Score Tests for Semiparametric Zero-inflated Poisson Models

Chin-Shang Li[1]

[1] Department of Public Health Sciences, Division of Biostatistics, University of California, Davis, USA

Correspondence: Chin-Shang Li, Department of Public Health Sciences, Division of Biostatistics, University of California, Davis, CA, USA. E-mail: cssli@ucdavis.edu

Received: March 31, 2012 Accepted: April 16, 2012 Online Published: June 15, 2012

doi:10.5539/ijsp.v1n2p1 URL: http://dx.doi.org/10.5539/ijsp.v1n2p1

Abstract

Count data sets often produce many zeros. It is sometimes potentially questionable to use a linear predictor to model the effect of a continuous covariate of interest in zero-inflated count data. To relax the restriction, Li (2011) proposed a semiparametric zero-inflated Poisson (ZIP) regression model by using fixed-knot cubic *basis* splines or *B*-splines to model the covariate effect, and used the likelihood ratio test to assess the validity of the linear relationship between the natural logarithm of the Poisson mean and the covariate. A score test is conducted to assess whether the extra proportion of zeros in the semiparametric ZIP regression model is equal to zero.

Keywords: *B*-spline, score test, semiparametric Poisson regression model, semiparametric zero-inflated Poisson regression model

1. Introduction

It is common to see count data with large numbers of zeros in many disciplines, e.g., biomedical studies, criminology, environmental economics, traffic accidents, et al. To handle count data with excess zeros, a so-called zero-inflated Poisson (ZIP) distribution is employed (Singh, 1963; Johnson, Kotz, & Kemp, 1992). The ZIP distribution is a mixture of a Poisson distribution and a degenerated distribution at zero as follows:

$$\begin{aligned} \mathrm{P}(Y=y;\lambda,\pi) &= \pi I_{\{y=0\}} + (1-\pi)\frac{\mathrm{e}^{-\lambda}\lambda^{y}}{y!},\ y=0,1,2,\ldots, \\ &= \left[\pi + (1-\pi)\mathrm{e}^{-\lambda}\right]^{I_{\{y=0\}}} \left[(1-\pi)\frac{\mathrm{e}^{-\lambda}\lambda^{y}}{y!}\right]^{I_{\{y>0\}}}. \qquad (1) \end{aligned}$$

Here $I_{\{\cdot\}}$ is the indicator function for an event. The $\pi \in [0,1]$ is a mixing weight to accommodate extra zeros. The ZIP distribution is reduced to a Poisson distribution when $\pi = 0$. The λ is the mean of the Poisson distribution.

One can think of the ZIP distribution in (1) as a population that consists of two parts: the proportion π consisting of subjects who are not at risk of an event of interest and the other part consisting of subjects who are at risk of the event and may have the event several times during a specific time period (Dietz & Böhning, 1997). The zeros from the first part are generally referred to as structural zeros and those from the Poisson distribution are called sampling zeros. This mixture distribution has become the foundation of much methodological development in zero-inflated count data analysis. Some authors have made the inferences on the existence of zero inflation in the count data (e.g., El-Shaarawi, 1985; van den Broek, 1995; Deng & Paul, 2000; Ridout, Hinde, & Demétrio, 2001; Jansakul & Hinde, 2002; Thas & Rayner, 2005); others have constructed various ZIP regression models. The seminal work on ZIP regression by Lambert (1992) was used to model the extra proportion of zeros π and the mean of the Poisson distribution λ simultaneously with linear predictors using the appropriate link functions, and the parametric ZIP regression model was applied to the manufacturing data. Many authors adopted this basic modeling structure, and a number of important extensions have been made (e.g., Welsh, Cunningham, Donnelly, & Lindenmayer, 1996; Shankar, Milton, & Mannering, 1997; Böhning, Dietz, Schlattmann, Mendonca, & Kirchner, 1999; Yau & Lee, 2001; Cheung, 2002; Hall & Zhang, 2004; Lu, Lin, & Shih, 2004; Min & Agresti, 2005; Hall & Wang, 2005; Hu, Li, & Lee, 2011). For example, Hu et al. (2001) applied the ZIP models to assess casualty risk of railroad-grade crossing crashes in Taiwan.

Each variant of these ZIP regression models has unique features, but modeling the effect of the covariate via a linear predictor is a common characteristic. Although it may be completely suitable to use a linear predictor in some applications, it may not be appropriate in other cases. Therefore, Li (2011) proposed a flexible procedure to model the covariate effect as a linear combination of fixed-knot cubic *basis*-spline or *B*-spline functions (Schoenberg, 1946; Curry & Schoenberg, 1966). Semiparametric analysis of (longitudinal) zero-inflated count data also has been proposed by, e.g., Lam, Xue, and Cheung (2006), Chiogna and Gaetan (2007), and Feng and Zhu (2011), but they did not conduct tests to assess the validity of a postulated parametric function for a covariate effect. For example, Chiogna and Gaetan (2007) proposed semiparametric zero-inflated Poisson models that use penalized regression splines to study the relationship between the number of indigo bunting and five land use predictors in an animal abundance study.

