Published in final edited form as:
J R Stat Soc Ser C Appl Stat. 2011 August ; 60(4): 559–574. doi:10.1111/j.1467-9876.2010.00756.x.

A Partially Linear Regression Model for Data from an Outcome-Dependent Sampling Design

Haibo Zhou,
University of North Carolina at Chapel Hill, USA

Jinhong You,
University of North Carolina at Chapel Hill, USA

Guoyou Qin, and
University of North Carolina at Chapel Hill, USA and Fudan University, CHINA

Matthew P. Longnecker
National Institute of Environmental Health Sciences, National Institutes of Health, Department of Health and Human Services, Research Triangle Park, NC, USA

Summary

The outcome dependent sampling scheme has been gaining attention in both the statistical literature and applied fields. Epidemiological and environmental researchers have been using it to select the observations for more powerful and cost-effective studies. Motivated by a study of the effect of *in utero* exposure to polychlorinated biphenyls on children's IQ at age 7, in which the effect of an important confounding variable is nonlinear, we consider a semi-parametric regression model for data from an outcome-dependent sampling scheme where the relationship between the response and covariates is only partially parameterized. We propose a penalized spline maximum likelihood estimation (PSMLE) for inference on both the parametric and the nonparametric components and develop their asymptotic properties. Through simulation studies and an analysis of the IQ study, we compare the proposed estimator with several competing estimators. Practical considerations of implementing those estimators are discussed.

Keywords

Outcome dependent sampling; Estimated likelihood; Semiparametric method; Penalized spline

1 Introduction

An outcome dependent sampling (ODS) design is an attempt to enhance study efficiency in a cost-effective way. Under an ODS design, the primary covariate, the exposure variable, is observed only on some subsets of the study subjects, conditional on the values of the response variable and possibly some other auxiliary covariates for the exposure. The principle motivation for ODS designs is to concentrate resources where there is the greatest amount of information. By allowing the selection probability of each individual in the ODS sample to be dependent on outcome, the investigators can enhance the efficiency and reduce the cost of the study.

Address for correspondence: Haibo Zhou, Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA., zhou@bios.unc.edu.

In a recent example that employed the ODS design(Gray et al. 2005), investigators were interested in how children’s IQ at 7 years of age is related to polychlorinated biphenyls (PCBs). The study subjects are children who were born into the Collaborative Perinatal Project (CPP) which is a prospective cohort designed to provide precise data for studies of a wide variety of neuropsychological outcomes and birth defects (Niswander and Gordon, 1972). Since it was too expensive to assay the PCB exposure for the entire study population of 44,075 subjects, the investigators decided to obtain exposure measurements for a sample that was sampled in an ODS way from the population based on the observed IQ scores (Gray et al., 2005). In the following, for simplicity, we refer to the data set including the PCB measurements (Gray et al., 2005) as the CPP data set.

Several authors have studied statistical inference for data from an ODS design. For example, Breslow and Holubkov (1997) developed maximum likelihood estimation of logistic regression coefficients for a hybrid two-phase design. Lawless, Kabfleisch and Wild (1999) considered a full semiparametric likelihood method. For a continuous outcome variable, Zhou et al. (2002) considered a general two-component ODS scheme where an overall simple random sample and additional supplementary samples are observed. Chatterjee, Chen and Breslow (2003) proposed a pseudoscore estimation method for regression problems with two-phase ODS. Breslow, McNeney and Wellner (2003) derived a large sample theory for semiparametric regression models. Weaver and Zhou (2005) proposed an estimated maximum likelihood method using the estimated likelihood technique (e.g., Carroll and Wand 1991, Pepe and Fleming 1991, Zhou and Pepe 1995) for incorporating additional information in the non-ODS sample. Wang and Zhou (2006) considered the case of an ordinal outcome variable with an auxiliary covariate. Zhou et al. (2007) further demonstrated the improved efficiency obtained by using the ODS design, and its applicability in a wide range of settings.

These existing methods are based on the assumption that the effect on the outcome of the covariates is linear. This assumption is chosen mainly for mathematical convenience. In practice, the true parametric relationship between the outcome and covariates is rarely known. For example, in the above mentioned epidemiological study (Gray et al. 2005), investigators were interested in identifying the relationship of the children’s IQ at age 7 to *in utero* exposure to PCBs, after adjusting for potential confounders, including the highest education level attained by the mother. Maternal education is often the strongest confounding factor in studies of environmental determinants of child IQ (Walkowiak et al. 1998; Angelsen et al. 2001; Bohm et al. 2002). The relation of maternal education to child IQ is not linear, with mother’s years in college having a much greater effect on child IQ than do years of education in primary and secondary school (e.g., Breslau et al. 2005; Oddy et al. 2003). Given the strength of this confounding factor, the manner in which education is modeled could affect the amount of bias in the coefficient for the exposure of interest. Thus, we were motivated to develop a partial linear method of modeling a covariate in the ODS setting.

Handling the nonparametric component in semiparametric models is generally challenging. One approach is to use nonparametric tools, e.g. the kernel estimator (Speckman 1988). This method is computationally intensive. Another method is to parametrize the nonparametric component using some flexible functions and then use some parametric tools, such as Fourier series approximation, Demmler-Reinsch series approximation, or wavelets. However, selecting the truncation parameter and allocating the knots in these methods can be challenging. An alternative approach is the penalized spline method, using a roughness penalty for the nonparametric regression function (e.g, Eilers and Marx 1996). The idea of a roughness penalty on splines is not new (e.g, O’Sullivan 1986), though the technique has become popular recently due to its effectiveness with penalized splines. With penalized splines, the number and location of knots are no longer crucial as long as the minimum number of knots is reached and the smoothing parameter is used to balance the goodness-of-fit and smoothness (Ruppert and Carroll 2000, Yu and Ruppert 2002, Wu and Yu 2004).

To model nonlinear covariate effects under the ODS sampling scheme, we consider a penalized spline maximum likelihood estimator (PSMLE) for the parametric and nonparametric components in a partially linear regression model. Using a simulation study and an analysis of the Collaborative Perinatal Project (CPP) data set, we present a case-study for comparing the PSMLE with several competing methods.

2 Penalized Spline Maximum Likelihood Estimation for ODS Design

2.1 Partial Linear Model and ODS Data Structure

To fix notation, let Y denote an outcome variable, and $\mathbf{X}$ and $\mathbf{Z}$ denote covariates. Let $f_{Y|X,Z}$ $(y|\mathbf{x}, \mathbf{z})$ be the conditional density of Y given $\mathbf{X} = \mathbf{x}$ and $\mathbf{Z} = \mathbf{z}$. We specify $f_{Y|X,Z}$ $(y|\mathbf{x}, \mathbf{z})$ through the following regression model

$$f_{Y|X,Z}(y|\mathbf{x},\mathbf{z})=m(y;\mathbf{x}^{\tau}\beta+\alpha(\mathbf{z})) \tag{1}$$

where $m(\cdot)$ is a known link function, $\boldsymbol{\beta}$ is an unknown parameter vector corresponding to $\mathbf{X}$ and $\alpha(\cdot)$ is an unknown function to be estimated. For simplicity, we write the conditional model as $f(y|\mathbf{x}, \mathbf{z}; \boldsymbol{\beta}, \alpha(\cdot))$. Model (1) can be viewed as a semiparametric model if $\alpha(\cdot)$ is unspecified. It is also referred to as a partial linear model since part of the covariate vector ($\mathbf{Z}$) is modeled as a nonlinear function.

We assume that the observed data are from an ODS sampling scheme (Zhou et al. 2002; Weaver and Zhou 2005). Specifically, assume that the domain of Y is union of K nonoverlapping intervals $C_k = (a_{k-1}, a_k]$, with a_k being known constants satisfying $a_0 \equiv -\infty < a_1 < a_2 < \ldots < a_K \equiv \infty$. The choice of the number of the intervals generally depends on the regions in the domain of the outcome variable which may contain great amount of information. A three Interval ODS scheme (K=3) can be selected for simplicity in practice (e.g., Gray et al. 2005). This will also be an over-representation of the tails of the distribution of the outcome that would be otherwise missing in a standard SRS scheme.

