Far East Journal of Theoretical Statistics

Published Online: December 2014
Available online at http://pphmj.com/journals/fjts.htm
Volume 49, Number 1, 2014, Pages 41-60

ASYMPTOTIC DISTRIBUTION OF PSEUDO-LIKELIHOOD RATIO STATISTIC FOR ZERO-INFLATED GENERALIZED LINEAR MODELS UNDER COMPLEX SAMPLING DESIGNS

Khyam Paneru[1,*] **and Hanfeng Chen**[2]

[1]Department of Mathematics
University of Wisconsin-Whitewater
Whitewater, WI 53190-1790, U. S. A.
e-mail: paneruk@uww.edu

[2]Department of Mathematics and Statistics
Bowling Green State University
U. S. A.

Abstract

Zero-inflated mixture (ZIM) regression for zero-inflated population in presence of many zero value responses has been developed by Paneru and Chen [22]. The ZIM regression addresses the issue of estimation problem in generalized linear models under complex probability sampling designs via a two-component mixture model where the non-zero component follows a parametric distribution. As a technical supplement to Paneru and Chen [22], this paper presents theoretical details and complete proof of asymptotic distribution of maximum pseudo-likelihood ratio test statistic. The proposed maximum pseudo-likelihood procedure is applied to a real data set to give both point and

Received: July 25, 2014; Revised: September 25, 2014; Accepted: October 15, 2014
2010 Mathematics Subject Classification: Primary 62F12; Secondary 62J12.
Keywords and phrases: generalized linear model, mixture models, pseudo-likelihood asymptotic distribution, zero-inflated regression model.

*Corresponding author

interval estimates of expected response at different "future" covariate values. It turns out that confidence intervals under the new pseudo-likelihood procedure are shorter than those obtained from the popular maximum likelihood procedure. Nice concave curves of likelihood ratio statistics under both procedures also visualize that the pseudo-likelihood procedure gives shorter confidence intervals.

1. Introduction

Zero-inflated regression model is referred to regression model in presence of many zero-value responses which cause the regression error distribution spiked at zero. The problem of zero-inflation in presence of many zero-value responses exists in many important application areas such as insurance, reliability, meteorology, auditing, ecology, queuing, and manufacturing. For examples, in modeling failure time, a large number of items produced may fail during installation (value of 0 for life time is recorded); in modeling waiting time, many customers may have zero waiting time in queue; in modeling defect counts, there is a relatively large number of zeros (non-defects) in an established manufacturing process. Different approaches and methods that exist in literature are focused on estimating population mean and developing regression models for zero-inflated population.

Kvanli et al. [14] discussed interval estimates of population mean using parametric mixture model, where the non-zero component follows a known probability distribution. Chen and Qin [7] and Chen et al. [6] proposed non-parametric empirical likelihood confidence intervals of zero-inflated population mean. Chen et al. [5] proposed pseudo-likelihood approach to address the issue of complex probability sampling designs to estimate the mean of zero-inflated population.

Lambert [15] introduced zero-inflated Poisson (ZIP) regression to model defect counts in manufacturing and addressed the issue of overdispersion due to excessive zero counts in Poisson distribution. ZIP regression model treats the data as a mixture of zeros and outcomes of a Poisson variate. Welsh et al.

[25] applied different regression models to zero-inflated count data with the application in abundance of rare species and illustrated a detail application with the abundance of Leadbeater's Possum in montane ash forests in South-eastern Australia. Since ordinary least square regression and generalized linear models perform poorly, Welsh et al. [25] suggested modeling response variable as a mixture of Bernoulli and Poisson distribution or Bernoulli and negative binomial distribution. To take care of possible serial correlation between repeated observations, Dobbie and Welsh [10] extended the two-component approach used by Welsh et al. [25]. Extensions and applications of ZIP regression can also be found in Hall [13], Böhning [2], Lee et al. [16], Böhning et al. [3], Ridout et al. [23], and Yau and Lee [27]. Alternatives to ZIP regression can be found in Cui and Yang [9] and Yau et al. [26]. As a special case of Welsh et al. [25] (assuming positive abundance has log-normally distributed error term), Fletcher et al. [12] described an approach that combines ordinary and logistic regression models for skewed data with many zeros.

Zero-inflated regression models existed in literature, such as those mentioned above, do not address the situation when the available data for analysis are obtained through complex probability sampling designs. In a recent paper, Paneru and Chen [22] proposed and developed zero-inflated mixture regression model to investigate estimation problems (both point and confidence interval) in generalized linear models associated with complex sampling designs. As a technical supplement to Paneru and Chen [22], this paper presents the theoretical details and complete proof of asymptotic distribution of pseudo-likelihood ratio statistic in zero-inflated mixture regression. As an application, the new procedure that addresses the issue of complex sampling designs is applied to real data with different "future" covariates and the results are compared to the popular maximum likelihood method.