The semiparametric ZIP regression model proposed by Li (2011) not only enhances fitting flexibility, but also can be used to assess the adequacy of a postulated linear relationship between the natural logarithm of the Poisson mean and the covariate. However, no tests have been proposed for the extra proportion of zeros π in the semiparametric ZIP regression model equal to 0. Motivated by this, we conduct a score test for $\pi = 0$. The score test has an advantage over the likelihood ratio and Wald tests because it only requires the parameter estimates under the null hypothesis $\pi = 0$, i.e., under the semiparametric Poisson regression model. It is noted that van den Broek (1995) proposed a score test for the extra proportion of zeros, comparing the parametric ZIP regression model with a constant proportion of excess zeros to a parametric Poisson regression model.

Section 2 introduces first briefly the semiparametric ZIP regression model (Li, 2011) and then the score test in detail. The practical use of the score test is illustrated with a real-life data set in Section 3. Some concluding remarks are given in Section 4.

2. Method

2.1 A Semiparametric ZIP Regression Model

Let Y be the event count random variable. Let W be a binary latent variable to indicate a subject's risk state: $W = 0$ if the subject is not at risk of an event; $W = 1$ if the subject is at risk of the event. Therefore, if $Y > 0$, $W = 1$, and if $Y = 0$, W is unobserved. Let $\boldsymbol{z} = (\boldsymbol{x}, u)$, where $\boldsymbol{x} = (x_1, \ldots, x_p)$, for $x_1 = 1$, is a vector of p covariates and u is a continuous covariate of interest.

In the parametric ZIP regression model proposed by Lambert (1992), both the mixing weight $\pi = \mathrm{P}(W = 0; \boldsymbol{z})$ and the Poisson mean $\lambda = E(Y|W = 1; \boldsymbol{z})$ are modeled as functions of $\boldsymbol{z}$. However, in this work we are concerned with the $\boldsymbol{z}$ that only affects the Poisson mean λ and not the probability parameter π. Hence, one can write the ZIP model as follows:

$$\begin{aligned}\mathrm{P}(Y = y; \boldsymbol{z}) &= \pi I_{\{y=0\}} + (1-\pi)\frac{\mathrm{e}^{-\lambda(\boldsymbol{z})}\,[\lambda(\boldsymbol{z})]^y}{y!},\ y = 0, 1, \ldots, \\ &= \left[\pi + (1-\pi)\mathrm{e}^{-\lambda(\boldsymbol{z})}\right]^{I_{\{y=0\}}}\left[(1-\pi)\frac{\mathrm{e}^{-\lambda(\boldsymbol{z})}[\lambda(\boldsymbol{z})]^y}{y!}\right]^{I_{\{y>0\}}}, \qquad (2)\end{aligned}$$

where $\lambda(\boldsymbol{z}) = E(Y|W = 1; \boldsymbol{z})$.

It can be derived easily from (2) that the first two moments of the ZIP distribution are $E(Y; \boldsymbol{z}) = (1-\pi)\lambda(\boldsymbol{z})$ and $Var(Y; \boldsymbol{z}) = E(Y; \boldsymbol{z}) + [\pi/(1-\pi)]E^2(Y; \boldsymbol{z})$. It can be seen from the variance formula that the ZIP model has the ability to account for data variation beyond that which is accommodated by the Poisson model. To extend the parametric ZIP regression model, Li (2011) assumed that the functional form of the effect of u is smooth but unknown, and the effects of $\boldsymbol{x}$ remain linear. The Poisson mean $\lambda(\boldsymbol{z})$ then can be written as

$$\ln[\lambda(\boldsymbol{z})] = \boldsymbol{x\beta} + g(u) \qquad (3)$$

through the canonical log link function. Here $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is a vector of unknown p regression coefficients for the $\boldsymbol{x} = (x_1, \ldots, x_p)$ with $x_1 = 1$. The g is an unspecified smooth function for the effect of u. The model in (3) is a semiparametric Poisson regression model, which can be considered a generalized partially linear model. Because the model in (3) contains both parametric and nonparametric components, Li (2011) referred to the model in (2) with the semiparametric Poisson regression model in (3) as a semiparametric ZIP regression model.

Let $(y_i, \boldsymbol{z}_i)$, $i = 1, \ldots, n$, be the data. The log-likelihood is then written as follows:

$$\ell(\boldsymbol{\beta}, g, \pi) = \sum_{i=1}^{n}\left\{I_{\{y_i=0\}}\ln\left[\pi + (1-\pi)\mathrm{e}^{-\lambda(\boldsymbol{z}_i)}\right] + I_{\{y_i>0\}}\ln\left[(1-\pi)\frac{\mathrm{e}^{-\lambda(\boldsymbol{z}_i)}\,[\lambda(\boldsymbol{z}_i)]^{y_i}}{y_i!}\right]\right\}. \qquad (4)$$

Because cubic splines provide the best compromise between smoothness and computational cost and the B-spline basis produces better-conditioned systems of equations than the truncated power basis and is more likely to have a numerically stable representation of a spline function, Li (2011) used the basis of cubic B-splines with q preselected knots to approximate the unspecified smooth function g in which the rth knot corresponds to the $\frac{r}{q+1}$th sample quantile of the distinct values of u_is.

