Semiparametric zero-inflated Bernoulli regression with applications

Chin-Shang Li\textsuperscript{a} and Minggen Lu\textsuperscript{b}

\textsuperscript{a}School of Nursing, The State University of New York, University at Buffalo, Buffalo, NY, USA; \textsuperscript{b}School of Community Health Sciences, University of Nevada, Reno, NV, USA

\textbf{ABSTRACT}

When the observed proportion of zeros in a data set consisting of binary outcome data is larger than expected under a regular logistic regression model, it is frequently suggested to use a zero-inflated Bernoulli (ZIB) regression model. A spline-based ZIB regression model is proposed to describe the potentially nonlinear effect of a continuous covariate. A spline is used to approximate the unknown smooth function. Under the smoothness condition, the spline estimator of the unknown smooth function is uniformly consistent, and the regression parameter estimators are asymptotically normally distributed. We propose an easily implemented and consistent estimation method for the variances of the regression parameter estimators. Extensive simulations are conducted to investigate the finite-sample performance of the proposed method. A real-life data set is used to illustrate the practical use of the proposed methodology. The real-life data analysis indicates that the prediction performance of the proposed semiparametric ZIB regression model is better compared to the parametric ZIB regression model.

\textbf{1. Introduction}

Logistic regression has been extensively used to study the relationship between a binary outcome variable and a set of covariates in many disciplines, including, e.g., biomedical studies, epidemiological studies, and social studies. See, e.g., [9,10] for a detailed survey on logistic regression. However, it is often seen that when analyzing a data set consisting of binary outcome data, the observed proportion of zeros is larger than expected under regular logistic regression. It is then often suggested to use a zero-inflated (ZI) Bernoulli distribution as an alternative to analyzing this type of binary outcome data. A ZIB distribution is a mixture of two Bernoulli distributions. The ZIB distribution can be thought of as a population that is composed of two latent groups: the susceptible group consisting of individuals who are at risk of an event of interest, and may have the event, and the non-susceptible (cured) group consisting of individuals who are not at risk of the event of interest.

\textbf{CONTACT} Chin-Shang Li csl2003@gmail.com

Supplemental data for this article can be accessed here. https://doi.org/10.1080/02664763.2021.1925228

© 2021 Informa UK Limited, trading as Taylor & Francis Group
Diop et al. [5] proposed a parametric ZIB regression model for this type of binary outcome data with linear predictors via logit links and applied it to analyze an example of data from a study of dengue fever to investigate the effects of covariates on dengue infection. Diop et al. [4] established the asymptotic properties of the maximum likelihood (ML) estimators of the parameters of a parametric ZIB regression model. In addition, for example, Hall [7] added the flexibility of a fixed-effects ZI Poisson (ZIP) model ([13]) and a fixed-effects ZI binomial model by incorporating random effects to accommodate the within-subject correlation and between-subject heterogeneity typical of repeated measures data, and applied these mixed-effects models to analyze an example from horticulture where data with excess zeros were collected in a repeated measures designed experiment. Also, see [16] for random effect models for repeated measures of ZI count data. Young et al. [20] utilized a ZIP regression model ([13]) and a zero-inflated negative binomial (ZINB) regression model to determine the predicted distributions of added and deleted housing units in a block, which are used to characterize the undercoverage and overcoverage, respectively. The used data were from the 2010 address canvassing operation. All the aforementioned authors assumed that the effects of covariates in a ZIB, ZIP, ZI binomial or ZINB regression model are usually modeled via a linear predictor function. However, sometimes it may not be completely appropriate to assume that a continuous covariate effect is linear or has some other parametric form. One can then use nonparametric regression approaches to relax the restrictive parametric assumption by using an unspecified smooth function to describe the possibly nonlinear effect of the continuous covariate. Thus, motivated to address this issue, we propose a semiparametric ZIB regression model, which is described in Section 2.1 and can enhance fitting flexibility, to fit ZI binary outcome data. As an application, the proposed semiparametric ZIB regression model can be applied to assess the adequacy of a postulated parametric ZIP regression model by comparing the prediction performance of these two models in terms of prediction error, normal quantile–quantile (Q–Q) plot of Pearson residuals, and likelihood ratio (LR) test, which are illustrated in Section 5.

In this work, for the proposed semiparametric ZIB regression model, the spline-based sieve M-estimation method (e.g. [14,15]) is used to estimate the unspecified smooth function for the effect of the continuous covariate. The spline coefficients for the chosen spline basis functions can completely describe the approximated unknown smooth function. The spline likelihood function can then be maximized to estimate simultaneously the regression parameters and the spline coefficients. The asymptotic variances of the estimators of the regression parameters can be estimated directly by using the spline estimation method. Additionally, for instance, Lam et al. [12] considered a semiparametric ZIP regression model for a possibly nonlinear relationship between the natural logarithm of the mean of the count variable and a particular continuous covariate when modeling ZI count data. They proposed a sieve ML estimation method for fitting the proposed semiparametric ZIP model and investigated asymptotic properties of the sieve ML estimators. He et al. [8] fit a doubly semiparametric ZIP model to ZI count data in which two partially linear link functions were assumed for both the mean of the Poisson component and the probability of zero. They used a sieve ML method to estimate both the regression parameters and the nonparametric functions, and studied asymptotic properties of the sieve ML estimators. To analyze longitudinal ZI count data, Feng and Zhu [6] considered a semiparametric ZIP mixed model that postulates a possible nonlinear relationship between the
natural logarithm of the mean of the count variable and a particular continuous covariate. They proposed a penalized log-likelihood function, and used the Monte Carlo expectation-maximization algorithm to estimate the parameters of the proposed semiparametric ZIP mixed model and established asymptotic properties of the resulting estimators.

The estimation method we use for the nonparametric part of the semiparametric ZIB regression model is different from those for the nonparametric part(s) of the aforementioned semiparametric ZIP regression models. We are the first to establish the large-sample properties of the resulting estimators of this proposed model based on the spline-based sieve M-estimation method. It can be shown that in our estimation method, the spline estimator of the unknown smooth function achieves the optimal convergence rate by taking into account the number of knots to increase at an appropriate rate. The spline-based sieve semiparametric ZIB regression model can be shown to achieve the asymptotic efficiency of the estimators of the regression parameters.

In Section 2, we introduce the semiparametric ZIB regression model and the estimation method for the parameters of this regression model. Section 3 describes the asymptotic results of the spline estimators and likelihood ratio inference, and proposes an easily implemented approach to consistently estimate the variances of the estimators of the regression parameters. In Section 4, extensive Monte Carlo simulation studies are conducted to investigate the finite-sample performance of the proposed methodology. Section 5 illustrates the practical use of the proposed methodology with a real-life data example. Some concluding remarks are given in Section 6.

2. Model and algorithm

2.1. Semiparametric ZIB regression model

Let \( Y \) be a binary response variable that indicates occurrence of the event of interest or not, e.g. whether or not catching one or more fish in a lake from a survey (i.e. \( Y = 1 \) if catching one or more fish, and \( Y = 0 \) otherwise). Assume \( U \) is a latent binary variable indicating an individual’s risk state: \( U = 1 \) if the individual is at risk of the event, e.g. susceptible to catching one or more fish (in other words, going fishing); \( U = 0 \) if the individual is not at risk of the event. Note that if \( Y = 0 \), then \( U \) is unknown (unobserved), and if \( Y = 1 \), then \( U = 1 \). Let \( x = (x_1, \ldots, x_p)^\top \) be a \( p \times 1 \) covariate vector, where \( x \) can be discrete, continuous, or a combination of both, and \( z \) a continuous covariate. Denote \( w = (x^\top, z)^\top \).

In a parametric ZIB regression model, both the susceptible probability \( \eta(w) \equiv \Pr(U = 1; w) \) and the probability of occurrence of the event for a susceptible individual \( \zeta(w) \equiv \Pr(Y = 1; w | U = 1) \) are linked to \( w \) via logit-linear predictors. The parametric ZIB regression model can then be written as follows:

\[
\Pr(Y = y; w) = [1 - \eta(w)]I_{y=0} + \eta(w)\zeta^y(w)[1 - \zeta(w)]^{1-y}, \quad y = 0, 1,
\]

\[
= \begin{cases} 
1 - \eta(w) + \eta(w)[1 - \zeta(w)], & \text{if } y = 0, \\
\eta(w)\zeta(w), & \text{if } y = 1.
\end{cases}
\]

(1)

Here, for a given \( w \), \( [1 - \eta(w)] \in [0, 1] \) is a mixing weight for the accommodation of extra zeros, and \( \zeta(w) \) is the mean of the Bernoulli distribution. \( I_{[1,]} \) is the indicator function. Note that the ZIB distribution is reduced to a regular Bernoulli distribution when \( \eta(w) = 1 \).
Although the use of a linear predictor may be completely appropriate in some applications, it may not be valid for other data situations. More specifically, when it is suspected that the effect of the continuous covariate \( z \) on the logarithm of the odds of the event for susceptible individuals is nonlinear, smoothing techniques can be used to approximate an unspecified smooth function that depicts the effect of \( z \). Motivated by this, we relax the restrictive parametric assumption in this parametric ZIB regression model by assuming that the functional form for the effect of \( z \) on the logarithm of the odds of the event for susceptible individuals is smooth but unknown, and the effects of \( x \) remain linear.

Let \( \psi \) be an unknown smooth function that is used to describe the effect of \( z \) on the logarithm of the odds of the event for susceptible individuals. We can then propose the semiparametric ZIB regression model that consists of the following logistic regression model for the probability of occurrence of the event of interest as follows:

\[
\logit[\eta(w; \beta)] = \log \left[ \frac{\Pr(U = 1; \beta, w)}{1 - \Pr(U = 1; \beta, w)} \right] = \tilde{x}^T \beta_1 + z \beta_2 = \tilde{w}^T \beta, \tag{2}
\]

where \( \tilde{x} = (1, x^T)^T \), \( \tilde{w} = (\tilde{x}, z)^T \), and \( \beta = (\beta_1^T, \beta_2)^T \) for \( \beta_1 = (\beta_{10}, \beta_{11}, \ldots, \beta_{1p})^T \), and the following semiparametric logistic regression model for the probability of occurrence of the event for a susceptible individual:

\[
\logit[\zeta(w; \alpha, \psi)] = \log \left[ \frac{\Pr(Y = 1; \alpha, \psi, w|U = 1)}{1 - \Pr(Y = 1; \alpha, \psi, w|U = 1)} \right] = x^T \alpha + \psi(z), \tag{3}
\]

where \( \alpha = (\alpha_1, \ldots, \alpha_p)^T \). To the best of our knowledge, the semiparametric ZIB regression model has not been proposed for modeling ZI binary outcome data. In this proposed model, a spline function is used to approximate the unknown smooth function that is used to describe the potentially nonlinear effect of \( z \) on the logarithm of the odds of the event of interest for a susceptible individual. With some algebra, one can express the probability of occurrence of the event of interest as follows:

\[
\pi(w; \alpha, \beta, \psi) = \Pr(Y = 1; \alpha, \beta, \psi, w) = \zeta(w; \alpha, \psi) \eta(w; \beta).
\]