We assume there exists a base population (of sample size N) that is a simple random sample of the underlying study population, on which we observe $\{Y,\mathbf{Z}\}$. The exposure variable $\mathbf{X}$ is only observed for a subset of this base population that is selected in an ODS way. In particular, observation of $\mathbf{X}$ comes from two components: First, we observe $\mathbf{X}$ on a simple random sample (SRS) of size n_0. Secondly, for each stratum defined by $\{Y \in C_k\}$, $k = 1, \ldots, K$, we observe $\mathbf{X}$ on a supplementary random sample of size n_k. The first component is sometimes omitted, so that $n_0 = 0$. Hence the data set where $\mathbf{X}$ is observed is

$$\begin{array}{ll} \{\mathbf{X}_i, i=1, \ldots, n_0\} & \text{SRS component and} \\ \{\mathbf{X}_i | Y_i \in C_k, i=1, \ldots, n_k\} & \text{Supplementary component}, k=1, \ldots, K. \end{array}$$

Let $n_V = n_0 + \sum_{k=1}^{K} n_k$ denote the size of the ODS subsample for which we observe $(Y,\mathbf{X},\mathbf{Z})$, and let $n_{\bar{V}} = N - n_V$ be the number of individuals for whom only $(Y,\mathbf{Z})$ is observed. To borrow some terms from the measurement error literature, we will refer to the n_V complete observations as the validation sample, and $n_{\bar{V}}$ incomplete observations as the nonvalidation sample. Let V represent the index set of all validation observations, and let $\bar{V}$ represent the index set of all nonvalidation observations. Further, let V_k and $\bar{V}_k$ represent the index sets for observations in the kth stratum ($Y \in C_k$) in validation and nonvalidation samples, respectively.

Of the 44,075 children from the previously mentioned CPP data set, 38,709 have complete data (no missing data other than PCB), which will represent our study population. Furthermore, a sample of size 1038 with measured PCB levels is obtained from the study population through the above ODS design. In particular, the domain of the outcome IQ was divided into 3 intervals, i.e., $C_1 = (-\infty, 82]$, $C_2 = (82, 110]$ and $C_3 = (110,\infty)$, where 82 and 110 equal to the mean of IQ (96) minus or plus one standard deviation of IQ (14). It was anticipated that a sampling design in which children with extreme IQ scores were oversampled would enhance the efficiency of the study relative to an SRS design of the same size. Therefore, in addition to a SRS sample of size 849, 81 children with $IQ < 82$ and 108 children with $IQ > 110$ are randomly selected from intervals C_1 and C_3. Thus, in the sampling notation, we have that $n_V = 1038$, $n_{\overline{V}} = 37671$, $a_1 = 82$, $a_2 = 110$, $n_0 = 849$, $n_1 = 81$, $n_2 = 0$ and $n_3 = 108$. Table 1 gives summary of the specific data structure.

When data are collected through the ODS scheme described above, several levels of information could be used for inference about **β** and α(·). The simplest possibility is to use only those observations that make up the SRS portion (if this exists) of the ODS design. Alternatively, one could try to use the complete data portion (V). Clearly, a more efficient estimate can be achieved if one uses all available data.

Using Bayes formula and the multinomial distribution for finite population sampling, Weaver and Zhou (2005) show that the full-information likelihood based on all available data is proportional to

$$\mathscr{L}(\beta,\alpha(\cdot),G_{X|Z}) \propto \left[\prod_{i\in V} f(Y_i|\mathbf{X}_i,\boldsymbol{Z}_i;\beta,\alpha(\cdot))\right]\left[\prod_{j\in\overline{V}} f_Y\,(Y_j|Z_j;\beta,\alpha(\cdot))\right]$$

where

$$f(Y_j|Z_j,\beta,\alpha(\cdot))=\int_{\mathcal{X}} f(Y_j|\mathbf{x},\boldsymbol{Z}_j;\beta,\alpha(\cdot))dG_{X|Z}(\mathbf{x}|\boldsymbol{Z}_j)$$

and $G_{X|Z}(\mathbf{x}|\mathbf{Z})$ is the conditional distribution of **X** given **Z**.

Since the sampling mechanism used to obtain **X** in the validation sample is not a simple random sample, we cannot use a simple global empirical distribution function to estimate $G_{X|Z}$. Proper accommodation for the ODS nature of the validation sample is needed. By the Law of Total Probability, the distribution function of $X|Z$ can be written as

$$G_{X|Z}(\mathbf{x}|\mathbf{z})=Pr\{\mathbf{X}\le\mathbf{x}|\mathbf{Z}=\mathbf{z}\}=\sum_{k=1}^{K} Pr\{Y\in C_k|\mathbf{Z}=\mathbf{z}\}Pr\{\mathbf{X}\le\mathbf{x}|Y\in C_k,\mathbf{Z}=\mathbf{z}\}.$$

Hence, we can estimate $G_{X|Z}(\mathbf{x}|\mathbf{z})$ by the kernel smoother (e.g., Nadaraya 1964, Watson 1964):

$$\widehat{G}_{X|Z}(\mathbf{x}|\mathbf{z})=\sum_{k=1}^{K}\widehat{G}_k(\mathbf{x}|\mathbf{z})\frac{\sum_{i\in V_k\cup\overline{V}_k} L(\mathbf{Z}_i-\mathbf{z})}{\sum_{i\in V\cup\overline{V}} L(\mathbf{Z}_i-\mathbf{z})}$$

with $\hat{G}_k(\mathbf{x}|\mathbf{z}) = \Sigma_{i\in V_k} I(\mathbf{X}_i \le \mathbf{x})L_h(\mathbf{Z}_i - \mathbf{z})/\Sigma_{i\in V_k} L_h(\mathbf{Z}_i - \mathbf{z})$, where $L_h(\cdot) = L(\cdot/h)/h$ and $h > 0$ is the bandwidth. $L(\cdot)$ is called the kernel function and is a piecewise smooth function satisfying $\int L(u)du = 1$. We use a standard normal density function in our computations. For further details on the kernel smoother see Eubank (1988).

Then we can obtain an estimator of $f(Y_j|Z_j, \boldsymbol{\beta}, \alpha(\cdot))$ as

$$\widehat{f}(Y_j|Z_j,\beta,\alpha(\cdot))=\sum_{k=1}^{K}\frac{\left(\sum_{i\in V_k} f(Y_j|\mathbf{X}_i,\mathbf{Z}_j;\beta,\alpha(\cdot))\;L\,(\mathbf{Z}_i-\mathbf{Z}_j))\right)\sum_{i\in V_k\cup\overline{V}_k} L(\mathbf{Z}_i-\mathbf{Z}_j)}{\sum_{i\in V_k} L(\mathbf{Z}_i-\mathbf{Z}_j)\sum_{i\in V\cup\overline{V}} L(\mathbf{Z}_i-\mathbf{Z}_j)}.$$

2.2 Penalized Spline for Modeling α(·)

For convenience of presentation, we assume $\mathbf{Z}_i$ is an univariate variable. The unknown function α(·) can be estimated by a penalized spline (Ruppert and Carroll 2000 and Ruppert 2002). Assume that

$$\alpha(z)=\delta+\delta_1 z+\ldots+\delta_m z^m+\sum_{k=1}^{\kappa}\delta_{m+k}(z-\vartheta_k)_+^m, \tag{2}$$

where $\{\vartheta_k\}_{k=1}^{\kappa}$ are spline knots. Model (2) uses the so-called truncated power function basis, though other bases (e.g., B-splines) could also be used. Define the spline coefficient vector $\boldsymbol{\delta} = (\delta, \delta_1,\ldots,\delta_{m+\kappa})^\tau$ and spline basis

$$\mathbf{B}(z)=(1,z,\ldots,z^m,(z-\vartheta_1)_+^m,\ldots,(z-\vartheta_\kappa)_+^m).$$