Let $B_1(u), \dots, B_{q+4}(u)$ be the cubic B-spline basis for the space of cubic splines with q preselected knots. For details of computing B-splines and their mathematical properties, see de Boor (2001). The cubic B-splines space includes a constant function, and the constant is given in the parametric component of the model (3), so to model the g one of the $q + 4$ B-spline basis functions needs to be dropped so that the resulting parametrization is of full rank. Any one of them can be dropped, but for convenience Li (2011) models the g as a linear combination of the first $K = q + 3$ fixed-knot cubic B-spline basis functions as follows:

$$g(u) = \sum_{k=1}^{K} b_k B_k(u), \tag{5}$$

where b_ks are the cubic B-spline coefficients to be estimated. Let $\boldsymbol{b} = (b_1, \dots, b_K)^{\mathrm{T}}$ and $\boldsymbol{B}_u = (B_1(u), \dots, B_K(u))$. The g in (5) then can be expressed as $g(u) = \boldsymbol{B}_u\boldsymbol{b}$ in vector notation. Therefore, the Poisson mean $\lambda(\boldsymbol{z})$ in (3) can be expressed as $\ln[\lambda(\boldsymbol{z})] = \boldsymbol{x\beta} + \boldsymbol{B}_u\boldsymbol{b}$, which can be written as

$$\ln[\lambda(\boldsymbol{z})] = \boldsymbol{A}_z\boldsymbol{\theta} \tag{6}$$

in a simpler vector notation, where $\boldsymbol{A}_z = (\boldsymbol{x}, \boldsymbol{B}_u)$ and $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{b}^{\mathrm{T}})^{\mathrm{T}}$. Consequently, the log-likelihood $\ell(\boldsymbol{\beta}, g, \pi)$ in (4) then becomes

$$\ell(\boldsymbol{\theta}, \pi) \;=\; \sum_{i=1}^{n} \left\{ I_{\{y_i=0\}} \ln\left[\pi + (1-\pi)\mathrm{e}^{-\lambda(z_i)}\right] + I_{\{y_i>0\}} \ln\left[(1-\pi)\frac{\mathrm{e}^{-\lambda(z_i)}[\lambda(z_i)]^{y_i}}{y_i!}\right]\right\}, \tag{7}$$

where $\lambda(\boldsymbol{z}_i) = \exp(\boldsymbol{A}_{z_i}\boldsymbol{\theta})$.

2.2 A Score Test

Possible tests for the null hypothesis $H_0 : \pi = 0$ are the likelihood ratio test, the Wald test and the score test. Because one needs to estimate the model parameters under the alternative hypothesis $\pi > 0$ while using the likelihood ratio and Wald tests, we consider the score test for $H_0 : \pi = 0$ because it has the advantage that we do not have to fit the semiparametric ZIP regression model but just a semiparametric Poisson regression model, which is the reduced model of the semiparametric ZIP regression model under $H_0 : \pi = 0$. Let $\tau = \frac{\pi}{1-\pi}$. Testing $H_0 : \pi = 0$ is then equivalent to testing $H_0 : \tau = 0$. With some algebra, one can write the log-likelihood $\ell(\boldsymbol{\theta}, \pi)$ in (7) as

$$\ell(\boldsymbol{\theta}, \tau) \;=\; \sum_{i=1}^{n} \left\{ -\ln(1+\tau) + I_{\{y_i=0\}} \ln\left[\tau + \mathrm{e}^{-\lambda(z_i)}\right] + I_{\{y_i>0\}} \left[-\lambda(z_i) + y_i\boldsymbol{A}_{z_i}\boldsymbol{\theta} - \ln(y_i!)\right]\right\}. \tag{8}$$

Based on the log-likelihood $\ell(\boldsymbol{\theta}, \tau)$ in (8), the score vector is $\boldsymbol{U}^{\mathrm{T}}(\boldsymbol{\theta}, \tau) = \left(\boldsymbol{U}_{\boldsymbol{\theta}}^{\mathrm{T}}(\boldsymbol{\theta}, \tau), \boldsymbol{U}_{\tau}(\boldsymbol{\theta}, \tau)\right)$ as follows:

$$\boldsymbol{U}_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \tau) = \frac{\partial\ell(\boldsymbol{\theta}, \tau)}{\partial\boldsymbol{\theta}} = \sum_{i=1}^{n} \left\{ I_{\{y_i=0\}} \frac{-\mathrm{e}^{-\lambda(z_i)}}{\tau + \mathrm{e}^{-\lambda(z_i)}}\lambda(z_i) + I_{\{y_i>0\}} \left[y_i - \lambda(z_i)\right]\right\} \boldsymbol{A}_{z_i}^{\mathrm{T}} \tag{9}$$

and

$$\boldsymbol{U}_{\tau}(\boldsymbol{\theta}, \tau) = \frac{\partial\ell(\boldsymbol{\theta}, \tau)}{\partial\tau} = \sum_{i=1}^{n} \left\{ \frac{-1}{1+\tau} + I_{\{y_i=0\}} \frac{1}{\tau + \mathrm{e}^{-\lambda(z_i)}}\right\}. \tag{10}$$