Assuming that \( \{y_i, w_i : i = 1, \ldots, n\} \) is a data set, we can express the log-likelihood function for the semiparametric ZIB regression model as follows:

\[
\ell_n(\alpha, \beta, \psi) = \sum_{i=1}^{n} \{y_i \log[\pi(w_i; \alpha, \beta, \psi)] + (1 - y_i) \log[1 - \pi(w_i; \alpha, \beta, \psi)]\}. \tag{4}
\]

Define \( S_n(T_n, l) \) as the class of splines of order \( l \geq 1 \) with the sequence of knots \( T_n \). The smooth function \( \psi \) is approximated by a spline function \( \psi_n \in S_n(T_n, l) \) that can be expressed as a linear combination of the \( B \)-spline basis functions as follows:

\[
\psi(z) \approx \psi_n(z) = \sum_{j=1}^{q_n} \gamma_j B_j(z); \tag{5}
\]

see [18] for more details. Let \( \mathbf{y} = (\gamma_1, \ldots, \gamma_{q_n})^T \). We use \( \psi_n(z) \) to replace \( \psi(z) \) in the log-likelihood function (4) to obtain the following spline log-likelihood function for
\(\mathbf{\tau} = (\mathbf{\alpha}^T, \mathbf{\beta}^T, \mathbf{y}^T)^T:\)

\[
\ell_n(\mathbf{\tau}) = \sum_{i=1}^{n} \{y_i \log[\pi(w_i; \mathbf{\tau})] + (1 - y_i) \log[1 - \pi(w_i; \mathbf{\tau})]\},
\]

where \(\pi(w; \mathbf{\tau}) = \zeta(w; \mathbf{\alpha}, \mathbf{y}) \eta(w; \mathbf{\beta})\) for \(\zeta(w; \mathbf{\alpha}, \mathbf{y})\) that can be expressed as follows:

\[
\zeta(w; \mathbf{\alpha}, \mathbf{y}) = \Pr(Y = 1; \mathbf{\alpha}, \mathbf{y}, w | U = 1) \nonumber = \frac{\exp[x^T\mathbf{\alpha} + \sum_{j=1}^{q} y_j B_j(z)]}{1 + \exp[x^T\mathbf{\alpha} + \sum_{j=1}^{q} y_j B_j(z)]}.
\]

The spline ML estimate \(\hat{\mathbf{\tau}} = (\hat{\mathbf{\alpha}}^T, \hat{\mathbf{\beta}}^T, \hat{\mathbf{y}}^T)^T\) of \(\mathbf{\tau}\) can be obtained by using the \texttt{optim} function with the method of L-BFGS-B ([3]) in \(\mathbb{R}\) to solve the system of score equations \(\partial \ell_n(\mathbf{\tau}) / \partial \mathbf{\tau} = 0\) because this L-BFGS-B algorithm can handle box (or bounds) constraints, i.e. each variable can be given a lower and/or upper bound. The obtained spline estimate of \(\psi(z)\) is then denoted by \(\hat{\psi}_n(z) = \sum_{j=1}^{q} \hat{\gamma}_j B_j(z)\).

### 2.2. Identifiability

To be able to estimate the parameters of the spline ZIB regression model, we need to verify the identifiability of this model. Let \(\mathcal{W}\) be the design space. Assume \(\mathcal{P} = \{\Pr(U = 1; \mathbf{\beta}, w) = \eta(w; \mathbf{\beta}) : 0 < \eta(w; \mathbf{\beta}) < 1 \text{ for } w \in \mathcal{W}\}\) is the space of susceptible probability functions. Let \(\mathcal{F} = \{\Pr(Y = y; \mathbf{\alpha}, \mathbf{y}, w | U = 1) = \zeta^*(w; \mathbf{\alpha}, \mathbf{y}) [1 - \zeta(w; \mathbf{\alpha}, \mathbf{y})]^y : \zeta(w; \mathbf{\alpha}, \mathbf{y}) > 0 \text{ for } w \in \mathcal{W} \text{ and } y = 0, 1\}\) be the class of conditional Bernoulli distributions of susceptible individuals. Denote the class of spline ZIB regression models by

\[
\mathcal{C} = \{\Pr(Y = y; \mathbf{\tau}, w) : \Pr(Y = y; \mathbf{\tau}, w) = [1 - \eta(w; \mathbf{\beta})]I_{y=0} + \eta(w; \mathbf{\beta}) \zeta^*(w; \mathbf{\alpha}, \mathbf{y}) [1 - \zeta(w; \mathbf{\alpha}, \mathbf{y})]^{1-y}, w \in \mathcal{W}, y = 0, 1, \eta(w; \mathbf{\beta}) \in \mathcal{P}, \zeta^*(w; \mathbf{\alpha}, \mathbf{y}) [1 - \zeta(w; \mathbf{\alpha}, \mathbf{y})]^y \in \mathcal{F}\}.
\]

Following [17], we define the identifiability of the spline ZIB regression model as follows:

**Definition 2.1:** The class of spline ZIB regression models \(\mathcal{C}\) is identifiable if for any two members in \(\mathcal{C}\) given by \(\Pr(Y = y; \mathbf{\tau}, w)\) and \(\Pr(Y = y; \mathbf{\tau}^*, w)\), then \(\Pr(Y = y; \mathbf{\tau}, w) = \Pr(Y = y; \mathbf{\tau}^*, w)\) if and only if \(\eta(w; \mathbf{\beta}) = \eta(w; \mathbf{\beta}^*)\) and \(\zeta(w; \mathbf{\alpha}, \mathbf{y}) = \zeta(w; \mathbf{\alpha}^*, \mathbf{y}^*)\) for \(w \in \mathcal{W}\) and \(y = 0, 1\).

To investigate the identifiability of the spline ZIB regression model, it is needed to show that \(\Pr(Y = y; \mathbf{\tau}, w) = \Pr(Y = y; \mathbf{\tau}^*, w)\) if and only if \(\eta(w; \mathbf{\beta}) = \eta(w; \mathbf{\beta}^*)\) and \(\zeta(w; \mathbf{\alpha}, \mathbf{y}) = \zeta(w; \mathbf{\alpha}^*, \mathbf{y}^*)\) for \(w \in \mathcal{W}\) and \(y = 0, 1\). Because the ‘if’ part is clearly true, we focus on ‘only if’. To show the ‘only if’ part, assume that \(\Pr(Y = y; \mathbf{\tau}, w) = \Pr(Y = y; \mathbf{\tau}^*, w)\). With some algebra, we can then have the ratio as follows:

\[
\frac{\eta(w; \mathbf{\beta})}{\eta(w; \mathbf{\beta}^*)} = \frac{I_{y=0} - \zeta^*(w; \mathbf{\alpha}^*, \mathbf{y}^*) [1 - \zeta(w; \mathbf{\alpha}^*, \mathbf{y}^*)]^{1-y}}{I_{y=0} - \zeta^*(w; \mathbf{\alpha}, \mathbf{y}) [1 - \zeta(w; \mathbf{\alpha}, \mathbf{y})]^{1-y}}.
\]

Because the left-hand side of the ratio (8) depends on \(w\) and the right-hand side of the ratio (8) is dependent on \(w\) and \(y\), the ratio is a positive function of \(w\), denoted by \(c(w)\), for
\( \mathbf{w} \in \mathcal{W} \). The quantities \( \eta(\mathbf{w}; \mathbf{\beta}^*) \) and \( \zeta(\mathbf{w}; \mathbf{\alpha}^*, \mathbf{\gamma}^*) \) must satisfy the following equations:

\[
\eta(\mathbf{w}; \mathbf{\beta}^*) = \eta(\mathbf{w}; \mathbf{\beta})/c(\mathbf{w}),
\]

\[
\zeta^y(\mathbf{w}; \mathbf{\alpha}^*, \mathbf{\gamma}^*)[1 - \zeta(\mathbf{w}; \mathbf{\alpha}^*, \mathbf{\gamma}^*)]^{1-y} = [1 - c(\mathbf{w})]I_{y=0} + c(\mathbf{w})\zeta^y(\mathbf{w}; \mathbf{\alpha}, \mathbf{\gamma})[1 - \zeta(\mathbf{w}; \mathbf{\alpha}, \mathbf{\gamma})]^{1-y}, \quad y = 0, 1,
\]

for \( \mathbf{w} \in \mathcal{W} \). If we can show \( c(\mathbf{w}) = 1 \) for \( \mathbf{w} \in \mathcal{W} \), then we show the identifiability of the spline ZIB regression model. Because from Equation (10) \( \zeta(\mathbf{w}; \mathbf{\alpha}^*, \mathbf{\gamma}^*) = c(\mathbf{w})\zeta(\mathbf{w}; \mathbf{\alpha}, \mathbf{\gamma}) \) as \( y = 0 \) or \( 1 \), it follows that \( \zeta(\mathbf{w}; \mathbf{\alpha}^*, \mathbf{\gamma}^*) = c(\mathbf{w})\zeta(\mathbf{w}; \mathbf{\alpha}, \mathbf{\gamma}) \) for \( \mathbf{w} \in \mathcal{W} \). If the spline ZIB regression model is not identifiable, then it is required that \( \eta(\mathbf{w}; \mathbf{\beta}^*) = \eta(\mathbf{w}; \mathbf{\beta})/c(\mathbf{w}) \) for some \( c(\mathbf{w}) \neq 1 \). Hence there exists \( \mathbf{\beta}^* = (\mathbf{\beta}_1^T, \mathbf{\beta}_2^T) = (\mathbf{\beta}_{10}^*, \mathbf{\beta}_{11}^*, \ldots, \mathbf{\beta}_{1p}^*, \mathbf{\beta}_2^*)^T \) such that \( \exp(\mathbf{\hat{w}}^T\mathbf{\beta}^*)/[1 + \exp(\mathbf{\hat{w}}^T\mathbf{\beta}^*)] = \exp(\mathbf{\hat{w}}^T\mathbf{\beta})/c(\mathbf{w})[1 + \exp(\mathbf{\hat{w}}^T\mathbf{\beta})] \) for all \( \mathbf{w} \in \mathcal{W} \). Now provided that \( \mathbf{w} \) takes at least \( p + 3 \) values and one of \( \mathbf{\beta}_{11}, \ldots, \mathbf{\beta}_{1p} \), and \( \mathbf{\beta}_2 \) is unequal to 0, there is no such \( \mathbf{\beta}^* \), i.e. \( \eta(\mathbf{w}; \mathbf{\beta}^*) \) cannot be expressed in a logistic form unless \( c(\mathbf{w}) = 1 \). Therefore, \( \Pr(Y = y; \mathbf{\bar{w}}, \mathbf{w}) \) is uniquely represented by \( \eta(\mathbf{w}; \mathbf{\beta}) \) and \( \zeta(\mathbf{w}; \mathbf{\alpha}, \mathbf{\gamma}) \) for \( \mathbf{w} \in \mathcal{W} \) and \( y = 0, 1 \). The identifiability follows.

### 2.3. Selection of knots

The Schwarz information criterion (SIC) ([19]) is employed to determine the optimal number of interior knots, \( m^* \), which minimizes the following SIC value:

\[
\text{SIC}(m_n) = -2\ell_n(\mathbf{\hat{\alpha}}, \mathbf{\hat{\beta}}, \mathbf{\hat{\gamma}}) + (m_n + l + 2p + 2)\log n.
\]

Here, \( m_n + l \) is the number of B-spline basis functions, and \( m_n + l + 2p + 2 \) is the number of free parameters. The SIC rather than the Akaike information criterion (AIC) ([1]) is used because if \( n \geq 8 \), then the SIC will tend to favor models of lower dimension than those chosen by the AIC ([11]). After \( m^* \) is determined, the locations of the \( m^* \) knots are chosen at equally spaced quantiles of the continuous covariate \( z \).

