Biostatistics (2016), **17**, 4, *pp.* 663–676
doi:10.1093/biostatistics/kxw015
Advance Access publication on March 22, 2016

Estimation of a partially linear additive model for data from an outcome-dependent sampling design with a continuous outcome

ZIWEN TAN, GUOYOU QIN*

Department of Biostatistics, School of Public Health and Key Laboratory of Public Health Safety, Fudan University, Shanghai 200032, China

gyqin@fudan.edu.cn

HAIBO ZHOU

Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

Summary

Outcome-dependent sampling (ODS) designs have been well recognized as a cost-effective way to enhance study efficiency in both statistical literature and biomedical and epidemiologic studies. A partially linear additive model (PLAM) is widely applied in real problems because it allows for a flexible specification of the dependence of the response on some covariates in a linear fashion and other covariates in a non-linear non-parametric fashion. Motivated by an epidemiological study investigating the effect of prenatal polychlorinated biphenyls exposure on children's intelligence quotient (IQ) at age 7 years, we propose a PLAM in this article to investigate a more flexible non-parametric inference on the relationships among the response and covariates under the ODS scheme. We propose the estimation method and establish the asymptotic properties of the proposed estimator. Simulation studies are conducted to show the improved efficiency of the proposed ODS estimator for PLAM compared with that from a traditional simple random sampling design with the same sample size. The data of the above-mentioned study is analyzed to illustrate the proposed method.

Keywords: Design-based sampling; Empirical likelihood; Polychlorinated biphenyls; Semi-parametric model.

1. Introduction

Epidemiologists and biostatisticians are now paying more attention to the outcome-dependent sampling (ODS) design recognizing that it is a cost-effective way to improve study efficiency. A general ODS scheme, e.g. the case–control study in Cornfield (1951), is a retrospective sampling scheme where the primary exposure or covariates are only observed on some subsets of the study population with a probability depending on the value of the outcome variable which can be either continuous or discrete. The principle idea of such a design is to concentrate resources where there is the greatest amount of information, thus

*To whom correspondence should be addressed.

making it more appealing than the traditional simple random sampling (SRS), to provide increased statistical power or equivalently, a reduced sample size given a fixed budget.

The ODS design is especially attractive when considering the high cost of assessing the exposure variable and the easy availability of the outcome. For example, in a recent epidemiological study employing an ODS design (Gray *and others*, 2005), investigators were interested in studying how in utero exposure to polychlorinated biphenyls (PCBs), which is represented by maternal third trimester serum PCB levels, is related to the offspring's intelligence quotient (IQ) at 7 years of age. The study subjects are children born into the Collaborative Perinatal Project (CPP); a prospective cohort was designed to identify determinants of a wide variety of neurological outcomes and birth defects (Niswander and Gordon, 1972). Because of the great expense associated with the blood serum assay to obtain the PCB levels of all 43 628 eligible children from the 55 908 pregnancies recruited from 12 U.S. study centers from 1959 to 1965 (Longnecker *and others*, 2001), the investigators chose to measure the PCB exposure for a sample that was assembled in an ODS way based on the observed IQ scores. That is, treating the eligible 43 628 children as a known population, an overall simple random sample of 1256 subjects and an additional 207 children from those with IQ scores either one or more standard deviations below or above the mean were included in the sample.

For analysis of data obtained through such a complex ODS design as described above, the usual methods like maximum likelihood, assuming identically and independently distributed data, and ordinary least squares, would derive inconsistent estimations (Holt *and others*, 1980). Unlike the case–control study where the sampling scheme can be ignored as the underlying model is logistic, the analysis for data from an ODS design with a continuous response must take into account that the ODS design is a biased sampling scheme to avoid biased estimates of the regression parameters. Several methods have been developed for this purpose. Commonly used methods include the weighted estimating equation approach by Horvitz and Thompson (1952), where the weights are inversely proportional to the probability of being sampled (inverse probability weighting), and another weighted pseudo-likelihood method by Holt *and others* (1980), requiring a correct specification of all the underlying distributions. These methods, however, only account for the sampling scheme in an approximate way. Zhou *and others* (2002) proposed a more efficient semi-parametric empirical likelihood method in which no distribution of the exposure variable was specified and the biased sampling design was more precisely reflected in the likelihood function. A pseudo-score estimation method was also considered by Chatterjee *and others* (2003) for ODS data, as well as the estimated likelihood method proposed by Weaver and Zhou (2005).

Recently, an increasing number of papers focusing on the efficiency gain and application of ODS designs, development of ODS designs and model fitting for the ODS data have been published. For example, in the framework of linear models, Zhou *and others* (2007) investigated the efficiency gain of the ODS designs under a wide range of the exposure distributions and illustrated the application of the ODS designs through 3 epidemiological real data. Xu and Zhou (2012) proposed a mixed regression model for a cluster base of a two-stage outcome-auxiliary-dependent sampling design with a continuous outcome. Zhou *and others* (2014) developed a two-phase probability-dependent sampling scheme. Ding *and others* (2012) studied the regression analysis of a summed missing data problem under an ODS design. Ding *and others* (2014) considered an ODS design for failure-time data with censoring. Moreover, the statistical inference of linear models has also been extended to the case of partially linear models. Zhou, Qin *and others* (2011) and Zhou, You *and others* (2011) both proposed a partially linear model for fitting the data from an ODS design. However, they only consider the model with a single non-parametric function, which may fail to meet with the multiple non-linearities that are commonly seen in practice. Furthermore, their models cannot be directly extended to one with multiple non-linearities due to the identifiability problem, which, in general, refers to the phenomenon that different non-parametric components can generate the same mean so that the true non-parametric components cannot be identified.

In the aforementioned study investigating the relation between prenatal PCB exposure and children's subsequent IQ performance at age 7, the dataset was analyzed to show that the PCB did not have a detrimental effect on IQ scores until reaching a higher serum level (Zhou, Qin *and others*, 2011). Meanwhile, another confounding variable, the highest education level of the mother at the child's birth (EDU), has also been detected to have a non-linear positive influence on IQ scores (Zhou, You *and others*, 2011), which agrees with previous research results that mother's years in college have a much greater effect on children's IQ than those in primary and secondary school do (Oddy *and others*, 2003; Breslau *and others*, 2005). Motivated by the findings above, a more reasonable model which can simultaneously incorporate both the PCB and EDU effects non-parametrically is desired. Therefore, we consider a partially linear additive model (PLAM) in this article to give a more thorough investigation of these relationships under the ODS scheme. In comparison with Zhou *and others* (2007) which focus on linear models, we here allow non-linear relationships between covariates and response. Furthermore, our proposed method can handle more than one non-linear function.