Let $\tilde{\boldsymbol{\theta}}$ be the maximum likelihood estimate of $\boldsymbol{\theta}$ under $H_0 : \tau = 0$ and $\tilde{\lambda}(\boldsymbol{z}_i) = \exp(\boldsymbol{A}_{z_i}\tilde{\boldsymbol{\theta}}_1)$. Then, (9) becomes

$$\begin{aligned}
\sum_{i=1}^{n} \left\{ -I_{\{y_i=0\}}\tilde{\lambda}(z_i) + I_{\{y_i>0\}} \left[y_i - \tilde{\lambda}(z_i)\right]\right\} \boldsymbol{A}_{z_i}^{\mathrm{T}} &= \sum_{i=1}^{n} \left\{ I_{\{y_i=0\}}y_i - I_{\{y_i=0\}}\tilde{\lambda}(z_i) + I_{\{y_i>0\}} \left[y_i - \tilde{\lambda}(z_i)\right]\right\} \boldsymbol{A}_{z_i}^{\mathrm{T}} \\
&= \sum_{i=1}^{n} \left(I_{\{y_i=0\}} + I_{\{y_i>0\}}\right)\left[y_i - \tilde{\lambda}(z_i)\right] \boldsymbol{A}_{z_i}^{\mathrm{T}} \\
&= \sum_{i=1}^{n} \left[y_i - \tilde{\lambda}(z_i)\right] \boldsymbol{A}_{z_i}^{\mathrm{T}} = \boldsymbol{0},
\end{aligned} \tag{11}$$

and (10) is

$$\sum_{i=1}^{n}\left[I_{\{y_i=0\}}\frac{1}{\mathrm{e}^{-\tilde{\lambda}(z_i)}}-1\right]=\sum_{i=1}^{n}\left[I_{\{y_i=0\}}\mathrm{e}^{\tilde{\lambda}(z_i)}-1\right]. \tag{12}$$

Thus,

$$\boldsymbol{U}^{\mathrm{T}}(\tilde{\boldsymbol{\theta}},0)=\left(\mathbf{0}^{\mathrm{T}},\sum_{i=1}^{n}\left[\frac{I_{\{y_i=0\}}}{\mathrm{e}^{-\tilde{\lambda}(z_i)}}-1\right]\right)=\left(\mathbf{0}^{\mathrm{T}},\sum_{i=1}^{n}\left[I_{\{y_i=0\}}\mathrm{e}^{\tilde{\lambda}(z_i)}-1\right]\right).$$

The second-order partial derivatives of $\ell(\boldsymbol{\theta},\tau)$ with respect to $\boldsymbol{\theta}$ and τ are

$$\frac{\partial^2\ell(\boldsymbol{\theta},\tau)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}}=\sum_{i=1}^{n}\left\{I_{\{y_i=0\}}\frac{-\mathrm{e}^{-\lambda(z_i)}\left[(1-\lambda(\boldsymbol{z}_i))\tau+\mathrm{e}^{-\lambda(z_i)}\right]}{[\tau+\mathrm{e}^{-\lambda(z_i)}]^2}\lambda(z_i)-I_{\{y_i>0\}}\lambda(\boldsymbol{z}_i)\right\}\boldsymbol{A}_{z_i}^{\mathrm{T}}\boldsymbol{A}_{z_i},$$

$$\frac{\partial^2\ell(\boldsymbol{\theta},\tau)}{\partial\boldsymbol{\theta}\partial\tau}=\sum_{i=1}^{n}\left\{I_{\{y_i=0\}}\frac{\mathrm{e}^{-\lambda(z_i)}}{[\tau+\mathrm{e}^{-\lambda(z_i)}]^2}\lambda(\boldsymbol{z}_i)\right\}\boldsymbol{A}_{z_i}^{\mathrm{T}},$$

and

$$\frac{\partial^2\ell(\boldsymbol{\theta},\tau)}{\partial\tau^2}=\sum_{i=1}^{n}\left\{\frac{1}{(1+\tau)^2}-I_{\{y_i=0\}}\frac{1}{[\tau+\mathrm{e}^{-\lambda(z_i)}]^2}\right\}.$$