### 3. Asymptotic results and estimation of variances

#### 3.1. Asymptotic results

Let \( \mathbf{\tau}_0 = (\mathbf{\theta}_0^T, \mathbf{\psi}_0^T)^T \) be the true value of \( \mathbf{\tau} = (\mathbf{\theta}^T, \mathbf{\psi})^T \), where \( \mathbf{\theta} = (\mathbf{\alpha}^T, \mathbf{\beta}^T)^T \). The large-sample properties of spline estimators are established with the following \( L_2 \)-norm:

\[
\|\mathbf{\tau}_1 - \mathbf{\tau}_2\|_2^2 = \|\mathbf{\theta}_1 - \mathbf{\theta}_2\|^2 + \|\mathbf{\psi}_1 - \mathbf{\psi}_2\|^2,
\]

where \( \| \cdot \| \) and \( \| \cdot \|_2 \) denote the Euclidean norm and usual \( L_2 \)-norm, respectively. The log-likelihood function of \( \mathbf{\tau} \) for a single observation \( \mathbf{v} = (y, \mathbf{w}^T)^T \) is given by

\[
\ell(\mathbf{\tau}; \mathbf{v}) = y\log[\pi(\mathbf{w}; \mathbf{\tau})] + (1 - y)\log[1 - \pi(\mathbf{w}; \mathbf{\tau})].
\]

Let \( \hat{\ell}(\mathbf{\tau}; \mathbf{v}) = \partial \ell(\mathbf{\tau}; \mathbf{v})/\partial \pi(\mathbf{w}; \mathbf{\tau}) \) denote the first derivative of \( \ell(\mathbf{\tau}; \mathbf{v}) \) with respect to \( \pi(\mathbf{w}; \mathbf{\tau}) \). The score functions for \( \mathbf{\theta} \) are then as follows:

\[
\hat{\ell}_{\mathbf{\theta}}(\mathbf{\tau}; \mathbf{v}) = \frac{\partial \ell(\mathbf{\tau}; \mathbf{v})}{\partial \mathbf{\theta}} = \left( \frac{\hat{\ell}_{\pi}(\mathbf{\tau}; \mathbf{v})\xi(\mathbf{w}; \mathbf{\tau})x}{\hat{\ell}_{\pi}(\mathbf{\tau}; \mathbf{v})\lambda(\mathbf{w}; \mathbf{\tau})}\right).
\]
Let $P(C_4)(C_2)$. The support of $\rho(r\eta)$. Specifically, in view of the property of $\eta$, required. To establish the identifiability of the model, condition $C_4$ is needed.

To calculate entropy in the proofs of the following theorems, conditions $C_2$ and $C_3$ are minimized the distance $E\frac{\eta}{\eta}$.

Condition $C_1$ is the standard assumption in semiparametric estimation.

Here, $\xi(w; \tau) = \zeta(w; \alpha, \psi)[1 - \zeta(w; \alpha, \psi)] \eta(w; \beta) + \eta(w; \beta) [1 - \eta(w; \beta)]$. Differentiating $\ell(\tau; v)$ along with the parametric submodel $\psi_t = \psi + th$ at $t = 0$ for $h$ with bounded variation yields the following score operator for $\psi$:

$$
\ell_\psi(\tau; v)[h] = \frac{d\ell(\tau, \psi; v)}{dt} \bigg|_{t=0} = \hat{\ell}(\tau; v)\xi(w; \tau)h.
$$

Let $H$ be the class of such $h$. For $h = (h_1, \ldots, h_{2p+2})^T \in H^{2p+2}$, denote $\ell_\psi(\tau; v)[h] = (\ell_\psi(\tau; v)[h_1], \ldots, \ell_\psi(\tau; v)[h_{2p+2}])^T$. It can be shown that there exists a unique $h^*$ that minimizes the distance $\rho(h) = \|\theta(\tau_0; v) - \hat{\ell}(\tau_0; v)[h]\|^2_2$ for $h \in H^{2p+2}$; see [2]. The efficient score functions for $\theta$ at $\tau = \tau_0$ are then as follows:

$$
\ell_\theta^*(\tau_0; v) = \ell_\theta(\tau_0; v) - \hat{\ell}(\tau_0; v)[h^*].
$$

Specifically, in view of the property of $h^*$, one can then express $h^*$ as

$$
h^* = \left(\begin{array}{c}
h_1^* \\
h_2^*
\end{array}\right) = \frac{1}{E_{v|z}[\mu_0^2 x|z]} \left(\begin{array}{c}
E_{v|z}[\mu_0^2 x|z] \\
E_{v|z}[\mu_0 v_0 \tilde{w}|z]
\end{array}\right).
$$

Here, $\mu_0 = \hat{\ell}(\tau_0; v)\xi(w; \tau_0)$ and $v_0 = \hat{\ell}(\tau_0; v)\lambda(w; \tau_0)$. Thus, we can have the following efficient score functions for $\theta$ at $\tau = \tau_0$:

$$
\ell_\theta^*(\tau_0; v) = \left(\begin{array}{c}
\mu_0 x - \mu_0 h_1^* \\
v_0 \tilde{w} - \mu_0 h_2^*
\end{array}\right).
$$

Here, $[E[\ell_\theta^*(\tau_0; v)]]^{\otimes 2}$ is the efficient information $I(\theta_0)$, where $a^{\otimes 2} = aa^T$ for any vector $a$.

To derive the large-sample properties of the spline estimator vector $\hat{\tau} = (\hat{\theta}^T, \hat{\psi})^T$, we require the following regularity conditions that are similar to those in [14]:

(C1) The true regression parameter vector $\theta_0$ is an interior point of a compact subset of $\mathbb{R}^{2p+2}$. The $r$th derivative of true function $\psi_0$, for $r \geq 3$, satisfies the Lipschitz condition on a compact set $\Xi$.

(C2) The support of $z$ is an interval within $\Xi$, and the second moment of $z$ is finite.

(C3) The fourth moment of $x$ is finite.

(C4) $P(x^T \alpha \neq x^T \alpha_0) > 0$ and $P(\tilde{w}^T \beta \neq \tilde{w}^T \beta_0) > 0$ for any $\alpha \neq \alpha_0$ and $\beta \neq \beta_0$.

Remark 3.1: Condition C1 is the standard assumption in semiparametric estimation. To calculate entropy in the proofs of the following theorems, conditions C2 and C3 are required. To establish the identifiability of the model, condition C4 is needed.

**Theorem 3.1 (Uniform Convergence and Convergence Rate):** Assume $q_n = O(n^v)$ for $1/(2r + 2) < v < 1/(2r)$ and conditions C1–C4 hold. Then, $\|\hat{\tau} - \tau_0\|_2 = O_p(n^{-\min(r,1-v)/(1-v)})$ and, hence, $\|\hat{\psi} - \psi_0\|_\infty = o_p(1)$. If $v = 1/(1+2r)$, then $O_p(n^{-\min(r,1-v)/(1+2r)}) = O_p(n^{-r/(1+2r)})$ that is the optimal rate of convergence in nonparametric regression.
\textbf{Theorem 3.2 (Asymptotic Normality):} Assume conditions C1–C4 hold and $I(\theta_0)$ is non-singular. Then, as $n \to \infty$, $n^{1/2}(\hat{\theta} - \theta_0)$ is asymptotically a normal distribution with mean 0 and covariance matrix $I^{-1}(\theta_0)$, denoted by

$$
n^{1/2}(\hat{\theta} - \theta_0) = n^{-1/2}I^{-1}(\theta_0) \sum_{i=1}^{n} \ell_{\theta}^*(\tau_i; v_i) + o_p(1) \to_d N(0, I^{-1}(\theta_0)).$$

\textbf{Remark 3.2:} In Theorem 3.1, $\hat{\psi}_{\theta}$ is shown to be uniformly convergent. The overall convergence rate of $\hat{\tau}$ is $n^{r/(1+2r)}$, but as shown in Theorem 3.2, the convergence rate of $\hat{\theta}$ is still $n^{1/2}$. $\hat{\theta}$ is shown to achieve the information bound and, hence, be efficient in semiparametric regression; see \cite{2}.

Assume that $\hat{\psi}_{\theta}$ is the spline estimator of $\psi$ for any $\theta$ in the neighborhood of $\hat{\theta}$. Let $pl(\theta) = \ell_n(\theta, \hat{\psi}_\theta) = \sum_{i=1}^{n} \ell(\theta, \hat{\psi}_\theta; v_i)$ be the profile log-likelihood for $\theta$ and $lrt(\theta_0) = 2pl(\hat{\theta}) - 2pl(\theta_0)$ the likelihood ratio test statistic for testing $\theta = \theta_0$. The asymptotic result of $lrt(\theta_0)$ is stated in the following theorem.

\textbf{Theorem 3.3 (Likelihood Ratio Inference):} Assume conditions C1–C4 hold. Then, as $n \to \infty$, $lrt(\theta_0)$ is asymptotically a chi-squared distribution with $2p + 2$ degrees of freedom, denoted by $lrt(\theta_0) \to_d \chi^2_{2p+2}$ under $H_0 : \theta = \theta_0$.

The proofs of Theorems 3.1–3.3 are sketched in the Web Appendix by using the proof skills of \cite{14}.

\subsection*{3.2. Estimation of variances}

Motivated by Zhang et al. \cite{21}, to estimate $I(\theta_0)$, we propose to use the empirical version of $I(\theta_0)$, expressed as follows:

$$
P_n[\hat{\ell}_\theta(\hat{\tau}; v) - \hat{\ell}_\psi(\hat{\tau}; v)[h^*]] \otimes 2,
\label{eq:12}
$$

where $P_n$ is the empirical measure. Again we apply the $B$-spline basis functions to appropriate $h^*$, i.e. $h^*_s = \sum_{j=1}^{q_n} \gamma_{j,s} B_j$, $s = 1, \ldots, 2p + 2$. By the definition of $h^*$, one can estimate the $B$-spline coefficient vector $\gamma_s = (\gamma_{1,s}, \ldots, \gamma_{q_n,s})^T$ by minimizing

$$
P_n \sum_{s=1}^{2p+2} \left[ \hat{\ell}_{\tau,s}(\hat{\tau}) - \sum_{j=1}^{q_n} \gamma_{j,s} \hat{\ell}_\psi(B_j) \right]^2,
\label{eq:13}
$$

where $\hat{\ell}_{\tau,s}(\hat{\tau})$ is the $s$th element of $\hat{\ell}_\theta(\hat{\tau})$. Denote by $\hat{\gamma}_{j,s}$ the estimator of $\gamma_{j,s}$. Let $\hat{h}^* = (\hat{h}_1^*, \ldots, \hat{h}^*_{2p+2})^T$, where $\hat{h}^*_s = \sum_{j=1}^{q_n} \hat{\gamma}_{j,s} B_j$, $s = 1, \ldots, 2p + 2$, and $B_n = (B_1, \ldots, B_{q_n})^T$. Applying standard least-squares calculation, one can express the empirical version of $I(\theta_0)$ in (12) as follows:

$$
P_n[\hat{\ell}_\theta(\hat{\tau}; v) - \hat{\ell}_\psi(\hat{\tau}; v)[\hat{h}^*]] \otimes 2 = \hat{I}_{\theta\theta} - \hat{I}_{\theta\psi} \hat{I}_{\psi}^{-1} \hat{I}_{\psi\theta},
\label{eq:14}
$$

where $\hat{I}_{\theta\theta} = P_n[\hat{\ell}_\theta(\hat{\tau}) \otimes 2$, $\hat{I}_{\theta\psi} = P_n[\hat{\ell}_\theta(\hat{\tau}) \hat{\ell}_\psi(\hat{\tau}) B_n]$, $\hat{I}_{\psi\theta} = \hat{I}_{\theta\psi}^T$, and $\hat{I}_{\psi\psi} = P_n[\hat{\ell}_\psi(\hat{\tau}) B_n^2 \otimes 2].$ Let $\hat{E}_{\theta\theta} = E_Y|w(\hat{I}_{\theta\theta})$, $\hat{E}_{\theta\psi} = E_Y|w(\hat{I}_{\theta\psi})$, $\hat{E}_{\psi\theta} = \hat{E}_{\theta\psi}^T$, and $\hat{E}_{\psi\psi} = E_Y|w(\hat{I}_{\psi\psi})$. It can
then be shown that in Theorem 3.4 $\mathbf{I}(\theta_0)$ is estimated consistently by the conditional expected information

$$\hat{E}_n = \hat{E}_{\theta\theta} - \hat{E}_{\theta \psi} \hat{E}_{\psi \psi}^{-1} \hat{E}_{\psi \theta}.$$  

