

## Semiparametric Analysis of Zero-Inflated Count Data

K. F. Lam,<sup>1,\*</sup> Hongqi Xue,<sup>2</sup> and Yin Bun Cheung<sup>3</sup>

<sup>1</sup>Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong

<sup>2</sup>Department of Mathematics, Graduate University of Chinese Academy of Sciences, Beijing, China

<sup>3</sup>MRC Tropical Epidemiology Group, London School of Hygiene and Tropical Medicine,  
Keppel Street, London, WC1E 7HT, U.K.

\* *email*: hrntlkf@hku.hk

**SUMMARY.** Medical and public health research often involve the analysis of count data that exhibit a substantially large proportion of zeros, such as the number of heart attacks and the number of days of missed primary activities in a given period. A zero-inflated Poisson regression model, which hypothesizes a two-point heterogeneity in the population characterized by a binary random effect, is generally used to model such data. Subjects are broadly categorized into the low-risk group leading to structural zero counts and high-risk (or normal) group so that the counts can be modeled by a Poisson regression model. The main aim is to identify the explanatory variables that have significant effects on (i) the probability that the subject is from the low-risk group by means of a logistic regression formulation; and (ii) the magnitude of the counts, given that the subject is from the high-risk group by means of a Poisson regression where the effects of the covariates are assumed to be linearly related to the natural logarithm of the mean of the counts. In this article we consider a semiparametric zero-inflated Poisson regression model that postulates a possibly nonlinear relationship between the natural logarithm of the mean of the counts and a particular covariate. A sieve maximum likelihood estimation method is proposed. Asymptotic properties of the proposed sieve maximum likelihood estimators are discussed. Under some mild conditions, the estimators are shown to be asymptotically efficient and normally distributed. Simulation studies were carried out to investigate the performance of the proposed method. For illustration purpose, the method is applied to a data set from a public health survey conducted in Indonesia where the variable of interest is the number of days of missed primary activities due to illness in a 4-week period.

**KEY WORDS:** Asymptotically efficient; Generalized partly linear model; Sieve maximum likelihood estimator; Zero-inflated Poisson regression model.

### 1. Introduction

Medical and public health research often involve the analysis of count data that exhibit a substantially large proportion of zeros, such as the number of heart attacks and the number of days of missing primary activities in a given period. By suggesting that zero-inflated count data may actually be generated according to a two-state process, a zero-inflated Poisson (ZIP) regression model is proposed by Mullahy (1986) which hypothesizes a two-point heterogeneity in the population. The ZIP regression model is widely applicable in diverse disciplines because of its flexibility. Lambert (1992) analyzed the number of defects in a manufacturing process; Welsh et al. (1996) examined the abundance of rare species; Shankar, Milton, and Mannering (1997) modeled accident frequencies; Bohning et al. (1999) analyzed data from the study of caries prevention in dental epidemiology; Street, Jones, and Furuta (1999) analyzed data on pharmaceutical utilization and expenditure in Russia; and Cheung (2002) analyzed data from a study of growth and development in medical research.

The approach simply makes use of a binary random effect to recognize explicitly the dichotomy of the population into two subpopulations, namely, the low-risk group leading

to structural zero counts and high-risk (or normal) group so that the counts can be modeled by a Poisson regression model. A major attraction of the ZIP regression model is the ability to pick up two different regimes from which the zero counts arise because zero counts can also be observed from subjects in the high-risk group, termed sample zeros, in addition to the structural zeros from the subjects in the low-risk group. The main aim is to identify the explanatory variables that have significant effects on (i) the probability that the subject is from the low-risk group by means of a logistic regression formulation; and (ii) the magnitude of the counts, given that the subject is from the high-risk group by means of a Poisson regression where the effects of the covariates are assumed to be linearly related to the natural logarithm of the mean of the counts, the link function in the generalized linear model setup. However the effects of some continuous covariates, such as the amount of dosages and age of the subjects, may not be linearly related to the link function. The effects may reach a plateau very quickly and remain at a very high level thereafter.

The work presented in this article is motivated by data from a public health survey conducted in Indonesia in 1997

(see Frankenberg and Thomas, 2000), where the variable of interest is the number of days people of working age missed their primary activities due to illness in the past 4 weeks. This outcome variable is important not only from a public health point of view but also from the viewpoint of economic evaluation of disease burdens. Parametric analyses on the relation between age and absenteeism have not shown consistent results (e.g., Shephard, 1999; Arola et al., 2003). Although it is biologically possible that the frequency and duration of illness may change smoothly with age, people at different ages are subject to the influence of different factors such as job attitude and insecurity of employment. Qualitative research suggested that these factors can affect the relation between age and absenteeism in unexpected ways (Cant, O'Loughlin, and Legge, 2001). In most large-scale public health surveys, the variable age is generally found to be insignificant when assuming a linear relationship in a parametric setup, mainly because the age effects may be very different among different age groups. A semiparametric regression analysis is extremely useful to explore the possibly nonlinear effect of age on the responses as it can help the policy makers to make up different policies targeted for different age groups so that the health policies can be made more cost effective.

We consider a semiparametric extension of the ZIP regression model by allowing a covariate to act nonlinearly to the link function of the Poisson regression model in this article. The semiparametric ZIP regression model combines a logistic formulation for the probability that an individual subject is drawn from the low-risk group and the generalized partly linear model for the Poisson counts given that the subject comes from the high-risk group. The model and the proposed sieve maximum likelihood estimator (MLE) will be discussed in Section 2. Asymptotic properties of the proposed estimators are studied in Section 3. Under some mild conditions, the estimators for the unknown parameters are shown to be asymptotically efficient and normally distributed. Two approaches to the estimation of the variance of the estimator are proposed. The method is applied to a data set from a public health survey conducted in Indonesia, where the variable of interest is the number of days of missing primary activities due to illness in a 4-week period. The results are reported in Section 4. Simulation studies were carried out to investigate the performance of the proposed method in Section 5. Some concluding remarks are provided in Section 6.