A PLAM can be regarded as a combination of a linear regression model and the generalized additive models (Hastie and Tibshirani, 1990). It allows for flexible specification of the dependence of the response on some covariates linearly and other covariates non-linearly. Coull *and others* (2001) adopted a PLAM to analyze the respiratory health and air pollution data. Liang *and others* (2008) applied PLAMs to study the relationship between environmental chemical exposures and semen quality. Carroll *and others* (2009) considered a PLAM for repeatedly measured data. Liu *and others* (2014) proposed a class of general partially linear additive transformation models for right-censored survival data. Although the PLAM is a powerful tool for data analysis and is widely applied in practice, its statistical inference under ODS designs has not been reported in the literature, possibly due to the complexity in development of theory and computation in the setting of biased ODS designs. Commonly used methods for PLAMs are generally developed for the data from SRS, such as the back-fitting method (Hastie and Tibshirani, 1990), marginal integration approach (Linton and Nielsen, 1995) and direct fitting approach based on low-rank smoothing (Hastie, 1996; Marx and Eilers, 1998). However, these methods cannot be directly applied to the ODS design because it is a biased sampling scheme, and a minor modification to these methods cannot accomplish this task. Therefore, we propose a penalized maximum likelihood method to make inference of the PLAM under an ODS scheme.

The rest of this article is organized as follows. In Section 2, we describe the data structure of the general ODS design and the PLAM, and derive the penalized likelihood function. In Section 3, we propose the inference method and present the asymptotic properties of the proposed estimator. Simulation studies are conducted in Section 4 to demonstrate the performance of the proposed ODS estimator in a PLAM, with a comparison to that derived from a traditional SRS design with the same sample size. In Section 5, a real data from the CPP study is analyzed to illustrate our proposed method. Finally, we give some discussions in Section 6.

2. Data structure, model, and the penalized log-likelihood

2.1 *Notation and ODS data structure*

In this section, we give a brief overview of the ODS design considered in Zhou *and others* (2002). Let Y denote a continuous outcome, the domain of which is assumed to be composed of K mutually exclusive intervals: $C_k = (a_{k-1}, a_k]$ with a_k being known constants satisfying $a_0 = -\infty < a_1 < a_2 < \cdots < a_K = \infty$. Thus, the study population is split into K strata by C_k, $k = 1, 2, \ldots, K$. In practice, it is found that $K = 3$ is sufficient to allow for the benefits of ODS designs. The data structure of the ODS sample consists of two components. First, an overall simple random sample of the population (the SRS sample) is taken with size n_0. Secondly, an additional simple random sample of size n_k from each of the K intervals of

Y (the supplemental samples) is extracted with a distinguishing, perhaps unknown selection probability. Then the total ODS sample size is $n = \sum_{k=0}^{K} n_k$.

To fix notations, let X_l for $l = 1, 2, \ldots, L$ define all the variables relating to the outcome in a non-linear way and $X = (X_1, X_2, \ldots, X_L)^T$, W is a $(p-1)$-dimensional vector of covariates relating to the outcome in a linear way. Then the ODS data structure can be summarized as follows:

SRS sample: $\{Y_{0j}, X_{0j}, W_{0j}\}$, $j = 1, 2, \ldots, n_0$;
Supplemental sample: $\{Y_{kj}, X_{kj}, W_{kj} \mid Y_{kj} \in C_k\}$, $k = 1, 2, \ldots, K$; $j = 1, 2, \ldots, n_k$.

2.2 *A partially linear additive model*

A general PLAM has the following form:

$$Y = \beta^*_{\text{int}} + \sum_{l=1}^{L} g^*_l(X_l) + W^T\beta + \epsilon, \tag{2.1}$$

where β^*_{int} is an intercept term, β is a $(p-1)$-dimensional vector of linear regression coefficients, $g^*_l(\cdot)$ are unknown smooth functions and ϵ is the random error. In this article, we adopt the penalized splines (P-spline) to handle the non-parametric functions. Under the working assumption that $g^*_l(\cdot)$ is an r_l-degree spline function with T_l fixed knots $\{t_1, \ldots, t_{T_l}\}$ for $l = 1, \ldots, L$, we have that $g^*_l(x_l) = \pi^{*T}(x_l)\alpha^*_l$, where $\pi^*(x_l) = (1, x_l, x_l^2, \ldots, x_l^{r_l}, (x_l - t_1)^{r_l}_+, \ldots, (x_l - x_{t_{T_l}})^{r_l}_+)^{\mathrm{T}}$ is an r_l-degree power spline basis with knots $\{t_1, \ldots, t_{T_l}\}$, $(x)^{r_l}_+ = x^{r_l} 1_{x \geqslant 0}$ and α^*_l is an $(r_i + T_l + 1)$-dimensional vector. Due to the requirement of identifiability, the $g^*_l(\cdot)$ in (2.1) are defined only up to an additive constant. Therefore, they can be replaced by centered functions $g_l(x_l) = g^*_l(x_l) - Eg^*_l(x_l)$, where $Eg^*_l(x_l)$ denotes the expectation of $g^*_l(x_l)$. Considering that $g_l(x_l)$ are spline functions, we then have $g_l(x_l) = g^*_l(x_l) - Eg^*_l(x_l) = \pi_l^T\alpha_l$, where $\pi_l^T = (x_l - Ex_l, x_l^2 - Ex_l^2, \ldots, x_l^{r_l} - Ex_l^{r_l}, \ldots, (x_l - t_1)^{r_l}_+ - E(x_l - t_1)^{r_l}_+, \ldots, (x_l - t_{T_l})^{r_l}_+ - E(x_l - t_{T_l})^{r_l}_+)^{\mathrm{T}}$ is a centered r_l-degree truncated power spline basis. Following Aerts *and others* (2002), the expectations can be replaced by their corresponding sample means in the computation. Therefore, for the requirement of identifiability, we will consider the inference of the following model in this article:

$$\begin{aligned} Y &= \beta_{\text{int}} + \sum_{l=1}^{L} g_l(X_l) + W^T\beta + \epsilon, \\ &= \beta_{\text{int}} + \sum_{l=1}^{L} \pi_l^T\alpha_l + W^T\beta + \epsilon, \\ &\doteq D^T\theta_0 + \epsilon, \end{aligned} \tag{2.2}$$

where $D = (\pi_1^T, \ldots, \pi_L^T, W^T, 1)^{\mathrm{T}}$ is a combined design matrix and $\theta_0 = (\alpha_1^T, \ldots, \alpha_L^T, \beta^T, \beta_{\text{int}})^{\mathrm{T}}$. We are interested in estimation of $g_l(\cdot), l = 1, \ldots, L$ and β.

In addition to the truncated power splines, there are alternative splines that can also be used to fit our model. We illustrate how to adapt B splines to our proposed method as an illustrative example in the supplementary material available at *Biostatistics* online. Another thing worth mentioning in practice is the possible interactions that can exist in various forms, such as linear–linear, linear–smooth and smooth–smooth interactions. Including interactions in the PLAM may have important implications depending on the research background. Yet, it is also a challenging topic and worth more consideration and discussion in the future.

