Statistics
in Medicine

Research Article

Received 24 May 2014, Accepted 11 November 2014 Published online 25 November 2014 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6382

Generalized partially linear single-index model for zero-inflated count data

Xiaoguang Wang,[a] Jun Zhang,[b] Liang Yu[a] and Guosheng Yin[c*†]

Count data often arise in biomedical studies, while there could be a special feature with excessive zeros in the observed counts. The zero-inflated Poisson model provides a natural approach to accounting for the excessive zero counts. In the semiparametric framework, we propose a generalized partially linear single-index model for the mean of the Poisson component, the probability of zero, or both. We develop the estimation and inference procedure via a profile maximum likelihood method. Under some mild conditions, we establish the asymptotic properties of the profile likelihood estimators. The finite sample performance of the proposed method is demonstrated by simulation studies, and the new model is illustrated with a medical care dataset. Copyright © 2014 John Wiley & Sons, Ltd.

Keywords: asymptotic normality; B-spline; generalized partially linear model; single-index model; zero-inflated count data

1. Introduction

Count data often arise in many scientific fields, such as medicine, economics, ecology, engineering, and sociology. Poisson regression models are the most commonly used approach to analyzing count data. However, such data often depart from the Poisson distribution because of the existence of some extreme observations that lead to more spread (a larger variance) than the mean, which is known as overdispersion. Often, this issue can be addressed by using sandwich variance estimation or incorporating an additional dispersion parameter under the Poisson regression model. A plausible alternative is to assume a negative binomial distribution for overdispersed count data, because the negative binomial distribution has a longer and heavier tail to accommodate the feature that the variance may exceed the mean. All of these models belong to the family of generalized linear models (GLMs); for a more comprehensive coverage of GLMs, see Nelder and Wedderburn [1].

In practice, the empirical count data often exhibit an excess number of zeros, which is greater than that would be expected from standard count distributions. Although overdispersion can be controlled to a certain extent, the usual GLMs are typically not sufficient for modeling a large number of zeros. To deal with count data with excessive zeros, the zero-inflated Poisson (ZIP) regression explicitly models the probability of zero counts [2, 3]: One component of the model captures the count of excessive zeros, and the other corresponds to nonzero counts based on the Poisson distribution. In particular, the hurdle model combines a left-truncated count part and a right-censored hurdle, so that all the zeros are assumed to be from the structural source; that is, there is no sampling zeros [2]. On the other hand, the zero-inflation model takes a slightly different approach, which is a mixture that combines a count component and a point mass at zero [3]. As a result, zero observations may come from two different origins: Sampling zeros are due to the usual Poisson (or negative binomial) distribution, while structural zeros are observed because of some specific structure in the data. Deng and Paul [4, 5] proposed score tests for both zero-inflation and

[a] *School of Mathematical Sciences, Dalian University of Technology, Dalian, Liaoning 116024, China*
[b] *Shen Zhen-Hong Kong Joint Research Center for Applied Statistical Sciences, College of Mathematics and Computational Science, Institute of Statistical Sciences, Shenzhen University, Shenzhen, China*
[c] *Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong*
**Correspondence to: Guosheng Yin, Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong*
† *E-mail: gyin@hku.hk*

overdispersion in the GLMs. A comprehensive overview of count data models in econometrics, including hurdle and zero-inflated models, is provided by Cameron and Trivedi [6].

To truly capture the underlying relationship between the response and covariates, various semiparametric regression models have been proposed to relax restrictive modeling assumptions. Lam *et al*. [7] extended the ZIP model to the semiparametric framework by replacing the linear regression with partially linear regression for the mean of the Poisson distribution, meanwhile maintaining linear regression in modeling the probability of zero. He *et al*. [8] considered partially linear regression functions for both the mean of Poisson distribution and the probability of zero, and they proposed a sieve maximum likelihood estimator (MLE) for the doubly semiparametric ZIP model.

Single-index models represent a broad class of semiparametric regression models, which reduce the dimensionality of multivariate predictors $\mathbf{X}$ to a univariate index $\mathbf{X}^T\boldsymbol{\alpha}$. The linear combination $\mathbf{X}^T\boldsymbol{\alpha}$ is cast in a nonparametric function $g(\mathbf{X}^T\boldsymbol{\alpha})$, where $g(\cdot)$ is an unknown and unspecified function [9–12]. In the GLM framework, Carroll *et al*. [13] extended the single-index model to the generalized partially linear single-index model (GPLSIM):

$$E(Y|\mathbf{X},\mathbf{W}) = \eta\left\{g(\mathbf{X}^T\boldsymbol{\alpha}) + \mathbf{W}^T\boldsymbol{\gamma}\right\},$$

where $\eta(\cdot)$ is a link function, and $\boldsymbol{\alpha} \in \mathbb{R}^{d_1}$ and $\boldsymbol{\gamma} \in \mathbb{R}^{d_2}$ are the vectors of unknown parameters. Without loss of generality, we assume that the first component of $\boldsymbol{\alpha}$ is positive and the parameter space of $\boldsymbol{\alpha}$ is $\left\{\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{d_1})^T, \|\boldsymbol{\alpha}\| = 1, \alpha_1 > 0\right\}$ for ensuring identifiability, where $\|\cdot\|$ denotes the Euclidean norm. The parameter space $\left\{(\alpha_1, \ldots, \alpha_{d_1})^T : \|\boldsymbol{\alpha}\| = 1, \alpha_1 > 0, \boldsymbol{\alpha} \in \mathbb{R}^{d_1}\right\}$ means that $\boldsymbol{\alpha}$ is on the boundary of a unit ball, which leads to a non-regular problem. To circumvent the difficulty, we first choose an identifiable parametrization that transforms the boundary of a unit ball in $\mathbb{R}^{d_1}$ to the interior of the unit ball in $\mathbb{R}^{d_1-1}$. By eliminating α_1, the parameter space for $\boldsymbol{\alpha}$ can be reconstructed as $\left\{\left(\left(1-\sum_{j=2}^{d_1}\alpha_j^2\right)^{1/2}, \alpha_2, \ldots, \alpha_{d_1}\right)^T : \sum_{j=2}^{d_1}\alpha_j^2 < 1\right\}$, and our estimation is based on the 'leave-one-out' parameter $\boldsymbol{\alpha}^{(1)} = (\alpha_2, \ldots, \alpha_{d_1})^T$. Here, we assume that the parameter space is not the entire real space in order to ensure that $\boldsymbol{\alpha}$ in model (1) can be uniquely defined, which is a commonly used assumption on the index parameter [13]. Model (1) unifies the traditional GLM and the single-index model, and thus renders more flexibility. When $d_1 = 1$, we obtain a generalized partially linear model [14]; and when $\boldsymbol{\alpha} = \mathbf{0}$, model (1) reduces to the generalized single-index model [12, 15]. Yu and Ruppert [16] developed a penalized spline estimation procedure for the partially linear single-index model (PLSIM). Xia and Härdle [17] proposed an estimator based on the minimum average conditional variance estimation. More recently, Liang *et al*. [18] proposed a profile least squares estimation method for the PLSIM.