## 2. Models and the Proposed Sieve MLE

Suppose that the counts,  $Y$ , are generated independently according to a zero-inflated Poisson distribution; the zeros are assumed to arise in two ways corresponding to distinct underlying states. The first state occurs with probability  $p$  and produces only zeros, while the other state occurs with probability  $(1 - p)$  and leads to a standard Poisson count with mean  $\lambda$  (see Jansakul and Hinde, 2002). In general, the zeros from the first state are called structural zeros and those from the Poisson distribution are called sampling zeros. This two-state process gives a simple two-component mixture distribution with probability mass function

$$P(Y = y) = \begin{cases} p + (1 - p)e^{-\lambda}, & y = 0, \\ (1 - p)\frac{e^{-\lambda}\lambda^y}{y!}, & y = 1, 2, \dots, \quad 0 \leq p \leq 1 \end{cases} \quad (1)$$

and  $E(Y) = (1 - p)\lambda$ .

To apply the ZIP model in practical modeling situations, Lambert (1992) suggested the following joint models for  $\lambda$  and  $p$ ,

$$\log(\lambda) = \beta^{*T}(1, X^T)^T \quad \text{and} \quad \log\left(\frac{p}{1 - p}\right) = \gamma^T Z, \quad (2)$$

where  $X = (X_1, \dots, X_{d_1})^T$  and  $Z = (1, Z_1, \dots, Z_{d_2})^T$  are vectors of observable covariates,  $\beta^* = (\beta_0, \beta_1, \dots, \beta_{d_1})^T$ ,  $\gamma = (\gamma_0, \dots, \gamma_{d_2})^T$  are  $(d_1 + 1)$ - and  $(d_2 + 1)$ -dimensional vectors of unknown regression parameters, respectively. In view of the possible extension to a zero-inflated model, it may be useful to consider a semiparametric link function for  $\lambda$ . The partly linear link function is one possibility giving the joint models with

$$\log(\lambda) = \beta^T X + g(T) \quad \text{and} \quad \log\left(\frac{p}{1 - p}\right) = \gamma^T Z, \quad (3)$$

where  $T$  is an observable continuous explanatory variable,  $g$  is an unknown smooth function;  $Z = (1, X^T, T)^T$ ,  $\beta = (\beta_1, \dots, \beta_{d_1})^T$  and  $\gamma = (\gamma_0, \dots, \gamma_{d_1+1})^T$  are  $(d_1)$ - and  $(d_1 + 2)$ -dimensional vectors of unknown regression parameters. In the following we refer to this new model as a semiparametric ZIP regression model. The ZIP model specified by (1) and (2) is a special case of the class of generalized linear mixed models, while that specified by (1) and (3) belongs to the class of generalized partly linear mixed models, an extension of the class of generalized partly linear models of Muller, Ronz, and Härdle (1997). Four basic assumptions, namely, C1–C4, for the arguments in the sequel to be valid are listed in the Appendix.

Let  $\theta = (\beta, \gamma, g)^T$  be the vector of unknown parameters of interest,  $\theta_0 = (\beta_0, \gamma_0, g_0)^T$  be the true value of  $\theta$ , and  $B = \{g \in C^r[0, 1] : -\infty < m_0 \leq g(t) \leq M_0 < +\infty, \forall t \in [0, 1]\}$ , where  $r = 1$  or  $2$  that determines the smoothness of the nonlinear function  $g$ , and that the boundary values  $m_0$  and  $M_0$  are known. The parameter space of  $\theta$  is thus  $\Theta = \{\theta : \beta \in A_1, \gamma \in A_2, g \in B\} = A_1 * A_2 * B$ , where  $A_1$  and  $A_2$  are bounded closed sets in  $R^{d_1}$  and  $R^{d_1+2}$ , respectively. Let  $W = (Y, X^T, T)^T$  and the density function of  $W$  is thus given by

$$Q(w, \theta) = \left[ I_{(y=0)} \left\{ p + (1 - p)e^{-e^{\beta^T x + g(t)}} \right\} + I_{(y>0)} (1 - p) \frac{e^{-e^{\beta^T x + g(t)}} e^{y(\beta^T x + g(t))}}{y!} \right] \varphi(x, t),$$

where  $\varphi$  is the joint density function of  $(X^T, T)$  and  $I_{(\cdot)}$  is the indicator function for the specified event. Let  $\tilde{W} = (W_1^T, \dots, W_n^T)^T$ , where  $W_1, \dots, W_n$  are independent and identically distributed copies of  $W$  with distribution  $P_\theta$ . The

log-likelihood contribution of observation  $i$  is denoted by  $\ell(\theta, w_i) = \log Q(w_i, \theta)$ .

To make inference about the unknown smooth function  $g$ , we propose to use the sieve method to approximate an infinite-dimensional parameter space  $\Theta$  by a series of finite-dimensional parameter spaces  $\Theta_n$  and hence to estimate the parameter on  $\Theta_n$  but not on  $\Theta$ . Let  $0 = t_0 < t_1 < \dots < t_m = 1$  define a partition of  $[0, 1]$  where  $m$  is the number of knots to be determined. Theoretically speaking, we choose  $m$  to be an integer that grows at rate  $O(n^k)$  for  $0 < k < 1$ .