Our spline model is $\alpha(z) = \boldsymbol{\delta}^\tau\mathbf{B}(z)$. Denote $\boldsymbol{\zeta} = (\boldsymbol{\beta}^\tau, \boldsymbol{\delta}^\tau)^\tau$. The PSMLE of $\hat{\zeta} = (\hat{\boldsymbol{\beta}}^\tau, \hat{\boldsymbol{\delta}}^\tau)^\tau$ is defined as $\boldsymbol{\zeta}$ that maximizes

$$Q_{\lambda,N}\,(\beta,\delta)=\widehat{\ell_F}\,(\beta,\delta)-\lambda N\delta^\tau\mathbf{D}\delta \tag{3}$$

where

$$\begin{aligned}
\widehat{\ell}_F\,(\beta,\delta)=&\sum_{i\in V}\ln\, f(Y_i|\mathbf{X}_i,\mathbf{Z}_i;\beta,\\
&\delta)+\sum_{j\in\overline{V}}\ln\,\widehat{f}(Y_j|\mathbf{X}_i,\\
&\mathbf{Z}_i;\beta,\\
&\delta)=\sum_{i\in V}\ln\, f(Y_i|\mathbf{X}_i,\\
&\mathbf{Z}_i;\beta,\\
&\delta)+\sum_{j\in\overline{V}}\ln\sum_{k=1}^{K}\frac{\left(\sum_{i\in V_k} f(Y_j|\mathbf{X}_i,\mathbf{Z}_j;\beta,\delta)\,L\,(\mathbf{Z}_i-\mathbf{Z}_j))\right)\sum_{i\in V_k\cup\overline{V}_k} L(\mathbf{Z}_i-\mathbf{Z}_j)}{\sum_{i\in V_k} L(\mathbf{Z}_i-\mathbf{Z}_j)\sum_{i\in V\cup\overline{V}} L(\mathbf{Z}_i-\mathbf{Z}_j)},
\end{aligned}$$

λ ≥ 0 is a smoothing parameter, and **D** is an appropriate positive semi-definite symmetric matrix such that $\delta^{\tau}\mathbf{D}\delta=\int_{\min(\mathbf{Z}_i)}^{\max(\mathbf{Z}_i)}[\alpha''(z)]^2dz$, which yields the usual quadratic integral penalty (Ruppert 2002).

We now describe some asymptotic results that are summarized in three theorems with outline proofs, in the Appendix.

First, under some regularity conditions, the proposed estimator is consistent and asymptotically normally distributed. Furthermore, suppose we are interested in testing a constraint on the parameters as in the hypothesis

$$H_0:\psi\,(\boldsymbol{\zeta})=0.$$

We define a likelihood ratio statistic $R(\boldsymbol{\zeta})$ that is based on the $Q_{\lambda,N}$ $(\boldsymbol{\beta}, \boldsymbol{\delta})$ as

$$R(\boldsymbol{\zeta})=2\left\{\max_{\boldsymbol{\zeta}} Q_{\lambda,N}(\boldsymbol{\zeta})-\max_{\boldsymbol{\zeta},\psi(\boldsymbol{\zeta})=0} Q_{\lambda,N}(\boldsymbol{\zeta})\right\},$$

where ψ(·) is a $q \times 1$ vector function ($q < p + m + \kappa + 1$). The likelihood ratio statistic is asymptotically distributed as χ_q^2 under H_0. Our likelihood ratio test (LRT) differs from that in Hastie and Tibshirani (1990) in the computation of the degrees of freedom. Specifically, the degrees of freedom for our proposed LRT is computed by the difference between the number of the parameters in the null model and the unrestricted model rather than the effective degrees of freedom in Hastie and Tibshirani (1990).

For the partial linear model, it is of particular interest to test whether the nonparametric function is linear. We can use the previously described likelihood ratio test to do this by re-expressing the $m + \kappa + 1$-dimensional vector **δ** as $\left(\delta_1^T,\delta_2^T\right)$, where $\boldsymbol{\delta}_1 = (\delta_{11}, \delta_{12})^T$ is a two-dimensional vector and $\boldsymbol{\delta}_2$ is a $m + \kappa - 1$ dimensional vector. We are then interested in testing the null hypothesis $H_0:\delta_2=\delta_2^0=(0,\ldots,0)^T$, i.e., $H_0:\psi(\boldsymbol{\zeta})=\delta_2=\delta_2^0=0$. Under H_0, $x^{\tau}\beta + \alpha(z) = x^{\tau}\boldsymbol{\beta} + \delta_{11} + \delta_{12}z$, i.e. the variable Z is related to response Y linearly.

2.3 Selection of smoothing Parameter, Knots and Penalty

To implement the proposed method in practice, it is desirable to have an automatic data-driven method for estimating the smoothing parameter λ. Generalized cross-validation

(GCV) is an attractive way to choose λ since it is computationally expedient and does not need a prior estimate of error variance. Following the conventional technique of penalized least squares (e.g., Ruppert 2002), we define

$$e(\lambda)=\mathrm{tr}\left[\{Q''_{N,\lambda}(\widehat{\zeta})\}^{-1}\ell''_{F}(\widehat{\zeta})\right]$$

where $\{Q''_{N,\lambda}(\widehat{\zeta})\}^{-1}\ell''_{F}(\widehat{\zeta})$ is the smoothing or hat matrix. In nonparametric regression, the trace of the smoothing matrix is often called the *degrees of freedom* of the fit. It has the rough interpretation as the equivalent number of parameters (Yu and Ruppert 2002). The GCV statistic is defined by

$$\mathrm{GCV}(\lambda)=-\frac{\mathrm{RSS}}{N\{1-e(\lambda)/N\}^{2}}$$

where RSS = $2\ell_F(\hat{\zeta})$ is the residual sum of squares corresponding to $\hat{\zeta}$, given λ. We select $\hat{\lambda}$ = $\mathrm{argmin}_{\lambda}\{\mathrm{GCV}(\lambda)\}$.

Since the complicated knot selection problem is reduced to the choice of a single smoothing parameter λ the selection of the number of knots and knot locations is no longer crucial for the penalized spline. Ruppert (2002), Wu and Yu (2004) and Yu (2008) have observed that the choice of the number of knots κ is not too important, provided it is large enough. Ruppert (2002) and Wu and Yu (2004) suggested choosing approximately min(n/4, 35) or min(n/4, 40) knots, respectively, where n is the number of distinct values of the sample of the nonparametric covariate. However, in many practical situations where the regression function is smooth and either monotonic or unimodal, 10 to 20 knots are very adequate, as suggested in Yu (2008). Given our choice of nonlinear function, the number of knots are chosen from 10 to 30 in our simulation which works quite well (see Figure 1). Given a fixed number of knots, Wu and Yu (2004) and Yu (2008) recommended that the knots are placed at equally-spaced sample quantiles of the index **Z**.

As in Wu and Yu (2004) and Yu (2008), we take a quadratic penalty of the form $\lambda\boldsymbol{\delta}^{T}\mathbf{D}\boldsymbol{\delta}$ in our paper. When the nonparametric function has discontinuity, the nonquadratic penalty functions may be a better choice. Ruppert and Carroll (1997) gave a general penalty of the form $\sum_{t=1}^{T}|\alpha_{r+t}|^{\gamma}, \gamma>0$, and pointed out that penalties with $\gamma \le 1$ can perform better than a quadratic penalty for discontinuous functions. Otherwise, the quadratic penalty is preferred.

