelope and drawing the
graph. (Continuous)

# use standardized deviance  residual


      mi.y <- min(as.integer(sres)) #1
      ma.y <- max(as.integer(c(sres,rmax)))+0.5 #1
      mi.x <- min(as.integer(nd)) - 0.5 #1
      ma.x <- max(as.integer(nd))+0.5 #1

      par(pty= "s")
      plot(nd, sres, xlab = "Half-normal scores",
           main = "Poisson model ",
           ylab = "standardizes Residuals", type = "n",
           axes = FALSE, ylim = c(mi.y, ma.y))

           axis(1,at=seq(mi.x, ma.x, by=0.5)) #x-axis
           axis(2,at=seq(mi.y, ma.y, by=0.25)) #y-axis

           points(nd, sres, pch=4, mkh=0.06)
           lines(nd, rmin, lty=1)
           lines(nd, rmean, lty=12)
           lines(nd, rmax, lty=1)

             }
```

```
# A function for calculating standardize residuals of Poisson
regression.



  hnp.pois<- function(object) {
    #  object : a diagnosted fitted model
      mu <- fitted(object)
      n <- length(mu)
      r.sim <- matrix(0,n,19)
      rmax <- 0
      rme <- 0
      minr <- 0
      ys <- 0


            y<-object$y
            hii<-lm.influence(object)$hat
            rdot<-resid(object)/sqrt(1-hii)
            rdot<-ifelse(is.na(rdot),0,rdot)
           sres <- sort(abs(rdot))

    #  Observed standardized Pearson


# ++++++++   Simulated envelop   ++++++++++++++


  for(i in 1:19)
           {
            ys <- rpois(n,mu)
            object$model[,1] <- ys
            ys.glm <- glm(object$model,family=poisson)
            hii<-lm.influence(ys.glm)$hat
            rdot<-resid(ys.glm)/sqrt(1-hii)
            rdot<-ifelse(is.na(rdot),0,rdot)
            r.sim[,i] <- rdot
            r.sim[,i] <- sort(abs(rdot))

 }
      resids <- list(sres = sres, r.sim = r.sim)
      resids

  }
```

```
#  The method of Fisher Scoring and ZIP Regression Models


fisher.ZIPW1 <- function(formula,X.mat)
 {
        S.E <- 0; y <- NULL; llikf <- 0
        y.pois <- glm(formula, family = poisson)
        y <- y.pois$y
         z<-ifelse(y==0,1,0)

      beta <- coefficients(y.pois)

     lamda <- fitted(y.pois)
     omega <- mean(z)

 # Calculate -2 x logL for a Poisson model

    plikf <- -2*sum(y*log(lamda)-lamda-lgamma(y+1))

 # Initial value

  # Calculate (I(y))

    ll.old<-abs(logLik(y.pois))
    ll.diff<-ll.old
    ll.diff<-ifelse(is.na(ll.diff),0,ll.diff)
    beta.omega <- c(beta,omega)
      i<-0
     while(ll.diff > 0.001) {


 # Gradient vectios
        d <- omega+((1-omega)*exp(-lamda))
      s11 <- -z*(1-omega)*exp(-lamda)*lamda
        s12<- (1-z)*(y-lamda)
        s1 <- (s11/d)+s12
        S1 <- t(X.mat)%*%s1   # for beta

      s2 <- z*(1-exp(-lamda)/d) - ((1-z)*(1/(1-omega)))
      S2 <- sum(s2)     # for omega


 # Creat a score vector
       S <- c(S1, S2)
# The method of Fisher Scoring and ZIP Regression Models
   (Continuous)

 # Create minus the 2nd derivative & information matrices

    c1 <- exp(-lamda)*((1-lamda)*omega+(1-omega)*
          exp(-lamda))*(1-omega)*lamda
    c2 <- omega+(1-omega)*exp(-lamda)
    c3 <- lamda*(1-omega)*(1-exp(-lamda))
    c  <- (c1/c2)+c3