In consideration of a balance between model flexibility and parsimony, we study the GPLSIM via a profile likelihood estimation procedure for zero-inflated count data. We apply the B-spline method to approximate the nonparametric single-index function $g(\cdot)$ and obtain its estimate via the nonparametric maximum likelihood method assuming that the parametric part is known. We then use a profile maximum likelihood method to estimate $(\boldsymbol{\alpha}, \boldsymbol{\gamma})$ while fixing $g(\cdot)$ at the estimated function.

The rest of the paper is arranged as follows. In Section 2, we propose the partially linear single-index ZIP model, and also, we develop the estimation procedure and the asymptotic properties of the estimators. Section 3 presents simulation studies to examine the finite sample performance of the proposed method. A real data example is used to illustrate the computational simplicity of our method in Section 4. We conclude the paper with a brief discussion in Section 5, and all the technical proofs of the asymptotic results are deferred to the Appendix.

2. Model estimation and asymptotics

2.1. Partially linear single-index ZIP model

Let Y denote the event count, and, to account for excessive zeros, we assume Y from a ZIP distribution. One part of the model occurs with probability p, producing structural zeros; and the other part occurs with probability $1-p$, leading to a standard Poisson count with mean λ. Hence, the zero-inflated model can be viewed as a mixture distribution of a Poisson and a binomial distribution, whose probability mass function is given by

$$\Pr(Y=y)=\begin{cases} p+(1-p)e^{-\lambda}, & y=0,\\ (1-p)\dfrac{e^{-\lambda}\lambda^{y}}{y!}, & y=1,2,\ldots, \end{cases} \tag{1}$$

where $0 < p < 1$ and $\lambda > 0$. The ZIP distribution in (1) reduces to a regular Poisson distribution when $p = 0$ and to a degenerate point mass at 0 when $p = 1$. In regression settings, both λ and p can accommodate covariates

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

where $\mathbf{X}$ and $\mathbf{Z}$ may share common covariates.

We extend model (2) to the partially linear single-index ZIP model by incorporating an unknown smooth function $g(\cdot)$:

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

where $\boldsymbol{\alpha}, \boldsymbol{\gamma}$, and $\boldsymbol{\beta}$ are unknown parameters, and the two sets of regressors $(\mathbf{X}, \mathbf{W})$ and $\mathbf{Z}$ may share common components. In contrast to (3), we can also link p via a PLSIM, while leaving λ in a usual linear model:

$$\log(\lambda)=\mathbf{Z}^T\boldsymbol{\beta} \quad \text{and} \quad \log\left(\frac{p}{1-p}\right)=g(\mathbf{X}^T\boldsymbol{\alpha})+\mathbf{W}^T\boldsymbol{\gamma}.$$

More generally, both λ and p can be associated with PLSIMs:

$$\log(\lambda)=g_\lambda\left(\mathbf{X}_1^T\boldsymbol{\alpha}_1\right)+\mathbf{W}_1^T\boldsymbol{\gamma}_1 \quad \text{and} \quad \log\left(\frac{p}{1-p}\right)=g_p\left(\mathbf{X}_2^T\boldsymbol{\alpha}_2\right)+\mathbf{W}_2^T\boldsymbol{\gamma}_2,$$

where $g_\lambda(\cdot)$ and $g_p(\cdot)$ are unknown and unspecified functions, and $(\mathbf{X}_1, \mathbf{W}_1)$ and $(\mathbf{X}_2, \mathbf{W}_2)$ are covariates.

2.2. *Estimation*

Without loss of generality, we focus on model (3):

$$\log(\lambda_i)=g\left(\mathbf{X}_i^T\boldsymbol{\alpha}\right)+\mathbf{W}_i^T\boldsymbol{\gamma} \quad \text{and} \quad \log\left(\frac{p_i}{1-p_i}\right)=\mathbf{Z}_i^T\boldsymbol{\beta}. \tag{4}$$

where $\{y_i, \mathbf{X}_i, \mathbf{W}_i, \mathbf{Z}_i; i = 1, \ldots, n\}$ is an independent and identically distributed sample from $(Y, \mathbf{X}, \mathbf{W}, \mathbf{Z})$. The estimation procedure can be easily adapted to other single-index models. Denote $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\beta})$ and $\boldsymbol{\alpha}^{(1)} = (\alpha_2, \cdots, \alpha_{d_1})^T$. Based on the observed data, the log-likelihood function can be expressed as

$$L(\boldsymbol{\theta})=\sum_{i=1}^{n}\left[I_{(y_i=0)}\log\left\{p_i+(1-p_i)e^{-\lambda_i}\right\}+I_{(y_i>0)}\left\{\log(1-p_i)-\lambda_i+y_i\log(\lambda_i)\right\}\right].$$