3 Simulation Studies

We investigate the small sample behavior of the proposed method through simulation studies, comparing the proposed estimator with several potential competing estimators. Mimicking the design of CPP study, we generate data according to the following regression model:

$$Y_i=\beta_1X_{1i}+\beta_2X_{2i}+\alpha\,(Z_i)+\varepsilon_i,\ i=1,\ldots,$$

where $X_{1i} \sim N(1, 0.25)$, $X_{2i} \sim N(0, 1)$, $\beta_1 = 1$, $\beta_2 = 1.5$, $Z_i = \xi_i + X_{1i}I(|X_{1i}| \le 1)$ with $\xi_i \sim U(0, 1)$, and $\xi_i \sim N(0, 1)$. We take (n_0, n_1, n_2, N) = (100, 25, 25, 600), (200, 25, 25, 500), (200, 50, 50, 1200) and $C_1 = (-\infty, \mu_y - \sigma_y)$ and $C_2 = (\mu_Y + \sigma_Y, \infty)$. We consider two choices for $\alpha(z)$ that represent nonlinear forms commonly observed in practice:

Case 1: $\alpha(z) = \sin(2\pi z)$,

Case 2: $\alpha(z) = (0.02 \exp(10(z - 1)))/(1 + \exp(8(z - 1.5)))$.

The first case has a cyclic pattern (Hickey et al. 1984, Strum and Pinsky 2006, Elkum et al. 2008) while the second has a flat response at the beginning and a sharp rise at the end of the range (e.g, Figure 1(a) and (c)).

Under each setting, we compare three estimators:

$\hat{\boldsymbol{\beta}}_P$: the proposed penalized spline maximum likelihood estimator.

$\hat{\boldsymbol{\beta}}_{HT}$: the modified Horvitz-Thompson weighted-likelihood method. The estimator of $\boldsymbol{\zeta} = (\boldsymbol{\beta}^\tau, \boldsymbol{\delta}^\tau)^\tau$ maximizes

$$\sum_{k=1}^{K} 1/(n_0/N+n_k/N_k) \sum_{i \in V_k} \ln f(Y_i|\mathbf{X}_i, \mathbf{Z}_i;\beta, \delta) - \lambda N \delta^\tau \mathbf{D} \delta$$

where $\lambda\delta^\tau\mathbf{D}\delta$ has the same definition as in (3).

$\hat{\boldsymbol{\beta}}_{BC}$: the modified Breslow-Cain pseudo-likelihood method. The estimator of $\boldsymbol{\zeta} = (\boldsymbol{\beta}^\tau, \boldsymbol{\delta}^\tau)^\tau$ maximizes

$$\sum_{i \in V} \left\{ \ln f(Y_i|\mathbf{X}_i, \mathbf{Z}_i;\beta, \delta) - \ln \sum_{k=1}^{K} 1/(n_0/N+n_k/N_k) Pr(Y_i \in C_k|\mathbf{X}_i) \right\} - \lambda N \delta^\tau \mathbf{D} \delta.$$

We take the number of knots as 15, 13 and 30 respectively corresponding to the sample size 600, 500 and 1200 in the simulations. The means, standard errors (SE), estimators of standard errors $(\widehat{\text{SE}})$ and coverages of 95% nominal confidence intervals (CI) were calculated from 1,000 independent runs. Table 2 lists results for the above estimators under various configurations.

Clearly, $\hat{\beta}_P$, $\hat{\beta}_{HT}$ and $\hat{\beta}_{BC}$ are approximately unbiased for both β_1 and β_2. The proposed estimator ($\hat{\beta}_P$) is always more efficient than $\hat{\beta}_{HT}$ and $\hat{\beta}_{BC}$. This supports the notion that taking the nonvalidation sample into account can improve the efficiency of estimation. The nominal 95% confidence intervals based on the proposed standard errors provide good coverage for the cases studied for $\hat{\beta}_P$, $\hat{\beta}_{HT}$, $\hat{\beta}_{BC}$.

Figure 1 (a) and (b) show the estimators of $\alpha(z) = \sin(2\pi z)$ and their corresponding pointwise SEs. Figure 1 (c) and (d) show the estimators of $\alpha(z) = (0.02 \exp(10(z-1)))/(1+\exp(8(z - 1.5)))$ and their corresponding pointwise SEs. From Figure 1, we see that $\hat{\alpha}_P (z)$, $\hat{\alpha}_{HT} (z)$ and $\hat{\alpha}_{HT}(z)$ are approximately unbiased. Furthermore, $\hat{\alpha}_P (z)$ has smallest pointwise SE among $\hat{\alpha}_{HT} (z)$ and $\hat{\alpha}_{BC}(z)$, suggesting that the proposed method is indeed a more efficient approach.

4 Analysis of the CPP Data

We analyze the CPP data to identify the relationship of the children's IQ at 7 years of age to *in utero* exposure to polychlorinated biphenyls (PCBs), after adjusting for potential confounders, including the highest education level attained by the mother (EDU).

Additional covariates in the analysis are socioeconomic status of the child's family (SES), the gender of the child (SEX, with female=1 and male=0) and the race of the child (RACE, with black=1 and other=0). To model the nonlinear effect of education noted in the Introduction, we consider the following partial linear model,

$$IQ=\beta_1 PCB+\beta_2 SES+\beta_3 RACE+\beta_4 SEX+\alpha(EDU)+\varepsilon$$

where $\alpha(EDU)$ is an unspecified function to be estimated along with β_i, $i = 1,\ldots,4$. To estimate the nonparametric function $\alpha(\cdot)$, we adopted a three-degree truncated power function basis $M(z)=(1, z, z^2, z^3, (z-\vartheta_1)_+^3, \ldots, (z-\vartheta_5)_+^3)^T$ with five fixed knots $\boldsymbol{\vartheta}_1,\ldots,\boldsymbol{\vartheta}_5$ selected as the equally spaced sample quantiles of EDU, i.e., 2,5,9,12,15. Under these specifications, the above model can be rewritten as $IQ = \beta_1 PCB+\beta_2 SES+\beta_3 RACE+\beta_4 SEX +M^T (EDU)\delta+\varepsilon$, where $\delta = (\delta_0,\ldots, \delta_8)^T$ is the parameter vector associated with the nonparametric function $\alpha(\cdot)$.

We analyzed the CPP data with the following methods using a penalized spline for $\alpha(EDU)$: the proposed method (P), the modified Horvitz-Thompson weighted likelihood method (HT), the modified Breslow-Cain pseudo-likelihood method (BC), and the MLE estimator based on the SRS sample (MLE-SRS). The smoothing parameter was selected as 0.0853 by the proposed GCV method. Additionally, for the proposed method, we also considered modeling $\alpha(EDU)$ as a linear, quadratic, or cubic function of EDU. Furthermore, we considered using a restricted cubic spline (Herndon and Harrell 1990) for $\alpha(EDU)$ and obtained the corresponding estimate through maximizing $\hat{\ell}_F (\boldsymbol{\beta}, \boldsymbol{\delta})$. The restricted cubic spline, which has a linearly constrained tails which is slightly different from the general cubic spline function and can be used directly to fit models without penalty.

The estimated $\hat{\alpha}(EDU)$ from the different methods and their corresponding 95% confidence intervals given in Figure 2. The fitted $\hat{\alpha}(EDU)$ tell a similar story in that there is a clear nonlinear trend present in all fitted lines. The most noticeable difference is the width of the confidence interval band, which indicates which method is more efficient. A careful inspection of the trend of $\hat{\alpha}(EDU)$ reveals that the rate of rise of $\hat{\alpha}(EDU)$ is much faster after around year 12 (i.e. after high school education). This agrees with the previous published results (e.g., Breslau et al., 2005; Oddy et al., 2003) that mother's years in college have a much greater effect on child IQ.

We conducted likelihood ratio tests for testing if the nonlinear fit of $\alpha(EDU)$ from the proposed method can be represented by a simple polynomial function. The following three tests on the form of $\alpha(EDU)$ are for linear, quadratic, and cubic functions, respectively.