More specifically, with the notations defined in Section 2, the consistent estimator of $\mathbf{I}(\theta_0)$ is expressed as follows:

$$\hat{E}_n = \frac{1}{n} \left[ \begin{pmatrix} X^T \hat{\xi}^2 \hat{V}^{-1} X & X^T \hat{\xi} \hat{V}^{-1} \hat{W} \\ \hat{W}^T \hat{\xi} \hat{V}^{-1} X & \hat{W}^T \hat{\xi} \hat{V}^{-1} \hat{W} \end{pmatrix} - \begin{pmatrix} X^T \hat{\xi}^2 \hat{V}^{-1} B \\ \hat{W}^T \hat{\xi} \hat{V}^{-1} B \end{pmatrix} \left( B^T \hat{\xi}^2 \hat{V}^{-1} B \right)^{-1} \begin{pmatrix} X^T \hat{\xi}^2 \hat{V}^{-1} B \\ \hat{W}^T \hat{\xi} \hat{V}^{-1} B \end{pmatrix}^T \right].$$  

(13)

Here, $\hat{\xi}$, $\hat{\lambda}$, and $\hat{V}$ represent $\xi$, $\lambda$, and $V$ evaluated at $\tau = \tilde{\tau}$, respectively. $\xi = \text{diag}\{\xi_1, \ldots, \xi_n\}$ for $\xi_i = \xi(w; \tilde{\tau}) = \zeta(w; \alpha, \gamma) \{1 - \zeta(w; \alpha, \gamma)\} \eta(w; \beta)$. $\lambda = \text{diag}\{\lambda_1, \ldots, \lambda_n\}$ for $\lambda_i = \lambda(w; \tilde{\tau}) = \zeta(w; \alpha, \gamma) \eta(w; \beta) \{1 - \eta(w; \beta)\}$. $V = \text{diag}\{\pi_1(1 - \pi_1), \ldots, \pi_n(1 - \pi_n)\}$ for $\pi_i = \pi(w; \tilde{\tau}) = \zeta(w; \alpha, \gamma) \eta(w; \beta)$, $i = 1, \ldots, n$. $X^T = (x_1^T, \ldots, x_n^T)$, $W^T = (\tilde{w}_1^T, \ldots, \tilde{w}_n^T)$. $B^T = (b_1^T, \ldots, b_n^T)$ for $b_i = (B_1(z_i), \ldots, B_q(z_i))^T$.

Theorem 3.4 (Estimation of Variances): Under conditions given in Theorem 3.2, the conditional expected information $\hat{E}_n$ is asymptotically consistent to $\mathbf{I}(\theta_0)$.

The approach can be employed to do statistical inference for the regression parameters in real applications. The proof of Theorem 3.4 is sketched in the Web Appendix by using the proof skills of [14].

4. Simulation study

Monte Carlo experiments were conducted to study the finite-sample performance of the proposed method. Two thousand replications were conducted for each experimental configuration. All the numerical experiments were performed with R-2.15.3 on a Windows 7 (64-bits) computer with 2.5 GHz Intel(R) i5-2520M CPU.

The data of binary outcome variable $Y$ were generated from the ZIB regression model that consists of the following two sub-regression models: the logistic regression model for the susceptible probability:

$$\text{logit}[\eta(w; \beta)] = \log \left[ \frac{\Pr(U = 1; \beta, w)}{1 - \Pr(U = 1; \beta, w)} \right] = \beta_{10} + \beta_{11} x_1 + \beta_{12} x_2 + \beta_{13} x_3 + \beta_2 z = \tilde{x}^T \beta_1 + \beta_2 z = \tilde{w}^T \beta,$$

(14)

where $\tilde{w} = (\tilde{x}^T, z)^T$ for $\tilde{x} = (1, x^T)^T$ and $\beta = (\beta_1^T, \beta_2)^T = (\beta_{10}, \beta_{11}, \beta_{12}, \beta_{13}, \beta_2)^T$, and the logistic regression model for the probability of occurrence of the event for susceptible
individuals:

\[
\logit [\zeta(w; \alpha, \psi)] = \log \left[ \frac{\Pr(Y = 1; \alpha, \psi, w|U = 1)}{1 - \Pr(Y = 1; \alpha, \psi, w|U = 1)} \right] = \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \psi(z) = x^T \alpha + \psi(z).
\]

(15)

Here, \( \alpha = (\alpha_1, \alpha_2, \alpha_3)^T \). \( \psi(z) \) is an unknown smooth function. Two functional forms for \( \psi(z) \) were considered.

The \( x_{i1} \)s and \( x_{i2} \)s, \( i = 1, \ldots, n = 500, 1000, \) or 1500, are values of design variables for the covariate with three nominal levels, where level 3 is used as a referent, generated from the multinomial distribution with probabilities 0.6, 0.2, and 0.2. The \( x_{i3} \)s were generated from the normal distribution with mean 1 and standard deviation 1. The \( z_i \)s were generated from the uniform \([−1, 1] \) distribution. Five scenarios were considered.

We set \( \alpha = (\alpha_1, \alpha_2, \alpha_3)^T = (-2.3, 2.5, -3.3)^T \) and \( \beta = (\beta_1^T, \beta_2^T) = (\beta_{10}, \beta_{11}, \beta_{12}, \beta_{13}, \beta_{14})^T = (-0.4, 1, 0.5, 1.1, -0.45)^T \) in the first and second scenarios, and \( \alpha = (2.5, -2, -2.8)^T \) and \( \beta = (-2, 1, 0.5, 1.3, -1.5)^T \) in the third and fourth scenarios. \( \psi(z) = 2 + 3 \sin(3z) \) and \( \psi(z) = 2.5 + 3z^2 - \log(z^2 + 1) + 2 \cos(2\pi z) \) were set in the first and third scenarios and the second and fourth scenarios, respectively. Under the parameter settings in the first and second scenarios, the average susceptible proportion was approximately 75% (i.e., the average non-susceptible proportion \( \approx 25\% \)), and approximately 25% and 37% of the susceptible individuals had the event of interest, respectively. Under the parameter settings in the third and fourth scenarios, the average susceptible proportion was approximately 50% (i.e., the average non-susceptible proportion \( \approx 50\% \)), and approximately 38% and 58% of the susceptible individuals had the event of interest, respectively. Under the parameter settings in the first four scenarios, we estimated the susceptible probabilities for an individual with \( w^*_1 = (x^*_1, z^*_1)^T = (0, 1, 0.5, 0.85)^T \), \( w^*_3 = (1, 0, 1, -0.38)^T \), and \( w^*_4 = (0, 1, 0.1, -0.7)^T \), which were \( \eta^*_1 = \eta(w^*_1; \beta) \approx 0.8824 \), \( \eta^*_2 = \eta(w^*_2; \beta) \approx 0.5665 \), \( \eta^*_3 = \eta(w^*_3; \beta) \approx 0.7047 \), and \( \eta^*_4 = \eta(w^*_4; \beta) \approx 0.4207 \), respectively. The event probabilities for a susceptible individual with \( w^*_1 = (x^*_1, z^*_1)^T = (x^*_1, x^*_2, x^*_3, z^*_1)^T = (1, 0, 0.5, 0.34)^T \), \( w^*_2 = (1, 0, -0.4, 0.27)^T \), \( w^*_3 = (1, 0, 0.5, -0.25)^T \), \( w^*_4 = (0, 1, 0.3, -0.25)^T \) were \( \zeta^*_1 = \zeta(w^*_1; \alpha, \psi) \approx 0.6471 \), \( \zeta^*_2 = \zeta(w^*_2; \alpha, \psi) \approx 0.7960 \), \( \zeta^*_3 = \zeta(w^*_3; \alpha, \psi) \approx 0.7417 \), and \( \zeta^*_4 = \zeta(w^*_4; \alpha, \psi) \approx 0.4953 \), respectively, which were estimated.

In the fifth scenario, another simulation study was conducted under the setting similar to that of the real example of fishing data given in Section 5. To this end, the data of \( x \) and \( z \) were generated from the empirical distribution of the data of number of persons and that of the data of length of stay (LOS), respectively. It is noted that the data of LOS were skewed to the left, 83.6% (209/250) of the data were less than 10 h, and 248 data points were between 0.004 and 35.592 h, but two data points were outliers and 65.34 and 71.036 h. The data of \( Y \) were then generated from the ZIB regression model in (14) and (15). Here, \( w = (x, z)^T \), \( \tilde{w} = (1, x, z)^T \) and \( \beta = (\beta_1^T, \beta_2^T) = (\beta_{10}, \beta_{11}, \beta_{12})^T \). \( \psi(\tilde{z}) \) is an unknown smooth function, where \( \tilde{z} \) is

\[
\tilde{z} = 2 \times \frac{z - 0.004}{71.036 - 0.004} - 1
\]

such that the range of \( \tilde{z} \) is from \(-1 \) to \( 1 \), where 0.004 and 71.036 are the minimum and maximum LOS, respectively. We set \( \alpha = 1.322 \), \( \beta = (-3.551, 1.400, 0.427)^T \), and \( \psi(\tilde{z}) = \psi(z) = 2 + 3 \sin(3z) \).
−3.5 + 1.95 \sin (1.2 \pi \hat{z}). Under the parameter settings, the average susceptible proportion was approximately 69% (i.e. the average non-susceptible proportion 31%), and approximately 58% of the susceptible subjects had the event of interest. Estimated were the susceptible probability for a subject with \( w_5^* = (x^*, z^*)^\top = (2, 4.1)^\top \), which was \( \hat{\eta}(w_5^*; \beta) \approx 0.7310 \) and the event probability for a susceptible subject with \( w_5^{**} = (x^{**}, z^{**})^\top = (2, 42.62)^\top \), which was \( \hat{\zeta}(w_5^{**}; \alpha, \psi) \approx 0.6175 \), where \( z^{**} = 42.62 \) corresponds to \( \hat{z}^{**} = 0.2 \).