2. Pseudo-likelihood Function in Zero-inflated Mixture (ZIM) Regression

As defined by Paneru and Chen [22], ZIM regression model is a mixture of zero and non-zero responses where the non-zero response y_i has probability density function $f(y_i; \mu_i, \sigma)$ with respect to a common measure μ, where $\mu_i = E(y_i | x_i) = \psi^{-1}(x_i' \beta)$ with a specific link function ψ and a structure parameter σ. The response y_i at covariate x_i follows the mixture model with pdf

$$g(y_i; \alpha, x_i' \beta, \sigma) = \alpha f(y_i; \psi^{-1}(x_i' \beta), \sigma) I(y_i \neq 0) + (1 - \alpha) I(y_i = 0), \quad (1)$$

where α is the unknown proportion of non-zero values. Model (1) can be viewed as a generalized linear ZIM regression model with one-to-one link function $\psi(\mu_i)$. Generalized linear model (GLM) is an extension to the linear regression model, and it allows regression analysis in more complex situations when classical assumptions for linear regression model do not hold. In generalized linear model, mean of response variable is related to the linear combination of predictors by a one-to-one mapping called the *link function*. For detailed discussion for generalized linear models, see Madsen and Thyregod [18], McCullagh and Nelder [19], and Nelder and Wedderburn [21].

Pseudo log-likelihood function for ZIM regression follows the ideas from Chen et al. [5]. Let the surveyed population $\mathcal{P}$ consists of N sampling units with the values $y_1, ..., y_N$ which are independently generated from super population $g(y_i; \alpha, x_i' \beta, \sigma)$ defined by the model (1). Thus, the log-likelihood function for surveyed population $\mathcal{P}$ is given by

$$\ell(\alpha, \beta, \sigma) = \sum_{i=1}^{N} \log g(y_i; \alpha, x_i' \beta, \sigma).$$

Let s be a random subset of n sampling units with values $y_1, ..., y_n$. Throughout the paper, assume that $m < n$, where m is the number of zero responses, and arrange the response y_i associated with x_i as

$$y_i \neq 0 \text{ for } 1 \leq i \leq n - m$$

$$= 0 \text{ for } n - m + 1 \leq n.$$

The estimate of $\ell(\alpha, \beta, \sigma)$, called *pseudo log-likelihood function* or simply called as *pseudo-likelihood function*, is defined by

$$\hat{\ell}(\alpha, \beta, \sigma) = \sum_{i \in s} w_i \log g(y_i; \alpha, x_i' \beta, \sigma),$$

where the sampling weights $(w_i, i = 1, ..., n)$ are chosen such that $E(\hat{\ell}) = \ell$. Here, E is the expectation under unequal probability sampling designs. When the inclusion probability for the ith unit is π_i, the sampling weight w_i can be chosen to be $w_i = 1/\pi_i$, for $i = 1, ..., n$. More detailed discussion about inclusion probabilities and sampling weights under complex sampling designs can be found in Chen et al. [5] and Chen and Sitter [4].

3. Pseudo-likelihood Ratio Statistic in ZIM Regression Models

Suppose x_0 is a "future" covariate value and one wishes to estimate the response mean at x_0 :

$$\tau = E(Y \mid X = x_0) = \alpha\mu(x_0'\beta) = \alpha\psi^{-1}(x_0'\beta).$$

For convenience, let us use $\lambda = (\tau, \beta, \sigma)$ to re-parametrize the model instead of (α, β, σ) and put $\hat{\ell}(\lambda) = \hat{\ell}(\alpha, \beta, \sigma)$. Let $\Omega_n(\lambda) = -\partial^2\hat{\ell}/\partial\lambda^2$ and $\Delta_n(\lambda) = \text{Var}(\partial\hat{\ell}/\partial\lambda)$, where $\Delta_n(\lambda)$ explains the variations due to the probability sampling design and the model (1) at true value λ_0 of λ. Rewrite $\Omega_n(\lambda)$ and $\Delta_n(\lambda)$ in the partition matrix form as

$$\Omega_n(\lambda) = \begin{pmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{pmatrix}_{(k+2)\times(k+2)} \quad \text{and} \quad \Delta_n(\lambda) = \begin{pmatrix} \delta_{11} & \delta_{12} \\ \delta_{21} & \delta_{22} \end{pmatrix}_{(k+2)\times(k+2)},$$

where δ_{22} and w_{22} are $(k+1)\times(k+1)$ submatrices.

Consider null hypothesis $H_0 : \tau = \tau_0$. Let λ_0 be the true value of λ, $\hat{\lambda}_0$ be the maximum pseudo-likelihood estimate of λ under the null model, and $\hat{\lambda}$ be the maximum pseudo-likelihood estimate of λ under the full model. Define

$$D_0 = -2\{\hat{\ell}(\lambda_0) - \hat{\ell}(\hat{\lambda}_0)\},$$

$$D_1 = -2\{\hat{\ell}(\lambda_0) - \hat{\ell}(\hat{\lambda})\},$$

$$Q_n(\lambda) = \Delta_n^{1/2}\left\{\Omega_n^{-1} - \begin{pmatrix} 0 & 0 \\ 0 & w_{22}^{-1} \end{pmatrix}\right\}\Delta_n^{1/2},$$

$$a_n^2(\lambda) = \{\mathrm{tr}(Q_n(\lambda))\}^{-1}, \tag{2}$$

where $\mathrm{tr}(\cdot)$ is the trace operator. Pseudo-likelihood ratio statistics for τ at $\tau = \tau_0$ is defined by

$$D(\tau_0) = a_n^2(D_1 - D_0),$$

i.e.,

$$D(\tau_0) = 2a_n^2\{\hat{\ell}(\hat{\lambda}) - \hat{\ell}(\hat{\lambda}_0)\}.$$