Using

$$E(I_{\{y_i=0\}})=P(Y_i=0)=\frac{\tau+\mathrm{e}^{-\lambda(z_i)}}{1+\tau}$$

and

$$E(I_{\{y_i>0\}})=1-E(I_{\{y_i=0\}})=1-\frac{\tau+e^{-\lambda(z_i)}}{1+\tau}=\frac{1-\mathrm{e}^{-\lambda(z_i)}}{1+\tau},$$

one can show that the expected Fisher information matrix

$$\boldsymbol{I}(\boldsymbol{\theta},\tau)=\begin{bmatrix}\boldsymbol{I}_{\boldsymbol{\theta\theta}}(\boldsymbol{\theta},\tau) & \boldsymbol{I}_{\boldsymbol{\theta}\tau}(\boldsymbol{\theta},\tau)\\ \boldsymbol{I}_{\boldsymbol{\theta}\tau}^{\mathrm{T}}(\boldsymbol{\theta},\tau) & I_{\tau\tau}(\boldsymbol{\theta},\tau)\end{bmatrix}$$

has the following entries:

$$\boldsymbol{I}_{\boldsymbol{\theta\theta}}(\boldsymbol{\theta},\tau)=-E\left[\frac{\partial^2\ell(\boldsymbol{\theta},\tau)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}}\right]=\sum_{i=1}^{n}\left\{\frac{-\mathrm{e}^{-\lambda(z_i)}\lambda(\boldsymbol{z}_i)\tau+\tau+\mathrm{e}^{-\lambda(z_i)}}{(1+\tau)[\tau+e^{-\lambda(z_i)}]}\right\}\lambda(\boldsymbol{z}_i)\boldsymbol{A}_{z_i}^{\mathrm{T}}\boldsymbol{A}_{z_i},$$

$$\boldsymbol{I}_{\boldsymbol{\theta}\tau}(\boldsymbol{\theta},\tau)=-E\left[\frac{\partial^2\ell(\boldsymbol{\theta},\tau)}{\partial\boldsymbol{\theta}\partial\tau}\right]=-\sum_{i=1}^{n}\left\{\frac{\mathrm{e}^{-\lambda(z_i)}}{(1+\tau)[\tau+\mathrm{e}^{-\lambda(z_i)}]}\right\}\lambda(\boldsymbol{z}_i)\boldsymbol{A}_{z_i}^{\mathrm{T}},$$

and

$$I_{\tau\tau}(\boldsymbol{\theta},\tau)=-E\left[\frac{\partial^2\ell(\boldsymbol{\theta},\tau)}{\partial\tau^2}\right]=\sum_{i=1}^{n}\left\{\frac{1-\mathrm{e}^{-\lambda(z_i)}}{(1+\tau)^2[\tau+\mathrm{e}^{-\lambda(z_i)}]}\right\}.$$

Therefore, the $\boldsymbol{I}(\tilde{\boldsymbol{\theta}},0)$ has the following entries:

$$\boldsymbol{I}_{\boldsymbol{\theta\theta}}(\tilde{\boldsymbol{\theta}},0)=\sum_{i=1}^{n}\tilde{\lambda}(z_i)\boldsymbol{A}_{z_i}^{\mathrm{T}}\boldsymbol{A}_{z_i}=\boldsymbol{A}^{\mathrm{T}}\mathrm{diag}(\tilde{\boldsymbol{\lambda}})\boldsymbol{A},$$

$$\boldsymbol{I}_{\theta\tau}(\tilde{\boldsymbol{\theta}}, 0) = -\sum_{i=1}^{n} \tilde{\lambda}(z_i)\boldsymbol{A}_{z_i}^{\mathrm{T}} = -\boldsymbol{A}^{\mathrm{T}}\tilde{\boldsymbol{\lambda}},$$

and

$$I_{\tau\tau}(\tilde{\boldsymbol{\theta}}, 0) = \sum_{i=1}^{n} \frac{1 - \mathrm{e}^{-\lambda(z_i)}}{\mathrm{e}^{-\lambda(z_i)}} = \sum_{i=1}^{n}[\mathrm{e}^{\lambda(z_i)} - 1].$$

Here $\boldsymbol{A} = (\boldsymbol{A}_{z_1}^{\mathrm{T}}, \ldots, \boldsymbol{A}_{z_n}^{\mathrm{T}})^{\mathrm{T}}$ is an $n \times (p + K)$ matrix. diag$(\tilde{\boldsymbol{\lambda}})$ is an $n \times n$ diagonal matrix with the (i, i)th entry $\tilde{\lambda}(z_i) = \exp(\boldsymbol{A}_{z_i}\tilde{\boldsymbol{\theta}})$, $i = 1, 2, \ldots, n$. Denote the inverse of $\boldsymbol{I}(\tilde{\boldsymbol{\theta}}, 0)$ by $\boldsymbol{M}(\tilde{\boldsymbol{\theta}}, 0) = \boldsymbol{I}^{-1}(\tilde{\boldsymbol{\theta}}, 0)$ that can be partitioned as

$$\boldsymbol{M}(\tilde{\boldsymbol{\theta}}, 0) = \begin{bmatrix} \boldsymbol{M}_{\theta\theta}(\tilde{\boldsymbol{\theta}}, 0) & \boldsymbol{M}_{\theta\tau}(\tilde{\boldsymbol{\theta}}, 0) \\ \boldsymbol{M}_{\theta\tau}^{\mathrm{T}}(\tilde{\boldsymbol{\theta}}, 0) & M_{\tau\tau}(\tilde{\boldsymbol{\theta}}, 0) \end{bmatrix}.$$