2.3 *The penalized log-likelihood function*

Denote $F(y\,|\,x, w; \theta)$ and $f(y\,|\,x, w; \theta)$ as the conditional cumulative distribution function and probability density function, respectively, for Y given X and W; $F_{X,W}(x, w)$ and $f_{X,W}(x, w)$ denote the joint cumulative distribution function and probability density function, respectively, of X and W. Then the likelihood for the observed data from an ODS design can be written as

$$\left\{\prod_{j=1}^{n_0} f(y_{0j}\,|\,x_{0j}, w_{0j}; \theta) f_{X,W}(x_{0j}, w_{0j})\right\}\left\{\prod_{k=1}^{K}\prod_{j=1}^{n_k} f(y_{kj}, x_{kj}, w_{kj}\,|\,y_{kj} \in C_k; \theta)\right\}, \tag{2.3}$$

where $f(y_{kj}, x_{kj}, w_{kj}|y_{kj} \in C_k; \theta)$ denotes the joint density function of $\{y_{kj}, x_{kj}, w_{kj}\}$ conditional on y_{kj} being in the interval C_k. By applying the Bayes formula, (2.3) can be written as

$$L_n(\theta) = \left\{\prod_{j=1}^{n_0} f(y_{0j}\,|\,x_{0j}, w_{0j}; \theta) f_{X,W}(x_{0j}, w_{0j})\right\} \times \prod_{k=1}^{K}\left\{\prod_{j=1}^{n_k} \frac{f(y_{kj}\,|\,x_{kj}, w_{kj}; \theta) f_{X,W}(x_{kj}, w_{kj})}{\int \psi_k(x, w; \theta)\, \mathrm{d}F_{X,W}(x, w)}\right\},$$

where $\psi_k(x, w; \theta) = (F(c_k\,|\,x, w; \theta) - F(c_{k-1}\,|\,x, w; \theta))$ for $k = 1, \ldots, K$. Define $\pi_k = \int \psi_k(x, w; \theta)$ $\mathrm{d}F_{X,W}(x, w)$, $k = 1, \ldots, K-1$, $\pi_K = 1 - \sum_{k=1}^{K-1} \pi_k$ and $p_{kj} = f_{X,W}(x_{kj}, w_{kj})$, $k = 0, \ldots, K$, $j = 1, \ldots, n_k$. We then have the log-likelihood as

$$l_n(\theta, \{p_{kj}\}, \{\pi_k\}) = l_{1n}(\theta) + \sum_{k=0}^{K}\sum_{j=1}^{n_k} \log p_{kj} - \sum_{k=1}^{K} n_k \log \pi_k, \tag{2.4}$$

where $l_{1n}(\theta) = \sum_{k=0}^{K}\sum_{j=1}^{n_k} \log f(y_{kj}|x_{kj}, w_{kj}; \theta)$ is a function only involving θ.

As mentioned earlier, we introduce the P-spline to estimate the non-parametric functions. After incorporating the penalty into (2.4), we obtain the following penalized log-likelihood function:

$$pl_n(\theta, \{p_{kj}\}, \{\pi_k\}; \{q_{s_l}\}) = l_{1n}(\theta) + \sum_{k=0}^{K}\sum_{j=1}^{n_k} \log p_{kj} - \sum_{k=1}^{K} n_k \log \pi_k - \frac{1}{2} n \theta^T \Psi \theta, \tag{2.5}$$

where $\Psi = \mathrm{diag}\{(\mathbf{0}^T_{r_1\times 1}, q_{s_1}\mathbf{1}^T_{T_1\times 1}, \ldots, \mathbf{0}^T_{r_L\times 1}, q_{s_L}\mathbf{1}^T_{T_L\times 1}, \mathbf{0}^T_{p\times 1})^T\}$ with $\{q_{s_l}\}$ being the smoothing parameters for the non-parametric functions and $\theta^T\Psi\theta$ is a common quadratic penalty function.

3. Estimation method and asymptotic properties

Based on the empirical likelihood approach proposed in Zhou *and others* (2002) with no assumptions on the underlying distribution of X and W, we propose a penalized maximum likelihood method to make inference of the PLAM for data from an ODS design.

We first profile the penalized log-likelihood in (2.5) over p_{kj} by fixing θ and obtain the empirical likelihood estimator function of p_{kj} over all distributions whose support contains the observed values of X and W. Then, with the restrictions $\sum_{k=0}^{K}\sum_{j=1}^{n_k} p_{kj}\{\psi_i(x_{kj}, w_{kj}; \theta) - \pi_i\} = 0$, $i = 1, \ldots, K-1$ and $\sum_{k=0}^{K}\sum_{j=1}^{n_k} p_{kj} = 1$, a Lagrange multiplier argument is used and the estimator of p_{kj} is derived as $\hat{p}_{kj} = 1/[n\{1 + \sum_{i=1}^{K-1} \lambda_i\{\psi_i(x_{kj}, w_{kj}; \theta) - \pi_i\}\}]$. Details of the derivation can be found in the supplementary material available at *Biostatistics* online. Substituting $\hat{p}_{kj}$ into (2.5), we then get the following profile penalized likelihood function:

$$ppl_n(\xi; \{q_{s_l}\}_{l=1}^L) = l_{1n}(\theta) - \sum_{k=0}^{K}\sum_{j=1}^{n_k} \log\{1 + v^T h(x_{kj}, w_{kj})\} - \sum_{k=0}^{K}\sum_{j=1}^{n_k} \log\{Q(x_{kj}, w_{kj})\}$$
$$- \sum_{k=1}^{K} n_k \log \pi_k - \frac{1}{2} n\theta^T \Psi \theta, \tag{3.1}$$