    I1 <- (t(X.mat) %*%diag(c[1:length(c)]))%*%X.mat #for beta

        i21 <- (1-exp(-lamda))^2/d
        i22 <- (1-exp(-lamda))/(1-omega)
        I2  <- i21+i22
        I2  <- sum(I2)   #for omega

      i12 <- -lamda*exp(-lamda)/d
      I12 <- t(X.mat)%*%i12   # for beta,omega

 # Creat a score vector and partitioned information matrix
      PI1 <- c(I12, I2)
      PI2 <- rbind(I1, t(I12))
      PI <- cbind(PI2, PI1)   # completed infomation matrix
      PI.inv <- solve(PI)
         S.E <- sqrt(PI.inv)
         S.E <-diag(S.E)

      beta.omega <- beta.omega + PI.inv%*%S
  # The fisher scoring method

         beta <- beta.omega[1:length(beta)]
        lamda <- exp(X.mat%*%beta) # Fitted vales
       omega1 <- beta.omega[length(beta.omega)]
         omega2<-ifelse(omega1<=0,0,omega1)
        omega2<-ifelse(omega1>=1,0.92,omega1)
         omega<-round(omega2,3)
            d <- omega+(1-omega)*exp(-lamda)

   ll.new<-sum(z*log(d)+(1-z)*(log(1-omega)-lamda+y*
            log(lamda)-lgamma(y+1)))
  ll.diff<-abs(ll.old-ll.new)
   ll.old<-ll.new
        i<-i+1
      }
      d <- omega+(1-omega)*exp(-lamda)
      ll<-sum(z*log(d)+(1-z)*(log(1-omega)-lamda+y*
          log(lamda)-lgamma(y+1)))

   result <-list(beta.zip=beta,omega=omega,
lamda.zip=lamda,SE.beta=S.E,log_like=ll,fitted.values=lamda,
model=y.pois$model,formula=y.pois$formula,iteration=i)
  return(result)
}
```

```
#  The method of Fisher Scoring and ZIP Regression Model
   With no covariate.

ZIP.bothconstant.var<- function(y) {
       lamda <-mean(y)
        nj<-c(table(y))
        n<-sum(nj)
        no<-nj[1]
        nj<-c(table(y))[2:length(nj)]
        j<-as.numeric(names(nj))

       lamda.old <-lamda
       lamda.diff<-lamda.old
       lamda.diff<-ifelse(is.na(lamda.diff),0,lamda.diff)

       i<-0  ; Omega<-NULL ;Lamda<-NULL

       while(lamda.diff > 0.001) {

       lamda<-((1-exp(-lamda))*sum(nj*j))/(sum(nj))

       lamda.new<-lamda
        lamda.diff<-abs(lamda.old-lamda.new)
        lamda.old<-lamda.new
         i<-i+1
       omega<-(no-n*exp(-lamda))/(n*(1-exp(- lamda)))
        }
 ll<-no*log(omega+(1-omega)*exp(-lamda))+(sum(nj*log((1-omega)
    *exp(-lamda)*lamda^j/factorial(j))))

# Create minus the 2nd derivative & information matrices

        d <- omega+(1-omega)*exp(-lamda)
       i21 <- n*(1-exp(-lamda))/((1-omega)*d) #for omega

       i12 <- (-exp(-lamda)/d)*n   # for lambda,omega

       I12 <- n*(((1-omega)/lamda)-((omega*(1-omega)*exp(-lamda))/d))
              # for lambda

 # Creat a  partitioned information matrix

       PI1 <- c(i12,i21)
       PI2 <- rbind(I12, t(i12))
       PI <- cbind(PI2, PI1)      # completed infomation matrix
       I.inv <- solve(PI)
       var<-diag(I.inv)
       var.lambda<-var[1:length(lamda)]
       var.omega<-var[length(var)]
       se<-sqrt(diag(I.inv))
       se<-round(se,4)
       se.lambda<-se[1:length(beta)]
       se.omega<-se[length(se)]

   result <-
list(lamda.zip=lamda,omega=omega,log_like=ll,var.omega=var.omega,
      se.omega=se.omega, se.lambda= se.lambda ,iteration=i)

return(result)
 }
```

```
# A function for creating a simulated envelope and drawing the
graph Half normal plot for  ZIP Regression Model

hnp.zip<-function(y.zip, X.mat)
      {
            source("function_ZIPW1.txt")

             y<-y.zip $model[[1]]
               n<-length(y)
             lamda<-y.zip$fitted.values ; omega<-y.zip$omega
             dr.sim<-matrix(0,n,19)
                 dminr<-0 ; drmax<-0 ; drmean<-0


             mu<-(1-omega)*lamda
                 odres<-(y-mu)
             odres<-odres/sqrt(mu*(1+(omega/(1-omega)*mu)))
             sres<-sort(sqrt(abs(odres)))