Let  $I_j(t) = I_{(t_{j-1} \leq t < t_j)}$  for  $1 \leq j \leq m - 1$  and  $I_m(t) = I_{(t_{m-1} \leq t \leq t_m)}$ . The unknown smooth function  $g$  is approximated by the piecewise linear function,

$$G_m(t; b) = \sum_{j=1}^m \left( \frac{b_j - b_{j-1}}{t_j - t_{j-1}} t - \frac{b_j t_{j-1} - b_{j-1} t_j}{t_j - t_{j-1}} \right) I_j(t),$$

with knots  $(t_0, b_0), \dots, (t_m, b_m)$ , where  $b = (b_0, \dots, b_m)^T$  is the vector of coefficients of  $G_m$ . Let  $B_n = \{G_m(t; b) : m_0 \leq b_i \leq M_0, 0 \leq i \leq m\}$ . For any  $\theta = (\beta, \gamma, g)^T \in \Theta$ , select  $b = (g(t_0), \dots, g(t_m))^T$ ,  $g_n(\cdot) = G_m(\cdot; b)$ . Let  $\pi_n \theta = (\beta, \gamma, g_n)^T \in \Theta_n$ , where  $\Theta_n = A_1 * A_2 * B_n$  is a product space. For any  $\theta_i \in \Theta$ ,  $i = 1, 2$ , define a pseudodistance  $d$  by  $d(\theta_1, \theta_2) = \|\beta_1 - \beta_2\| + \|\gamma_1 - \gamma_2\| + \|g_1 - g_2\|_2$ , and with some simple derivations, it can be shown easily that  $d(\pi_n \theta, \theta) \leq \|g - g_n\|_\infty = O(n^{-rk}) \rightarrow 0$  as  $n \rightarrow \infty$ , where  $\|a\|$  is the Euclidean norm of the vector  $a$ ,  $\|f\|_\infty = \sup_t |f(t)|$  is the supremum norm of a function  $f$ , and for  $X$  being distributed according to the probability measure  $P$ ,  $\|f(X)\|_2 = (\int f^2 dP)^{\frac{1}{2}}$  is the  $L_2(P)$  norm of a function  $f$ . Hence  $\Theta_n = A_1 * A_2 * B_n$  could be chosen as a sieve space of  $\Theta$ . Furthermore, if we select  $L_n(\theta, \tilde{W}) = \frac{1}{n} \sum_{i=1}^n \ell(\theta, w_i)$  as the empirical criterion function, then  $\hat{\theta}_n = (\hat{\beta}_n, \hat{\gamma}_n, \hat{g}_n)^T = \arg \sup_{\theta \in \Theta_n} \{L_n(\theta, \tilde{W})\}$  is the sieve MLE for  $\theta_0$  (see Grenander, 1981 and Shen and Wong, 1994).

### 3. Asymptotic Properties of Sieve MLEs

The asymptotic properties of the proposed sieve MLE are studied in this section. For simplicity, we only present the key results here. One may refer to Lam, Xue, and Cheung (2006) for the proof of the theorems that can also be downloaded from <http://web.hku.hk/~hrntlkf/zipreg.pdf>

In addition to the four assumptions C1–C4, five mild conditions, namely, A1–A5, needed to establish the asymptotic properties of  $\hat{\theta}_n$  are also listed in the Appendix.

**THEOREM 1 (strong consistency):** *Suppose assumptions C1–C4 hold, and condition A1 is satisfied, then we have  $d(\hat{\theta}_n, \theta_0) \rightarrow 0$  almost surely under  $P_{\theta_0}$ .*

**THEOREM 2 (rate of convergence):** *Suppose assumptions C1–C4 hold, and condition A1 is satisfied, then we have  $d(\hat{\theta}_n, \theta_0) = O_p(\max\{n^{-(1-k)/2}, n^{-rk}\})$ . If we select  $k = 1/(1 + 2r)$  for  $r = 1$  or  $2$ , then  $d(\hat{\theta}_n, \theta_0)$  achieves a convergence rate  $O_p(n^{-\frac{r}{1+2r}})$ . Moreover the convergence rate of the nonparametric part is  $n^{-\frac{r}{1+2r}}$ , which is  $n^{-\frac{1}{3}}$  for  $r = 1$  (i.e.,  $g$  is first-order continuously differentiable) and  $n^{-\frac{2}{3}}$  for  $r = 2$  (i.e.,  $g$  is second-order continuously differentiable) that are, respectively, the optimal*

*rates of convergence in density estimation and nonparametric regression with the usual smoothness assumptions.*

Let  $E_0$  be the expectation with respect to  $P_{\theta_0}$ . We denote  $\xi(\theta, W) = \beta^T X + g(T)$ ,  $D(\xi) = \frac{\partial \xi}{\partial \xi} \ell(\theta, w) = -I_{(y=0)} \frac{(1-p)e^{\xi} - e^{-\xi}}{p + (1-p)e^{-\xi}} + I_{(y>0)}(y - e^{\xi})$ ,  $C_1 = E_0 D^2(\xi(\theta_0, W))$ ,  $J(W) = C_1^{-1} D^2(\xi(\theta_0, W)) Q(W, \theta_0)$ , where  $J$  is a density function and we let  $E_J$  be the expectation with respect to  $J$ .

Let the score functions for  $\beta$  and  $\gamma$  be

$$\begin{aligned} \dot{\ell}_\beta(\theta, W) &= \frac{\partial \ell}{\partial \beta} = \frac{\partial \ell}{\partial \xi} \frac{\partial \xi}{\partial \beta} = D(\xi) X, \\ \dot{\ell}_\gamma(\theta, W) &= \frac{\partial \ell}{\partial \gamma} = \frac{\partial \ell}{\partial p} \frac{\partial p}{\partial \gamma} = \dot{\ell}_p \frac{Z e^{\gamma^T Z}}{(1 + e^{\gamma^T Z})^2} = \dot{\ell}_p p(1-p) Z. \end{aligned}$$

The efficient score function and Fisher information matrix of the estimator for the unknown parameter of the model can be obtained explicitly by the following theorem.