The point estimate for the parameter τ of interest is defined to be the maximum pseudo-likelihood estimate $\hat{\tau}$ that maximizes $\hat{\ell}$. The confidence intervals for the parameter of interest τ can be constructed via the pseudo-likelihood ratio statistic. The limiting distribution of pseudo-likelihood ratio statistic under the hypothesis $H_0 : \tau = \tau_0$ is required to determine a critical value for constructing the confidence interval. It will be shown below that $D(\tau_0)$ has a χ_1^2 distribution when the true value of τ is τ_0. Consequently,

$100(1-\gamma)\%$ confidence interval for τ is then given by $I_\gamma = \{\tau : D(\tau) \le C_\gamma\}$, where the critical value C_γ is determined approximately by χ^2_1 distribution with a preset level $1-\gamma$.

4. Asymptotic Distribution of Pseudo-likelihood Ratio Statistic

We proceed with some regularity conditions.

RC1. The probability density function f for non-zero observations satisfies the regularity conditions specified in Serfling [24, p. 144].

RC2. The maximum pseudo-likelihood estimate $\hat{\lambda}$ is consistent, that is,

$$\lim_{n\to\infty} P(|\hat{\lambda} - \lambda| \le \varepsilon) = 1 \text{ for all } \varepsilon > 0.$$

RC3. Let s_i be the first order partial derivative of $\log g(y_i;\, \tau,\, x_i'\beta,\, \sigma)$ with respect to $\lambda = (\tau,\, \beta,\, \sigma)$ given by

$$s_i = \frac{\partial}{\partial\lambda} \log g(y_i;\, \tau,\, x_i'\beta,\, \sigma).$$

So,

$$\frac{\partial\hat{\ell}}{\partial\lambda} = \frac{\partial}{\partial\lambda} \sum_{i\in s} w_i \log g(y_i;\, \tau,\, x_i'\beta,\, \sigma)$$

$$= \sum_{i\in s} w_i s_i$$

and

$$\Delta_n(\lambda) = \text{Var}\left(\frac{\partial\hat{\ell}}{\partial\lambda}\right)$$

$$= \text{Var}\left(\sum_{i\in s} w_i s_i\right).$$

Then, as $n \to \infty$,

$$\Delta_n^{-1/2}\left(\sum_{i\in s} w_i s_i\right) \to N(0,\, I)$$

in distribution, where I is the identity matrix.

RC4. Let q_i be the negative of second order partial derivative of $\log g(y_i;\, \tau,\, x_i'\beta,\, \sigma)$ with respect to $\lambda = (\tau,\, \beta,\, \sigma)$ given by

$$\begin{aligned} q_i &= -\frac{\partial^2}{\partial\lambda^2}\log g(y_i;\, \tau,\, x_i'\beta,\, \sigma) \\ &= -\frac{\partial s_i}{\partial\lambda}. \end{aligned}$$

So,

$$\begin{aligned} \Omega_n(\lambda) &= -\frac{\partial^2\hat{\ell}}{\partial\lambda^2} \\ &= -\frac{\partial^2}{\partial\lambda^2}\sum_{i\in s} w_i \log g(y_i;\, \tau,\, x_i'\beta,\, \sigma) \\ &= \sum_{i\in s} w_i q_i. \end{aligned}$$

Then, for any positive definite matrix Ω, as $n \to \infty$, $N^{-1}\Omega_n \to \Omega$ in probability, that is,

$$\lim_{n\to\infty} P(|\, N^{-1}\Omega_n - \Omega\,| \le \varepsilon) = 1 \text{ for all } \varepsilon > 0.$$

RC5. As $n \to \infty$, each of the matrices $a_n^2 Q_n(\lambda_0)$ and $a_n^2\delta_{22}w_{22}^{-1}$ converges in probability to a positive definite matrix.

RC6. Consistency and asymptotic normality of the maximum likelihood estimates in generalized linear regression models hold true as specified in Fahrmeir and Kaufmann [11].

Note. RC1- RC5 are similar to those in Chen et al. [5]. RC6 is needed for generalized linear models. Under identity link function, similar to linear regression model, RC6 can be replaced by the asymptotic property of regression estimates provided in Lehmann [17].

Theorem 1. *Assume regularity conditions RC*1-*RC*6. *Then when the true value of* τ *is* τ_0, *as* $n \to \infty$, $D(\tau_0) \to \chi_1^2$ *in distribution*.

We will need a few lemmas to prove Theorem 1.

Lemma 1 (Rank of a partition matrix)**.** *Let matrix*

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}.$$

(i) *If A is nonsingular*, *then* $Rank(M) = Rank(A) + Rank(D - CA^{-1}B)$.