where $h(x_{kj}, w_{kj}) = (h_1(x_{kj}, w_{kj}), \ldots, h_{K-1}(x_{kj}, w_{kj}))^{\mathrm{T}}$, $h_i(x_{kj}, w_{kj}) = \{\psi_i(x_{kj}, w_{kj}; \theta) - \pi_i\}/\{Q(x_{kj}, w_{kj})\}$, $i = 1, \ldots, K-1$, and $Q(x_{kj}, w_{kj}) = \sum_{i=1}^{K-1} \psi_i(x_{kj}, w_{kj}; \theta) n_i/(n\pi_i) + n_0/n$. As the Lagrange multipliers $\{\lambda_i\}_{i=1}^{K-1}$ are not centered at zero due to the biased nature of the ODS design, we re-parameterize them to $v_i = \lambda_i - n_i/(n\pi_i)$, $i = 1, \ldots, K-1$ and define $\xi = (\theta^T, \pi^T, v^T)^{\mathrm{T}}$ with $\pi = (\pi_1, \ldots, \pi_{K-1})^{\mathrm{T}}$ and $v = (v_1, \ldots, v_{K-1})^{\mathrm{T}}$. Finally, a Newton–Raphson iterative procedure is applied to obtain the maximizer of the profile penalized likelihood function in (3.1), namely our proposed estimator $\hat{\xi}$ for ξ_0. Note that the smoothing parameters are required to be selected and can be realized through the generalized cross-validation (GCV). In the simulation study and real data analysis, we select the smoothing parameter $\{q_{s_l}\}$ through grid search. To be more specific, for each q_{s_l}, 15 equally spaced (based on the log scale) points across the closed interval $(10^{-6}, 10^7)$ are selected as the grids. For each possible combination of $\{q_{s_l}\}$ values on the grids, we calculate its GCV score and choose the one yielding the smallest GCV score as the selected set of values for $\{q_{s_l}\}$. Borrowing the idea of Qu and Li (2006), the GCV score is defined as

$$\mathrm{GCV}\left(\{q_{s_l}\}_{l=1}^L\right) = \frac{1}{n}\tilde{l}_n \Bigg/ \left(1 - \frac{1}{n} df\right)^2,$$

where $\tilde{l}_n = ppl_n(\xi; \{q_{s_l}\}_{l=1}^L) + (1/2)n\theta^T \Psi \theta$, $\mathrm{df} = \mathrm{trace}\{G(\{q_{s_l}\}_{l=1}^L)\}$ is the effective degrees of freedom with $G\left(\{q_{s_l}\}_{l=1}^L\right) = (\partial^2/\partial\theta\partial\theta^T \tilde{l}_n - n\Psi)^{-1} \partial^2/\partial\theta\partial\theta^T \tilde{l}_n$. Then $\{\hat{q}_{s_l}\}_{l=1}^L = \mathrm{argmin}_{\{q_{s_l}\}} \mathrm{GCV}(\{q_{s_l}\})$.

The following theorem summarizes the asymptotic properties for the proposed estimator of the PLAM from an ODS design.

THEOREM 1 Assume that Conditions (C.1)–(C.4) hold. (i) If smoothing parameters satisfy $q_{s_l} = o(1)$, $l = 1, \ldots, L$, then $\hat{\xi}$ converges to ξ_0 with probability one. (ii) If smoothing parameters satisfy $q_{s_l} = o(1/\sqrt{n})$, $l = 1, \ldots, L$, then $\sqrt{n}(\hat{\xi} - \xi_0) \to N(0, \Omega)$.

The conditions and sketch of the proof is given in the supplementary material available at *Biostatistics* online.

Define $\partial/\partial\xi \left\{ppl_n(\xi; \{q_{s_l}\}_{l=1}^L)\right\} = \sum_{k=0}^K \left\{\sum_{j=1}^{n_k} g_{kj}(y_{kj}, x_{kj}, w_{kj}; \xi)\right\}$ and $(1/n)V_n\left(\xi; \{q_{s_l}\}_{l=1}^L\right) = (1/n)\partial^2/\partial\xi\partial\xi^T\{ppl_n(\xi; \{q_{s_l}\}_{l=1}^L)\}$. The covariance matrix Ω can be consistently estimated by $\hat{\Omega} = \hat{A}^{-1}\hat{B}\hat{A}^{-1}$, where $\hat{A} = (1/n)V_n\left(\hat{\xi}; \{q_{s_l}\}_{l=1}^L\right)$ and $\hat{B} = (1/n)\sum_{k=0}^K \sum_{j=1}^{n_k} g_{kj}\left(y_{kj}, x_{kj}, w_{kj}; \hat{\xi}\right)\ g_{kj}^T \left(y_{kj}, x_{kj}, w_{kj}; \hat{\xi}\right)$.

4. SIMULATION STUDIES

In this section, we conduct simulation studies to evaluate the finite-sample behavior of our proposed estimator. For all simulations, we generate 1000 simulated datasets and each ODS sample consists of 800 individuals. Two different non-parametric functions are considered in the PLAM, one being monotonic and the other being unimodal within threshold regions. In the simulation, we adopt a 3-degree centered truncated power spline basis and choose 10 fixed knots, which are selected as the equally spaced sample quantiles, for each of the two functions, as suggested by Yu and Ruppert (2002). Simulation results based on B splines are included in the supplementary material available at *Biostatistics* online.

We assume the following PLAM in the simulation study:

$$Y = \beta_0 + g_1(X_1) + g_2(X_2) + W^T\beta + \epsilon, \tag{4.1}$$

where $g_1(X_1) = \Phi(4X_1) - 0.5$ with $\Phi(\cdot)$ being a standard normal distribution function, $g_2(X_2) = -1.1\sin(\pi X_2)$, X_1 and X_2 are independently generated from a normal distribution with mean zero and standard deviation 0.25, and W is generated from a normal distribution with mean 1 and standard deviation 0.4. The random error ϵ is normally distributed with mean zero and variance $\sigma_0^2 = 0.3$ and we take $\beta_0 = \beta = 1$.

The ODS design in the simulation study is similar to the design of the CPP study. The total sample is composed of an overall SRS sample n_0 and two equal-sized supplemental samples n_1 and n_3 from the two tail parts of the 3 strata of $Y(n_1 = n_3, n_2 = 0)$ divided by cut points $\mu_Y \pm a\sigma_Y$. Here a is a constant determining the location of the cut points for creating supplemental samples given μ_Y and σ_Y. The proportion of the SRS sample in the ODS is calculated as $\rho = n_0/n$, $n = n_0 + n_1 + n_3$. For this simulation study, 100 000 data are generated as the population data and the mean and standard deviation of the generated population Y's are calculated as μ_Y and σ_Y, respectively. Furthermore, we investigate the effect of different a as well as the varying allocation of ODS sample size between the SRS sample and supplemental samples, on the estimation efficiency by considering scenarios with three different values of $a(0.8, 1.0\ \&\ 1.2)$ and $\rho(0.4, 0.5\ \&\ 0.6)$ after referring to Zhou *and others* (2007), namely $n_0 = 320$, $n_1 = n_3 = 240$, $n_0 = 400$, $n_1 = n_3 = 200$ and $n_0 = 480$, $n_1 = n_3 = 160$.