**THEOREM 3 (efficient score function and Fisher information matrix):** *In addition to the model assumptions C1–C4, suppose conditions A1 and A2 are satisfied, then the efficient score function of  $(\beta, \gamma)$  is given by*

$$\begin{pmatrix} U_\beta^* \\ U_\gamma^* \end{pmatrix} = \begin{pmatrix} D(\xi(\theta_0, W))\{X - E_J(X|T)\} \\ \dot{\ell}_\gamma(\theta_0, W) - D(\xi(\theta_0, W)) E_J \left( \frac{\dot{\ell}_\gamma(\theta_0, W)}{D(\xi(\theta_0, W))} \middle| T \right) \end{pmatrix},$$

the Fisher information matrix is

$$\begin{aligned} H(\beta_0, \gamma_0) &= E_{P_{\theta_0}} \left( (U_\beta^*, U_\gamma^*)^T (U_\beta^*, U_\gamma^*) \right) \\ &= C_1 E_J \left( \begin{pmatrix} X - E_J(X|T) \\ \frac{\dot{\ell}_\gamma(\theta_0, W)}{D(\xi(\theta_0, W))} - E_J \left( \frac{\dot{\ell}_\gamma(\theta_0, W)}{D(\xi(\theta_0, W))} \middle| T \right) \end{pmatrix} \right)^{\otimes 2}, \end{aligned}$$

and the matrix  $H$  is positive definite with its every component being bounded.

**THEOREM 4 (asymptotic normality and efficiency):** *In addition to the model assumptions C1–C4, suppose the conditions A1–A5 are satisfied, then*

$$\sqrt{n}(\hat{\beta}_n - \beta_0, \hat{\gamma}_n - \gamma_0)^T \xrightarrow{d} N(0, H^{-1}(\beta_0, \gamma_0)).$$

Hence  $(\hat{\beta}_n, \hat{\gamma}_n)^T$  is asymptotically efficient and normally distributed.

An important issue in statistical inference is to provide a reasonable variance estimate of the estimator  $(\hat{\beta}_n, \hat{\gamma}_n)$ . Two approaches are suggested in this article. The first approach uses the inverse of the semiparametric Fisher information matrix. Similar to Huang (1996), we let  $\mu_1(\theta_0, T) = E_J(X|T)$  and  $\mu_2(\theta_0, T) = E_J(\frac{\dot{\ell}_\gamma(\theta_0, W)}{D(\xi(\theta_0, W))} | T)$ . By Theorem 4 and some

reasonable estimator  $\hat{\theta}_n$  for  $\theta_0$ , we can estimate  $H(\beta_0, \gamma_0)$  by

$$\hat{H}_n(\hat{\beta}_n, \hat{\gamma}_n) = \frac{1}{n} \sum_{i=1}^n D^2(\xi(\hat{\theta}_n, W_i)) \times \left( \begin{array}{c} X_i - \mu_{1n}(\hat{\theta}_n, T_i) \\ \frac{\dot{\ell}_\gamma(\hat{\theta}_n, W_i)}{D(\xi(\hat{\theta}_n, W_i))} - \mu_{2n}(\hat{\theta}_n, T_i) \end{array} \right)^{\otimes 2}.$$

However, the estimation of  $\mu_1(\theta_0)$  and  $\mu_2(\theta_0)$  is very complicated due to the existence of the conditional expectations. One approach is to make use of the kernel method, which is quite cumbersome. If we can ignore the dependence between  $X$  and  $T$  and replace the conditional expectation  $E_J(\cdot | T)$  by  $E_J(\cdot)$ ,  $\mu_1(\theta_0)$  and  $\mu_2(\theta_0)$  can be estimated by

$$\mu_{1n}(\hat{\theta}_n) = \frac{\sum_{i=1}^n D^2(\xi(\hat{\theta}_n, W_i)) X_i}{\sum_{i=1}^n D^2(\xi(\hat{\theta}_n, W_i))} \quad \text{and} \\ \mu_{2n}(\hat{\theta}_n) = \frac{\sum_{i=1}^n D(\xi(\hat{\theta}_n, W_i)) \dot{\ell}_\gamma(\hat{\theta}_n, W_i)}{\sum_{i=1}^n D^2(\xi(\hat{\theta}_n, W_i))},$$

respectively. As remarked by Xue, Lam, and Li (2004), it is not unreasonable to assume independence of  $X$  and  $T$  in practical applications that the differences in the variance estimates between the kernel method and assumed independence approach are not alarming even when  $X$  and  $T$  are correlated. The performance of this assumed independence approach is very good in general.

Without the independence assumption, the second approach uses the inverse of the observed information matrix based on the log-likelihood function. Treating  $b$  as a finite-dimensional nuisance parameter, the log-likelihood function  $\ell(\beta, \gamma, g, w)$  in Section 2 is now approximated by  $\ell(\beta, \gamma, b, w)$  with  $g(t)$  replaced by  $G_m(t; b)$ . Then the observed joint information matrix  $I(\beta, \gamma, b_0)$  is given by

$$I(\beta, \gamma, b) = \begin{pmatrix} -\frac{\partial^2 \ell}{\partial \beta^2} & -\frac{\partial^2 \ell}{\partial \beta \partial \gamma} & -\frac{\partial^2 \ell}{\partial \beta \partial b} \\ -\frac{\partial^2 \ell}{\partial \gamma \partial \beta} & -\frac{\partial^2 \ell}{\partial \gamma^2} & -\frac{\partial^2 \ell}{\partial \gamma \partial b} \\ -\frac{\partial^2 \ell}{\partial b \partial \beta} & -\frac{\partial^2 \ell}{\partial b \partial \gamma} & -\frac{\partial^2 \ell}{\partial b^2} \end{pmatrix}.$$

The variance-covariance matrix of  $(\hat{\beta}_n, \hat{\gamma}_n, \hat{b})$  is approximated by  $I^{-1}(\hat{\beta}_n, \hat{\gamma}_n, \hat{b})/n$  and the variance-covariance matrix of  $(\hat{\beta}_n, \hat{\gamma}_n)$  can be obtained by deleting the last  $(m+1)$  rows and columns of  $I^{-1}(\hat{\beta}_n, \hat{\gamma}_n, \hat{b})/n$ . This approach is computationally more difficult to implement as it involves determining and computing the second derivatives of a high-dimensional complex function and the inversion of a possibly ill-conditioned matrix. However, this approach is more appropriate when the

independence assumption between  $X$  and  $T$  is questionable. The performance of the two approaches is compared in a simulation study and is reported in Section 5.