The derivatives of a function with respect to $\boldsymbol{\alpha}^{(1)}$ can be readily obtained by the chain rule. Let $\mathbf{J} = \partial\boldsymbol{\alpha}/\partial\boldsymbol{\alpha}^{(1)}$ be the Jacobian matrix of size $d_1 \times (d_1 - 1)$:

$$\mathbf{J}=\begin{pmatrix} -\boldsymbol{\alpha}^{(1)T}/\sqrt{1-\|\boldsymbol{\alpha}^{(1)}\|^2} \\ \mathbf{I}_{d_1-1} \end{pmatrix},$$

where $\mathbf{I}_{d_1-1}$ is the identity matrix of size $d_1 - 1$. Let $\mathbf{B}(t) = (B_1(t), \ldots, B_{k_n+v+1}(t))^T$ be a set of B-spline basis functions of order $v + 1$ with k_n quasi-uniform internal knots. We approximate the single-index function $g(\cdot)$ via B-spline basis functions, $g(t) \approx \sum_{s=1}^{k_n+v+1} B_s(t)a_s$, where $\mathbf{a} = (a_1, \ldots, a_{k_n+v+1})^T$ is the vector of spline coefficients [19].

To estimate both the parametric and nonparametric components of the model, we employ a two-step procedure. First, given $\boldsymbol{\theta}$, we use the B-spline to approximate the nonparametric function $g(\cdot)$, which leads to the nonparametric log-likelihood function:

$$\begin{aligned}\log\{L(\mathbf{a})\} = \sum_{i=1}^{n}\Bigg[& I_{(y_i=0)}\log\left\{\frac{\exp\left(\mathbf{Z}_i^T\boldsymbol{\beta}\right)}{1+\exp\left(\mathbf{Z}_i^T\boldsymbol{\beta}\right)} + \frac{\exp\left(-\exp\left(\mathbf{B}\left(\mathbf{X}_i^T\boldsymbol{\alpha}\right)^T\mathbf{a}+\mathbf{W}_i^T\boldsymbol{\gamma}\right)\right)}{1+\exp\left(\mathbf{Z}_i^T\boldsymbol{\beta}\right)}\right\} \\ & + I_{(y_i>0)}\left\{\log\left(\frac{1}{1+\exp\left(\mathbf{Z}_i^T\boldsymbol{\beta}\right)}\right) - \exp\left(\mathbf{B}\left(\mathbf{X}_i^T\boldsymbol{\alpha}\right)^T\mathbf{a}+\mathbf{W}_i^T\boldsymbol{\gamma}\right)\right. \\ & \left. + y_i\left(\mathbf{B}\left(\mathbf{X}_i^T\boldsymbol{\alpha}\right)^T\mathbf{a}+\mathbf{W}_i^T\boldsymbol{\gamma}\right)\right\}\Bigg].\end{aligned} \tag{5}$$

Let $\hat{\mathbf{a}}$ denote the maximizer of (5), and thus, we can estimate $g(\cdot)$ with $\hat{g}(t) = \mathbf{B}(t)^T\hat{\mathbf{a}}$. Next, we maximize the profile log-likelihood by plugging in $\hat{g}(\cdot)$:

$$\begin{aligned}\log\{L(\boldsymbol{\theta})\} = \sum_{i=1}^{n}\Bigg[& I_{(y_i=0)}\log\left\{\frac{\exp\left(\mathbf{Z}_i^T\boldsymbol{\beta}\right)}{1+\exp\left(\mathbf{Z}_i^T\boldsymbol{\beta}\right)} + \frac{\exp\left(-\exp\left(\hat{g}\left(\mathbf{X}_i^T\boldsymbol{\alpha}\right)+\mathbf{W}_i^T\boldsymbol{\gamma}\right)\right)}{1+\exp\left(\mathbf{Z}_i^T\boldsymbol{\beta}\right)}\right\} \\ & + I_{(y_i>0)}\left\{\log\left(\frac{1}{1+\exp\left(\mathbf{Z}_i^T\boldsymbol{\beta}\right)}\right) - \exp\left(\hat{g}\left(\mathbf{X}_i^T\boldsymbol{\alpha}\right)+\mathbf{W}_i^T\boldsymbol{\gamma}\right)\right. \\ & \left. + y_i\left(\hat{g}\left(\mathbf{X}_i^T\boldsymbol{\alpha}\right)+\mathbf{W}_i^T\boldsymbol{\gamma}\right)\right\}\Bigg],\end{aligned} \tag{6}$$

to obtain the estimator $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}})$. As a result, the estimation procedure iteratively updates the estimate of the nonparametric component $\hat{g}(\cdot)$ (or namely, $\hat{\mathbf{a}}$) by the fixed-point algorithm by Cui *et al*. [20] and those of the parametric components $\hat{\boldsymbol{\theta}}$ by the Newton–Raphson algorithm. Specifically, our iterative algorithm proceeds as follows.

Step 1. Fit the usual linear ZIP model to the data, and take the MLE of $\boldsymbol{\theta}$ as the initial value $\hat{\boldsymbol{\theta}}_{[0]}$. Because of the constraint on $\boldsymbol{\alpha}$, the initial value of $\boldsymbol{\alpha}$ is renormalized as $\hat{\boldsymbol{\alpha}}_{[0]}/\|\hat{\boldsymbol{\alpha}}_{[0]}\|$.
Step 2. At the jth iteration, plug $\hat{\boldsymbol{\theta}}_{[j]}$ into the nonparametric log-likelihood (5), and obtain its maximizer $\hat{\mathbf{a}}_{[j]}$.
Step 3. Plug $\hat{\mathbf{a}}_{[j]}$ into the profile log-likelihood (6), and obtain the updated MLE $\hat{\boldsymbol{\theta}}_{[j+1]}$ under the constraint $\|\hat{\boldsymbol{\alpha}}\| = 1$.
Step 4. Iterate steps 2 and 3 until a predetermined convergence criterion is met.