  # ++++++++   Simulated envelop   ++++++++++++++
      for(i in 1:19)
           {
  # Generate ZIP data
                 ybin <- rbinom(n,1, omega)
                 y1 <- rpois(n,lamda)
                 y <- 0*ybin + (1-ybin)*y1

               model<-y.zip$model ; model[[1]]<-y

            formula<-y.zip$formula

                z<-ifelse(ys==0,1,0)
                ysim.zip<-fisher.ZIPW1(formula,X.mat)

            lamda.sim <- ysim.zip$fitted.values
            omega.sim <- ysim.zip$omega

 # Simulate standardize deviance residual

             mu1<-(1-omega.sim)*lamda.sim
          drsim<-(y-mu1)
          drsim<-drsim/sqrt(mu1*(1+(omega.sim/
                (1- omega.sim)*mu1)))
          dr.sim[,i]<-sort(sqrt(abs(drsim)))


 }
```

# A function for creating a simulated envelope and drawing the graph Half normal plot for ZIP Regression Model (continuous)

```
#  envelope
          for(i in 1:nrow(dr.sim)){
         drmax[i]<-max(dr.sim[i,])
         drmean[i]<-mean(dr.sim[i,])
         dminr[i]<-min(dr.sim[i,])
         }

  # Half normal plot


        n<-length(sres)
       i <- seq(1, n, by = 1 )
      nd <- qnorm((i + n - 0.125)/(2*n + 0.5))

      mi.y <- min(as.integer(sres)) #1
      ma.y <- max(as.integer(c(sres,drmax)))+0.5 #1
      mi.x <- min(as.integer(nd)) - 0.5 #1
      ma.x <- max(as.integer(nd))+1 #1

        par(pty= "s")
      plot(nd, sres, xlab = "Half-normal scores", main = "ZIP
model ",ylab = "Standardizes Deviance Residual ", type = "n",
axes = FALSE, ylim = c(mi.y, ma.y))


            axis(1,at=seq(mi.x, ma.x, by=0.5)) #x-axis

            axis(2,at=seq(mi.y, ma.y, by=0.25)) #y-axis



            points(nd, sres, pch=4, mkh=0.06)

      lines(nd, dminr, lty=1)

      lines(nd, drmean, lty=12)

      lines(nd, drmax, lty=1)

}
```

```
 # A function for creating a simulated envelope and drawing the
graph Half normal plot for  ZIP Regression Model  with no
covariate.


hnp.zip.conts<-function(y.zip) {

     source("function_ZIP_bothconstant.txt")


             y<-y.zip$y
             n<-length(y)

 lambda<-y.zip$lamda.zip ; omega<-y.zip$omega
             dr.sim<-matrix(0,n,19)
        dminr<-0 ; drmax<-0 ; drmean<-0


                    mu<-(1-omega)*lambda
                 odres<-(y-mu)

            odres<-odres/sqrt(mu*(1+(omega/(1-omega)*mu)))
             sres<-sort(sqrt(abs(odres)))


  # ++++++++   Simulated envelop   ++++++++++++++


       for(i in 1:19)
           {

  # Generate ZIP data
                 ybin <- rbinom(n,1, omega)
                 y1 <- rpois(n,lambda)
                 y <- 0*ybin + (1-ybin)*y1

            ysim.zip<-ZIP.bothconstant(y)


            lambda.sim <- ysim.zip$lamda.zip
             omega.sim <- ysim.zip$omega
             omega.sim <- ifelse(omega.sim<=0,0.05,omega.sim)

 # Simulate standardize deviance residual


            mu1<-(1-omega.sim)*lambda.sim
          drsim<-(y-mu1)
          drsim<-drsim/sqrt(mu1*(1+(omega.sim/
                (1-omega.sim)*mu1)))

      dr.sim[,i]<-sort(sqrt(abs(drsim)))

                }
 # A function for creating a simulated envelope and drawing the
graph Half normal plot for ZIP Regression Model with no
covariate.(continuous)

  #  envelope


          for(i in 1:nrow(dr.sim)){
         drmax[i]<-max(dr.sim[i,])
         drmean[i]<-mean(dr.sim[i,])
         dminr[i]<-min(dr.sim[i,])