#### 4. An Example

The number of knots  $m$  and their locations  $t_i (i = 0, \dots, m)$  need to be determined in a practical application. They are known to depend on the smoothness of the nonparametric function  $g$  and probably the distribution of  $T$ , namely,  $\Phi$ . A large-scale simulation is carried out to investigate the effect of the choices of  $m$  and their locations with different  $g$  and  $\Phi$ . It is found that unless  $\Phi$  is highly skewed, it is generally adequate to make use of the uniform partition with  $t_i = i/m (i = 0, \dots, m)$ . Therefore we shall confine ourselves to this set of partitions in the following discussion. Of course, more turning points in  $g$  require a larger value of  $m$  for better approximation. As shown in Xue et al. (2004), the minimum acceptable value of  $m$  can be determined reasonably well by means of the Akaike information criterion (AIC) given by

$$\text{AIC} = -2nL_n(\hat{\theta}_n, \tilde{W}_n) + 2K,$$

where  $K$  is the number of parameters to be estimated. The simulation results are very similar to that reported by Xue et al. (2004) in the analysis of current status data, and hence will not be discussed in detail here. The empirical results suggest that the AIC works well in the current setup.

The zero-inflated semiparametric Poisson regression model specified by (1) and (3) is fitted to data from a public health survey conducted in Indonesia in 1997 (see Frankenberg and Thomas, 2000). A total of 16,176 subjects from 6638 randomly selected households participated in the study. The variable of interest  $Y$  is the number of days of missed primary activities due to illness in the past 4 weeks self-reported by the respondent. The explanatory variables are gender ( $X_1 = 1$  for female and 0 otherwise), per capita annual household income ( $X_2$  in thousands), household hygiene index ( $X_3$  ranges from 0 to 5, i.e., from best to worst), and age ( $T$  ranges from 18 to 60 years old). The interviewers considered five different household hygiene conditions, that is, surrounded by human and animal waste, surrounded by piles of trash, surrounded by stagnant water, a stable under/next to the house, and well maintained and cleaned-up. The household hygiene index  $X_3$  indicates the number of conditions scored negatively by the interviewers. One observation is selected at random from each household to avoid the problem of possible within-household correlation. Moreover we only include subjects of working age (18–60 years) in the analysis, leading to a sample of size  $n = 5700$  of which 5330 were zero counts.

We postulate that the variable age has a nonlinear relationship on the natural logarithm of the mean number of days of missed primary activities due to illness among subjects in the high-risk group. We categorized the subjects with structural zeros as being from the low-risk group with others being from the high-risk group. We assume that this proportion of low-risk subjects may be affected by the explanatory variables  $Z = (1, X^T, T)^T$ . Models with  $m$  from 1 to 10 with  $t_i = i/m$  for  $i = 0, \dots, m$  were fitted to the data. The standard error estimates by the two methods are very similar and we only report those obtained by the first convenient approach. The estimation results are highly consistent among all models

**Table 1**  
*Estimates for the survey data*

|                       | ( <i>m</i> = 3)  | ( <i>m</i> = 4)  | ( <i>m</i> = 6)  | ( <i>m</i> = 9)  |
|-----------------------|------------------|------------------|------------------|------------------|
| AIC                   | 955.89           | 951.33           | 957.03           | 957.60           |
| $\hat{\gamma}_0$ (SE) | 3.5985 (0.1397)  | 3.5918 (0.1398)  | 3.5952 (0.1397)  | 3.5908 (0.1398)  |
| $\hat{\gamma}_1$ (SE) | -1.7023 (0.1152) | -1.7006 (0.1153) | -1.7013 (0.1153) | -1.6992 (0.1154) |
| $\hat{\gamma}_2$ (SE) | 0.8400 (0.9675)  | 0.8443 (0.9678)  | 0.8431 (0.9674)  | 0.8424 (0.9676)  |
| $\hat{\gamma}_3$ (SE) | -0.5257 (0.2272) | -0.5230 (0.2272) | -0.5241 (0.2272) | -0.5226 (0.2272) |
| $\hat{\gamma}_4$ (SE) | -0.3837 (0.2011) | -0.3770 (0.2012) | -0.3807 (0.2011) | -0.3781 (0.2012) |
| $\hat{\beta}_1$ (SE)  | 0.3326 (0.0271)  | 0.3377 (0.0271)  | 0.3360 (0.0271)  | 0.3477 (0.0272)  |
| $\hat{\beta}_2$ (SE)  | -0.2941 (0.2834) | -0.2839 (0.2788) | -0.2784 (0.2788) | -0.2972 (0.2783) |
| $\hat{\beta}_3$ (SE)  | 0.2600 (0.0446)  | 0.2826 (0.0450)  | 0.2745 (0.0447)  | 0.2850 (0.0449)  |
| $\hat{b}_0$           | 0.9262           | 0.7843           | 0.8501           | 0.9953           |
| $\hat{b}_1$           | 1.2329           | 1.2792           | 1.1148           | 0.7631           |
| $\hat{b}_2$           | 1.1665           | 1.0798           | 1.2560           | 1.3213           |
| $\hat{b}_3$           | 1.2998           | 1.2830           | 1.0922           | 1.2172           |
| $\hat{b}_4$           |                  | 1.2274           | 1.2234           | 1.0905           |
| $\hat{b}_5$           |                  |                  | 1.2466           | 1.0739           |
| $\hat{b}_6$           |                  |                  | 1.2671           | 1.2325           |
| $\hat{b}_7$           |                  |                  |                  | 1.2463           |
| $\hat{b}_8$           |                  |                  |                  | 1.2735           |
| $\hat{b}_9$           |                  |                  |                  | 1.2041           |

fitted, and the results for  $m = 3, 4, 6,$  and  $9$  are summarized in Table 1. According to the AIC, a model with  $m = 4$  or more is suggested.