The smooth function \( \psi(\cdot) \) was approximated by a cubic spline. The number of interior knots was determined based on the SIC value (11). After the number of interior knots was determined, the quantile method was used to choose the knot locations.

Tables 1–5 present the summary statistics for the estimates of parameters given in scenarios 1–5, respectively, from the ZIB regression model by a cubic spline, including sample mean (mean), bias, standard deviation (SD), and mean squared error (MSE). The simulation results indicated that the biases of spline estimates were very small except estimation of \( \zeta(\hat{w}_5^{**}; \alpha, \psi) \approx 0.6175 \) in scenario 5 because \( z^{**} = 42.62 \) (i.e. \( \hat{z}^{**} = 0.2 \)) in \( \hat{w}_5^{**} = (x^{**}, z^{**})^\top = (2, 42.62)^\top \) is between 35.592 and 65.34, but there are no data of LOS between 35.592 and 65.34 in the original data set of LOS. The standard deviations of the estimates were decreased at a rate of \( n^{-1/2} \) as \( n \) was increased.

We also evaluated the performance of the proposed standard error method for the spline estimates of \( (\alpha^\top, \beta^\top)^\top \) or \( (\alpha, \beta^\top)^\top \) based on the conditional expected information (13). Tables 1–5 show the standard errors of the spline estimates of \( (\alpha^\top, \beta^\top)^\top \) or \( (\alpha, \beta^\top)^\top \) given in scenarios 1–5, respectively. It was found for scenarios 1–4 that overall the proposed standard error estimation method worked reasonably well. The averages of estimated standard errors were all close to the corresponding Monte Carlo standard deviations of the estimators. The empirical coverage probabilities were not statistically significantly different from the nominal coverage probability 95\% except for \( \alpha_1 \) and \( \beta_{13} \) (\( n = 500 \) in scenario 1), \( \alpha_1 \) and \( \alpha_3 \) (\( n = 500 \) in scenario 2), \( \beta_{12} \) (\( n = 1500 \) in scenario 2), \( \alpha_2, \alpha_3, \) and \( \beta_{13} \) (\( n = 500 \) in scenario 3), \( \alpha_1, \alpha_2, \beta_{10}, \beta_{11}, \beta_{13}, \) and \( \beta_2 \) (\( n = 1000 \) in scenario 3), \( \alpha_3, \beta_{12}, \) and \( \beta_{13} \) (\( n = 1500 \) in scenario 3), \( \alpha_3 \) and \( \beta_{13} \) (\( n = 500 \) in scenario 4), \( \alpha_3 \) (\( n = 1000 \) in scenario 4), and \( \beta_{10} \) and \( \beta_{11} \) (\( n = 1500 \) in scenario 4). These empirical coverage probabilities tended to be closer to this nominal coverage probability when the sample size was increased. The performances for scenario 5 were overall little worse compared to scenarios 1–4 because the data of \( z \) were generated from the empirical distribution of the data of LOS in which the data of LOS were skewed to the left, 83.6\% (209/250) of the data are less than 10 h, and 248 data points were between 0.004 and 35.592 h, but two data points were outliers and 65.34 and 71.036 h.

Figures 1–5 depict the pointwise mean estimates of \( \psi \) and the corresponding 2.5th and 97.5th percentiles obtained from 2000 Monte Carlo samples for scenarios 1–5, respectively. The fitted curves followed closely \( \psi \), which indicates that there was little bias. The lower and upper limits of confidence intervals were also reasonably close to \( \psi \); this reveals that the variation in the estimates was small. The variability was decreased when the sample size was increased. In summary, the Monte Carlo study shows that the proposed semiparametric ZIB regression model is a practical model for analyzing ZI binary outcome data when the continuous covariate has a possibly nonlinear effect on the logarithm of the odds of the event for susceptible individuals.
Table 1. [Scenario 1: the average non-susceptible proportion $\approx 25\%$ when $\beta = (\beta^\top_1, \beta_2)^\top = (\beta_{10}, \beta_{11}, \beta_{12}, \beta_{13}, \beta_2)^\top = (-0.4, 1, 0.5, 1, -0.45)^\top$; $\approx 25\%$ of the susceptible individuals having the event when $\alpha = (\alpha_1, \alpha_2, \alpha_3)^\top = (-2.3, 2.5, -3.3)^\top$ and $\psi(z) = 2 + 3 \sin(3z)]$ Summary of parameter estimation for simulation study. $\eta^*_1 = \eta(w^*_1; \beta) \approx 0.8824$ for $w^*_1 = (x^*_1, x^*_2, x^*_3, z^*)^\top = (1, 0, 1, -0.7)^\top$. $\zeta^*_1 = \zeta(w^*_1; \alpha, \psi) \approx 0.6471$ for $w^*_1 = (x^*_1, x^*_2, x^*_3, z^*)^\top = (1, 0, 0.5, 0.34)^\top$.

|       | Mean | Bias   | SD    | MSE   | Mean | Bias   | SD    | MSE   | Mean | Bias   | SD    | MSE   |
|-------|------|--------|-------|-------|------|--------|-------|-------|------|--------|-------|-------|
| $\alpha_1$ | -2.4485 | -0.1485 | 1.0362 | 0.6919 | -2.4024 | -0.1024 | 0.6972 | 0.5146 | -2.3912 | -0.0912 | 0.6919 | 0.5388 |
| $\alpha_2$ | 2.7791 | 0.2791 | 1.0910 | 1.2683 | 2.6690 | 0.1690 | 0.6972 | 0.5146 | 2.5731 | 0.0731 | 0.5346 | 0.2912 |
| $\alpha_3$ | -3.6041 | -0.3041 | 0.9207 | 0.9403 | -3.5203 | -0.2203 | 0.6127 | 0.4239 | -3.4367 | -0.1367 | 0.4739 | 0.2432 |
| $\beta_{10}$ | -0.2546 | 0.1454 | 0.7278 | 0.5509 | -0.3609 | 0.0391 | 0.4171 | 0.1755 | -0.3832 | 0.0168 | 0.3243 | 0.1054 |
| $\beta_{11}$ | 1.1925 | 0.1925 | 0.8882 | 0.8260 | 1.0352 | 0.0352 | 0.5229 | 0.2746 | 1.0194 | 0.0194 | 0.4036 | 0.1632 |
| $\beta_{12}$ | 0.4280 | -0.0720 | 0.7600 | 0.5828 | 0.4862 | -0.0138 | 0.4493 | 0.2021 | 0.4934 | -0.0066 | 0.3587 | 0.1287 |
| $\beta_{13}$ | 1.3695 | 0.2695 | 0.6807 | 0.5360 | 1.1806 | 0.0806 | 0.3791 | 0.1502 | 1.1451 | 0.0451 | 0.2911 | 0.0868 |
| $\beta_2$ | -0.6537 | -0.2037 | 0.7655 | 0.6275 | -0.5144 | -0.0644 | 0.4438 | 0.2011 | -0.4684 | -0.0184 | 0.3414 | 0.1169 |
| $\eta^*_1$ | 0.8746 | -0.0280 | 0.1293 | 0.0168 | 0.8665 | -0.0159 | 0.0993 | 0.0101 | 0.8682 | -0.0142 | 0.0856 | 0.0075 |
| $\zeta^*_1$ | 0.6191 | -0.0280 | 0.1616 | 0.0269 | 0.6325 | -0.0147 | 0.1221 | 0.0151 | 0.6248 | -0.0223 | 0.1003 | 0.0106 |
|       | Mean.se | SD.se | CP | Mean.se | SD.se | CP | Mean.se | SD.se | CP |
| $\alpha_1$ | 0.9576 | 0.3146 | 93.4% | 0.6529 | 0.1488 | 94.7% | 0.5221 | 0.0974 | 95.4% |
| $\alpha_2$ | 0.9802 | 0.3020 | 94.6% | 0.6551 | 0.1324 | 94.7% | 0.5147 | 0.0759 | 94.8% |
| $\alpha_3$ | 0.8451 | 0.3243 | 94.0% | 0.5700 | 0.1461 | 95.1% | 0.4462 | 0.0890 | 94.7% |
| $\beta_{10}$ | 0.6326 | 0.2646 | 95.3% | 0.3991 | 0.0980 | 95.3% | 0.3162 | 0.0539 | 95.0% |
| $\beta_{11}$ | 0.8066 | 0.3633 | 95.7% | 0.4977 | 0.0919 | 95.0% | 0.3986 | 0.0544 | 95.1% |
| $\beta_{12}$ | 0.6844 | 0.2094 | 94.8% | 0.4386 | 0.0727 | 95.0% | 0.3495 | 0.0411 | 94.6% |
| $\beta_{13}$ | 0.5710 | 0.3051 | 96.1% | 0.3504 | 0.0956 | 95.1% | 0.2775 | 0.0523 | 94.6% |
| $\beta_2$ | 0.6536 | 0.2791 | 95.0% | 0.4129 | 0.0994 | 95.2% | 0.3271 | 0.0540 | 94.0% |

Sample mean (Mean), bias, standard deviation (SD), and mean squared error (MSE) of the semiparametric zero-inflated Bernoulli regression model by a cubic spline. Results of variance study. Sample mean of estimated standard errors (mean.se), standard deviation of estimated standard errors (SD.se), and coverage probability of 95% confidence interval (CP) of $(\alpha^\top, \beta^\top)^\top$ obtained from 2000 Monte Carlo samples with sample size 500, 1000 or 1500.
### Table 2. Scenario 2: the average non-susceptible proportion $\approx 25\%$ when $\beta = (\beta_1, \beta_2)^T = (\beta_{10}, \beta_{11}, \beta_{12}, \beta_{13})^T = (-0.4, 1, 0.5, -0.45)^T; \approx 37\%$ of the susceptible individuals having the event when $\alpha = (\alpha_1, \alpha_2, \alpha_3)^T = (-2.3, 2.5, -3.3)^T$ and $\psi(z) = 2.5 + 3z^2 - \log(z/2 + 1) + 2 \cos(2\pi z)$] Summary of parameter estimation for simulation study. $\eta_2^* = \eta(w_2^*; \beta) \approx 0.5665$ for $w_2^* = (x_1^*, x_2^*, z_1)^T = (0, 1, 0.5, 0.85)^T$. $\zeta_2^* \approx 0.7960$ for $w_2^* = (x_1^*, x_2^*, z_1)^T = (1, 0, -0.4, 0.27)^T$.