         }

# Half normal plot


       n<-length(sres)
      i <- seq(1, n, by = 1 )
     nd <- qnorm((i + n - 0.125)/(2*n + 0.5))

      mi.y <- min(as.integer(sres)) #1
      ma.y <- max(as.integer(c(sres,drmax)))+0.5 #1
      mi.x <- min(as.integer(nd)) - 0.5 #1
      ma.x <- max(as.integer(nd))+1 #1

        par(pty= "s")
      plot(nd, sres, xlab = "Half-normal scores", main = "ZIP
model",ylab = "Standardizes Pearson Residual ", type = "n", axes
= FALSE,ylim = c(mi.y, ma.y))


            axis(1,at=seq(mi.x, ma.x, by=0.5)) #x-axis
            axis(2,at=seq(mi.y, ma.y, by=0.25)) #y-axis


            points(nd, sres, pch=4, mkh=0.06)
      lines(nd, dminr, lty=1)
      lines(nd, drmean, lty=12)
      lines(nd, drmax, lty=1)


}
```

```
# The score test

score_test<-function(y) {

            nj<-c(table(y))
             n<-sum(nj)
            no<-nj[1]
            nj<-c(table(y))[2:length(nj)]
             j<-as.numeric(names(nj))

      ymean<-sum(y)/n
     y.pois<-glm(y~1,family=poisson)
         mu<-ymean
          p<-exp(-mu)

            S1<- (no-n*p)^2
            S2<-(n*p*(1-p))-(n*ymean*(p^2))
             S<-S1/S2

       result<-list(S=S)
      return(result)

     }
```

```
# The Likeilhood Ratio Test

LRT_test<-function(y){

            nj<-c(table(y))
            n<-sum(nj)
          n0<-nj[1]
          nj<-c(table(y))[2:length(nj)]
           j<-as.numeric(names(nj))

        ymean<-sum(y)/n
       y.pois<-glm(y~1,family=poisson)

      source("function_ZIP_bothconstant.txt")
       y.zip<-ZIP.bothconstant(y)
      lamda1<-y.zip$lamda.zip

      l1<- n0*log(n0/n)+(n-n0)*(log(ymean/lamda1)-lamda1)
      l2<-n*ymean*(log(lamda1)+1-log(ymean))

      LRT<-2*(l1+l2)

 result<-list(LRT=LRT)
 return(result)
}
```

```
# The chi-square test

chiq_test<-function(y) {

             nj<-c(table(y))
              n<-sum(nj)
             no<-nj[1]
              j<-as.numeric(names(nj))

           C<-j
           O<-nj
       ymean<-sum(y)/n
           E<-dpois(C,ymean)*n

      chi<-sum((O-E)^2/E)

    result<-list(chi=chi)
    return(result)

        }
```

```
# The Confidence interval test

CI_test<-function(y,level){


             nj<-c(table(y))
              n<-sum(nj)
             no<-nj[1]
             nj<-c(table(y))[2:length(nj)]
              j<-as.numeric(names(nj))

          ymean<-sum(y)/n
              z<-qnorm(level,0,1,lower.tail = TRUE)

        source("function_ZIP_bothconstant.txt")

           y.zip<-ZIP.bothconstant(y)

          lamda1<-y.zip$lamda.zip

      ci1<- ymean+z*sqrt((ymean+ymean*(lamda1-ymean))/n)
       CI<-1-(ci1/lamda1)

       result<-list(CI=CI)
     return(result)

       }
```

```
# The Cochran test

C_test<-function(y){

       nj<-c(table(y))
        n<-sum(nj)
       no<-nj[1]

    ymean<-sum(y)/n

        c1<- (no-n*exp(-ymean))
        c2<-sqrt(n*exp(-ymean)*(1-exp(-ymean)
            -ymean*exp(-ymean)))
         c<-c1/c2

       result<-list(c=c)
       return(result)

       }
```

```
# The Wald test

wald.zip<-function(y){
      ybar<-mean(y)
      nj<-c(table(y))
       n<-sum(nj)
       n0<-nj[1]
      source("function_ZIP_bothconstant.txt")
       y.zip<-ZIP.bothconstant(y)

      lamda<-y.zip$lamda.zip
      ome1<-y.zip$omega
      up1<-n0-n*exp(-lamda)
      down1<-n*(1-exp(-lamda))
      ome2<-up1/down1

# Compute the Wald test
      up2<-n0*ybar*(n0-n*exp(-lamda)*(lamda-ybar))
      down2<-n^2*lamda*((1-exp(-lamda))*(n0-(n*exp(-lamda)*
             (lamda-ybar)))-(n*lamda*exp(-2*lamda)))
      Var<-up2/down2
      W2<-ome2^2/Var
#Compute Wald test from formula
      up3<-(n0-n*exp(-lamda))^2*lamda*((1-exp(-lamda))*
           (n0-(n*exp(-lamda)*(lamda-ybar)))
           -(n*lamda*exp(-2*lamda)))
    down3<-n0*ybar*(1-exp(-lamda))^2*(n0-n*exp(-lamda)*
           (lamda-ybar))
      W3<-up3/down3
  result <- list(Wald.test2=W2,Wald.test3=W3)
  return(result)
 }
```