Test 1: $H_0 : \alpha(EDU) = \delta_0 + \delta_1 EDU$,

Test 2: $H_0 : \alpha(EDU) = \delta_0 + \delta_1 EDU + \delta_2 EDU^2$,

Test 3: $H_0 : \alpha(EDU) = \delta_0 + \delta_1 EDU + \delta_1 EDU^2 + \delta_3 EDU^3$.

The test statistic for the Test 1–3 are: for linear $\alpha(EDU)$, $T1=314.40>\chi^2_{0.95}(7)=14.07$ with p< 0.001; for quadratic $\alpha(EDU)$, $T2=94.90>\chi^2_{0.95}(6)=12.59$ with p< 0.001; and for cubic $\alpha(EDU)$, $T3=65.97>\chi^2_{0.95}(5)=11.07$ with p< 0.001, respectively. These results suggest that, although the cubic fit in Figure 2(e) may be sufficiently close to the fully nonparametric fit in Figure 2(a) for practical purposes, there is still statistical evidence suggesting that $\hat{\alpha}(EDU)$ may be more complex than a cubic function.

The parameter estimates from the six methods are presented in Table 4 which also includes the analysis with a linear effect for EDU using the Zhou et al. (2002) method which is based on the ODS data only. Overall, the point estimates from the above methods are similar. The most obvious difference across the methods is that the standard error estimates from the proposed methods (the P, cubic and restricted cubic spline) are much smaller for the covariates. This reflects the fact that these three methods utilized the real values of the covariates in the entire study cohort while the others only used the fraction as weight in the inference. In addition, we computed the values of the penalized log-likelihood function (3) for these three methods, which are $Q_p = -150441.651$, $Q_c = -150474.636$, $Q_R = -150452.816$ respectively corresponding to the P, cubic and restricted cubic spline methods, indicating that the P method is more suitable for this CPP data set than the other two methods. However, for practical purposes, the restricted cubic spline method is a viable alterative in this case.

5 Discussion

In this paper we proposed a semiparametric regression model to analyze data obtained by outcome dependent sampling. We only partially parametrize the relationship between the response variable and the covariates. By combining the estimated semiparametric likelihood and penalized spline techniques, we propose a penalized spline maximum likelihood estimation method for the key parametric and nonparametric components. The resulting estimators were shown to be consistent and asymptotically normal. In practice, our penalized spline maximum likelihood estimation offers a few additional advantages. For example, as a direct approach, penalized spline maximum likelihood estimates can be obtained through standard penalized maximum likelihood estimation. The algorithm is efficient and convergence is fast; moreover, as a global smoothing method, the penalized spline maximum likelihood approach yields a parsimonious model, which is convenient for inference and forecasting.

The smoothing parameter λ is used to balance goodness-of-fit and smoothness. Compared with the restricted cubic spline, use of the penalized spline can help avoid undersmoothing in some cases, e.g., large number of knots are needed when the nonparametric function has many local mimima and maxima. We only focused on the setting where **Z** is univariate. When **Z** is bivariate or multivariate, we can approximate the unknown function $\alpha(\cdot)$ by bivariate or multivariate basis functions (Ruppert, Wand and Carroll 2003). In addition, when **Z** is bivariate or multivariate, to avoid excess dimensionality one can use structural nonparametric regression models such as a varying-coefficient model, additive model, or a single-index model. We believe the proposed penalized spline maximum likelihood estimation can be extended to these corresponding semiparametric regression models without significant modification.

Acknowledgments

This work is supported by a grant from National Institute of Health (R01CA 79949) (for Zhou, You and Qin) and in part by the intramural research program of the NIH, National Institute of Environmental Health Sciences (for Longnecker). Qin's work is also partially supported by the NSF of China (10801039).

Appendix

Outline of Proofs for the Main Results

The asymptotic properties of the proposed estimators based on the estimated penalized likelihood (3) and model (2) are summarized in the following theorems.

Theorem 1 *Under some regularity conditions* (Weaver and Zhou 2005), *if the smoothing parameter* $\lambda = o(1)$, then $\hat{\zeta}$ *is a strong consistent estimator of* ζ.

Theorem 2 *Under same regularity conditions above, if the smoothing parameter* $\lambda = o(N^{-1/2})$, *then* $\hat{\zeta}$ *has asymptotic distribution,*

$$\sqrt{N}(\widehat{\zeta}-\zeta)\xrightarrow{D}N_{p+m+\kappa+1}(0,\Omega)\quad as\ N\to\infty,$$

with

$$\mathbf{\Omega}=\mathbf{I}^{-1}(\zeta)+\sum_{k=1}^{K}\frac{\pi_k^2}{\rho_k\rho_V+\pi_k\rho_0\rho_V}\mathbf{I}^{-1}(\zeta)\sum\nolimits_k(\zeta)\mathbf{I}^{-1}(\zeta)$$

where $d_k = \pi_k(1-\rho_0\rho_V)-\rho_k\rho_V$, $\pi_k = \pi_k(\boldsymbol{\beta}, \alpha(\cdot), G_{X,Z})$, $\rho_k = \lim_{N\to\infty} n_k/n_V$ *and* $\rho_V = \lim_{N\to\infty} n_V/N$,

$$\mathbf{I}(\zeta)=-\sum_{k=0}^{K}\rho_k\rho_V E_k\left[\frac{\partial^2\log f(Y|\mathbf{X},\mathbf{Z};\zeta)}{\partial\zeta\partial\zeta^\tau}\right]-\sum_{k=1}^{K}d_kE_k\left[\frac{\partial^2\log f(Y|\mathbf{Z};\zeta)}{\partial\zeta\partial\zeta^\tau}\right],$$

E_k *denotes expectation conditional on* $Y \in C_k$,

$$\sum\nolimits_k(\boldsymbol{\zeta})=Var_{X,Z|Y\in C_k}\left\{\sum_{k_1=1}^{K}d_{k_1}E_{Y|Y\in C_k}[\mathbf{M}_{\mathbf{X},\mathbf{Z}}(Y;\boldsymbol{\zeta})]\right\}$$

and

$$\mathbf{M}_{\mathbf{X},\mathbf{Z}}(Y;\boldsymbol{\zeta})=\frac{\partial f(Y|\mathbf{X},\mathbf{Z};\boldsymbol{\zeta})/\partial\boldsymbol{\zeta}}{f(Y|\mathbf{Z};\boldsymbol{\zeta})}-\frac{\partial f(Y|\mathbf{Z};\boldsymbol{\zeta})/\partial\boldsymbol{\zeta}}{(f(Y|\mathbf{Z};\boldsymbol{\zeta}))^2}\times f(Y|\mathbf{X},\mathbf{Z};\boldsymbol{\zeta}).$$

A consistent estimator for $\mathbf{\Omega}$ can be constructed using sample quantities.

Theorem 3 *Under same regularity conditions above, if the smooth parameter* $\lambda = o(N^{-1/2})$, *then for testing the null hypothesis* $H_0 : \psi(\zeta) = 0$, *we have* $R \to \chi_q^2$.