|       | $n = 500$ |       | $n = 1000$ |       | $n = 1500$ |
|-------|-----------|-------|-----------|-------|-----------|
|       | Mean    | Bias  | SD       | MSE   | Mean    | Bias  | SD       | MSE   | Mean    | Bias  | SD       | MSE   |
| $\alpha_1$ | -2.3381 | -0.0381 | 0.8368 | 0.7018 | -2.3550 | -0.0550 | 0.5587 | 0.3151 | -2.3608 | -0.0608 | 0.4461 | 0.2027 |
| $\alpha_2$ | 2.6099  | 0.1099 | 0.9193 | 0.8572 | 2.5838  | 0.0838 | 0.6074 | 0.3760 | 2.5693  | 0.0693 | 0.4908 | 0.2457 |
| $\alpha_3$ | -3.4382 | -0.1382 | 0.7313 | 0.5539 | -3.4114 | -0.1141 | 0.4897 | 0.2528 | -3.3954 | -0.0954 | 0.3746 | 0.1949 |
| $\beta_{10}$ | -0.3455 | 0.0545 | 0.4799 | 0.2333 | -0.3809 | 0.0191 | 0.2888 | 0.0837 | -0.3975 | 0.0025 | 0.2344 | 0.0549 |
| $\beta_{11}$ | 1.1155  | 0.1155 | 0.6734 | 0.4668 | 1.0337  | 0.0337 | 0.3978 | 0.1594 | 1.0264  | 0.0264 | 0.3247 | 0.1061 |
| $\beta_{12}$ | 0.4673  | -0.0327 | 0.5691 | 0.3249 | 0.4824  | -0.0176 | 0.3584 | 0.1287 | 0.5009  | 0.0009 | 0.3052 | 0.0931 |
| $\beta_{13}$ | 1.2884  | 0.1884 | 0.4512 | 0.2390 | 1.1674  | 0.0674 | 0.2681 | 0.0764 | 1.1408  | 0.0408 | 0.2030 | 0.0429 |
| $\beta_2$ | -0.4595 | -0.0095 | 0.3575 | 0.1279 | -0.4620 | -0.0120 | 0.2210 | 0.0490 | -0.4549 | -0.0049 | 0.1755 | 0.0308 |
| $\eta_2^*$ | 0.5866  | 0.0202 | 0.1100 | 0.0125 | 0.5709  | 0.0044 | 0.0726 | 0.0053 | 0.5703  | 0.0038 | 0.0580 | 0.0034 |
| $\zeta_2^*$ | 0.7999  | 0.0039 | 0.1350 | 0.0182 | 0.8023  | 0.0063 | 0.0966 | 0.0094 | 0.7988  | 0.0028 | 0.0751 | 0.0057 |

Sample mean (Mean), bias, standard deviation (SD), and mean squared error (MSE) of the semiparametric zero-inflated Bernoulli regression model by a cubic spline. Results of variance study. Sample mean of estimated standard errors (Mean.se), standard deviation of estimated standard errors (SD.se), and coverage probability of 95% confidence interval (CP) of $(\alpha_1, \beta_1^T)^T$ obtained from 2000 Monte Carlo samples with sample size 500, 1000 or 1500.
Table 3. [Scenario 3:] the average non-susceptible proportion $\approx 50\%$ when $\beta = (\beta_1^T, \beta_2)^T = (\beta_{10}, \beta_{11}, \beta_{12}, \beta_{13}, \beta_{2})^T = (-2, 1, 0.5, 1.3, -1.5)^T$; $\approx 38\%$ of the susceptible individuals having the event when $\alpha = (\alpha_1, \alpha_2, \alpha_3)^T = (2.5, -2, -2.8)^T$ and $\psi(z) = 2 + 3 \sin(3z)]$. Summary of parameter estimation for simulation study. $\eta^*_3 = \eta(w^*_3; \beta) \approx 0.7047$ for $w^*_3 = (x^*_1, x^*_2, x^*_3)^T$. $\zeta^* = \zeta(w^*_3; \alpha, \psi) \approx 0.7417$ for $w^*_3 = (x^*_1, x^*_2, x^*_3)^T = (1, 0, 0.5, -0.25)^T$.

|        | $n = 500$ |        |        |        |        |        |        |        |        |        |        |        |        |        |        |
|--------|----------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| $\alpha_1$ | 2.7187  | 0.2187 | 1.1024 | 1.2632 | 2.6642 | 0.1642 | 0.7671 | 0.6154 | 2.6245 | 0.1245 | 0.6500 | 0.4380 |
| $\alpha_2$ | -1.9309 | 0.0691 | 1.5854 | 2.5183 | -1.9766 | 0.0234 | 1.1329 | 1.2841 | -1.9824 | 0.0176 | 0.9686 | 0.9385 |
| $\alpha_3$ | -2.8301 | 0.0301 | 0.7642 | 0.5849 | -2.8622 | 0.0622 | 0.5246 | 0.2791 | -2.8597 | 0.0597 | 0.4352 | 0.1767 |
| $\beta_{10}$ | -1.7636 | 0.2364 | 0.8185 | 0.7258 | -1.9018 | 0.0982 | 0.4906 | 0.2503 | -1.9383 | 0.0617 | 0.4087 | 0.1078 |
| $\beta_{11}$ | 0.7637  | -0.2363 | 0.8224 | 0.7322 | 0.8997  | -0.1003 | 0.4936 | 0.2537 | 0.9325  | -0.0675 | 0.4160 | 0.1776 |
| $\beta_{12}$ | 0.6968  | 0.1968 | 1.9646 | 2.2778 | 0.5564  | 0.0564 | 0.8909 | 0.7969 | 0.4966  | -0.0034 | 0.7060 | 0.4985 |
| $\beta_{13}$ | 1.6514  | 0.3514 | 0.4794 | 0.3532 | 1.4094  | 0.1094 | 0.2715 | 0.0857 | 1.3661  | 0.0661 | 0.2113 | 0.0490 |
| $\beta_{2}$ | -1.8803 | -0.3803 | 0.6661 | 0.5883 | -1.6215 | -0.1215 | 0.3644 | 0.1476 | -1.5650 | -0.0650 | 0.2956 | 0.0916 |
| $\eta^*_3$ | 0.7743  | 0.0695 | 0.0981 | 0.0145 | 0.7285  | 0.0238 | 0.0715 | 0.0057 | 0.7179  | 0.0131 | 0.0597 | 0.0037 |
| $\zeta^{**}$ | 0.7099  | -0.0318 | 0.1479 | 0.0229 | 0.7513  | 0.0095 | 0.1020 | 0.0105 | 0.7597  | 0.0180 | 0.0860 | 0.0077 |
|        | Mean.se  | SD.se  | CP         |        | Mean.se  | SD.se  | CP         |        | Mean.se  | SD.se  | CP         |        |
| $\alpha_1$ | 0.6111  | 0.3329 | 95.9%      |        | 0.6766  | 0.1443 | 96.1%      |        | 0.6216  | 0.0935 | 94.9%      |        |
| $\alpha_2$ | 1.5933  | 0.4680 | 96.7%      |        | 1.1589  | 0.2291 | 96.1%      |        | 0.9463  | 0.1513 | 95.0%      |        |
| $\alpha_3$ | 0.7655  | 0.2403 | 93.4%      |        | 0.5507  | 0.1221 | 95.8%      |        | 0.4473  | 0.0760 | 96.2%      |        |
| $\beta_{10}$ | 0.7886  | 0.2926 | 95.3%      |        | 0.5175  | 0.1179 | 96.3%      |        | 0.4131  | 0.0742 | 95.5%      |        |
| $\beta_{11}$ | 0.7919  | 0.2760 | 95.4%      |        | 0.5219  | 0.1100 | 96.3%      |        | 0.4172  | 0.0694 | 94.8%      |        |
| $\beta_{12}$ | 1.4180  | 0.9292 | 94.8%      |        | 0.8994  | 0.2255 | 95.0%      |        | 0.7144  | 0.1298 | 96.4%      |        |
| $\beta_{13}$ | 0.4606  | 0.1799 | 96.5%      |        | 0.2777  | 0.0641 | 97.2%      |        | 0.2187  | 0.0371 | 96.7%      |        |
| $\beta_{2}$ | 0.6086  | 0.2243 | 96.0%      |        | 0.3733  | 0.0792 | 97.4%      |        | 0.2953  | 0.0482 | 95.8%      |        |

Sample mean (Mean), bias, standard deviation (SD), and mean squared error (MSE) of the semiparametric zero-inflated Bernoulli regression model by a cubic spline. Results of variance study. Sample mean of estimated standard errors (mean.se), standard deviation of estimated standard errors (SD.se), and coverage probability of 95% confidence interval (CP) of $(\alpha^T, \beta^T)^T$ obtained from 2000 Monte Carlo samples with sample size 500, 1000 or 1500.
Table 4. [Scenario 4: the average non-susceptible proportion \( \approx 50\% \) when \( \beta = (\beta_1^T, \beta_2)^T = (\beta_{10}, \beta_{11}, \beta_{12}, \beta_{13}, \beta_2)^T = (-2, 1, 0.5, 1.3, -1.5)^T; \approx 58\% \) of the susceptible individuals having the event when \( \alpha = (\alpha_1, \alpha_2, \alpha_3)^T = (2.5, -2, -2.8)^T \) and \( \psi(z) = 2.5 + 3z^2 - \log(z/2 + 1) + 2 \cos(2\pi z) \)] Summary of parameter estimation for simulation study. \( \eta^*_4 = \eta(w^*_4; \beta) \approx 0.4207 \) for \( w^*_4 = (x^*_1, x^*_2, x^*_3, z^*_1)^T = (0, 1, 0.3, -0.25)^T \). \( \zeta^{**}_4 = \zeta(w^{**}_4; \alpha, \psi) \approx 0.4953 \) for \( w^{**}_4 = (x^{**}_1, x^{**}_2, x^{**}_3, z^{**})^T = (0, 1, 0.3, -0.25)^T \).

| Parameter | Mean | Bias | SD | MSE | Mean | Bias | SD | MSE | Mean | Bias | SD | MSE |
|-----------|------|------|----|-----|------|------|----|-----|------|------|----|-----|
| \( \alpha_1 \) | 2.4560 | -0.0440 | 0.8547 | 0.7325 | 2.4586 | -0.0414 | 0.6260 | 0.3151 | 2.4673 | -0.0327 | 0.5016 | 0.2527 |
| \( \alpha_2 \) | -1.9754 | 0.0246 | 1.1273 | 1.2714 | -1.9692 | 0.0308 | 0.7822 | 0.3760 | -1.9885 | 0.0115 | 0.6584 | 0.4336 |
| \( \alpha_3 \) | -2.5858 | 0.2142 | 0.6092 | 0.4170 | -2.6834 | 0.1166 | 0.4436 | 0.2528 | -2.7295 | 0.0705 | 0.3942 | 0.1269 |
| \( \beta_{10} \) | -1.9055 | 0.0945 | 0.5605 | 0.3231 | -1.9502 | 0.0498 | 0.3605 | 0.1325 | -1.9725 | 0.0275 | 0.2693 | 0.0733 |
| \( \beta_{11} \) | 0.8517 | -0.1483 | 0.5799 | 0.3583 | 0.9287 | -0.0713 | 0.3718 | 0.1433 | 0.9590 | -0.0410 | 0.2792 | 0.0796 |
| \( \beta_{12} \) | 0.7184 | 0.2184 | 1.0035 | 1.0548 | 0.5684 | 0.0684 | 0.5797 | 0.3408 | 0.5260 | 0.0260 | 0.4562 | 0.2088 |
| \( \beta_{13} \) | 1.5949 | 0.2949 | 0.3794 | 0.2309 | 1.4145 | 0.1145 | 0.2104 | 0.0574 | 1.3624 | 0.0624 | 0.1630 | 0.0305 |
| \( \beta_2 \) | -1.5701 | 0.0701 | 0.3348 | 0.1170 | -1.5266 | 0.0266 | 0.2046 | 0.0426 | -1.5103 | 0.0103 | 0.1608 | 0.0260 |
| \( \eta_4 \) | 0.5096 | 0.0889 | 0.1729 | 0.0378 | 0.4592 | 0.0386 | 0.1158 | 0.0149 | 0.4391 | 0.0184 | 0.0939 | 0.0092 |
| \( \zeta^{**}_4 \) | 0.4354 | -0.0599 | 0.2118 | 0.0484 | 0.4775 | -0.0178 | 0.1671 | 0.0094 | 0.4949 | -0.0004 | 0.1392 | 0.0194 |