(ii) *If D is nonsingular*, *then* $Rank(M) = Rank(D) + Rank(A - BD^{-1}C)$.

Proof. See Abadir and Magnus [1] for detail. □

Lemma 2 (Inverse of a partition matrix)**.** *Let matrix*

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}.$$

(i) *If A and* $E = D - CA^{-1}B$ *are nonsingular*, *then*

$$M^{-1} = \begin{pmatrix} A^{-1} + A^{-1}BE^{-1}CA^{-1} & -A^{-1}BE^{-1} \\ -E^{-1}CA^{-1} & E^{-1} \end{pmatrix}.$$

(ii) *If D and* $F = A - BD^{-1}C$ *are nonsingular*, *then*

$$M^{-1} = \begin{pmatrix} F^{-1} & -F^{-1}BD^{-1} \\ -D^{-1}CF^{-1} & D^{-1} + D^{-1}CF^{-1}BD^{-1} \end{pmatrix}.$$

Proof. See Abadir and Magnus [1] for details. □

Lemma 3 (Limit of $a_n^2 Q_n$)**.** *Limit of* $a_n^2 Q_n$ *is idempotent with rank* 1.

Proof. As defined in equation (2), we have

$$a_n^2(\lambda) = \{\mathrm{tr}(Q_n(\lambda))\}^{-1}$$

and

$$Q_n = \Delta_n^{1/2}\left\{\Omega_n^{-1} - \begin{pmatrix} 0 & 0 \\ 0 & w_{22}^{-1} \end{pmatrix}\right\}\Delta_n^{1/2}.$$

Trace of $a_n^2 Q_n$ is given by

$$\begin{aligned} \mathrm{tr}(a_n^2 Q_n) &= \mathrm{tr}\{(\mathrm{tr}(Q_n))^{-1} Q_n\} \\ &= \{\mathrm{tr}(Q_n)\}^{-1}\mathrm{tr}(Q_n) \\ &= 1. \end{aligned}$$

Using Lemma 2, the inverse of Ω_n is given by

$$\begin{aligned} \Omega_n^{-1} &= \begin{pmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{pmatrix}^{-1} \\ &= \begin{pmatrix} F^{-1} & -F^{-1}w_{12}w_{22}^{-1} \\ -w_{22}^{-1}w_{21}F^{-1} & w_{22}^{-1} + w_{22}^{-1}w_{21}F^{-1}w_{12}w_{22}^{-1} \end{pmatrix} \end{aligned}$$

provided w_{22} and $F = w_{11} - w_{12}w_{22}^{-1}w_{21}$ are nonsingular. So,

$$\Omega_n^{-1} - \begin{pmatrix} 0 & 0 \\ 0 & w_{22}^{-1} \end{pmatrix} = \begin{pmatrix} F^{-1} & -F^{-1}w_{12}w_{22}^{-1} \\ -w_{22}^{-1}w_{21}F^{-1} & w_{22}^{-1}w_{21}F^{-1}w_{12}w_{22}^{-1} \end{pmatrix}.$$

Using Lemma 1,

$$\begin{aligned} &\mathrm{Rank}\left(\Omega_n^{-1} - \begin{pmatrix} 0 & 0 \\ 0 & w_{22}^{-1} \end{pmatrix}\right) \\ &= \mathrm{Rank}\begin{pmatrix} F^{-1} & -F^{-1}w_{12}w_{22}^{-1} \\ -w_{22}^{-1}w_{21}F^{-1} & w_{22}^{-1}w_{21}F^{-1}w_{12}w_{22}^{-1} \end{pmatrix} \end{aligned}$$

$$= \text{Rank}(F^{-1}) + \text{Rank}(w_{22}^{-1}w_{21}F^{-1}w_{12}w_{22}^{-1} - w_{22}^{-1}w_{21}F^{-1}FF^{-1}w_{12}w_{22}^{-1})$$

$$= \text{Rank}(F^{-1}) + \text{Rank}(w_{22}^{-1}w_{21}F^{-1}w_{12}w_{22}^{-1} - w_{22}^{-1}w_{21}F^{-1}w_{12}w_{22}^{-1})$$

$$= 1 + 0$$

$$= 1. \qquad \square$$

Under the full model, $\text{Rank}(\Delta_n) = k + 2$. So,

$$\text{Rank}(Q_n) \le \min(1,\ k + 2).$$

From regularity condition RC5, each of the matrices $a_n^2 Q_n$ and $a_n^2 \delta_{22} w_{22}^{-1}$ converges to a positive definite matrix. So, the limit of $a_n^2 Q_n$ exists with $\text{tr}(a_n^2 Q_n) = 1$ and $\text{Rank}(a_n^2 Q_n) = 1$. This proves that limit of $a_n^2 Q_n$ is idempotent with rank 1.