Proof of Theorem 1. Let the full-information log likelihood be

$$\ell_F(\beta,\delta)=\sum_{k=0}^{K}\sum_{i\in V_k}\ln f(Y_i|\mathbf{X}_i,\mathbf{Z}_i;\beta,\delta)+\sum_{k=1}^{K}\sum_{j\in\overline{V}_k}\int_{\mathcal{X}}f(Y_j|Z_j,\mathbf{x};\beta,\delta)dG_{X|Z}(\mathbf{x}|\mathbf{Z}_j).$$

Then

$$\frac{\partial Q_{N,\lambda}(\beta,\delta)}{\partial\zeta}=\frac{1}{N}\left(\frac{\partial\widehat{\ell_F}(\beta,\delta)}{\partial\zeta}-\frac{\partial\ell_F(\beta,\delta)}{\partial\zeta}\right)+\frac{1}{N}\frac{\partial\ell_F(\beta,\delta)}{\partial\zeta}-2\lambda\text{blockdiag}(\mathbf{0}_{p\times p},\mathbf{D})(\mathbf{0}_p^\tau,\delta^\tau)^\tau.$$

According to the proof of Theorem 3.1 in Weaver (2001) it holds that

$$\frac{1}{N}\left(\frac{\partial\widehat{\ell_F}(\beta,\delta)}{\partial\zeta}-\frac{\partial\ell_F(\beta,\delta)}{\partial\zeta}\right)\to_p0\ \ \text{and}\ \ \frac{1}{N}\frac{\partial\ell_F(\beta,\delta)}{\partial\zeta}\to_p0\ \ \text{as}\ N\to\infty.$$

Therefore, combining the fact that λ = $o(1)$ we have $\partial Q_{N,\lambda}(\boldsymbol{\beta}, \boldsymbol{\delta})/\partial\zeta \rightarrow_p 0$ as $N \rightarrow \infty$. In addition,

$$\frac{\partial^2 Q_{N,\lambda}(\beta,\delta)}{\partial\zeta\partial\zeta^\tau}=\frac{1}{N}\left(\frac{\partial^2\widehat{\ell_F}(\beta,\delta)}{\partial\zeta\partial\zeta^\tau}-\frac{\partial^2\ell_F(\beta,\delta)}{\partial\zeta\partial\zeta^\tau}\right)+\frac{1}{N}\frac{\partial^2\ell_F(\beta,\delta)}{\partial\zeta\partial\zeta^\tau}-2\lambda\text{blockdiag}(\mathbf{0}_{p\times p},\mathbf{D}).$$

According to the proof of Theorem 3.1 in Weaver (2001) it holds that

$$\frac{1}{N}\left(\frac{\partial^2\widehat{\ell_F}(\beta,\delta)}{\partial\zeta\partial\zeta^\tau}-\frac{\partial^2\ell_F(\beta,\delta)}{\partial\zeta\partial\zeta^\tau}\right)\rightarrow_p\mathbf{0}_{(p+m+\kappa+1)\times(p+m+\kappa+1)}\ \text{ and }-\frac{1}{N}\frac{\partial^2\ell_F(\beta,\delta)}{\partial\zeta\partial\zeta^\tau}\rightarrow_p\mathbf{I}(\boldsymbol{\zeta})$$

uniformly for $\zeta \in \Theta$ as $N \rightarrow \infty$. Therefore, combining the fact that λ = $o(1)$ we have

$$\frac{\partial^2 Q_{N,\lambda}(\beta,\delta)}{\partial\zeta\partial\zeta^\tau}\rightarrow_p\mathbf{I}(\boldsymbol{\zeta})\ \text{ uniformly for }\boldsymbol{\zeta}\in\Theta\ \text{ as }N\rightarrow\infty.$$

Thus, if we let $\mathbf{f}_N(\boldsymbol{\zeta})=\frac{\partial Q_{N,\lambda}(\beta,\delta)}{\partial\zeta}$ we can apply Lemma 3.3 in Weaver (2001) to conclude that $\widehat{\boldsymbol{\zeta}}=\mathbf{f}_N^{(-1)}(\mathbf{0})$ exists in the set Θ with probability approaching one as $N \rightarrow \infty$, and since the size of Θ is arbitrarily small, that $\hat{\zeta} \rightarrow_p \zeta$. Furthermore, the sequence of estimators $\{\hat{\zeta}\}$ is unique in the sense that any other sequence $\{\bar{\zeta}\}$ such that $\frac{\partial Q_{N,\lambda}(\bar{\beta},\bar{\delta})}{\partial\zeta}=0$ and $\bar{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}$ must lie outside of the set Θ with probability going to one as $N \rightarrow \infty$.

Proof of Theorem 2. For consistent estimator $\hat{\zeta}$, using a first order Taylor expansion near ζ yields that

$$\mathbf{0}=\frac{\partial Q_{N,\lambda}(\beta,\delta)}{\partial\zeta}|_{\widehat{\zeta}}=\frac{\partial Q_{N,\lambda}(\beta,\delta)}{\partial\zeta}|_{\zeta}+\frac{\partial^2 Q_{N,\lambda}(\beta,\delta)}{\partial\zeta\partial\zeta^\tau}|_{\zeta^*}(\widehat{\zeta}-\zeta),$$

where ζ* is a vector between $\hat{\zeta}$ and ζ. By standard manipulation, we have

$$\sqrt{N}(\widehat{\zeta}-\zeta)=\left\{-\frac{1}{N}\frac{\partial^2 Q_{N,\lambda}(\beta^*,\delta^*)}{\partial\zeta\partial\zeta^\tau}\right\}^{-1}\left\{\frac{1}{\sqrt{N}}\frac{\partial Q_{N,\lambda}(\beta,\delta)}{\partial\zeta}\right\}.$$

Thus, to prove the asymptotic normality of $\sqrt{N}(\widehat{\zeta}-\zeta)$, we need only show $1/\sqrt{N}\frac{\partial Q_{N,\lambda}(\beta,\delta)}{\partial\zeta}$ has an asymptotic normal distribution and that $-1/\sqrt{N}\frac{\partial^2 Q_{N,\lambda}(\beta^*,\delta^*)}{\partial\zeta\partial\zeta^\tau}$ converges in probability to an invertible matrix.

According to the definition of $Q_{N,\lambda}(\boldsymbol{\beta}, \boldsymbol{\delta})$ we have

$$
\begin{aligned}
&\sqrt{N}\frac{\partial Q_{N,\lambda}(\beta,\delta)}{\partial \zeta}\\
&\quad=\frac{1}{\sqrt{N}}\sum_{i\in V_0}\frac{\partial f(Y_i|\mathbf{X}_i,\mathbf{Z}_i;\zeta)/\partial\zeta}{f(Y_i|\mathbf{X}_i,\mathbf{Z}_i;\zeta)}\\
&\quad+\frac{1}{\sqrt{N}}\sum_{k=1}^{K}\sum_{i\in V_k}\frac{\partial f(Y_i|\mathbf{X}_i,\mathbf{Z}_j;\zeta)/\partial\zeta}{f(Y_i|\mathbf{X}_i,\mathbf{Z}_i;\zeta)}\\
&\quad+\frac{1}{\sqrt{N}}\sum_{k=1}^{K}\sum_{j\in\overline{V}_k}\frac{\int_{\chi}\frac{\partial f(Y_j|\mathbf{x},\mathbf{Z}_j;\zeta)}{\partial\zeta}dG_{X|Z}(\mathbf{x}|\mathbf{Z}_j)}{\int_{\chi}f(Y_j|\mathbf{x},\mathbf{Z}_j;\zeta)dG_{X|Z}(\mathbf{x}|\mathbf{Z}_j)}\\
&\quad+\frac{1}{\sqrt{N}}\sum_{k=1}^{K}\sum_{j\in\overline{V}_k}\left\{\frac{\int_{\chi}\frac{\partial f(Y_j|\mathbf{x},\mathbf{Z}_j;\zeta)}{\partial\zeta}d\widehat{G}_{X|Z}(\mathbf{x}|\mathbf{Z}_j)}{\int_{\chi}f(Y_j|\mathbf{x},\mathbf{Z}_j;\zeta)d\widehat{G}_{X|Z}(\mathbf{x}|\mathbf{Z}_j)}-\frac{\int_{\chi}\frac{\partial f(Y_j|\mathbf{x},\mathbf{Z}_j;\zeta)}{\partial\zeta}dG_{X|Z}(\mathbf{x}|\mathbf{Z}_j)}{\int_{\chi}f(Y_j|\mathbf{x},\mathbf{Z}_j;\zeta^*)dG_{X|Z}(\mathbf{x}|\mathbf{Z}_j)}\right\}\\
&\quad-2\lambda\text{blockdiag}(\mathbf{0}_{p\times p},\mathbf{D})(\mathbf{0}_p^{\tau},\delta^{\tau})^{\tau}
\end{aligned}
$$

Note that $\lambda = o(N^{-1/2})$. Therefore, we have $2\lambda\text{blockdiag}(\mathbf{0}_{p\times p},\mathbf{D})\left(\mathbf{0}_p^{\tau},\delta^{\tau}\right)^{\tau}=o(N^{-\frac{1}{2}})$. Thus, by the same argument as the proof of Theorem 3.2 ofWeaver (2001), we can show that Theorem 2 holds.