By using the formula of inverse of (partitioned) matrices and the fact of $\tilde{\boldsymbol{\lambda}} = \mathrm{diag}(\tilde{\boldsymbol{\lambda}})\boldsymbol{A}\boldsymbol{e}_{p+K}$, where $\boldsymbol{e}_{p+K}$ is a $(p + K) \times 1$ vector that has a 1 as the first element and has the other elements equal to zero, we can have

$$\begin{aligned}
M_{\tau\tau}(\tilde{\boldsymbol{\theta}}, 0) &= \left[I_{\tau\tau}(\tilde{\boldsymbol{\theta}}_1, 0) - \boldsymbol{I}_{\theta\tau}^{\mathrm{T}}(\tilde{\boldsymbol{\theta}}, 0)\boldsymbol{I}_{\theta\theta}^{-1}(\tilde{\boldsymbol{\theta}}, 0)\boldsymbol{I}_{\theta\tau}(\tilde{\boldsymbol{\theta}}, 0)\right]^{-1} \\
&= \left\{\sum_{i=1}^{n}\left[\mathrm{e}^{\tilde{\lambda}(z_i)} - 1\right] - \tilde{\boldsymbol{\lambda}}^{\mathrm{T}}\boldsymbol{A}\left[\boldsymbol{A}^{\mathrm{T}}\mathrm{diag}(\tilde{\boldsymbol{\lambda}})\boldsymbol{A}\right]^{-1}\boldsymbol{A}^{\mathrm{T}}\tilde{\boldsymbol{\lambda}}\right\}^{-1} \\
&= \left\{\sum_{i=1}^{n}\left[\mathrm{e}^{\tilde{\lambda}(z_i)} - 1\right] - \boldsymbol{e}_{p+K}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\mathrm{diag}(\tilde{\boldsymbol{\lambda}})\boldsymbol{A}\left[\boldsymbol{A}^{\mathrm{T}}\mathrm{diag}(\tilde{\boldsymbol{\lambda}})\boldsymbol{A}\right]^{-1}\boldsymbol{A}^{\mathrm{T}}\mathrm{diag}(\tilde{\boldsymbol{\lambda}})\boldsymbol{A}\boldsymbol{e}_{p+K}\right\}^{-1} \\
&= \left\{\sum_{i=1}^{n}\left[\mathrm{e}^{\tilde{\lambda}(z_i)} - 1\right] - \boldsymbol{e}_{p+K}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\mathrm{diag}(\tilde{\boldsymbol{\lambda}})\boldsymbol{A}\boldsymbol{e}_{p+K}\right\}^{-1} \\
&= \left\{\sum_{i=1}^{n}\left[\mathrm{e}^{\tilde{\lambda}(z_i)} - 1\right] - \mathbf{1}^{\mathrm{T}}\tilde{\boldsymbol{\lambda}}\right\}^{-1} \\
&= \left\{\sum_{i=1}^{n}\left[\mathrm{e}^{\tilde{\lambda}(z_i)} - 1\right] - n\bar{y}\right\}^{-1},
\end{aligned}$$

where from (11) we have used $\sum_{i=1}^{n}[y_i - \tilde{\lambda}(z_i)] = 0$ that is equivalent to $\mathbf{1}^{\mathrm{T}}\tilde{\boldsymbol{\lambda}} = \sum_{i=1}^{n} y_i = n\bar{y}$ for $\bar{y} = \sum_{i=1}^{n} y_i/n$. Therefore, the score statistic to test $H_0 : \tau = 0$ is