Proof of Theorem 1. As defined by equation (2), pseudo-likelihood ratio statistic $D(\tau_0) = a_n^2(D_1 - D_0)$. Similarly, D_0, D_1, Q_n and a_n^2 are defined by equation (2). We have $\Omega_n = -\dfrac{\partial^2 \hat{\ell}}{\partial \lambda^2}$ and $\dfrac{\partial \hat{\ell}}{\partial \lambda} = \sum_{i \in s} w_i s_i$. Expanding $\sum_{i \in s} w_i s_i(\lambda_0)$ about $\lambda_0 = \hat{\lambda}$, we get

$$\sum_{i \in s} w_i s_i(\lambda_0) = \sum_{i \in s} w_i s_i(\hat{\lambda}) + \{-\Omega_n(\hat{\lambda})\}(\lambda_0 - \hat{\lambda}) + o_p\{N(\lambda_0 - \hat{\lambda})\}$$

$$= 0 + \Omega_n(\hat{\lambda})(\hat{\lambda} - \lambda_0) + o_p\{N(\lambda_0 - \hat{\lambda})\}$$

$$= \Omega_n(\hat{\lambda})(\hat{\lambda} - \lambda_0) + o_p\{N(\lambda_0 - \hat{\lambda})\}$$

$$= \{\Omega_n(\lambda_0) + o_p\{N(\hat{\lambda} - \lambda_0)\}\}(\hat{\lambda} - \lambda_0) + o_p\{N(\lambda_0 - \hat{\lambda})\}$$

$$= \Omega_n(\lambda_0)(\hat{\lambda} - \lambda_0) + o_p\{N(\lambda_0 - \hat{\lambda})\}.$$

So,

$$\Omega_n(\lambda_0)(\hat{\lambda} - \lambda_0) + o_p\{N(\lambda_0 - \hat{\lambda})\} = \sum_{i \in s} w_i s_i(\lambda_0), \tag{3}$$

where the order of the reminder term is insured by the regularity condition RC4. Thus,

$$\Omega_n(\lambda_0)(\hat{\lambda} - \lambda_0) = \sum_{i \in s} w_i s_i(\lambda_0)\{1 + o_p(1)\}.$$

Hence, we get

$$\hat{\lambda} - \lambda_0 = \Omega_n^{-1}(\lambda_0) \sum_{i \in s} w_i s_i(\lambda_0)\{1 + o_p(1)\}. \tag{4}$$

Now expanding $\hat{\ell}(\lambda_0)$ at $\lambda_0 = \hat{\lambda}$,

$$\begin{aligned}
\hat{\ell}(\lambda_0) &= \hat{\ell}(\hat{\lambda}) + \frac{\partial \hat{\ell}(\lambda_0)}{\partial \lambda_0}\bigg|_{\lambda_0 = \hat{\lambda}} (\lambda_0 - \hat{\lambda}) \\
&\quad + \frac{1}{2}(\lambda_0 - \hat{\lambda})' \frac{\partial^2 \hat{\ell}(\lambda_0)}{\partial \lambda_0^2}\bigg|_{\lambda_0 = \hat{\lambda}} (\lambda_0 - \hat{\lambda}) + o_p\{N(\lambda_0 - \hat{\lambda})\} \\
&= \hat{\ell}(\hat{\lambda}) + \sum_{i \in s} w_i s_i(\hat{\lambda})(\lambda_0 - \hat{\lambda}) \\
&\quad + \frac{1}{2}(\lambda_0 - \hat{\lambda})' \{-\Omega_n(\hat{\lambda})\}(\lambda_0 - \hat{\lambda}) + o_p\{N(\lambda_0 - \hat{\lambda})\}.
\end{aligned}$$

So,

$$\begin{aligned}
&\hat{\ell}(\lambda_0) - \hat{\ell}(\hat{\lambda}) \\
&= 0 - \frac{1}{2}(\lambda_0 - \hat{\lambda})' \Omega_n(\hat{\lambda})(\lambda_0 - \hat{\lambda}) + o_p\{N(\lambda_0 - \hat{\lambda})\} \\
&= -\frac{1}{2}(\lambda_0 - \hat{\lambda})' (\Omega_n(\lambda_0) + o_p\{N(\hat{\lambda} - \lambda_0)\})(\lambda_0 - \hat{\lambda}) + o_p\{N(\lambda_0 - \hat{\lambda})\} \\
&= -\frac{1}{2}(\lambda_0 - \hat{\lambda})' \Omega_n(\lambda_0)(\lambda_0 - \hat{\lambda}) + o_p\{N(\lambda_0 - \hat{\lambda})\}.
\end{aligned} \tag{5}$$

Hence,

$$D_1 = -2\{\hat{\ell}(\lambda_0) - \hat{\ell}(\hat{\lambda})\}$$

$$= (\hat{\lambda} - \lambda_0)' \Omega_n(\lambda_0)(\hat{\lambda} - \lambda_0)\{1 + o_p(1)\}$$

$$= \left(\sum_{i \in s} w_i s_i\right)' \Omega_n^{-1} \Omega_n \Omega_n^{-1} \left(\sum_{i \in s} w_i s_i\right) + o_p(1) \text{ (by using equation (4))}$$

$$= \left(\sum_{i \in s} w_i s_i\right)' \Omega_n^{-1} \left(\sum_{i \in s} w_i s_i\right) + o_p(1), \tag{6}$$

where s_i and Ω_n are evaluated at $\lambda = \lambda_0$. Similarly, expanding $\sum_{i \in s} w_i s_i(\lambda_0)$ and $\hat{\ell}(\lambda_0)$ about $\lambda_0 = \hat{\lambda}_0$,

$$D_0 = \left(\sum_{i \in s} w_i s_i\right)' \begin{pmatrix} 0 & 0 \\ 0 & w_{22}^{-1} \end{pmatrix} \left(\sum_{i \in s} w_i s_i\right) + o_p(1). \tag{7}$$