2.3. Asymptotic properties

We denote the true parameter vector as $\boldsymbol{\theta}_0 = (\boldsymbol{\alpha}_0, \boldsymbol{\gamma}_0, \boldsymbol{\beta}_0)$ and $\boldsymbol{\theta}_0^{(1)} = \left(\boldsymbol{\alpha}_0^{(1)}, \boldsymbol{\gamma}_0, \boldsymbol{\beta}_0\right)$, and also denote $\Lambda = \mathbf{X}^T\boldsymbol{\alpha}, \mathbf{m}_x(\Lambda) = E(\mathbf{X}|\Lambda), \mathbf{m}_w(\Lambda) = E(\mathbf{W}|\Lambda)$, and $\mathbf{m}_z(\Lambda) = E(\mathbf{Z}|\Lambda)$. To study the large-sample properties of the proposed estimators, we first present some regularity conditions:

C1. Covariates $\mathbf{X}, \mathbf{W}$, and $\mathbf{Z}$ have bounded supports.
C2. The true parameter $\boldsymbol{\theta}_0^{(1)}$ is an interior point of its parameter space.
C3. The ℓth-order derivative of $g(\cdot)$ is bounded for $\ell \geqslant 2$, and it is not a constant on the support of $\mathbf{X}^T\boldsymbol{\alpha}$.
C4. All of the conditional expectations $\mathbf{m}_x, \mathbf{m}_w$, and $\mathbf{m}_z$ with respect to Λ are ℓth-order continuous.

These conditions are commonly used for PLSIMs. In particular, condition C3 ensures the approximation of B-spline to $g(\cdot)$.

Theorem 1
Under the regularity conditions C1–C3, $\hat{\boldsymbol{\theta}}$ converges in probability to the true parameter $\boldsymbol{\theta}_0$.

The proof is straightforward by using Theorem 5.1 in the work of Ichimura [12] and thus omitted here. To ensure the consistency, it is commonly assumed that the likelihood function has a unique maximum. Hereafter, we denote $\mathbf{A}^{\otimes 2} = \mathbf{A}\mathbf{A}^T$ for a matrix $\mathbf{A}$ and suppose the dimensions of $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ to be d_2 and d_3, respectively.

Theorem 2
Under the regularity conditions C1–C4,

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} N_{d_1+d_2+d_3}(\mathbf{0}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma} = \mathrm{Diag}(\mathbf{J}, \mathbf{I}_{d_2+d_3})\mathbf{D}^{-1}\mathrm{Diag}(\mathbf{J}, \mathbf{I}_{d_2+d_3})^T, \mathbf{D} = E(\boldsymbol{\Psi}^{\otimes 2})$, and $\boldsymbol{\Psi}$ is defined in the Appendix.

Here, the inverse matrix $\mathbf{D}^{-1}$ denotes the Moore–Penrose inverse when $\mathbf{D}$ is not a full-rank matrix. The proof of Theorem 2 is mainly based on verification of three conditions in Newey [21], which is briefly outlined in the Appendix. To estimate the covariance matrix $\boldsymbol{\Sigma}$, we replace the unknown quantities by their respective empirical counterparts.

3. Numerical studies

We conduct simulation studies to evaluate the finite sample performance of our proposed estimation method for the partially linear single-index ZIP model. We are interested in the behavior of the estimator for the parametric part as well as that for the nonparametric single-index function. As for selecting the order and the number of knots for the spline, we first specify the order and quasi-uniform knots, and then the number of knots is automatically produced. Because higher-order splines would induce more complicated interactions and collinearity among the variables in the model, we suggest using cubic splines ($v = 3$) in practice. The number of knots can be identified using an adaptive Bayesian information criterion (BIC):

$$\mathrm{BIC}(k_n) = L(\boldsymbol{\theta}) + \frac{\log n}{2n}(k_n + v + 1 + 2d_1 + 2d_2).$$

We consider two different configurations with the partially linear single index in the λ or p regression components.

Example 1 (λ Single-index model)
We replicate 200 datasets, each consisting of $n = 500$ observations, from the model

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

where $\mathbf{X} = (X_1, X_2)^T$ with X_1 and X_2 independently from Unif$[0, 1]$, $\mathbf{W} = (W_1, W_2)^T$ with Bernoulli random variables W_1 and W_2 taking a value of 0 or 1 with an equal probability of 0.5, and $\mathbf{Z} = (\mathbf{X}^T, \mathbf{W}^T)^T$. The true parameter values are $(\alpha_1, \alpha_2, \gamma_1, \gamma_2) = \left(1/\sqrt{2}, 1/\sqrt{2}, 0.8, 1.2\right)$ and $(\beta_1, \beta_2, \beta_3, \beta_4) = (-0.5, 0.5, -0.5, 0.5)$. The true nonparametric function is $g(u) = 2\exp(-3u^2)$.

Table I summarizes the estimation results for the parametric component of the model, including the average parameter estimate, bias, asymptotic standard deviation, and mean squared error (MSE). Clearly, the biases of all the parameter estimates are negligible, and the MSEs are small, which demonstrate the feasibility of our estimation procedure. Figure 1 shows the nonparametric estimate of the single-index

Table I. Simulation results for the parametric component of the λ single-index model.