$$\begin{aligned}
S(\tilde{\boldsymbol{\theta}}, 0) &= \boldsymbol{U}^{\mathrm{T}}(\tilde{\boldsymbol{\theta}}, 0)\boldsymbol{I}^{-1}(\tilde{\boldsymbol{\theta}}, 0)\boldsymbol{U}(\tilde{\boldsymbol{\theta}}, 0) \\
&= \left[\sum_{i=1}^{n}\left(I_{\{y_i=0\}}\mathrm{e}^{\tilde{\lambda}(z_i)} - 1\right)\right] M_{\tau\tau}(\tilde{\boldsymbol{\theta}}, 0)\left[\sum_{i=1}^{n}\left(I_{\{y_i=0\}}\mathrm{e}^{-\tilde{\lambda}(z_i)} - 1\right)\right] \\
&= \left[\sum_{i=1}^{n}\left(I_{\{y_i=0\}}\mathrm{e}^{\tilde{\lambda}(z_i)} - 1\right)\right]\left\{\sum_{i=1}^{n}\left[\mathrm{e}^{\tilde{\lambda}(z_i)} - 1\right] - n\bar{y}\right\}^{-1}\left[\sum_{i=1}^{n}\left(I_{\{y_i=0\}}\mathrm{e}^{\tilde{\lambda}(z_i)} - 1\right)\right] \\
&= \frac{\left[\sum_{i=1}^{n}\left(I_{\{y_i=0\}}\mathrm{e}^{\tilde{\lambda}(z_i)} - 1\right)\right]^2}{\sum_{i=1}^{n}\left[\mathrm{e}^{\tilde{\lambda}(z_i)} - 1\right] - n\bar{y}},
\end{aligned}$$

which has an asymptotic chi-squared distribution with 1 degree of freedom under $H_0 : \tau = 0$.

3. Example

To illustrate the practical use of the score test, we use the data set from a study of the attendance behavior of 316 high school juniors at two schools, which is available at the website http://www.ats.ucla.edu/stat/mplus/dae/poissonreg.dat. The response variable is the number of days of absence. The predictors include gender of the student and standardized test scores in mathematics and language arts. Let

$\boldsymbol{z} = (\boldsymbol{x}, u)$ for $\boldsymbol{x} = (x_1, x_2, x_3) = (1, I_{\{\text{male}\}}, \text{standardized language arts score})$ and u = standardized mathematics score (SMS). Here the $I_{\{\text{male}\}}$ is a binary indicator of student gender, which is 1 if male; 0 otherwise. Among the 316 SMSs, there were 77 distinct SMSs.

To assess whether the SMS has a linear effect on the natural logarithm of the expected number of days of absence, Li (2011) fitted the proposed semiparametric ZIP regression model and the parametric ZIP regression model to the attendance behavior data. The likelihood ratio test results revealed that the proposed semiparametric ZIP regression model had a better fit than the parametric ZIP regression model, i.e., the SMS had a statistically significantly nonlinear effect on the natural logarithm of the expected number of days of absence. To apply the score test for zero-inflation in the semiparametric ZIP regression model (Li, 2011), i.e., test $H_0 : \pi = 0$, because the semiparametric ZIP regression model (Li, 2011) under $H_0 : \pi = 0$ becomes the semiparametric Poisson regression model, we fit this semiparametric Poisson regression model, using 15 knots, to this data set in which the functional form of the effect of SMS, $g(u)$, is smooth but unknown that is modeled as $g(u) = \sum_{k=1}^{K} b_k B_k(u)$, where $K = 18$. More specifically, we fit the following semiparametric Poisson regression model to the data:

$$\ln[\lambda(\boldsymbol{z})] = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \sum_{k=1}^{K} b_k B_k(u).$$

The p-value of the score test using 15 knots is less than 0.0001, so it rejects $H_0 : \pi = 0$, which gives evidence that there are too many zeros observed in the study data.

4. Concluding Remarks

A score test is conducted to evaluate whether there is an extra proportion of zeros in the semiparametric ZIP regression model (Li, 2011). The advantage of the score test for zero-inflation over the likelihood ratio and Wald tests is to only fit the semiparametric Poisson regression model. We focus only on the case in which the conducted score test is for zero-inflation in the semiparametric ZIP regression model (Li, 2011) that is used to describe a relationship between the natural logarithm of the Poisson mean and a continuous covariate by a linear combination of fixed-knot cubic B-splines, but the conducted score test also can be used for zero-inflation in a semiparametric ZIP regression model that is used to depict relationships between the natural logarithm of the Poisson mean and several continuous covariates of interest. We assume the functional form for the effects of these covariates to be smooth but unknown. Their effects can be incorporated in an additive fashion, and each effect is modeled as a linear combination of fixed-knot cubic B-splines.

Acknowledgment

The author is grateful to the editor and a referee whose helpful comments improved the presentation. This publication was made possible by Grant Number UL1 RR024146 from the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH), and NIH Roadmap for Medical Research.

References

Böhning, D., Dietz, E., Schlattmann, P., Mendonca, L., & Kirchner, U. (1999). The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. *J. R. Stat. Soc., Ser. A, 62*, 195-209. http://dx.doi.org/10.1111/1467-985X.00130

Cheung, Y. B. (2002). Zero-inflated models for regression analysis of count data: a study of growth and development. *Statist. Med., 21*, 1461-1469. http://dx.doi.org/10.1002/sim.1088

Chiogna, M., & Gaetan, C. (2007). Semiparametric zero-inflated Poisson models with application to animal abundance studies. *Environmetrics, 18*, 303-314. http://dx.doi.org/10.1002/env.830

Curry, H. B., & Schoenberg, I. J. (1966). On Pólya frequency functions IV: The fundamental splines and their limits. *J. Analyse Math., 17*, 71-107. http://dx.doi.org/10.1007/BF02788653

de Boor, C. (2001). *A practical guide to splines*. Revised edition. Applied Mathematical Sciences 27. New York, NY: Spring-Verlag.