The results from Table 1 suggest that males are significantly less likely to belong to the high-risk group at the 5% level according to the asymptotic normally distributed properties of the estimator  $\hat{\gamma}$ . Moreover, males in the high-risk group are found to have a significantly smaller mean number of days of missed primary activities than the females in the high-risk group. A better hygiene condition significantly increases the probability that the subject is at low risk. Better hygiene condition also significantly reduces members' mean number of days of missed primary activities even though they are from the high-risk group. Annual household income has no effect on either of the two processes. There is a tendency that younger subjects are less likely to belong to the high-risk group. This is significant at the 0.05 level of significance when using a one-tail test, but not when using a two-tail test. Moreover, from the estimate of  $b$ , it is easy to see that subjects aged between 18 and 25 years from the high-risk group have smaller mean  $\lambda$ , but the age effect has been very stable among those aged 25 years or older, showing that young subjects in the high-risk group resumed normal activities at a faster rate even if they are ill, but the illness duration was more or less the same among the subjects aged between 25 and 60 years.

It is obvious from the above analysis that the estimator for the nonparametric component is sensitive enough to capture the pattern of the possibly nonlinear relationship. The method is quite useful in this context as it can also provide some information on the cutoff values if it is desired to categorize the individuals into two or more age groups to provide with different treatments when making up certain health policies for the general population. Moreover, policies can be made more cost effective as a tailor-made policy can be targeted

to a particular age group, but not to the general population because individuals from different age groups may have different levels of vulnerability and exposure to different behavioral factors. Taking this survey as an example, a fully parametric ZIP regression model has also been fitted to the data and as expected, the variable age has no effect whatsoever on the log of the mean among subjects in the high-risk group. However if only subjects aged 18–25 years are included in the fully parametric analysis, the mean number of days of missed primary activities significantly increases with age among subjects in the high-risk group. This increasing trend is also detected in the analysis of the full data set using the proposed semiparametric ZIP regression model. It is important to suppress this increasing trend in  $g(t)$  as this has a lifelong effect with  $g(t)$  retaining at a high level for  $t > 25$ . One reason for this increasing trend from a sociological perspective is that young subjects are more tempted to establish unhealthy habits, such as smoking, when they first join the workforce (Trinidad et al., 2004). If this is the case, proper health education targeted to subjects in this age group will hopefully pull down  $g(t)$  to a much lower level for all  $t$  and not just for  $t \leq 25$ . An important consequence is that the financial burden and the loss in productivity due to the disease can be substantially reduced.

**5. Simulation**

A simulation study is carried out to study the performance of the proposed sieve MLE. To better connect with the data example, the design of the simulation study mimics the data example in the last section. Recall that the explanatory variables are gender ( $X_1$ ), per capita annual household income ( $X_2$  in thousands), household hygiene index ( $X_3$  ranges from 0 to 5 from best to worst), and age ( $T$  ranges from 18 to 60 years old). For simplicity we assume that

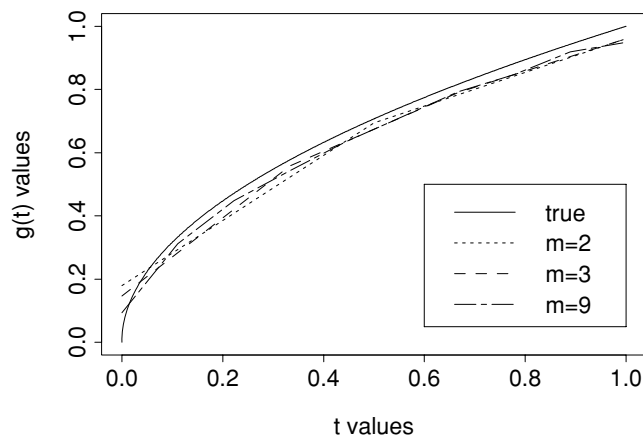
**Table 2**  
Simulation results

|                                      | ( $m = 2$ )      | ( $m = 3$ )      | ( $m = 6$ )      | ( $m = 9$ )      |
|--------------------------------------|------------------|------------------|------------------|------------------|
| AIC                                  | 405.055          | 405.375          | 408.17           | 411.195          |
| $\hat{\gamma}_0$ (ESD)               | 3.5014 (0.1665)  | 3.5002 (0.1667)  | 3.5005 (0.1665)  | 3.4996 (0.1668)  |
| (SE <sub>1</sub> , SE <sub>2</sub> ) | (0.1687, 0.1684) | (0.1687, 0.1684) | (0.1686, 0.1684) | (0.1687, 0.1684) |
| $\hat{\gamma}_1$ (ESD)               | -2.0027 (0.1069) | -2.0030 (0.1070) | -2.0037 (0.1071) | -2.0034 (0.1072) |
| (SE <sub>1</sub> , SE <sub>2</sub> ) | (0.1103, 0.1097) | (0.1103, 0.1097) | (0.1102, 0.1097) | (0.1102, 0.1097) |
| $\hat{\gamma}_2$ (ESD)               | 0.0021 (0.1467)  | 0.0020 (0.1467)  | 0.0018 (0.1468)  | 0.0019 (0.1467)  |
| (SE <sub>1</sub> , SE <sub>2</sub> ) | (0.1447, 0.1443) | (0.1447, 0.1443) | (0.1447, 0.1443) | (0.1447, 0.1443) |
| $\hat{\gamma}_3$ (ESD)               | -0.4972 (0.1484) | -0.4792 (0.1484) | -0.4975 (0.1484) | -0.4973 (0.1486) |
| (SE <sub>1</sub> , SE <sub>2</sub> ) | (0.1456, 0.1452) | (0.1456, 0.1452) | (0.1456, 0.1452) | (0.1456, 0.1452) |
| $\hat{\gamma}_4$ (ESD)               | -0.4993 (0.1445) | -0.4970 (0.1448) | -0.4961 (0.1448) | -0.4957 (0.1448) |
| (SE <sub>1</sub> , SE <sub>2</sub> ) | (0.1462, 0.1462) | (0.1462, 0.1464) | (0.1462, 0.1466) | (0.1462, 0.1465) |
| $\hat{\beta}_1$ (ESD)                | 0.5264 (0.0554)  | 0.5249 (0.0556)  | 0.5214 (0.0564)  | 0.5225 (0.0558)  |
| (SE <sub>1</sub> , SE <sub>2</sub> ) | (0.0663, 0.0655) | (0.0663, 0.0655) | (0.0664, 0.0654) | (0.0665, 0.0656) |
| $\hat{\beta}_2$ (ESD)                | 0.0104 (0.0573)  | 0.0098 (0.0578)  | 0.0079 (0.0581)  | 0.0082 (0.0584)  |
| (SE <sub>1</sub> , SE <sub>2</sub> ) | (0.0632, 0.0625) | (0.0632, 0.0626) | (0.0634, 0.0627) | (0.0635, 0.0628) |
| $\hat{\beta}_3$ (ESD)                | 0.5081 (0.0604)  | 0.5070 (0.0603)  | 0.5048 (0.0596)  | 0.5054 (0.0599)  |
| (SE <sub>1</sub> , SE <sub>2</sub> ) | (0.0649, 0.0642) | (0.0649, 0.0642) | (0.0650, 0.0643) | (0.0652, 0.0645) |