Proof of Theorem 3. Note that when $\lambda = o(N^{-1/2})$, the penalty term in $R(\zeta)$ tends to zero with a rate of $o(N^{-1/2})$. Then through the similar procedure for proof of the asymptotic distribution of the classical likelihood ratio statistics, Theorem 3 can be obtained. We have omitted the details here.

References

Angelsen NK, Vik T, Jacobsen G, Bakketeig LS. Breast feeding and cognitive development at age 1 and 5 years. Arch Dis Child. 2001; 85:183–188. [PubMed: 11517096]

Bohm B, Katz-Salamon M, Institute K, Smedler AC, Lagercrantz H, Forssberg H. Developmental risks and protective factors for influencing cognitive outcome at 5 1/2 years of age in very-low-birthweight children. Dev Med Child Neurol. 2002; 44:508–516. [PubMed: 12206615]

Breslau N, Paneth N, Lucia VC, Paneth-Pollak R. Maternal smoking during pregnancy and offspring IQ. Int J Epidemiol. 2005; 34:1047–1053. [PubMed: 16085682]

Breslow N, McNeney B, Wellner JA. Large sample theory for semiparametric regression models with two-phase, outcome dependent sampling. Ann. Statist. 2003; 31:1110–1139.

Breslow N, Holubkov R. Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling. J. Roy. Statist. Soc., B. 1997; 59:447–461.

Carroll RJ, Wand MP. Semiparametric estimation in Logistic measurement error model. Journal of the Royal Statistics Society, B. 1991; 53:573–585.

Chatterjee N, Chen Y, Breslow N. A pseudoscore estimator for regression problems with two-phase sampling. J. Amer. Statist. Assoc. 2003; 98:158–168.

Eilers PHC, Marx BD. Flexible smoothing with B-spline and penalties. Statist. Science. 1996; 11:89–102.

Elkum NB, Myles JD, Kumar P. Analyzing biological rhythms in clinical trials. Contemp Clin Trials. 2008; 29:720–726. [PubMed: 18571991]

Gray KA, Longnecker MP, Klebanoff MA, Brock JW, Zhou H, Needham L. In Utero exposure to background levels of Polychlorinated Biphenls and cognitive functioning among school-aged children. Am J Epidemiology. 2005; 162:17–26.

Hastie, TJ.; Tibshirani, RJ. Generalized Additive Models. London: Chapman and Hall; 1990.

Herndon JE, Harrell RE. The restricted cubic spline hazard model. Communications in Statistics: Theory and Methods. 1990; 19:639–694.

Hickey DS, Kirkland JL, Lucas SB, Lye M. Analysis of circadian rhythms by fitting a least squares sine curve. Comput Biol Med. 1984; 14:217–223. [PubMed: 6723267]

Hurvich CM, Simonoff JS, Tsai CL. Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. J. R. Stat. Soc. Ser. B. 1998; 60:271–293.

Lawless JF, Kabfleisch JD, Wild CJ. Semiparametric methods for response-selective and missing data problems in regression. J. Roy. Statist. Soc., B. 1999; 61:413–438.

Oddy WH, Kendall GE, Blair E, De Klerk NH, Stanley FJ, Landau LI, Silburn S, Zubrick S. Breast feeding and cognitive development in childhood: a prospective birth cohort study. Paediatr Perinat Epidemiol. 2003; 17:81–90. [PubMed: 12562475]

O'Sullivan F. A statistical perspective on ill-posed inverse problems. With comments and a rejoinder by the author. Statist. Sci. 1986; 1:502–527.

Pepe MS, Fleming TR. A nonparametric method for dealing with mismeasured covariate data. J. Amer. Statist. Assoc. 1991; 86:108–113.

Ruppert D. Selecting the number of knots for penalized splines. J. Comput. Graph. Statist. 2002; 11:735–757.

Ruppert D, Carroll R. Spatially-adaptive penalties for spline fitting. Austr. and New Zeal. J. Statist. 2000; 42:205–223.

Ruppert, D.; Wand, MP.; Carroll, RJ. Semiparametric regression. Cambridge University Press; 2003.

Speckman P. Kernel smoothing in partial linear models. J. Roy. Statist. Soc. Ser. B. 1988; 50:413–436.

Strum DP, Pinsky MR. Modeling ischemia-induced dyssynchronous myocardial contraction. Anesth Analg. 2006; 103:846–853. [PubMed: 17000791]

Wang X, Zhou H. Semiparametric Empirical Likelihood for ODS with Ordinal Outcome Variable. Biometrics. 2006; 62:1149–1160. [PubMed: 17156290]

Walkowiak J, Altmann L, Kramer U, Sveinsson K, Turfeld M, Weishoff-Houben M, Winneke G. Cognitive and sensorimotor functions in 6-year-old children in relation to lead and mercury levels: adjustment for intelligence and contrast sensitivity in computerized testing. Neurotoxicol Teratol. 1998; 20:511–521. [PubMed: 9761589]

Weaver, MA. Unpublished doctoral dissertation. University of North Carolina; 2001. Semiparametric methods for continuous outcome regression models with covariate data from an outcome-dependent subsample.

Weaver MA, Zhou H. An Estimated Likelihood Method for Continuous Outcome Regression Models With Outcome-Dependent Sampling. J. Amer. Statist. Assoc. 2005; 100:459–469.

Wu, Z.; Yu, Y. Single-index varying coefficient models with dependent data. University of Cincinnati: Working paper; 2004.

Yu, Y. Penalized spline estimation for generalized partially linear single-index models. University of Cincinnati: Working paper; 2008.

Yu Y, Ruppert D. Penalized spline estimation for partially linear single-index models. J. Amer. Statist. Assoc. 2002; 97:1042–1054.

Zhou H, Pepe MS. Auxiliary covariate data in failure time regression. Biometrika. 1995; 85:139–149.

Zhou H, Chen J, Cai J. Random effects logistic regression analysis with auxiliary covariates. Biometrics. 2002; 58:352–360. [PubMed: 12071408]

Zhou H, Chen J, Rissanen TH, Korrick SA, Hu H, Salonen JT, Longnecker MP. Outcome-dependent sampling: an efficient sampling and inference procedure for studies with a continuous outcome. Epidemiology. 2007; 18:461–468. [PubMed: 17568219]

Zhou H, Weaver MA, Qin J, Longnecker MP, Wang MC. A semiparametric empirical likelihood method for data from an outcome-dependent sampling scheme with a continuous outcome. Biometrics. 2002; 58:413–421. [PubMed: 12071415]

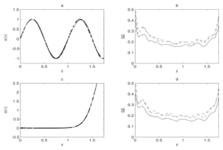

Figure 1.

The estimated function α(·) in Cases 1 and 2 in the simulation studies. Left column: $\hat{\alpha}_P$ (z) (solid line), $\hat{\alpha}_{HT}$ (z) (dashed line), $\hat{\alpha}_{BC}$ (z) (dash-dotted line) and true α(z)(bold dotted line). Right column: The corresponding SEs of these estimators. Note that the notation α(·) represents α(z).