For all simulations presented here, we compared the proposed estimator based on the ODS sample including both the overall SRS sample and the supplemental samples (P estimator) with three other estimators. One estimator is a penalized maximum likelihood estimator (PMLE) based on the SRS portion of the ODS sample (V-estimator). Another estimator is a PMLE based on an SRS sample with the same sample size as the ODS sample (S-estimator). The third estimator is derived in a similar way but by treating the ODS sample as if it was an SRS sample (M-estimator). The PMLE is derived from the penalized log-likelihood function for simple random sample under normal distribution, $pl_n(\theta; \{q_{s_l}\}) = l_{1n}(\theta) - n\theta^T\Psi\theta/2$, where $l_{1n}(\theta) = -\sum_{j=1}^n \log(2\pi\sigma^2) - (y_j - D_j^T\theta)^2/(2\sigma^2)$ and Ψ is the penalty matrix defined as (2.5) in Section 2. Similar GCV scores to that in Section 3 can be defined for the PMLE method and the same selection procedure of the smoothing parameters as that used by the proposed method can be also applied to the PMLE method. For the ODS sample, the penalized log-likelihood function is defined as (2.5) in Section 2.

For the non-parametric part, we computed the average of the mean square error (AMSE) of the estimated non-parametric functions $\hat{g}_1$ over 801 equal spaced grid points on the interval $[-0.75, 0.75]$ (the mean of X_1 minus and plus 3 times the standard deviation of X_1) over 1000 replications, as well as those of $\hat{g}_2$. Besides, the coverage probability of the 95% nominal confidence interval (CI) at 3 points, namely -0.75, 0, 0.75, for g_1 and g_2, respectively, are also investigated.

To facilitate the comparison, we calculated a relative average mean square error (RMSE) for the estimator of the non-parametric part which is defined as $\text{AMSE}(\hat{g}_Q)/\text{AMSE}(\hat{g}_\text{P})$, where $\hat{g}_Q$ and $\hat{g}_\text{P}$ denote the Q and P estimator, respectively, for either of the two non-parametric functions g_1 and g_2, and Q represents the P, S, V, and M estimator. Similarly, the coverage probability of the 95% nominal CI and a relative MSE (RMSE) defined as $\text{MSE}(\hat{\beta}_Q)/\text{MSE}(\hat{\beta}_\text{P})$ are also calculated for the estimator of the parametric part $\hat{\beta}$.

The simulation results are summarized in Table 1. From the table, we can find that the P estimator has the smallest AMSE (average MSE over 801 grid points) for both $\hat{g}_1$ and $\hat{g}_2$, and the smallest MSE for $\hat{\beta}$, among all the estimators compared since all the RMSEs for the S, V, and M estimators are >1.

Table 1. *Simulation results for the PLAM in the study*

a	Methods	RMSE($\hat{g}_1$)	RMSE($\hat{g}_2$)	RMSE($\hat{\beta}$)	$CP_{\hat{g}_1}(-0.75, 0, 0.75)$	$CP_{\hat{g}_2}(-0.75, 0, 0.75)$	$CP(\hat{\beta})$
				$n_0 = 320, n_1 = n_3 = 240$			
0.8	P	1.0000	1.0000	1.0000	(0.830, 0.948, 0.840)	(0.831, 0.906, 0.820)	0.953
	V	4.8290	4.4746	2.8397	(0.766, 0.947, 0.761)	(0.747, 0.860, 0.761)	0.942
	S	1.0619	1.0769	1.1167	(0.821, 0.931, 0.830)	(0.815, 0.874, 0.829)	0.938
	M	1.1763	2.7745	4.8363	(0.825, 0.951, 0.815)	(0.748, 0.886, 0.746)	0.587
1.0	P	1.0000	1.0000	1.0000	(0.831, 0.936, 0.818)	(0.813, 0.910, 0.847)	0.935
	V	4.6509	4.4903	2.6667	(0.752, 0.929, 0.744)	(0.748, 0.858, 0.756)	0.932
	S	1.1548	1.1432	1.0548	(0.817, 0.924, 0.821)	(0.801, 0.863, 0.820)	0.932
	M	1.2503	3.4196	6.1534	(0.824, 0.931, 0.815)	(0.676, 0.861, 0.707)	0.424
1.2	P	1.0000	1.0000	1.0000	(0.813, 0.928, 0.825)	(0.827, 0.913, 0.841)	0.927
	V	4.4107	5.0147	2.6605	(0.725, 0.920, 0.744)	(0.735, 0.860, 0.723)	0.933
	S	1.0304	1.0951	1.0287	(0.817, 0.924, 0.821)	(0.801, 0.863, 0.820)	0.932
	M	1.2797	5.9730	8.1981	(0.820, 0.926, 0.832)	(0.646, 0.898, 0.608)	0.319
				$n_0 = 400, n_1 = n_3 = 200$			
0.8	P	1.0000	1.0000	1.0000	(0.835, 0.943, 0.838)	(0.835, 0.898, 0.838)	0.950
	V	3.2948	3.4415	2.1254	(0.759, 0.922, 0.775)	(0.768, 0.870, 0.796)	0.952
	S	1.1313	1.1370	1.1232	(0.821, 0.931, 0.830)	(0.815, 0.874, 0.829)	0.938
	M	1.1764	2.3350	4.3450	(0.832, 0.947, 0.840)	(0.794, 0.889, 0.784)	0.624
1.0	P	1.0000	1.0000	1.0000	(0.829, 0.921, 0.832)	(0.834, 0.894, 0.831)	0.945
	V	2.9937	3.1717	2.3145	(0.766, 0.917, 0.762)	(0.757, 0.856, 0.764)	0.939
	S	1.1685	1.1463	1.1494	(0.817, 0.924, 0.821)	(0.801, 0.863, 0.820)	0.932
	M	1.2906	3.4226	6.5563	(0.818, 0.923, 0.811)	(0.731, 0.871, 0.743)	0.432
1.2	P	1.0000	1.0000	1.0000	(0.831, 0.935, 0.819)	(0.819, 0.890, 0.836)	0.944
	V	3.3235	3.2038	2.1598	(0.781, 0.919, 0.773)	(0.788, 0.847, 0.761)	0.937
	S	1.1690	1.0921	1.0488	(0.817, 0.924, 0.821)	(0.801, 0.863, 0.820)	0.932
	M	1.3773	4.6348	8.3850	(0.830, 0.934, 0.814)	(0.676, 0.861, 0.697)	0.293
				$n_0 = 480, n_1 = n_3 = 160$			
0.8	P	1.0000	1.0000	1.0000	(0.823, 0.945, 0.827)	(0.822, 0.878, 0.832)	0.953
	V	2.2548	2.3630	1.8352	(0.776, 0.936, 0.780)	(0.787, 0.846, 0.791)	0.949
	S	1.1434	1.0801	1.1654	(0.817, 0.924, 0.821)	(0.801, 0.863, 0.820)	0.932
	M	1.1586	1.8934	3.6940	(0.821, 0.948, 0.816)	(0.786, 0.862, 0.794)	0.682
1.0	P	1.0000	1.0000	1.0000	(0.844, 0.922, 0.829)	(0.809, 0.880, 0.847)	0.939
	V	2.3113	2.2494	1.8696	(0.791, 0.942, 0.785)	(0.778, 0.839, 0.779)	0.948
	S	1.1240	1.0684	1.1937	(0.817, 0.924, 0.821)	(0.801, 0.863, 0.820)	0.932
	M	1.2205	2.5597	5.8704	(0.846, 0.923, 0.823)	(0.745, 0.866, 0.773)	0.470
1.2	P	1.0000	1.0000	1.0000	(0.831, 0.937, 0.843)	(0.832, 0.900, 0.830)	0.944
	V	2.2539	2.5031	1.8372	(0.800, 0.927, 0.783)	(0.761, 0.866, 0.789)	0.951
	S	1.1432	1.1148	1.0911	(0.817, 0.924, 0.821)	(0.801, 0.863, 0.820)	0.932
	M	1.3580	3.5656	8.3516	(0.821, 0.939, 0.825)	(0.733, 0.872, 0.736)	0.290