Note that under $H_0 : \tau = \tau_0$, the first component of $\lambda_0 - \hat{\lambda}_0$ is equal to zero. So there are three null submatrices in the partition of Ω_{-1}^n in equation (7). Using equations (6) and (7), pseudo-likelihood ratio statistic is given by

$$D = a_n^2(D_1 - D_0)$$

$$= a_n^2 \left(\sum_{i \in s} w_i s_i\right)' \left\{\Omega_n^{-1} - \begin{pmatrix} 0 & 0 \\ 0 & w_{22}^{-1} \end{pmatrix}\right\} \left(\sum_{i \in s} w_i s_i\right) + o_p(1)$$

$$= a_n^2 \left(\Delta_n^{-1/2} \sum_{i \in s} w_i s_i\right)' \Delta_n^{1/2} \left\{\Omega_n^{-1} - \begin{pmatrix} 0 & 0 \\ 0 & w_{22}^{-1} \end{pmatrix}\right\} \Delta_n^{1/2} \left(\Delta_n^{-1/2} \sum_{i \in s} w_i s_i\right) + o_p(1)$$

$$= \left(\Delta_n^{-1/2} \sum_{i \in s} w_i s_i\right)' (a_n^2 Q_n) \left(\Delta_n^{-1/2} \sum_{i \in s} w_i s_i\right) + o_p(1). \tag{8}$$

From regularity condition RC3, as $n \to \infty$,

$$\Delta_n^{-1/2}\left(\sum_{i\in s} w_i s_i\right) \to N(0,\, I)$$

in distribution, where I is the identity matrix. Using this limiting distribution and result of Lemma 3 (limit of $a_n^2 Q_n$ is idempotent with rank 1) in equation (8), the limiting distribution of pseudo-likelihood ratio statistic can be obtained. So as $n \to \infty$,

$$D(\tau_0) \to \chi_1^2 \tag{9}$$

in distribution. This completes the proof for Theorem 1. □

Remark. The main idea of the above proof is same as in Chen et al. [5].

5. Estimation of a_n^2

From equation (2), we have $a_n^2 = \{\mathrm{tr}(Q_n)\}^{-1}$ and

$$Q_n = \Delta_n^{1/2}\left\{\Omega_n^{-1} - \begin{pmatrix} 0 & 0 \\ 0 & w_{22}^{-1} \end{pmatrix}\right\}\Delta_n^{1/2}.$$

So, a_n^2 can be written as

$$a_n^2 = \frac{1}{\mathrm{tr}(\Omega_n^{-1}\Delta_n) - \mathrm{tr}(w_{22}^{-1}\delta_{22})}, \tag{10}$$

where w_{22} and δ_{22} are submatrices of Ω_n and Δ_n, respectively. Thus, a_n^2 can be estimated through the estimates of $\Omega_n(\lambda_0)$ and $\Delta_n(\lambda_0)$. Since $\hat{\lambda}$ is the maximum pseudo-likelihood estimate of λ, a suitable estimate of $\Omega_n(\lambda_0)$ is $\hat{\Omega}_n(\lambda_0) = \Omega_n(\hat{\lambda})$. Similarly, to estimate $\Delta_n(\lambda_0)$, express $\Delta_n(\lambda_0)$ as

$$\Delta_n(\lambda) = \mathrm{Var}\left(\frac{\partial \hat{\ell}}{\partial \lambda}\right)$$

$$= \mathrm{Var}\left(\sum_{i \in s} w_i s_i(\lambda)\right)$$

$$= E\left\{\mathrm{Var}\left(\sum_{i \in s} w_i s_i(\lambda) | \mathcal{P}\right)\right\} + \mathrm{Var}\left\{E\left(\sum_{i \in s} w_i s_i(\lambda) | \mathcal{P}\right)\right\}. \qquad (11)$$

First term. The estimate of the first term in equation (11) can be obtained from Sen-Yates-Grundy estimate described in Cochran [8]. So, the estimate of this term is given by

$$V = \sum_{i \in s} \sum_{j > i} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} (w_i s_i - w_j s_j)(w_i s_i - w_j s_j)',$$

where $s_i = s_i(\hat{\lambda})$ and π_{ij} is the joint inclusion probability under the sampling design.