Sample mean (Mean), bias, standard deviation (SD), and mean squared error (MSE) of the semiparametric zero-inflated Bernoulli regression model by a cubic spline. Results of variance study. Sample mean of estimated standard errors (mean.se), standard deviation of estimated standard errors (SD.se), and coverage probability of 95% confidence interval (CP) of \((\alpha^T, \beta^T)^T\) obtained from 2000 Monte Carlo samples with sample size 500, 1000 or 1500.
Table 5. [Scenario 5: the average non-susceptible proportion $\approx 31\%$ when $\beta = (\beta_1^T, \beta_2)^T = (\beta_{10}, \beta_{11}, \beta_2)^T = (-3.551, 1.4, 0.427)^T$; $\approx 58\%$ of the susceptible individuals having the event when $\alpha = 1.322$ and $\psi(\bar{z}) = -3.5 + 1.95 \sin(1.2\pi \bar{z})$, where $\bar{z} = 2 \times \{z - 0.004\}/(71.036 - 0.004) - 1\] Summary of parameter estimation for simulation study. $\eta^*_5 = \eta(w^*_5; \beta) \approx 0.7310$ for $w^*_5 = (x^*, z^*)^T = (2, 4.1)^T$. $\zeta^{**}_5 = \zeta(w^{**}_5; \alpha, \psi) \approx 0.6175$ for $w^{**}_5 = (x^{**}, z^{**})^T = (2, 42.62)^T$.

|        | $n = 500$       |       |       |       | $n = 1000$       |       |       |       | $n = 1500$       |       |       |       |
|--------|-----------------|-------|-------|-------|-----------------|-------|-------|-------|-----------------|-------|-------|-------|
| $\alpha$ | 1.2822         | -0.0398 | 0.2097 | 0.0456 | 1.2846         | -0.0374 | 0.1432 | 0.0219 | 1.2888         | -0.0332 | 0.1173 | 0.0149 |
| $\beta_{10}$ | -4.3403        | -0.7893 | 0.8667 | 1.3741 | -4.1303        | -0.5793 | 0.5720 | 0.6628 | -4.0685        | -0.5175 | 0.4772 | 0.4955 |
| $\beta_{11}$ | 1.6303         | 0.2303 | 0.3208 | 0.1559 | 1.5497         | 0.1497 | 0.2084 | 0.0659 | 1.5235         | 0.1235 | 0.1676 | 0.0433 |
| $\beta_2$ | 0.6161         | 0.1891 | 0.2842 | 0.1165 | 0.5300         | 0.1030 | 0.2079 | 0.0538 | 0.4977         | 0.0707 | 0.1820 | 0.0381 |
| $\eta^*_5$ | 0.7698         | 0.0388 | 0.1390 | 0.0208 | 0.7351         | 0.0041 | 0.1202 | 0.0145 | 0.7174         | -0.0136 | 0.1131 | 0.0130 |
| $\zeta^{**}_5$ | 0.4030         | -0.2145 | 0.1766 | 0.0772 | 0.3999         | -0.2176 | 0.1338 | 0.0652 | 0.4025         | -0.2150 | 0.1102 | 0.0584 |

|        | Mean, Bias, SD, MSE | Mean, Bias, SD, MSE | Mean, Bias, SD, MSE |
|--------|---------------------|---------------------|---------------------|
|        | Mean, SD, CP        | Mean, SD, CP        | Mean, SD, CP        |
| $\alpha$ | 0.2394, 0.0478, 96.3% | 0.1743, 0.0325, 97.4% | 0.1449, 0.0259, 98.0% |
| $\beta_{10}$ | 1.1339, 0.1406, 93.0% | 0.7857, 0.0773, 93.0% | 0.6468, 0.0485, 93.3% |
| $\beta_{11}$ | 0.3695, 0.1128, 96.5% | 0.2427, 0.0703, 96.1% | 0.1908, 0.0455, 95.2% |
| $\beta_2$ | 0.3778, 0.1367, 98.0% | 0.2711, 0.0815, 98.0% | 0.2274, 0.0652, 97.5% |

Sample mean (Mean), bias, standard deviation (SD), and mean squared error (MSE) of the semiparametric zero-inflated Bernoulli regression model by a cubic spline. Results of variance study. Sample mean of estimated standard errors (mean.se), standard deviation of estimated standard errors (SD.se), and coverage probability of 95% confidence interval (CP) of $(\alpha, \beta^T)^T$, based on 2000 Monte Carlo samples with sample size 500, 1000 or 1500.
Figure 1. [Scenario 1: the average non-susceptible proportion $\approx 25\%$ when $\beta = (-0.4, 1, 0.5, 1.1, -0.45)^T$; $\approx 32\%$ of the susceptible individuals having the event when $\alpha = (-2.3, 2.5, -3.3)^T$ and $\psi(z) = 2 + 3\sin(3z)$] Estimates and corresponding 95% pointwise confidence intervals of $\psi(z) = 2 + 3\sin(3z)$. The solid curves are the true function. The dashed curves are the average cubic-spline fits and the long dashed curves are the corresponding the 2.5th and 97.5th percentiles of the estimates obtained from 2000 Monte Carlo samples.

5. Real example: fishing data

We illustrate the practical use of the proposed methodology with the fishing data set available at https://github.com/rmcelreath/rethinking/blob/master/data/Fish.csv. This data set consists of 250 groups that went to a state park. Each group was asked how many fish they caught, whether or not they used live bait, how many children were in the group, how many people were in the group (persons), how long they stayed (LOS), and whether or not they brought a camper to the park. To illustrate the practicality of the proposed methodology, we are mainly interested in whether or not they caught one or more fish, denoted by $Y$, which is the binary outcome variable, and investigating the effects of persons, denoted by $x$, and LOS, denoted by $z$. 
Figure 2. [Scenario 2: the average non-susceptible proportion \( \approx 25\% \) when \( \beta = (-0.4, 1, 0.5, 1.1, -0.45)^T \); \( \approx 44\% \) of the susceptible individuals having the event when \( \alpha = (-2.3, 2.5, -3.3)^T \) and \( \psi(z) = 2.5 + 3z^2 - \log(z/2 + 1) + 2\cos(2\pi z) \)]

Estimates and corresponding 95% pointwise confidence intervals of \( \psi(z) = 2.5 + 3z^2 - \log(z/2 + 1) + 2\cos(2\pi z) \). The solid curves are the true function. The dashed curves are the average cubic-spline fits and the long dashed curves are the corresponding the 2.5th and 97.5th percentiles of the estimates obtained from 2000 Monte Carlo samples.

In this study, some groups did not even fish, so these groups had systematically zero fish. Among the 250 groups, 142 groups (56.8%) did not catch fish. To describe the potentially nonlinear effect of LOS \( (z) \) on the logarithm of the odds of catching one or more fish for a group going fishing, we fit the proposed semiparametric ZIB regression model to the fishing data, which consists of the logistic regression model for the probability that a group went fishing \( \text{logit}[\eta(\mathbf{w} ; \beta)] = \mathbf{w}^T \beta \), where \( \mathbf{w} = (x, z)^T \), \( \hat{\mathbf{w}} = (1, x, z)^T \) and \( \beta = (\beta_1, \beta_2, \beta_3)^T \), and the logistic regression model for the probability of catching one or more fish for the group going fishing \( \text{logit}[\zeta(\mathbf{w} ; \alpha, \psi)] = \alpha x + \psi(z) \). The unknown smooth function \( \psi(z) \) was approximated by a cubic spline with 1 interior knot that was selected based on the SIC value (11). The average of the estimated probabilities, \( \eta(\mathbf{w}_i ; \hat{\beta}) \)'s, of going finishing was 0.722. The average of the estimated probabilities, \( \zeta(\mathbf{w}_i ; \hat{\alpha}, \hat{\psi}_i) \)'s, of catching one or more fish for groups going fishing was 0.601, where \( \hat{\psi}_i = \hat{\psi}(z_i) \). Table 6
Figure 3. [Scenario 3:] the average non-susceptible proportion ≈ 50% when $\beta = (-2, 1, 0.5, 1.3, -1.5)^T$; ≈ 53% of the susceptible individuals having the event when $\alpha = (2.5, -2, -2.8)^T$ and $\psi(z) = 2 + 3\sin(3z)$] Estimates and corresponding 95% pointwise confidence intervals of $\psi(z) = 2 + 3\sin(3z)$. The solid curves are the true function. The dashed curves are the average cubic-spline fits and the long dashed curves are the corresponding the 2.5th and 97.5th percentiles of the estimates obtained from 2000 Monte Carlo samples.

gives the results of fishing data analysis by fitting the semiparametric ZIB regression model and a parametric ZIB regression model. To compare the prediction performance of these two models in terms of prediction error, we used the cross-validation technique to estimate the prediction errors of these two models. The 25-fold cross-validation estimates of prediction errors of the semiparametric ZIB regression model and the parametric ZIB regression model were 0.151 and 0.154, respectively. This indicates that the prediction performance of the proposed semiparametric ZIB regression model used to fit the fishing data is slightly better than that of the parametric ZIB regression model. For model diagnostics and goodness-of-fit, Figure 6 provides normal Q–Q plots of Pearson residuals for semiparametric and parametric ZIB regression models. It is seen from the two plots that semiparametric ZIB regression model gives a slightly better fit.
Figure 4. Scenario 4: the average non-susceptible proportion $\approx 50\%$ when $\beta = (-2, 1, 0.5, 1.3, -1.5)^T$; $\approx 67\%$ of the susceptible individuals having the event when $\alpha = (2.5, -2, -2.8)^T$ and $\psi(z) = 2.5 + 3z^2 - \log((z/2 + 1) + 2 \cos(2\pi z))$. Estimates and corresponding 95\% pointwise confidence intervals of $\psi(z) = 2.5 + 3z^2 - \log((z/2 + 1) + 2 \cos(2\pi z))$. The solid curves are the true function. The dashed curves are the average cubic-spline fits and the long dashed curves are the corresponding the 2.5th and 97.5th percentiles of the estimates obtained from 2000 Monte Carlo samples.

To statistically and quantitatively assess which one of the parametric and semiparametric ZIB regression models provides a better fit to the fishing data, i.e. test the null hypothesis of the parametric ZIB regression model versus the alternative hypothesis of the semiparametric ZIB regression model, we compared their log-likelihoods by using the LR test. The log-likelihoods of the parametric and semiparametric ZIB regression models were $-103.4975$ and $-99.8085$, respectively, and the observed value of the LR test statistic was $7.378$ with the corresponding $p$-value of 0.061 based on the asymptotic chi-square distribution with 3 degrees of freedom. The $p$-value indicates that the null hypothesis of the parametric ZIB regression model for this data set is not rejected at the significance level of 0.05, but is rejected at the significance level of 0.1.
Figure 5. [Scenario 5]: the average non-susceptible proportion $\approx 31\%$ when $\beta = (-3.551, 1.4, 0.427)^T$; $\approx 58\%$ of the susceptible individuals having the event when $\alpha = 1.322$ and $\psi(\tilde{z}) = -3.5 + 1.95 \sin(1.2\pi \tilde{z})$, where $\tilde{z} = 2 \times \{(z - 0.004)/(71.036 - 0.004)\} - 1$] Estimates and corresponding 95% pointwise confidence intervals of $\psi(\tilde{z}) = -3.5 + 1.95 \sin(1.2\pi \tilde{z})$. The solid curves are the true function. The dashed curves are the average cubic-spline fits and the long dashed curves are the corresponding the 2.5th and 97.5th percentiles of the estimates obtained from 2000 Monte Carlo samples.