	α_1	α_2	γ_1	γ_2	β_1	β_2	β_3	β_4
True	0.7071	0.7071	0.8000	1.2000	−0.5000	0.5000	−0.5000	0.5000
Estimate	0.6946	0.6901	0.8096	1.1976	−0.6013	0.5755	−0.5141	0.4956
Bias	−0.0125	−0.0170	0.0096	−0.0024	−0.1013	0.0755	−0.0141	−0.0044
Standard deviation	0.1474	0.1408	0.1054	0.0994	0.3457	0.3609	0.2351	0.2624
Mean squared error	0.0218	0.0200	0.0112	0.0098	0.1292	0.1353	0.0552	0.0685

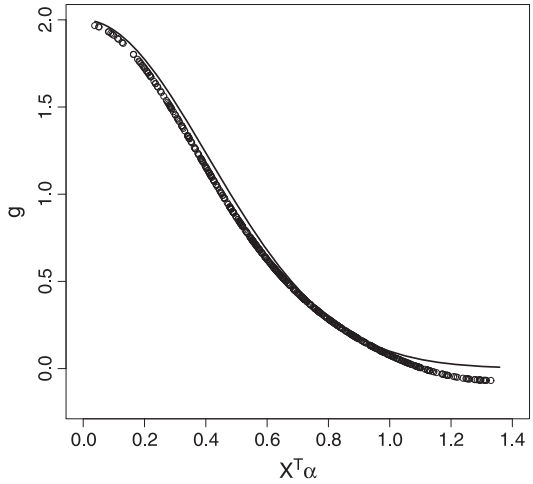

Figure 1. The estimated curve of the single-index function $g(\cdot)$ denoted by the circles, with the true function denoted by the solid line.

Table II. Simulation results for the parametric component of the p single-index model.

	α_1	α_2	γ_1	γ_2	β_1	β_2	β_3	β_4
True	0.7071	0.7071	0.8000	1.2000	−0.5000	0.5000	−0.5000	0.5000
Estimate	0.7071	0.7071	0.7836	1.2213	−0.5334	0.5039	−0.4933	0.4523
Bias	0.0001	0.0000	−0.0164	0.0213	−0.0334	0.0039	0.0067	−0.0477
Standard deviation	0.0051	0.0051	0.4560	0.4120	0.2510	0.2311	0.2673	0.2407
Mean squared error	0.0000	0.0000	0.2061	0.1685	0.0635	0.0529	0.0708	0.0596

function $g(\cdot)$ in conjunction with the curve of the true function. Except for the right-tail part of the curve, our B-spline estimator for the single-index function is close to the true curve, and thus, it is able to characterize the nonlinear shape.

Example 2 (p single-index model)
We also consider accommodating the partially linear single-index in the logistic model:

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

while the rest of the setups are the same as those of Example 1. Table II summarizes the estimation results, which leads to the same conclusions as those in Example 1. In addition, Figure 2 shows that the nonparametric estimate for the single-index function $g(\cdot)$ can capture the main pattern of the true function.

4. Example

To understand how Americans use and pay for health services, the National Medical Expenditure Survey (NMES) was conducted in 1987 and 1988. Under the household survey of the NMES, more than 38,000 individuals in 15,000 households across the U.S.A. were interviewed quarterly about their health insurance coverage, the services they used, and the cost and source of payments of those services. The data were verified by cross-checking information provided by survey respondents with providers of health-care services. In addition to health-care data, NMES also provided information on health status, employment, social-demographic characteristics, and economic status. Following Deb and Trivedi [22], we consider a subsample of individuals of ages 66 years and older who were all covered by Medicare (with a total of 4406 observations).

The response variable is the number of visits to a physician in an office setting. The continuous covariates of interest include hospital stays (X_1), the number of chronic conditions including cancer,

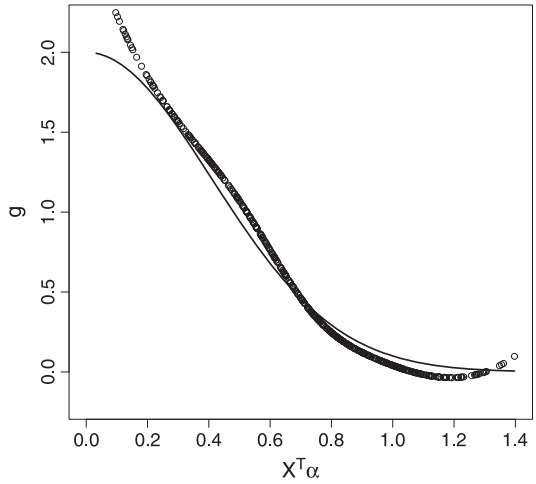

Figure 2. The estimated curve of the single-index function $g(\cdot)$ denoted by the circles, with the true function denoted by the solid line.

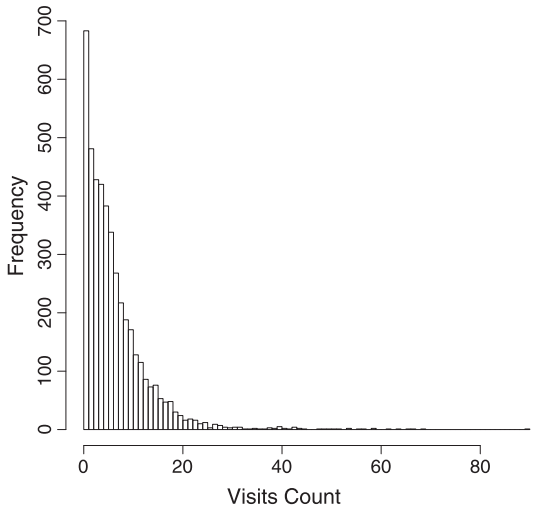

Figure 3. Histogram of visit counts in the medical care data.

heart attack, gall bladder problems, emphysema, arthritis, diabetes, and other heart diseases (X_2), and years of education (X_3), and the categorical covariates include status of excellent health ($W_1 = 1$ if self-perceived health is excellent, and 0 otherwise), status of poor health ($W_2 = 1$ if self-perceived health is poor, and 0 otherwise), private insurance coverage ($W_3 = 1$ if the person is covered by private health insurance, and 0 otherwise) and Medicaid ($W_4 = 1$ if the person is covered by Medicaid, and 0 otherwise). The count of visits ranges from 0 to 89, with a mean of 5.77 and variance of 45.69. Clearly, the variance is much larger than the mean, indicating potential overdispersion. As shown in Figure 3, the data contain more zero visits (the percentage of the zero count is $683/4406 = 15.5\%$) than a typical Poisson distribution. The partially linear single-index ZIP model is employed to study the relationship between the number of visits and covariates. By using all continuous variables to form the single index, we examine the λ single-index model:

$$\log(\lambda) = g\left(\sum_{i=1}^{3} \alpha_i X_i\right) + \sum_{j=1}^{4} \gamma_j W_j,$$

$$\log\left(\frac{p}{1-p}\right) = \sum_{i=1}^{3} \beta_i X_i + \sum_{j=1}^{4} \beta_{j+3} W_j.$$