Second term. Let

$$t_i = \begin{cases} 1, & \text{if } i \in s,\ i = 1, ..., N, \\ 0, & \text{otherwise.} \end{cases}$$

Since the inclusion probability of *i*th unit is π_i, $E(t_i) = \pi_i = 1/w_i$. So,

$$E\left\{\sum_{i \in s} w_i s_i(\lambda) | \mathcal{P}\right\} = E\left(\sum_{i=1}^{N} t_i w_i s_i\right)$$

$$= \sum_{i=1}^{N} \pi_i w_i s_i(\lambda)$$

$$= \sum_{i=1}^{N} s_i(\lambda).$$

Thus, the second term in equation (11) is given by

$$\mathrm{Var}\left\{E\left(\sum_{i\in s} w_i s_i(\lambda)\,|\,\mathcal{P}\right)\right\} = \mathrm{Var}\left(\sum_{i=1}^{N} s_i(\lambda)\right),$$

which is equal to N times of the Fisher information matrix of the model (1) at $\lambda = \lambda_0$. So, it can be estimated from the observed Fisher information matrix and the estimate is $\hat{\Omega}_n(\lambda_0) = \Omega_n(\hat{\lambda})$. Hence, the estimate of $\Delta_n(\lambda_0)$ is given by $\hat{\Delta}_n = V + \hat{\Omega}_n$. Finally, the estimate of a_n^2 is given by

$$\hat{a}_n^2 = \frac{1}{\mathrm{tr}(\hat{\Omega}_n^{-1}\hat{\Delta}_n) - \mathrm{tr}(\hat{w}_{22}^{-1}\hat{\delta}_{22})}.$$

6. Application

ZIM regression can be applied in both continuous-type and count-type responses. As an application to the real data, log-normal model developed in Paneru and Chen [22] is applied to a commonly used data about inpatient charge of patients in a hospital. The data set consists of 483 observations with approximately 88% zero value responses. Further detail about this data can be found in Murray et al. [20] and Zhou and Cheng [28]. Inpatient charge (in hundreds of dollars) is considered as response variable and health status score (in a scale of 0-100) is considered as an explanatory variable. Observations are divided into two strata according to the gender of patients. Sampling weights for two strata (male and female) are $w_m = 1/0.584$ and $w_f = 1/0.416$. These sampling weights are calculated according to the observed proportions, where the observed proportions of males and females are 58.4% and 41.6%, respectively.

Both point and interval estimates of the parameter of interest τ, the expected inpatient charge (in hundreds of dollars), for different "future" health status scores are presented in Table 1. For example, for a health status score of 20, the "future" covariate vector is denoted by $x_0 = (1, 20)'$. For

comparison purpose, both point and interval estimates of τ are calculated using pseudo-likelihood function and ordinary likelihood function (a popular method in applications). The 95% confidence intervals show that confidence intervals under new procedure via pseudo-likelihood function are narrower than those based on the ordinary likelihood function. At a "future" covariate vector $x_0 = (1, 80)'$, graphs of pseudo-likelihood ratio statistics and likelihood ratio statistics are presented in Figure 1. Both graphs have a nice concave upward shape with the vertex at the pseudo MLE $\hat{\tau}$.

Table 1. Point estimate and 95% confidence interval for τ (in hundreds of dollars) at "future" health status scores of 20, 40, 60 and 80

"future" covariate vector	Method	Lower bound	$\hat{\tau}$	Upper bound
$x_0 = (1, 20)'$	**Pseudo MLE**	12.725	19.137	29.721
	Ordinary MLE	10.027	17.719	34.690
$x_0 = (1, 40)'$	**Pseudo MLE**	14.417	20.450	29.630
	Ordinary MLE	11.914	19.155	33.960
$x_0 = (1, 60)'$	**Pseudo MLE**	14.747	21.853	32.780
	Ordinary MLE	12.225	20.708	38.775
$x_0 = (1, 80)'$	**Pseudo MLE**	13.920	23.353	39.382
	Ordinary MLE	11.145	22.387	49.948

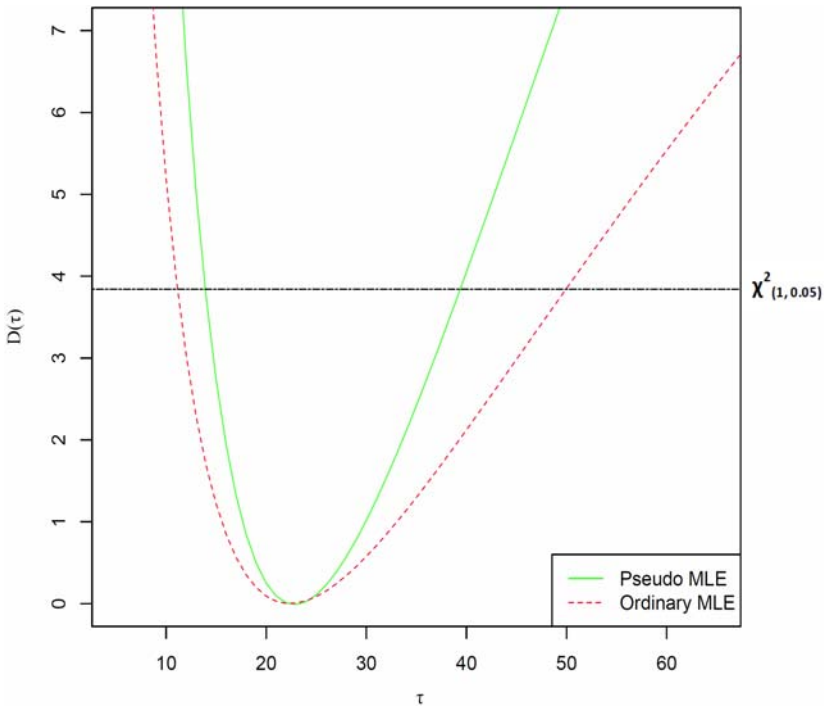

Figure 1. Graph of likelihood ratio statistic $D(\tau)$ vs τ at health status score of 80 such that $x_0 = (1, 80)'$.