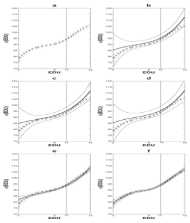

Figure 2.

The estimated function α(·) on EDU for CPP data. Plot a: the curve obtained by proposed method; Plot b: the curve obtained by HT method; Plot c: the curve obtained by BC method; Plot d: the curve obtained by MLE method based on the SRS sample; Plot e: the curve obtained by the proposed method applied to the model considering α(·) as a cubic function. f: the curve obtained by the restricted cubic spline maximum method conducted by maximizing $\hat{\ell}_F\,(\boldsymbol{\beta}, \boldsymbol{\delta})$. Note that for plots (b)–(f), the curve by the proposed penalized spline method is also plotted as background using dashed lines.

Table 1

Number of subjects in the CPP dataset according to subgroup membership.

		Supplemental Sample			
	SRS	C_1 **= (−∞, 82]**	C_2 **= (82, 110]**	C_3 **= (110,∞)**	**Total**
V	849	81	0	108	1038
$\bar{V}$	–	6045	25860	5766	37671
Total	849	6126	25860	5874	38709

Table 2

The finite sample performance for the parametric components.

			β_1				β_2			
$\alpha(z)$	**(n_0, n_1, n_2, N)**	**Methods**	**Mean**	**SE**	**$\widehat{\text{SE}}$**	**CI**	**Mean**	**SE**	**$\widehat{\text{SE}}$**	**CI**
$\sin(2\pi z)$	(100,25,25,600)	β_P	0.987	0.127	0.130	0.947	1.499	0.059	0.058	0.952
		β_{HT}	1.007	0.186	0.183	0.952	1.502	0.081	0.080	0.945
		β_{BC}	1.007	0.178	0.180	0.951	1.501	0.076	0.078	0.942
	(200,25,25,500)	β_P	0.984	0.111	0.099	0.953	1.500	0.057	0.056	0.951
		β_{HT}	0.989	0.138	0.134	0.947	1.498	0.064	0.064	0.949
		β_{BC}	0.994	0.143	0.139	0.948	1.499	0.061	0.063	0.954
	(200,50,50,1200)	β_P	0.992	0.101	0.097	0.955	1.499	0.048	0.047	0.949
		β_{HT}	0.995	0.138	0.134	0.948	1.503	0.062	0.059	0.951
		β_{BC}	0.995	0.130	0.132	0.946	1.502	0.060	0.058	0.952
$\dfrac{0.02\exp(10(z-1))}{1+\exp(8(z-1.5))}$	(100,25,25,600)	β_P	1.003	0.124	0.120	0.946	1.499	0.056	0.054	0.952
		β_{HT}	1.004	0.172	0.176	0.952	1.005	0.078	0.079	0.945
		β_{BC}	1.002	0.175	0.178	0.943	1.504	0.072	0.076	0.943
	(200,25,25,500)	β_P	0.993	0.112	0.106	0.954	1.497	0.051	0.048	0.955
		β_{HT}	0.994	0.137	0.134	0.947	1.501	0.060	0.064	0.951
		β_{BC}	1.002	0.130	0.137	0.951	1.502	0.065	0.063	0.953
	(200,50,50,1200)	β_P	0.997	0.094	0.092	0.955	1.498	0.042	0.040	0.949
		β_{HT}	0.993	0.129	0.124	0.947	1.500	0.057	0.053	0.953
		β_{BC}	0.992	0.124	0.122	0.946	1.499	0.058	0.054	0.947

Note: $\hat{\boldsymbol{\beta}}_P$: the proposed penalized spline maximum likelihood estimation. $\hat{\boldsymbol{\beta}}_{HT}$: the Horvitz-Thompson's weighted-likelihood method. $\hat{\boldsymbol{\beta}}_{BC}$: the Breslow-Cain pseudo-likelihood method.

Table 3

Simple summary statistics for the CPP data set.

		IQ	EDU	SES	RACE	SEX	PCB
Population	Mean	96.0	10.6	4.7	0.5	0.5	
$N = 38709$	SE	14.3	2.4	2.2			
	Min/Max	56.0/153.0	0.0/18.0	0.0/9.5			
SRS	Mean	95.4	10.7	4.7	0.5	0.5	3.1
$n_0 = 849$	SE	14.0	2.3	2.1			1.9
	Min/Max	59.0/142.0	1.0/18.0	0.3/9.3			0.3/16.3
Left Tail	Mean	74.3	9.5	3.5	0.8	0.5	2.7
$n_1 = 81$	SE	6.3	2.3	1.6			1.4
	Min/Max	56.0/82.0	3.0/14.0	0.5/8.0			0.8/8.3
Right Tail	Mean	118.8	12.9	6.9	0.1	0.4	3.7
$n_3 = 108$	SE	6.4	2.4	2.0			2.2
	Min/Max	111.0/145.0	7.0/18.0	1.3/9.3			0.8/17.6

Table 4

Analysis results for the CPP data set.

	PCB	SES	RACE	SEX	EDU
$\hat{\beta}_P$	0.114	1.434	−7.677	−0.069	see Figure 2(a)
SE(β)	0.106	0.039	0.137	0.123	
95% CI	(−0.094, 0.322)	(1.358,1.510)	(−7.946, −7.409)	(−0.310,0.172)	
$\hat{\beta}_{HT}$	0.300	1.067	−8.008	−1.004	see Figure 2(b)
SE(β)	0.188	0.220	0.773	0.700	
95% CI	(−0.068,0.669)	(0.635,1.498)	(−9.523, −6.492)	(−2.375,0.368)	
$\hat{\beta}_{BC}$	0.375	1.056	−8.309	−0.808	see Figure 2(c)
SE(β)	0.180	0.221	0.771	0.687	
95% CI	(0.022,0.728)	(0.623,1.489)	(−9.820, −6.799)	(−2.155,0.538)	
$\hat{\beta}_{MLE\text{-}SRS}$	.0.247	0.979	−7.736	−0.755	see Figure 2(d)
SE(β)	0.227	0.256	0.903	0.837	
95% CI	(−0.198, 0.692)	(0.477,1.480)	(−9.506, −5.966)	(−2.396,0.885)	
$\hat{\beta}_{Cublic}$	0.199	1.459	−7.684	−0.069	see Figure 2(e)
SE(β)	0.103	0.039	0.137	0.123	
95% CI	(−0.002, 0.400)	(1.383, 1.534)	(−7.952, −7.416)	(−0.311, 0.173)	
$\hat{\beta}_{RCS}$	0.117	1.440	−7.677	−0.069	see Figure 2(f)
SE(β)	0.105	0.039	0.137	0.123	
95% CI	(−0.088, 0.323)	(1.364, 1.516)	(−7.945, −7.409)	(−0.311, 0.173)	
$\hat{\beta}_{linear}$	0.27	1.03	−10.24	−0.73	1.52
SE(β)	0.20	0.24	0.86	0.77	0.22
95% CI	(−0.12, 0.66)	(0.56,1.50)	(−11.93, −8.55)	(−2.24,0.78)	(1.09, 1.95)

NOTE: $\hat{\beta}_P$: the proposed penalized spline maximum likelihood method;

$\hat{\beta}_{HT}$: the Horvitz-Thompson’ weighted likelihood method;

$\hat{\beta}_{BC}$: the Breslow-Cain pseudo-likelihood method;

$\hat{\beta}_{MLE\text{-}SRS}$: the MLE method based on the SRS sample;

$\hat{\beta}_{RCS}$: the restricted cubic spline method conducted by maximizing $\hat{\ell}_F(\boldsymbol{\beta}, \boldsymbol{\delta})$;

$\hat{\beta}_{Cubic}$: the proposed method applied to the model considering $\alpha(\cdot)$ as a cubic function;

$\hat{\beta}_{Linear}$: the results from Zhou et al. (2002) which considered a linear model.