Table III shows the estimated coefficients based on 5 and 7 degrees of freedom (d.f. $= k_n + v + 1$), with the corresponding standard errors in parentheses. The estimation results of the linear ZIP model are also presented as a benchmark for comparison. Because the partially linear single-index ZIP model is more flexible than the usual linear ZIP model, the estimated variances are generally expected to be reduced to some extent. All the standard errors corresponding to the γs and βs are much smaller than those from the linear ZIP model, while the opposite is true for the αs because they are inside a nonparametric function. We can also see that the estimates are relatively stable for different degrees of freedom. All signs of the estimates for d.f. = 5 and d.f. = 7 are the same, and the estimators and their standard errors are close. Based on the BIC, the optimal d.f. is 7. Based on the p-values from the two models, we conclude that the effects of most covariates are significant and consistent across the proposed partially linear single-index ZIP model and the usual linear ZIP model, except for $\hat{\beta}_3$, that is associated with the factor years of education. Based on the proposed model, the factor years of education significantly affects the structural zeros in the number of visits to a physician, while the usual linear ZIP model shows the effect of years of education to be insignificant. Both models conclude that the status of excellent health does not have a significant impact on the structural zeros. Figure 4 exhibits an obvious nonlinear pattern between the estimated single-index function $g(\cdot)$ and $X^T\hat{\boldsymbol{\alpha}}$, which behaves like a quadratic curve. Alternatively, if other single-index models are employed to analyze the medical care data, estimation results can be

Table III. Estimated coefficients, standard errors in parentheses, and p-values under the proposed partially linear single-index ZIP model with 5 or 7 degrees of freedom (d.f.) for the medical care data.

	Partially linear single-index ZIP model				
Estimates	d.f. = 5	d.f. = 7	p-value	ZIP model	p-value
$\hat{\alpha}_1$	0.779 (0.024)	0.786 (0.027)	<0.001	0.110 (0.005)	<0.001
$\hat{\alpha}_2$	0.527 (0.013)	0.547 (0.015)	<0.001	0.147 (0.007)	<0.001
$\hat{\alpha}_3$	0.340 (0.011)	0.287 (0.011)	<0.001	0.039 (0.008)	<0.001
$\hat{\gamma}_1$	−0.301 (0.013)	−0.288 (0.013)	<0.001	−0.144 (0.031)	<0.001
$\hat{\gamma}_2$	0.242 (0.006)	0.231 (0.006)	<0.001	0.427 (0.018)	<0.001
$\hat{\gamma}_3$	0.160 (0.007)	0.166 (0.007)	<0.001	1.780 (0.008)	<0.001
$\hat{\gamma}_4$	0.221 (0.009)	0.210 (0.009)	<0.001	1.280 (0.021)	<0.001
$\hat{\beta}_1$	−0.161 (0.001)	−0.138 (0.002)	<0.001	−0.122 (0.061)	0.044
$\hat{\beta}_2$	−0.590 (0.003)	−0.566 (0.003)	<0.001	−0.530 (0.058)	<0.001
$\hat{\beta}_3$	−0.081 (0.007)	−0.130 (0.008)	<0.001	−0.069 (0.047)	0.142
$\hat{\beta}_4$	0.082 (0.035)	0.053 (0.035)	0.125	−0.090 (0.164)	0.582
$\hat{\beta}_5$	−0.542 (0.011)	−0.574 (0.011)	<0.001	−0.695 (0.176)	<0.001
$\hat{\beta}_6$	−2.001 (0.003)	−1.953 (0.003)	<0.001	−1.988 (0.060)	<0.001
$\hat{\beta}_7$	−1.318 (0.025)	−1.413 (0.025)	<0.001	−1.492 (0.164)	<0.001

ZIP, zero-inflated Poisson.

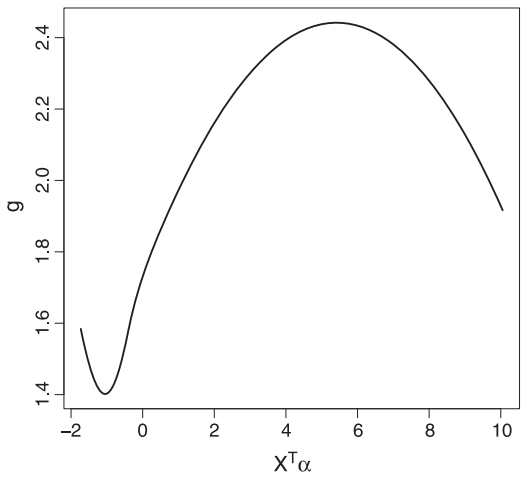

Figure 4. The estimated curve of $g(\cdot)$ for the medical care data.

obtained and the estimated single-index functions also behave nonlinearly. Because of similarity, we omit further discussion.

5. Concluding remarks

Analysis of zero-inflated count data with multiple covariates poses an important practical problem. In this article, we investigate the semiparametric ZIP model with the PLSIM in the Poisson component or the Bernoulli component. The proposed two-step estimation method achieves consistency and asymptotic normality properties under mild conditions. Although a nonparametric single-index function is involved, the profile estimation procedure facilitates the parametric estimates well via the B-spline approximation. Simulation studies and one real data example illustrate that the proposed method can effectively deal with the ZIP data. It remains a subtle issue on which covariates should be selected to form the single index. In our approach, we construct the single index with all the continuous covariates, while certain goodness-of-fit and model-checking techniques are warranted along this direction. Furthermore, if both the Poisson and logistic regression models involve single indices, more cautions need to be taken regarding how to form these single indices.