Notes: RMSE($\hat{g}_i$), relative mean squared error for $\hat{g}_i$; RMSE($\hat{\beta}$), relative MSE for $\hat{\beta}$; $CP(\hat{\beta})$, coverage probability of 95% nominal CI for $\hat{\beta}$; $CP_{\hat{g}_i}(-0.75, 0, 0.75)$, coverage probability of 95% nominal CI of $\hat{g}_i$ at -0.75, 0, 0.75, $i = 1, 2$; P, proposed estimator; V, the estimator based on the SRS portion of the ODS design; S, the estimator based on an equal-sized SRS sample as the ODS design; M, the estimator treating ODS samples as SRS.

Also, the nominal 95% CIs based on the estimated standard errors for the regression coefficients β are found to provide good coverage. The coverage probabilities of $g_1(\cdot)$ and $g_2(\cdot)$ are higher at point 0 than those at point -0.75 and 0.75, since more data are collected around point 0 and thus can fit better. With a larger sample, the coverage probabilities are expected to be improved. Note that the ODS design is a biased sampling design, so the M method which treats the ODS sample as if it was an SRS sample may result in inconsistent estimates. The obviously high RMSE($\hat{\beta}$) and low coverage probability of β using the M method exhibit evidence that ignoring the ODS design can result in biased estimation. However, unlike the case of a linear model or non-linear model with single monotonic non-linear function, in the ODS design of the CPP study which concentrates more resources on the tails of the outcome, we do not recognize an obvious pattern that the efficiency gain from the ODS design is higher with a low proportion of subjects in the SRS sample or with a high concentration of sampling at the tails of the outcome. This may be due to the introduction of more than one non-linear function simultaneously in the PLAM where these functions would possibly twist with each other. Therefore, how to obtain more informative supplemental samples and improve the efficiency of the ODS design for the case of additive models is an interesting topic and deserves further study in the future.

5. Analysis of the CPP data

As mentioned in the introduction section of the article, our motivating example is the CPP study where primary exposure (maternal pregnancy serum level of PCB) was found to have a non-linear relationship with the offspring's subsequent IQ performance at age 7, so was another confounding variable EDU. We then apply our proposed PLAM accounting for both of the two possibly non-linear functions to the dataset, hoping to draw a more complete picture of the relation between prenatal PCB exposure and the children's IQ scores.

In our analysis, we use the Weschler Intelligence Scale for children at 7 years of age (IQ), which had been observed for the entire CPP population, as the outcome variable and the prenatal PCB exposure level, measured through a serum assay with a relatively high cost, as the exposure variable. Additional confounding variables include the socioeconomic status of the children's family (SES), the gender (SEX), and race (RACE) of the children, and their mother's education (EDU).

As only 1038 of the 1463 subjects obtained in Gray *and others* (2005) were observed to have complete data on all covariates, the ODS data structure for our CPP dataset becomes as follows: an overall SRS sample of 849 subjects, and two supplemental samples with 108 or 81 subjects from the children in the CPP population whose IQ scores are at least 1 standard deviation (14) above or below the mean (96), respectively. The description of the dataset is given in Table 2. Note that we ignore any possible multicenter effect in the analysis presented here.

We then consider the following PLAM for the dataset:

$$\mathrm{IQ} = \beta_0 + g_1(\mathrm{PCB}) + g_2(\mathrm{EDU}) + \beta_1\,\mathrm{SES} + \beta_2\,\mathrm{RACE} + \beta_3\,\mathrm{SEX} + \epsilon, \tag{5.1}$$

where ϵ is a normal error with zero mean. Based on the previous partially linear studies, we adopt a 2-degree centered truncated power function with 10 fixed knots selected as the equally spaced sample quantiles of PCB level (1.16, 1.70, 2.17, 2.61, 3.05, 3.52, 4.06, 4.63, 5.54, and 6.88) to estimate the non-parametric function $g_1(\cdot)$ for PCB. And for estimation of $g_2(\cdot)$ for EDU, we adopt a 3-degree centered power spline with 5 fixed knots selected as the equally spaced sample quantiles of EDU (3, 6, 9, 12, and 15).

We then apply our proposed method to fit this model with the two smoothing parameters q_{s_1} and q_{s_2}, both chosen to be 10^7 by the GCV method, along with 2 other methods, the V method and the M method, as comparisons. Estimates of the regression coefficients are given in Table 3, and the estimates of the non-parametric functions $g_1(\cdot)$ and $g_2(\cdot)$ are presented in Figure 1. Example codes can be found in the supplementary material available at *Biostatistics* online.

Table 2. *Description of variables in the CPP analysis dataset* ($N_{\text{total}} = 1038$)

	Continuous variables					
	MEAN	STD	$Q1$	$Q3$	MIN	MAX
IQ	96.23	16.09	84.00	108.00	56.00	145.00
PCB (ug/L)	3.16	1.93	1.88	3.86	0.25	17.61
EDU (year)	10.86	2.44	9.00	12.00	1.00	18.00
SES	4.84	2.20	3.30	6.30	0.30	9.30
	Categorical variables					
	N	Percent (%)			N	Percent (%)
SEX				RACE		
1 = Female	523	50.39		1 = Black	506	48.75
0 = Male	515	49.61		0 = Other	532	51.25

Notes: For continuous variables: MEAN, mean of the variable; STD, standard deviation of the variable; $Q1$, 25% percentile of the variable; $Q3$, 75% percentile of the variable; MIN, minimum value of the variable; MAX, maximum value of the variable. For categorical variables: N: the number of subjects with the specific value of the variable; Percent: N/N_{total}.