the explanatory variables are independently distributed. Let  $U_{ij}$  ( $i = 1, \dots, n$ ;  $j = 1, \dots, 4$ ) be randomly generated independently according to the uniform distribution on  $[0, 1]$ . Then  $X_{i1}$  takes up the value 1 (female) if  $U_{i1}$  exceeds 0.5 (by assuming that the ratio of male and female is the same) or the value 0 otherwise.  $X_{i3}$  will take up the integral part of  $U_{i3} \times 6$  to represent the hygiene index;  $X_{i2}$  and  $T$  are generated according to some suitable transformation of  $U_{i2}$  and  $U_{i4}$  using the respective minimum values and the ranges observed in the data. For example,  $T_i = 18 + (60 - 18) * U_{i4}$ . The nonparametric function  $g$  is chosen to be  $g(t) = t^{1/2}$  to mimic the real situations for simplicity, and the regression parameters are set at  $\gamma_0 = (3.50, -2.00, 0.00, -0.50, -0.50)^T$  and  $\beta_0 = (0.50, 0.00, 0.50)^T$  such that they are not too far away from their respective estimates. The probability that the subject is from the high-risk group falls in the range from 0.62 to 0.97. In this study, 500 samples, each of size  $n = 5000$ , are generated. As in the last section, we choose the location of the knots at  $t_i = i/m$  for  $i = 0, \dots, m$ . A pilot simulation study indicates that 2 and 3 are the modal choices of  $m$  determined by the AIC, which agrees with that of Xue et al. (2004). For illustration purpose, we simply fix the values of  $m$  at 2, 3, 6, and 9 in the simulation in order to determine the effects of choosing the number of knots  $m$  greater than that determined by the AIC.

We are mainly interested in the behavior of the estimator for the parametric component, namely,  $(\gamma, \beta)$ . Therefore the mean of the AICs, the mean of the estimates  $(\hat{\theta})$ , the empirical standard deviation (ESD( $\hat{\theta}$ )), the mean of the estimated standard errors (SE<sub>1</sub>( $\hat{\theta}$ )) using the first convenient approach, and the mean of the estimated standard errors based on the observed log-likelihood method (SE<sub>2</sub>( $\hat{\theta}$ )) of the 500 estimates of  $\theta$  are recorded and summarized in Table 2. It is obvious that the estimator works extremely well and the means of the estimates are very close to their respective true values. Moreover the estimated standard errors using the two approaches are virtually indifferent, and they closely resemble their respec-

tive empirical standard deviations. The mean estimates of the  $b_j$ 's for  $m = 2, 3$ , and 9, together with the true  $g(t)$ , are plotted in Figure 1. Note that the estimates of  $b_j$ 's for  $m = 6$  have been omitted because it overlaps substantially with that of  $m = 9$ . It is obvious that the estimator for the nonparametric part is able to capture the shape of  $g(t)$  reasonably well. Xue et al. (2004) reported a small bias at the boundary values of  $T = 0$  and 1, which decreases with increasing values of  $m$  using a similar approximation to the unknown function  $g$  in the analysis of current status data. This phenomenon is not so severe in the current situation and the estimates are quite close to the true value at the  $m$  knots in general, but slight underestimation of  $g(t)$  is observed, but is very mild in general. From the above simulation studies, the performance of the proposed estimation method is highly satisfactory in practice, and the AIC-type criterion provides a decent guideline to the choice of  $m$ .



**Figure 1.** Plot of the estimated  $g(t)$  with various values of  $m$ .

## 6. Concluding Remarks

This article considers the semiparametric analysis of zero-inflated count data that combines a logistic regression formulation for the probability of being in the low-risk group leading to structural zeros and the Poisson counts from the high-risk group according to a semiparametric regression model. The ZIP regression model specified in (1) and (3) is particularly useful when the effects of the covariate  $T$  do not act linearly on the link function of the mean of the response variable  $Y$ . Models with different numbers of knots should be fitted to the data to ensure that the estimates are stable before making a premature decision. The AIC is generally sensible enough to determine an adequate model to provide good approximation to the nonlinear function  $g$ . Simulation studies not reported here indicate that the locations of the knots do not have much of an effect on the estimates and that, with a fixed  $m$ , a partition with  $t_i = i/m$  ( $i = 0, \dots, m$ ) is a good candidate for practical use.