We also used a parametric bootstrap approach to approximate the distribution of the LR test statistic to test the null hypothesis of the parametric ZIB regression model for the fishing data set, denoted by $\{(y_i, w_i^T), i = 1, \ldots, 250\}$. The steps of the parametric bootstrap approach are described as follows:

Step 1. Generate a bootstrap sample $y_i^*$ from the fitted parametric ZIB regression model. Step 2. For each bootstrap sample $\{(y_i^*, w_i^T), i = 1, \ldots, 250\}$, fit the parametric ZIB regression model and the semiparametric ZIB regression model with a cubic spline with 1 interior knot to the bootstrap sample to obtain their log-likelihoods to compute the value of the bootstrap LR test statistic.
Figure 6. Normal quantile–quantile (Q–Q) plots of Pearson residuals for semiparametric and parametric zero-inflated Bernoulli (ZIB) regression models.

Table 6. Applications of semiparametric zero-inflated Bernoulli (ZIB) regression model and parametric ZIB regression model to fishing data.

| Model            | Incidence part | Susceptible part |
|------------------|----------------|------------------|
|                  | Semiparametric ZIB regression | Logistic regression |
|                  | Parametric ZIB regression | Logistic regression |
| Variable         | Estimate  | SE   | P-value | Estimate  | SE   | P-value |
| Intercept        | -11.394  | 3.934 | 0.004   | -2.213  | 0.499 | < 0.001 |
| Persons (x)      | 1.322    | 0.158 | < 0.001 | 1.230    | 0.235 | < 0.001 |
| LOS (z)          | 12.318   | 4.030 | 0.002   | 0.029    | 0.022 | 0.188   |

Abbreviations: For semiparametric ZIB regression, semiparametric logistic regression in the incidence part, semiparametric logistic regression model for the probability of catching one or more fish for a group going fishing; logistic regression in the susceptible part, logistic regression model for the probability that a group went fishing. For parametric ZIB regression, logistic regression in the incidence part, logistic regression model for the probability of catching one or more fish for a group going fishing; logistic regression in the susceptible part, logistic regression model for the probability that a group went fishing. Estimate, estimate of coefficient; SE, standard error; Persons, the number of people in a group; Length of stay (LOS), length a group stayed at the state park. Log-likelihood of the fitted semiparametric ZIB regression model (a cubic spline with 1 interior knot) = -99.8085. Log-likelihood of the fitted parametric ZIB regression model = -103.4975.

Step 3. Repeat Steps 1 and 2 independently 2000 times to obtain 2000 values of the bootstrap LR test statistic.

Step 4. Because the observed value of the LR test statistic was 7.378 with the corresponding p-value of 0.067 based on the empirical distribution of 2000 values of the bootstrap LR test statistic, it indicates that the null hypothesis of the parametric ZIB regression model for this data set is not rejected at the significance level of 0.05, but is rejected.
Figure 7. The estimate with 95% pointwise confidence intervals of $\psi(\text{length of stay})$ obtained with a cubic spline of 1 interior knot. The solid curve is the estimated function. The dashed curves are the corresponding 95% confidence limits based on the 2.5th and 97.5th percentiles of the estimates obtained from 2000 bootstrap samples.

at the significance level of 0.1. Thus, based on the two $p$-values, 0.061 and 0.067, the semiparametric ZIB regression model is marginally better than the parametric ZIB regression model for fitting the fishing data at the significance level of 0.05.

It can be seen from Table 6 that the probability that a group went fishing was increased when the number of people in the group and the length of the group stay was increased. The probability of catching one or more fish for a group going fishing was increased when the number of people in the group was increased. Figure 7 depicts the estimate of the unknown smooth function $\psi$ of LOS with 95% pointwise confidence intervals obtained from 2000 bootstrap samples when fitting the semiparametric ZIB regression model with the cubic spline of 1 interior knot. It can be seen from Figure 7 that the estimate of $\psi$ indicates that the probability of catching one or more fish for a group going fishing was not necessarily increased when the length of the group stay was increased. A plausible explanation is that the group stayed longer in the park does not necessarily imply that the group spent more time fishing if the group went fishing. Some groups that had a long stay in the park only spent a short amount of time fishing if they went fishing. Additionally, based on the analyses of the prediction error, normal Q–Q plot, and LR test, the proposed semiparametric ZIB regression is (slightly) better than the parametric model, but the nonlinear relationship with LOS is obvious. A plausible explanation of this phenomenon is again that the data of LOS were skewed to the left, 83.6% (209/250) of the data were less than 10 h, and 248 data points were between 0.004 and 35.592 h, but two data points were outliers and 65.34 and 71.036 h.

6. Conclusion

A semiparametric ZIB regression model has been proposed to fit binary outcome data with excess zeros and to depict the potentially nonlinear effect of a continuous covariate on the logarithm of the odds of an event of interest for a susceptible individual. An unknown
smooth function, which is used to describe this potentially nonlinear effect, is approximated by a spline function that can be expressed as a linear combination of B-spline basis functions. The SIC ([19]) has been used to determine the optimal number of interior knots, and the locations of the knots have been chosen by the quantile method. The asymptotic properties of the spline estimators have been established. A direct and consistent variance estimation method based on the least-squares method has also been proposed. The proposed methodology can be easily extended to describe the effects of several continuous covariates of interest. The functional form for each covariate effect is assumed to be smooth but unknown. We can incorporate the effects of these covariates in an additive fashion, and splines are used to model each covariate effect.

An extensive simulation study has been conducted to investigate the finite-sample performance of the proposed methodology. The standard deviations of the estimates are decreased at a rate of \( n^{-1/2} \). For all sample sizes considered in the simulation study, the empirical coverage probabilities tend to be closer to the nominal probability 95%, and the behaviors of the empirical coverage probabilities are similar. Overall, the proposed standard error method for the spline estimates of the regression parameters works reasonably well. The averages of estimated standard errors are all close to the corresponding Monte Carlo standard deviations of the estimators. The fitted curves follow closely the true curve, and the lower and upper limits of confidence intervals are also close to the true function. The simulation study shows that the proposed semiparametric ZIB regression model is a practical model for analyzing ZI binary outcome data when a continuous covariate has a possibly nonlinear effect on the logarithm of the odds of an event of interest for susceptible individuals. Finally, the fishing data set has been used to illustrate the practicality of the proposed methodology. The fishing data analysis indicates that the prediction performance of the proposed semiparametric ZIB regression model is better than that of the parametric ZIB regression model.

**Acknowledgments**

The authors are very grateful for helpful comments and suggestions from two referees and an Associate Editor that improved the presentation. This work was presented by C.S. Li at the 3rd International Conference on Econometrics and Statistics, National Chung Hsing University, Taiwan, ROC. http://cmstatistics.org/RegistrationsV2/EcoSta2019/viewSubmission.php?in=233&token=orpno0-o2pq935r956469q97p5n819rs1.

**Disclosure statement**

No potential conflict of interest was reported by the authors.

**References**

[1] H. Akaike, *A new look at the statistical model identification*, IEEE. Trans. Automat. Control 19 (1974), pp. 716–723.

[2] P.J. Bickel, C.A. Klaassen, Y. Ritov, J. Klaassen, J.A. Wellner, and Y.A. Ritov, *Efficient and Adaptive Estimation for Semiparametric Models*, Johns Hopkins University Press, Baltimore, 1993.

[3] R.H. Byrd, P. Lu, J. Nocedal, and C. Zhu, *A limited memory algorithm for bound constrained optimization*, SIAM J. Sci. Comput. 16 (1995), pp. 1190–1208.
[4] A. Diop, A. Diop, and J.F. Dupuy, *Maximum likelihood estimation in the logistic regression model with a cure fraction*, Electron. J. Stat. 5 (2011), pp. 460–483.

[5] A. Diop, A. Diop, and J.F. Dupuy, *Simulation-based inference in a zero-inflated Bernoulli regression model*, Commun. Stat. Simul. Comput. 45 (2016), pp. 3597–3614.

[6] J. Feng and Z. Zhu, *Semiparametric analysis of longitudinal zero-inflated count data*, J. Multivar. Anal. 102 (2011), pp. 61–72.

[7] D.B. Hall, *Zero-inflated Poisson and binomial regression with random effects: A case study*, Biometrics 56 (2000), pp. 1030–1039.

[8] X. He, H. Xue, and N.Z. Shi, *Sieve maximum likelihood estimation for doubly semiparametric zero-inflated Poisson models*, J. Multivar. Anal. 101 (2010), pp. 2026–2038.

[9] D.W. Hosmer, S. Lemeshow, and R.X. Sturdivant, *Applied Logistic Regression*, 3rd ed., John Wiley & Sons, New York, 2013.

[10] D.G. Kleinbaum and M. Klein, *Logistic Regression*, 3rd ed., Springer Verlag, New York, 2010.

[11] A.B. Koehler and E.S. Murphree, *A comparison of the Akaike and Schwarz criteria for selecting model order*, J. R. Stat. Soc. Ser C 37 (1988), pp. 187–195.

[12] K.F. Lam, H. Xue, and Y.B. Cheung, *Semiparametric analysis of zero-inflated count data*, Biometrics 62 (2006), pp. 996–1003.

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

[14] M. Lu and C.S. Li, *Spline-based semiparametric estimation of a zero-inflated Poisson regression single-index model*, Ann. Inst. Stat. Math. 68 (2016), pp. 1111–1134.

[15] M. Lu and D. Loomis, *Spline-based semiparametric estimation of partially linear Poisson regression with single-index models*, J. Nonparametr. Stat. 25 (2013), pp. 905–922.

[16] Y. Min and A. Agresti, *Random effect models for repeated measures of zero-inflated count data*, Stat. Model. 5 (2005), pp. 1–19.

[17] R.A. Redner and H.F. Walker, *Mixture densities, maximum likelihood and the EM algorithm*, SIAM Rev. 26 (1984), pp. 195–239.

[18] L. Schumaker, *Spline Functions: Basic Theory*, Wiley & Son, New York, 1981.

[19] G. Schwarz, *Estimating the dimension of a model*, Ann. Stat. 6 (1978), pp. 461–464.

[20] D.S. Young, A.M. Raim, and N.R. Johnson, *Zero-inflated modelling for characterizing coverage errors of extracts from the US Census Bureau’s Master Address File*, J. R. Stat. Soc. Ser. A 180 (2017), pp. 73–97.

[21] Y. Zhang, L. Hua, and J. Huang, *A spline-based semiparametric maximum likelihood estimation method for the Cox model with interval-censored data*, Scand. J. Stat. 37 (2010), pp. 338–354.