Table 3. *Estimates for the CPP analysis dataset*

	$\hat{\beta}_{\text{P}}$	SE($\hat{\beta}_{\text{P}}$)	95%CI($\hat{\beta}_{\text{P}}$)	$\hat{\beta}_{\text{V}}$	SE($\hat{\beta}_{\text{V}}$)	95%CI($\hat{\beta}_{\text{V}}$)	$\hat{\beta}_{\text{M}}$	SE($\hat{\beta}_{\text{M}}$)	95%CI($\hat{\beta}_{\text{M}}$)
Intercept	95.27	1.32	(92.68, 97.86)	95.36	1.54	(92.34, 98.38)	95.62	1.59	(92.51, 98.73)
PCB	See Figure 1			See Figure 1			See Figure 1		
EDU	See Figure 1			See Figure 1			See Figure 1		
SES	1.04	0.22	(0.61,1.47)	0.97	0.26	(0.46,1.47)	1.25	0.26	(0.74,1.76)
RACE	−8.28	0.76	(−9.77, −6.80)	−7.90	0.90	(−9.67, −6.14)	−10.14	0.92	(−11.94, −8.33)
SEX	−0.82	0.68	(−2.16, 0.52)	−0.76	0.84	(−2.40, 0.87)	−0.97	0.80	(−2.55, 0.60)

Notes: $\hat{\beta}_{\text{P}}$, $\hat{\beta}_{\text{V}}$, and $\hat{\beta}_{\text{M}}$ denote the estimates obtained by the P, V, and M methods, respectively; SE($\hat{\beta}_{\text{P}}$), SE($\hat{\beta}_{\text{V}}$), and SE($\hat{\beta}_{\text{M}}$) are the estimated standard errors of corresponding estimators; 95% CI($\hat{\beta}_{\text{P}}$), 95% CI($\hat{\beta}_{\text{V}}$), and 95% CI($\hat{\beta}_{\text{M}}$) are the 95% CIs for the corresponding estimators.

From the left panel in Figure 1, we can see that the IQ score is related to the PCB level non-linearly. The P estimator of the non-parametric function g_1 shows that the relationship is positive in the lower range of PCB level, and then, after reaching a high point when the PCB level is about 7.44 μg/L, a decreasing trend of the estimator is revealed. While most individuals only own a relatively low PCB level, this result should not be over-interpreted because of the possible existence of other uncaptured confounding variables like a higher fish intake during pregnancy mentioned in Qin and Zhou (2011), which can both relate to a higher serum level of PCBs and a higher IQ in offspring.

The right panel in Figure 1 presents a similar story as with Zhou, You *and others* (2011) that there exists a clear non-linear positive relation between the mother's education level (EDU) and child's IQ score. The P estimate of the function g_2 indicates that EDU has a much greater influence on children's IQ around year 12 (i.e. after higher school education level), which completely agrees with previous results (Oddy *and others*, 2003; Breslau *and others*, 2005).

Both panels in Figure 1 show that the P method yields smaller estimated variances and narrower 95% CIs than the V method.

From Table 3, it is obvious that the P method provides narrower 95% CIs than the V and M methods. The point estimates by the M method show noticeable difference with those by the P and V methods, which is

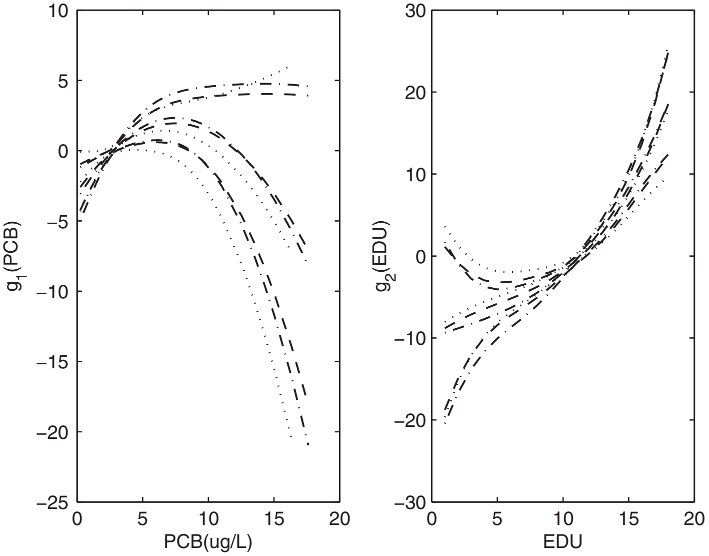

Fig. 1. The estimated function of g_1 on PCB (μg/L) and g_2 on EDU. Thicker dashed curves, dotted curves and dot-dashed curves show the estimates and its corresponding estimated CIs obtained by the P, V, and M methods, respectively. P, proposed estimator; V, the estimator based on the SRS portion of the ODS design; M, the estimator treating ODS samples as SRS samples.

possibly due to the inconsistency of the M method. The conclusions by these three methods are consistent. The socioeconomic status of the child's family is found to have a positive relation with the child's IQ score, and blacks tend to have a worse IQ performance than children of other races. But no evidence has been shown that there exists an association between the child's sex and IQ.

6. Discussion

In this article, we innovatively introduced a PLAM for data obtained through an ODS design with a continuous outcome. To make inference under this biased sampling scheme, we proposed a penalized maximum likelihood method. Simulation studies shows that the proposed method yields more efficient estimates than those obtained from an SRS design with the same sample size across all the scenarios we considered, indicating that applying an ODS scheme in practice would actually reduce the study costs while achieving the same statistical efficiency by requiring a smaller sample size than the SRS design. This advantage is especially meaningful for a budget-limited study where measurement of the main exposure variable is really expensive, yet the outcome is relatively easy to obtain.

Unlike previous findings in the linear model or partially linear model with a single monotonic non-linear function, we do not recognize an obvious pattern that the efficiency gain from the ODS design is higher with a low proportion of subjects in the SRS sample and high concentration of sampling at the extreme tails of the outcome. This may be due to the introduction of more than one non-linear functions simultaneously in the PLAM where these functions would possibly twist with each other. Linear model theory shows that the variance of regression coefficient estimator is inversely proportional to the summed squares of observed X values. In general, there are less X values falling at its distributional tails than at the center and an efficiency gain can be expected if more X's are sampled at the tails. In a simple setting as a linear model or non-linear model with single monotonic non-linear function, if we sample the response Y at its two distributional tails, the observed exposure values X are also more likely to occur at its distributional tails. Therefore, we would expect an obvious efficiency gain from sampling at the tails of the outcome; the larger the proportion of sampling at tails, the more X values would occur at its distributional tails, and thus the more efficiency gain we would expect. However, this may not be the case when we allow for more than one non-linear, non-monotonic function in the model. Sampling more at the tails of Y does not necessarily indicate more covariates occurring at their tails. We are therefore not surprised when observing an unobvious efficiency gain when a larger proportion of sampling at tails is applied. Further studies are needed to explore how to obtain more informative supplemental samples of the ODS design under the framework of additive models.

SUPPLEMENTARY MATERIAL

Supplementary material is available at http://biostatistics.oxfordjournals.org.