Appendix A

A1. Proof of Theorem 2

Under the conditions of Theorem 2, we establish the asymptotic normality of $\hat{\boldsymbol{\theta}}$ by using a general result of Newey [21]. Denote $\hat{\Lambda} = \mathbf{X}^T\hat{\boldsymbol{\alpha}}$ and $\Lambda_0 = \mathbf{X}^T\boldsymbol{\alpha}_0$. By (4), we have $p = e^{\mathbf{Z}^T\boldsymbol{\beta}}/(1+e^{\mathbf{Z}^T\boldsymbol{\beta}})$ and $\lambda = e^{g(\mathbf{X}^T\boldsymbol{\alpha})+\mathbf{W}^T\boldsymbol{\gamma}}$. Let $\mathbf{h}(\Lambda) = g'(\Lambda)\mathbf{J}^T(\mathbf{X}-\mathbf{m}_x(\Lambda))$. Denote

$$\boldsymbol{\Psi} \equiv \boldsymbol{\Psi}(g,\mathbf{h},\mathbf{m}_x,\mathbf{m}_w,\mathbf{m}_z,\boldsymbol{\alpha},\boldsymbol{\gamma},\boldsymbol{\beta},Y,\mathbf{X},\mathbf{W},\mathbf{Z}) = \begin{pmatrix}\boldsymbol{\Psi}_1\\ \boldsymbol{\Psi}_2\\ \boldsymbol{\Psi}_3\end{pmatrix},$$

where

$$\begin{aligned}
\boldsymbol{\Psi}_1 &= I_{(Y=0)}(p-1)\lambda e^{-\lambda}\mathbf{h}(\Lambda)/(p+(1-p)e^{-\lambda}) + I_{(Y>0)}(Y-\lambda)\mathbf{h}(\Lambda),\\
\boldsymbol{\Psi}_2 &= I_{(Y=0)}\lambda(p-1)e^{-\lambda}(\mathbf{W}-\mathbf{m}_w(\Lambda))/(p+(1-p)e^{-\lambda}) + I_{(Y>0)}(Y-\lambda)(\mathbf{W}-\mathbf{m}_w(\Lambda)),\\
\boldsymbol{\Psi}_3 &= I_{(Y=0)}(1-p)\left\{p\mathbf{Z}-pe^{-\lambda}\mathbf{Z}-\lambda e^{-\lambda}\mathbf{m}_z(\Lambda)\right\}/(p+(1-p)e^{-\lambda}) + I_{(Y>0)}\{(Y-\lambda)\mathbf{m}_z(\Lambda)-p\mathbf{Z}\}.
\end{aligned}$$

For any given $g^*, \mathbf{h}^*, \mathbf{m}_x^*, \mathbf{m}_w^*$, and $\mathbf{m}_z^*$, define

$$\begin{aligned}
&\boldsymbol{\psi}(g^*-g,\mathbf{h}^*-\mathbf{h},\mathbf{m}_x^*-\mathbf{m}_x,\mathbf{m}_w^*-\mathbf{m}_w,\mathbf{m}_z^*-\mathbf{m}_z,\boldsymbol{\alpha},\boldsymbol{\gamma},\boldsymbol{\beta},Y,\mathbf{X},\mathbf{W},\mathbf{Z})\\
&= \frac{\partial\boldsymbol{\Psi}}{\partial g}(g^*-g) + \frac{\partial\boldsymbol{\Psi}}{\partial\mathbf{h}}^T(\mathbf{h}^*-\mathbf{h}) + \frac{\partial\boldsymbol{\Psi}}{\partial\mathbf{m}_x}^T\left(\mathbf{m}_x^*-\mathbf{m}_x\right) + \frac{\partial\boldsymbol{\Psi}}{\partial\mathbf{m}_w}^T\left(\mathbf{m}_w^*-\mathbf{m}_w\right) + \frac{\partial\boldsymbol{\Psi}}{\partial\mathbf{m}_z}^T\left(\mathbf{m}_z^*-\mathbf{m}_z\right),
\end{aligned} \tag{A.1}$$

where the partial derivatives are the Fréchet partial derivatives and their expectations are zeros after algebraic simplification. Accordingly,

$$\begin{aligned}
&\|\boldsymbol{\Psi}(g^*,\mathbf{h}^*,\mathbf{m}_x^*,\mathbf{m}_w^*,\mathbf{m}_z^*,\boldsymbol{\alpha},\boldsymbol{\gamma},\boldsymbol{\beta},Y,\mathbf{X},\mathbf{W},\mathbf{Z}) - \boldsymbol{\Psi}(g,\mathbf{h},\mathbf{m}_x,\mathbf{m}_w,\mathbf{m}_z,\boldsymbol{\alpha},\boldsymbol{\gamma},\boldsymbol{\beta},Y,\mathbf{X},\mathbf{W},\mathbf{Z})\\
&\quad - \boldsymbol{\psi}(g^*-g,\mathbf{h}^*-\mathbf{h},\mathbf{m}_x^*-\mathbf{m}_x,\mathbf{m}_w^*-\mathbf{m}_w,\mathbf{m}_z^*-\mathbf{m}_z,\boldsymbol{\alpha},\boldsymbol{\gamma},\boldsymbol{\beta},Y,\mathbf{X},\mathbf{W},\mathbf{Z})\|_s\\
&= O\left(\|g^*-g\|_s^2 + \|\mathbf{h}^*-\mathbf{h}\|_s^2 + \|\mathbf{m}_x^*-\mathbf{m}_x\|_s^2 + \|\mathbf{m}_w^*-\mathbf{m}_w\|_s^2 + \|\mathbf{m}_z^*-\mathbf{m}_z\|_s^2\right),
\end{aligned}$$

where $\|\cdot\|_s$ denotes the Sobolev norm. This leads to Assumption 5.1 (i) in the work of Newey [21]. Assumption 5.2 in the work of Newey [21] holds by the expression of (A.1). Moreover, the result

$$E\boldsymbol{\psi}(g^* - g, \mathbf{h}^* - \mathbf{h}, \mathbf{m}_x^* - \mathbf{m}_x, \mathbf{m}_w^* - \mathbf{m}_w, \mathbf{m}_z^* - \mathbf{m}_z, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\beta}, Y, \mathbf{X}, \mathbf{W}, \mathbf{Z}) = \mathbf{0},$$

leads to Assumption 5.3 in the work of Newey [21].