```
# Simulation study for distribution of the Wald test

wald.test<-function(n,lamda,level=0.95){

      R<-3000
      W<-NULL

     source("function_ZIP_bothconstant.txt")

      for(i in 1:R){

      y<-rpois(n,lamda)

     source("function_ZIP_bothconstant.txt")
      y.zip<-ZIP.bothconstant(y)

      ybar<-mean(y)
      nj<-c(table(y))
       n<-sum(nj)
      n0<-nj[1]

       omega<-y.zip$omega
      lamda1<-y.zip$lamda.zip

#Compute Wald test

   up<-(n0-n*exp(-lamda1))^2*lamda1*((1-exp(-lamda1))*
       (n0-(n*exp(-  lamda1)*(lamda1-ybar)))-(n*lamda1*
       exp(-2*lamda1)))
 down<-n0*ybar*(1-exp(-lamda1))^2*(n0-n*exp(-lamda1)*
       (lamda1-ybar))



 Wald<-up/down

    W<-c(W,Wald)

 }
      wald <- sum(W>=qchisq(level,1))/R

      wald<-round(wald,3)

result<-list(omega=omega,wald.test=wald)
return(result)

 }
```

```
# The power of the six tests

power.test<-function(n,lamda,ome,level=0.95){


   source("function_ZIP_bothconstant.txt")
   source("function_LRT.txt")
   source("function_scoretest.txt")
   source("function_chi_test.txt")
   source("function_CI_test.txt")
   source("function_ctest.txt")



            W<-NULL;  L<-NULL ; S<-NULL;
          Chi<-NULL ; CI<-NULL ; C<NULL;  y <- NULL



for(i in 1:R){

       ybin <- rbinom(n,1,ome)
         y1 <- rpois(n, lambda)
          y <- 0*ybin + (1-ybin)*y1


 y.zip<-ZIP.bothconstant(y)
      ybar<-mean(y)
        nj<-c(table(y))
         n<-sum(nj)
        n0<-nj[1]


      omega<-y.zip$omega
      omega<-ifelse(omega<= 0,0, omega)

#Compute Wald test
   lamda1<-y.zip$lamda.zip

      up<-(n0-n*exp(-lamda1))^2*lamda1*((1-exp(-lamda1))*
          (n0-(n*exp(-lamda1)*(lamda1-ybar)))-(n*lamda1*
           exp(-2*lamda1)))

     down<-n0*ybar*(1-exp(-lamda1))^2*(n0-n*exp(-lamda1)*
          (lamda1-ybar))

     Wald<-up/down

      W <- c(W,Wald)
# The power of the six tests (continuous)

# various test for compare between Poisson and ZIP models

    LRT<-LRT_test(y)
      L<-c(L,LRT)


  score<-score_test(y)
      S<-c(S,score)


  chisq<-chiq_test(y)
    Chi<-c(Chi,chisq)


   conf<-CI_test(y,level)
     CI<-c(CI,conf)

  cochran<-C_test(y)
        C<-c(C,cochran)

     }

        W.power<-sum(W>=qchisq(level,1))/R
        W.power<-round(W.power,3)
        L.power<-sum(L>=qchisq(level,1))/R
        L.power<-round(L.power,3)
        S.power<-sum(S>=qchisq(level,1))/R
        S.power<-round(S.power,3)
        Chi.power<-sum(Chi>=qchisq(level,5))/R
        Chi.power<-round(Chi.power,3)
        CI.power<-sum(CI>0)/R
        CI.power<-round(CI.power,3)
        C.power<-sum(C>=qnorm(level))/R
        C.power<-round(C.power,3)

result<-
list(Wald.power=W.power,LRT.power=L.power,score.power=S.power,
chi.power=Chi.power,CI.power=CI.power,Cochran.power=C.power)
result

}
```

VITAE

Name Miss Saranya Numna

Student ID 5010220131

Educational Attainment

Degree	Name of Institution	Year of Graduation
B.Sc. in Ed	Prince of Songkla University	2006

Scholarship Awards during Enrolment

Teaching Assistant from Faculty of Scince, Prince of Songkla University, 2007-2009.

List of Publication and Proceeding

Saranya Numna and Naratip Jansakul. 2009. Test Statistics for Zero-inflated Poisson model. 2552 - 21-22 2552 501-508