ACKNOWLEDGEMENTS

The authors thank the associate editor and two reviewers for constructive suggestions that largely improved the presentation of this article. *Conflict of Interest:* None declared.

FUNDING

G.Q. was supported by The National Natural Science Foundation of China (11371100) and H.Z. by National Institute of Health grants R01ES 021900 and P01CA142538.

REFERENCES

AERTS, M., CLAESKENS, G. AND WAND, M. P. (2002). Some theory for penalized spline generalized additive models. *Journal of Statistical Planning and Inference* **103**, 455–470.

BRESLAU, N., PANETH, N. AND LUCIA, V. C. (2005). Paneth–Pollak R. Maternal smoking during pregnancy and offspring IQ. *International Journal of Epidemiology* **34**, 1047–1053.

CARROLL, R., MAITY, A., MAMMEN, E. AND YU, K. (2009). Efficient semiparametric marginal estimation for the partially linear additive model for longitudinal/clustered data. *Statistics in Biosciences* **1**, 10–31.

CHATTERJEE, N., CHEN, Y. H. AND BRESLOW, N. E. (2003). A pseudoscore estimator for regression problems with two-phase sampling. *Journal of the American Statistical Association* **98**, 158–168.

CORNFIELD, J. (1951). A method of estimating comparative rates from clinical data; applications to cancer of the lung, breast, and cervix. *Journal of the National Cancer Institute* **11**, 1269–1275.

COULL, J. T., NOBRE, A. C. AND FRITH, C. D. (2001). The noradrenergic alpha2 agonist clonidine modulates behavioural and neuroanatomical correlates of human attentional orienting and alerting. *Cerebral Cortex* **11**, 73–84.

DING, J. L., LIU, Y. Y., PEDEN, D. B., KLEEBERGER, S. R. AND ZHOU, H. B. (2012). Regression analysis for a summed missing data problem under an outcome-dependent sampling scheme. *Canadian Journal of Statistics* **40**, 282–303.

DING, J. L., ZHOU, H. B., LIU, Y. Y., CAI, J. W. AND LONGNECKER, M. P. (2014). Estimating effect of environmental contaminants on women's subfecundity for the MoBa study data with an outcome-dependent sampling scheme. *Biostatistics* **15**, 636–650.

Gray, K. A., Klebanoff, M. A., Brock, J. W., Zhou, H. B., Darden, R., Needham, L. and Longnecker, M. P. (2005). In utero exposure to background levels of polychlorinated biphenyls and cognitive functioning among school-age children. *American Journal of Epidemiology* **162**, 17–26.

Hastie, T. (1996). Pseudosplines. *Journal of the Royal Statistical Society Series B* **58**, 379–396.

Hastie, T. and Tibshirani, R. (1990) *Generalized Additive Models*, 1st edition. London; New York: Chapman and Hall.

Holt, D., Smith, T. M. F. and Winter, P. D. (1980). Regression-analysis of data from complex surveys. *Journal of the Royal Statistical Society Series A* **143**, 474–487.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47**, 663–685.

Liang, H., Thurston, S. W., Ruppert, D., Apanasovich, T. and Hauser, R. (2008). Additive partial linear models with measurement errors. *Biometrika* **95**, 667–678.

Linton, O. and Nielsen, J. P. (1995). A kernel-method of estimating structured nonparametric regression-based on marginal integration. *Biometrika* **82**, 93–100.

Liu, L., Li, J. B. and Zhang, R. Q. (2014). General partially linear additive transformation model with right-censored data. *Journal of Applied Statistics* **41**, 2257–2269.

Longnecker, M. P., Klebanoff, M. A., Zhou, H. B. and Brock, J. W. (2001). Association between maternal serum concentration of the DDT metabolite DDE and preterm and small-for-gestational-age babies at birth. *Lancet* **358**, 110–114.

Marx, B. D. and Eilers, P. H. C. (1998). Direct generalized additive modeling with penalized likelihood. *Computational Statistics and Data Analysis* **28**, 193–209.

Niswander, K. R. and Gordon, M. (1972). National institute of neurological diseases and stroke. *The Women and their Pregnancies; The Collaborative Perinatal Study of the National Institute of Neurological Diseases and Stroke*. National Institute of Health; Washington: For sale by the Supt. of Docs., U.S. Govt. Print. Off.

Oddy, W. H., Kendall, G. E., Blair, E., de Klerk, N. H., Stanley, F. J., Landau, L. I., Silburn, S. and Zubrick, S. (2003). Breast feeding and cognitive development in childhood: a prospective birth cohort study. *Paediatric and Perinatal Epidemiology* **17**, 81–90.

Qin, G. Y. and Zhou, H. B. (2011). Partial linear inference for a 2-stage outcome-dependent sampling design with a continuous outcome. *Biostatistics* **12**, 506–520.

Qu, A. and Li, R. (2006). Quadratic inference functions for varying coefficient models with longitudinal data. *Biometrics* **62**, 379–391.

Weaver, M. A. and Zhou, H. B. (2005). An estimated likelihood method for continuous outcome regression models with outcome-dependent sampling. *Journal of the American Statistical Association* **100**, 459–469.

Xu, W. L. and Zhou, H. B. (2012). Mixed effect regression analysis for a cluster-based two-stage outcome-auxiliary-dependent sampling design with a continuous outcome. *Biostatistics* **13**, 650–664.

Yu, Y. and Ruppert, D. (2002). Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association* **97**, 1042–1054.

Zhou, H. B., Chen, J. W., Rissanen, T. H., Korrick, S. A., Hu, H., Salonen, J. T. and Longnecker, M. P. (2007). Outcome-dependent sampling: an efficient sampling and inference procedure for studies with a continuous outcome. *Epidemiology* **18**, 461–468.

Zhou, H. B., Qin, G. Y. and Longnecker, M. P. (2011). A partial linear model in the outcome-dependent sampling setting to evaluate the effect of prenatal PCB exposure on cognitive function in children. *Biometrics* **67**, 876–885.

Zhou, H. B., Weaver, M. A., Qin, J., Longnecker, M. P. and Wang, M. C. (2002). A semiparametric empirical likelihood method for data from an outcome-dependent sampling scheme with a continuous outcome. *Biometrics* **58**, 413–421.

Zhou, H., Xu, W. L., Zeng, D. L. and Cai, J. W. (2014). Semiparametric inference for data with a continuous outcome from a two-phase probability-dependent sampling scheme. *Journal of the Royal Statistical Society Series B* **76**, 197–215.

Zhou, H. B., You, J. H., Qin, G. Y. and Longnecker, M. P. (2011). A partially linear regression model for data from an outcome-dependent sampling design. *Journal of the Royal Statistical Society Series C* **60**, 559–574.

[*Received August 15, 2015; revised February 15, 2016; accepted for publication February 17, 2016*]