Applying Corollary 6.21 of Schumaker [19], we have

$$\begin{aligned}
\hat{g}(u) - g(u) &= o_p(n^{-1/4}),\\
\hat{g}'(u) - g'(u) &= o_p(n^{-1/4}),\\
\hat{\mathbf{m}}_x(u) - \mathbf{m}_x(u) &= o_p(n^{-1/4}),\\
\hat{\mathbf{m}}_w(u) - \mathbf{m}_w(u) &= o_p(n^{-1/4}),\\
\hat{\mathbf{m}}_z(u) - \mathbf{m}_z(u) &= o_p(n^{-1/4}),
\end{aligned}$$

which all hold uniformly. These results imply that $\hat{\mathbf{h}} - \mathbf{h} = o_p(n^{-1/4})$, and thus, Newey's Assumption 5.1 (ii) holds.

After verifying Assumptions 5.1–5.3 in the work of Newey [21], we apply Newey's Lemma 5.1 and find that $\hat{\boldsymbol{\theta}}^{(1)} = (\hat{\boldsymbol{\alpha}}^{(1)}, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}})$ has the same limit distribution as the solution to the equation

$$\sum_{i=1}^{n} \boldsymbol{\Psi}(g, \mathbf{h}, \mathbf{m}_x, \mathbf{m}_w, \mathbf{m}_z, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\beta}, Y_i, \mathbf{X}_i, \mathbf{W}_i, \mathbf{Z}_i) = \mathbf{0}. \tag{A.2}$$

Furthermore, it is easy to show that the solution to (A.2) has the following asymptotic distribution:

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}^{(1)} - \boldsymbol{\theta}_0^{(1)}\right) \xrightarrow{D} N_{d_1+d_2+d_3-1}(0, \mathbf{D}^{-1}). \tag{A.3}$$

The asymptotic normality of $\hat{\boldsymbol{\theta}}$ has the same limit distribution as described in the statement of Theorem 2 with a simple application of the multivariate delta method to (A.3), as the first component of $\hat{\boldsymbol{\alpha}}$ is $\hat{\boldsymbol{\alpha}}_1 = (1 - \|\hat{\boldsymbol{\alpha}}^{(1)}\|^2)^{1/2}$. Hence, we complete the proof. □

Acknowledgements

We thank two referees, the associate editor, and the editor for their many insightful comments that substantially improved this paper. Wang's research was partially supported by the National Natural Sciences Foundation of China (NSFC) grants 11101063, 11471065, and 11371077, Zhang's research by NSFC Tianyuan fund for Mathematics 11326179 and NSFC grant 11401391, and Yin's research by grant 705613 from the Research Grants Council of Hong Kong.

References

1. Nelder JL, Wedderburn RWM. Generalized linear model. *Journal of the Royal Statistical Society. Series A* 1972; **135**: 370–384.
2. Mullahy J. Specification and testing of some modified count data models. *Journal of Econometrics* 1986; **33**:341–365.
3. Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* 1992; **34**: 1–14.
4. Deng D, Paul SR. Score tests for zero-inflation in generalized linear models. *Canadian Journal of Statistics* 2000; **27**: 563–570.
5. Deng D, Paul SR. Score tests for zero-inflation and over-dispersion in generalized linear models. *Statistica Sinica* 2005; **15**:257–276.
6. Cameron AC, Trivedi PK. *Regression Analysis of Count Data* Second Edition. Cambridge University Press: Cambridge, 2013.
7. Lam KF, Xue H, Cheung YB. Semiparametric analysis of zero-inflated count data. *Biometrics* 2006; **62**:996–1003.
8. He X, Xue H, Shi NZ. Sieve maximum likelihood estimation for doubly semiparametric zero-inflated Poisson models. *Journal of Multivariate Analysis* 2010; **101**:2026–2038.
9. Stoker TM. Consistent estimation of scaled coefficients. *Econometrica* 1986; **54**:1461–1481.
10. Härdle W, Stoker TM. Investigating smooth multiple regression by method of average derivatives. *Journal of the American Statistical Association* 1989; **84**:986–995.
11. Li KC. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association* 1991; **86**: 316–342.
12. Ichimura H. Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics* 1993; **58**:71–129.

13. Carroll RJ, Fan J, Gijbels I, Wand MP. Generalized partially linear single-index models. *Journal of the American Statistical Association* 1997; **92**:477–489.
14. Severini TA, Staniswalis JG. Quasilikelihood estimation in semiparametric models. *Journal of the American Statistical Association* 1994; **89**:501–511.
15. Härdle W, Hall P, Ichimura H. Optimal smoothing in single-index models. *Annals of Statistics* 1993; **21**:157–178.
16. Yu Y, Ruppert D. Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association* 2002; **97**:1042–1054.
17. Xia Y, Härdle W. Semi-parametric estimation of partially linear single-index models. *Journal of Multivariate Analysis* 2006; **97**:1162–1184.
18. Liang H, Liu X, Li R, Tsai CL. Estimation and testing for partially linear single index models. *Annals of Statistics* 2010; **38**:3811–3836.
19. Schumaker LL. *Spline Functions: Basic Theory*. Wiley: New York, 1981.
20. Cui X, Härdle W, Zhu L. The EFM approach for single-index models. *Annals of Statistics* 2011; **39**:1658–1688.
21. Newey WK. The asymptotic variance of semiparametric estimators. *Econometrica* 1994; **62**:1349–1382.
22. Deb P, Trivedi PK. Demand for medical care by the elderly: a finite mixture approach. *Journal of Applied Econometrics* 1997; **12**:313–336.