The performance of the proposed estimator has been shown to be highly satisfactory, and the estimator for the nonlinear function  $g$  is sensitive enough to capture the true shape or pattern of  $g$ . When the sample size is large, it is conjectured that the proposed model can further be extended to incorporate another unknown smooth function, say  $g_1(t)$ , to the logit of  $p$  so that

$$\log\left(\frac{p}{1-p}\right) = \gamma^T z + g_1(t).$$

The estimation can be made possible with a slight modification of the proposed sieve MLE and the theorems developed in this article may also be valid with slight modifications. However, the performance of the estimator is yet to be explored as the estimates may not be very stable numerically.

Clustered zero-inflated count data often arise in general public health household survey. For future research, one may consider a simple extension of the proposed semiparametric model to accommodate clustered zero-inflated count data by means of a random effects approach. The estimation will be straightforward but the properties of the estimator need to be studied.

## ACKNOWLEDGEMENTS

We would like to thank the editor, an associate editor, and two referees for their constructive comments that substantially improved the contents and presentation of the article. The work described in this article was substantially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (project HKU 7425/03M) and a grant from the National Natural Science Foundation of China (project 10301035).

## REFERENCES

Arola, H., Pitkanen, M., Nygard, C. H., Huhtala, H., and Manka, M. L. (2003). The connection between age, job control and sickness absences among Finnish food workers. *Occupational Medicine* **53**, 229–230.

Bohning, D., Dietz, E., Schlattmann, P., Mendonca, L., and Kirchner, U. (1999). The zero-inflated Poisson model and

the decayed, missing and filled teeth index in dental epidemiology. *Journal of the Royal Statistical Society, Series A* **162**, Part 2, 195–209.

Cant, R., O'Loughlin, K., and Legge, V. (2001). Sick leave—Cushion or entitlement? A study of age cohorts' attitudes and practices in two Australian workplaces. *Work* **17**, 39–48.

Cheung, Y. B. (2002). Zero-inflated models for regression analysis of count data: A study of growth and development. *Statistics in Medicine* **21**, 1461–1469.

Frankenberg, E. and Thomas, D. (2000). *The Indonesian Family Life Survey: Study Design and Results from Waves 1 and 2*. Santa Monica, California: RAND.

Grenander, U. (1981). *Abstract Inference*. New York: Wiley.

Huang, J. (1996). Efficient estimation for the proportional hazards model with interval censoring. *Annals of Statistics* **24**, 540–568.

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

Lam, K. F., Xue, H., and Cheung, Y. B. (2006). *Semiparametric analysis of zero-inflated count data*. Research Report, Volume 424, Department of Statistics and Actuarial Science, The University of Hong Kong.

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

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics* **33**, 341–365.

Muller, M., Ronz, B., and Härdle, W. (1997). Computer-assisted semiparametric generalized linear models. *Computational Statistics* **12**, 153–172.

Shankar, V., Milton, J., and Mannering, F. (1997). Modeling accident frequencies as zero-altered probability processes: An empirical inquiry. *Accident Analysis and Prevention* **29**, 829–837.

Shen, X. T. and Wong, W. H. (1994). Convergence rate of sieve estimates. *Annals of Statistics* **22**, 580–615.

Shephard, R. J. (1999). Age and physical work capacity. *Experimental Aging Research* **25**, 331–343.

Street, A., Jones, A., and Furuta, A. (1999). Cost-sharing and pharmaceutical utilisation and expenditure in Russia. *Journal of Health Economics* **18**, 459–472.

Trinidad, D. R., Gilpin, E. A., Lee, L., and Pierce, J. P. (2004). Has there been a delay in the age of regular smoking onset among African Americans? *Annals of Behavioral Medicine* **28**, 152–157.

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

Xue, H., Lam, K. F., and Li, G. (2004). Sieve maximum likelihood estimator for semiparametric regression models with current status data. *Journal of the American Statistical Association* **99**, 346–356.

Received October 2004. Revised January 2006.

Accepted January 2006.

## APPENDIX

The assumptions C1–C4 needed for the arguments in Sections 2 and 3 to be valid are listed below:

- C1.  $\beta \in A_1$ ,  $\gamma \in A_2$ , where  $A_1$  and  $A_2$  are bounded closed sets in  $R^{d_1}$  and  $R^{d_1+2}$ , respectively;  $X$  is bounded;  $T \in [0, 1]$ .
- C2.  $g \in C^r[0, 1]$  for  $r = 1$  or  $2$ .
- C3. We denote the joint density of  $(X^T, T)$  by  $\varphi(x, t)$  and the marginal distribution function of  $T$  by  $\Phi$ . Moreover,  $\varphi$  and  $\Phi$  do not depend on  $(\beta, \gamma, g)$ .
- C4.  $\max_{1 \leq j \leq m} \{\Phi(t_j) - \Phi(t_{j-1})\} \leq Cn^{-k}$  for some constant  $C$  and  $0 < k < 1$ .

The conditions A1–A5 needed for the theorems in Section 3 to be valid are listed below:

- A1.  $E(X - E(X|T))^{\otimes 2} > 0$ .
- A2.  $\theta_0$  is an interior point of  $\Theta$ . In other words,  $(\beta_0, \gamma_0)$  is an interior point of  $A_1 * A_2$ , and  $m_0 < g_0 < M_0$ .
- A3. The distribution function  $\Phi$  is known and the joint density function  $\varphi(x, t)$  (see C3) is second-order continuously differentiable in  $t$  with a bounded derivative.
- A4. The partition of  $[0, 1]$  is restricted so that  $\min_j \{\Phi(t_j) - \Phi(t_{j-1})\} = O(n^{-k'})$  with  $k \leq k' < \frac{1-k}{2}$  for  $\frac{1}{5} < k < \frac{1}{3}$ , or  $k \leq k' < 2k$  for  $\frac{1}{8} < k \leq \frac{1}{5}$ , where  $\Phi$  is the distribution function of  $T$  as defined in C3.
- A5. The unknown function  $g$  is a smooth function and at least second-order continuously differentiable.