References

[1] K. M. Abadir and J. R. Magnus, Matrix Algebra, Cambridge University Press, 2005.

[2] D. Böhning, Zero-inflated Poisson models and C. A. MAN: a tutorial collection of evidence, Biom. J. 40 (1998), 833-843.

[3] D. Böhning, E. Dietz, P. Schlattmann, L. Mendonca and U. Kirchner, The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology, J. Roy. Statist. Soc. Ser. A 162 (1999), 195-209.

[4] J. Chen and R. Sitter, A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys, Statist. Sinica 9 (1999), 385-406.

[5] H. Chen, J. Chen and S. Chen, Confidence intervals for the mean of a population containing many zero values under unequal-probability sampling, Canad. J. Statist. 38 (2010), 582-597.

[6] J. Chen, S. Chen and J. N. K. Rao, Empirical likelihood confidence intervals for the mean of a population containing many zero values, Canad. J. Statist. 31 (2003), 53-68.

[7] S. X. Chen and J. Qin, Empirical likelihood-based confidence intervals for data with possible zero observations, Statist. Probab. Lett. 65 (2003), 29-37.

[8] W. G. Cochran, Sampling Techniques, 3rd ed., Wiley, New York, 1977.

[9] Y. Cui and W. Yang, Zero-inflated generalized Poisson regression mixture model for mapping quantitative trait loci underlying count trait with many zeros, J. Theoret. Biol. 256 (2009), 276-285.

[10] M. J. Dobbie and A. H. Welsh, Modelling correlated zero-inflated count data, Aust. N. Z. J. Stat. 43 (2001), 431-444.

[11] L. Fahrmeir and H. Kaufmann, Consistency and asymptotic normality of the maximum likelihood estimate in generalized linear models, Ann. Statist. 13 (1985), 342-368.

[12] D. Fletcher, D. MacKenzie and E. Villouta, Modelling skewed data with many zeros: a simple approach combining ordinary and logistic regression, Environ. Ecol. Stat. 12 (2005), 45-54.

[13] D. B. Hall, Zero-inflated Poisson and binomial regression with random effects: a case study, Biometrics 56 (2000), 1030-1039.

[14] A. H. Kvanli, Y. K. Shen and L. Y. Deng, Construction of confidence intervals for the mean of a population containing many zero values, J. Bus. Econom. Statist. 16 (1998), 362-368.

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

[16] A. H. Lee, K. Wang and K. K. W. Yau, Analysis of zero-inflated Poisson data incorporating extent of exposure, Biom. J. 43 (2001), 963-975.

[17] E. L. Lehmann, Elements of Large-sample Theory, Springer, New York, 1999.

[18] H. Madsen and P. Thyregod, Introduction to General and Generalized Linear Models, Chapman & Hall/CRC, 2011.

[19] P. McCullagh and J. A. Nelder, Generalized Linear Model, Chapman & Hall/CRC, 1989.

[20] M. D. Murray, L. E. Harris, J. M. Overhage, X. Zhou, G. J. Eckert, F. E. Smith, N. N. Buchanan, F. D. Wolinsky, C. J. McDonald and W. M. Tierney, Failure of computerized treatment suggestions to improve health outcomes of outpatients with uncomplicated hypertension: results of a randomized controlled trial, Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy 24 (2004), 324-337.

[21] J. A. Nelder and R. W. Wedderburn, Generalized linear models, J. Roy. Statist. Soc. Ser. A 135 (1972), 370-384.

[22] K. Paneru and H. Chen, Regression analysis under complex probability sampling designs in presence of many zero-value responses, Adv. Appl. Stat. 40 (2014), 1-29.

[23] M. Ridout, J. Hinde and C. G. B. Demétrio, A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives, Biometrics 57 (2001), 219-223.

[24] R. J. Sering, Approximation Theorems of Mathematical Statistics, Wiley, New York, 1980.

[25] A. H. Welsh, R. B. Cunningham, C. F. Donnelly and D. B. Lindenmayer, Modelling the abundance of rare species: statistical models for counts with extra zeros, Ecological Modelling 88 (1996), 297-308.

[26] K. K. W. Yau, K. Wang and A. H. Lee, Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros, Biom. J. 45 (2003), 437-452.

[27] K. K. W. Yau and A. H. Lee, Zero-inflated Poisson regression with random effects to evaluate an occupational injury prevention programme, Stat. Med. 20 (2001), 2907-2920.

[28] X. Zhou and H. Cheng, A computer program for estimating the re-transformed mean in heteroscedastic two-part models, Computer Methods and Programs in Biomedicine 90 (2008), 210-216.