Deng, D., & Paul, S. R. (2000). Score tests for zero inflation in generalized linear models. *The Canadian Journal of Statistics, 28*, 563-570. http://dx.doi.org/10.2307/3315965

Dietz, K. & Böhning, D. (1997). The use of two-component mixture models with one completely or partly known component. *Computational Statistics, 12*, 219-234.

El-Shaarawi, A. H. (1985). Some goodness-of-fit methods for the Poisson plus added zeros distribution. *Applied and Environmental Microbiology, 49*, 1304-1306. http://aem.asm.org/content/49/5/1304

Feng, J., & Zhu, Z. (2011). Semiparametric analysis of longitudinal zero-inflated count data. *Journal of Multivariate Analysis, 102*, 61-72. http://dx.doi.org/10.1016/j.jmva.2010.08.001

Hall, D. B., & Wang, L. (2005). Two-component mixtures of generalized linear mixed effects models for cluster correlated data. *Statistical Modelling, 5*, 21-37. http://dx.doi.org/10.1191/1471082X05st090oa

Hall, D. B., & Zhang, Z. (2004). Marginal models for zero inflated cluster data. *Statistical Modelling, 4*, 161-180. http://dx.doi.org/10.1191/1471082X04st076oa

Hu, S. R., Li, C. S., & Lee, C. K. (2011). Assessing casualty risk of railroad-grade crossing crashes using zero-inflated Poisson models. *Journal of Transportation Engineering, 137*, 527-536. http://dx.doi.org/10.1061/(ASCE)TE.1943-5436.0000243

Jansakul, N., & Hinde, J. P. (2002). Score tests for zero-inflated Poisson models. *Computational Statistics & Data Analysis, 40*, 75-96. http://dx.doi.org/10.1016/S0167-9473(01)00104-9

Johnson, N. L., Kotz, S., & Kemp, A. W. (1992). *Univariate Discrete Distributions*. New York, NY: Wiley & Sons.

Lam, K. F., Xue, H., & Cheung, Y. B. (2006). Semiparametric analysis of zero-inflated count data. *Biometrics, 62*, 996-1003. http://dx.doi.org/10.1111/j.1541-0420.2006.00575.x

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics, 34*, 1-14. http://dx.doi.org/10.2307/1269547

Li, C. S. (2011). A lack-of-fit test for parametric zero-inflated Poisson models. *Journal of Statistical Computation and Simulation, 81*, 1081-1098. http://dx.doi.org/10.1080/00949651003677410

Lu, S. E., Lin, Y., & Shih, W. C. J. (2004). Analyzing excessive no changes in clinical trials with clustered data. *Biometrics, 60*, 257-267. http://dx.doi.org/10.1111/j.0006-341X.2004.00155.x

Min, Y., & Agresti, A. (2005). Random effect models for repeated measures of zero-inflated count data. *Statistical Modelling, 5*, 1-19. http://dx.doi.org/10.1191/1471082X05st084oa

Ridout, M., Hinde, J., & Demétrio, G. (2001). A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. *Biometrics, 57*, 219-223. http://dx.doi.org/10.1111/j.0006-341X.2001.00219.x

Schoenberg, I. J. (1946). Contributions to the problem of approximation of equidistant data by analytic functions. *Quart. Appli. Math., 4*, 45-99; 112-141.

Shankar, V., Milton, J., & Mannering, F. (1997). Modeling accident frequencies as zero-altered probability processes: An empirical inquiry. *Accident Analysis and Prevention, 29*, 829-837. http://dx.doi.org/10.1016/S0001-4575(97)00052-3

Singh, S. (1963). A note on inflated Poisson distribution. *Journal of the Indian Statistical Association, 1*, 140-144.

Thas, O., & Rayner, J. C. W. (2005). Smooth tests for the zero-inflated Poisson distribution. *Biometrics, 61*, 808-815. http://dx.doi.org/10.1111/j.1541-0420.2005.00351.x

van den Broek, J. (1995). A score test for zero inflation in a Poisson distribution. *Biometrics, 51*, 738-743. http://dx.doi.org/10.2307/2532959

Welsh, A. H., Cunningham, R. B., Donnelly, C. F., & Lindenmayer, D. B. (1996). Modelling the abundance of rare species: Statistical models for counts with extra zeros. *Ecological Modelling, 88*, 297-308. http://dx.doi.org/10.1016/0304-3800(95)00113-1

Yau, K. K. W., & Lee, A. H. (2001). Zero-inflated Poisson regression with random effects to evaluate an occupational injury prevention programme. *Statist. Med., 20*, 2907-2920. http://dx.doi.org/10.1002/sim.860